Heterogeneity in Impacts of School Characteristics on Student Learning in Developing Countries: Evidence from Peruvian and Vietnamese Panel Data

May, 2013

Incomplete First Draft: Please Do Not Cite

Paul Glewwe
Department of Applied Economics
University of Minnesota

Sofya Krutikova
Young Lives Programme
University of Oxford

Caine Rolleston
Young Lives Programme
University of Oxford

I. Introduction and Motivation

1. Most, if not almost all, economists agree on the importance of human capital, especially formal education, in determining a country's standard of living.

2. Much progress has been made on getting almost all children in developing countries enrolled in primary school, and most enrolled in secondary school.

3. Yet there is sobering evidence that many children in those countries are not learning very much from their time in school. This is seen in Table 1.

Table 1: Student Performance on Math Tests (TIMSS) for Selected Countries

	Grade 8 math score	% advanced (625+)	% very low (<400)
South Korea	613	47%	1%
Japan	570	27%	3%
Russia	539	14%	5%
USA	509	7%	8%
England	507	8%	12%
Thailand	427	2%	38%
Chile	416	1%	43%
Indonesia	386	0%	57%
Ghana	331	0%	79%
Botswana (gr.9)	397	0%	50%
S. Africa (gr.9)	352	1%	76%

Source: TIMSS 2011 International Results in Mathematics

A score of 400 corresponds to "some basic mathematical knowledge. Students can add and subtract whole numbers. They have some recognition of parallel and perpendicular lines, familiar geometric shapes, and coordinate maps."

While 1% to 12% of students in developed countries have **very low scores**, in **developing countries the range is from 38% to 79%**.

Yet at the same time there are **a few students in developing countries who do very well**, as shown in the second column of Table 1.

This paper focuses on this "heterogeneity" of learning outcomes in developing countries.

Why is there such heterogeneity in learning outcomes?

Banerjee and Duflo (2011, pp.89ff.) argue that:

"… teachers still start from the premise that their mandate remains to prepare the best students for the difficult exams that, in most developing countries, act as a gateway either to the last years of [secondary] school or to college…The teacher ignores the children who have fallen behind…"

This paper examines whether this pattern of differential school effectiveness is found in two developing countries, Peru and Vietnam.

More generally, **this paper assesses the contribution of four distinct factors that could explain learning gaps** between advantaged and disadvantaged primary school students in Peru and Vietnam:

a) The child and household **characteristics** that increase learning, such as parental education, are higher among the more advantaged groups.

b) The **impacts** of the child and household characteristics that increase learning are stronger for advantaged children.

c) More advantaged children "**sort**" into better schools.

d) **Learning increases** due to school characteristics are **higher for advantaged children** (relative to disadvantaged children) ***within*** **schools**. This last factor is the one stressed by Banerjee and Duflo.

II. Data

The **Young Lives study**, a multi-country longitudinal study of child poverty in developing countries that tracks approximately 3,000 children in each of four countries: Ethiopia, India (Andhra Pradesh), Peru and Vietnam.

In all four countries two cohorts of children are followed; one consists of children born in 1994-5 and the other follows children born in 2000-01. **This study uses data only from the "younger cohort"**, and focuses on the data **from Peru and Vietnam**.

In all 4 countries, approximately 100 children who were **aged 6-18 months in 2002** were randomly selected from each of 20 sites in each country to form the "younger cohort" of **approximately 2,000** very young children **in each country**.

Data are statistically representative at the site level, but strictly speaking are not representative at the national level because the 20 sites in each country were purposively selected to represent diversity within each country.

The Young Lives study collects data at both the **household level** and the **school level**.

Analysis Using School Level Data

Three rounds of household data have been collected to date (in **2002**, **2006-7** and **2009**) and **school-level data** were collected in **2011-12** from the schools of a sub-sample of the younger cohort children in Peru and Vietnam.

In 2012 these children were aged **about 11 years**. They usually started school in 2006.

The school surveys include **assessment tests in reading comprehension** (Spanish or Vietnamese) **and mathematics**, as well as key indicators of school quality.

The school level analysis employs data from **548 younger cohort children in Peru** and **1131 in Vietnam**, for whom linked school and household data are available.

The school surveys in Peru and Vietnam were conducted in October-December, 2011.

In **Vietnam**, the sample targeted **all younger cohort children in Grade 5** of primary school in the 2011-12 school year, so it sampled all schools attended by these children

In **Peru**, a **subset** of the younger cohort children was tested. A few schools refused to participate. The children included are **mostly in Grade 4**, while **some** are **in grade 5**. We measure children's academic skills using mathematics and reading comprehension tests that were administered as part of the school survey in each country.

In Vietnam, both tests were developed by the Ministry of Education.

In Peru, the tests were developed by GRADE (*Grupo de Análisis para el Desarrollo*).

Test data were subjected to item-response analysis using a three-parameter item-response theory (IRT) model to recover **estimates of underlying/latent performance**.

The **scores in both countries** were **re-scaled to have a mean of 500 and a standard deviation of 100**, for ease of interpretation. Note, however, that this common scaling **does not allow direct comparison of** scores between **these two countries**.

Analysis Using 2009 (Round 3) Household Level Data

Using only the 2011-12 school level data, we lose some of the sample from Vietnam and most of the sample from Peru:

- In **Vietnam**, we lose younger cohort children who are not in Grade 5.
- In **Peru**, we lose children who were not in the subset of schools that were randomly selected for the school survey.

To increase the sample, we also present results using the tests administered in both countries in Round 3 (2009).

Peru Round 3 Sample

Of original 2000 younger cohort children, 1523 have both a school code and a math test score in 2009. They are spread across 593 schools.

Keeping only schools with both rich & poor kids leaves **593 children in 71 schools**.

Our measures of **children's background characteristics** for both countries include child, parental and household characteristics. **Table 2** summarizes these variables:

Table 2: Measures of child background characteristics included in the analysis

Child characteristics	Age (in months), gender and ethnicity Mother tongue Height for age (z-scores) Dummies for youngest child, oldest child and only child Dummies for attended crèche and attended preschool Hours of work CDA Rasch score
Parent characteristics	Mother and father in the household Carer has no, primary or lower secondary education (dummies) Carer expects child to be in school/professional occupation at age 20 Caregiver's position in Cantril's ladder of life
Household characteristics	Location (rural/urban) Sex of head Head with no, primary or lower secondary education (dummies) Wealth index

III. Methodology and Estimation Issues

A. Equations of Interest.

With only a small loss of generality, **divide the children's lives** from birth to their current age (10-11 years old) **into three time periods**, each of which are denoted by t:

t = 1: **First 1-2 years** of life (corresponds to Round 1 data)

t = 2: **Ages 2-5**, which ends when child enrols in primary school (Round 2 data)

t = 3: **Ages 6-11**, the primary school years (Round 3 and school survey data)

The **equation for cognitive skills at end of time period 2**, S_2, is:

$$S_2 = S_2(N_1, N_2, PE, PT_1, PT_2, PRE_2, AI) \qquad (1)$$

That is, these cognitive skills are a function of:

- Early childhood nutrition in period 1 (N_1) and preschool nutrition in period 2 (N_2)
- Parental education (PE)
- Parental time spent with the child during periods 1 and 2 (PT_1 and PT_2)
- Pre-school attendance (including day care) in time period 2 (PRE_2)
- "Innate ability" (IA)

This is a **structural equation for skill formation BEFORE** child starts **primary school**.

The **equation for skills in time period 3** (denoted by S_3) is:

$$S_3 = S_3(S_2; N_1, N_2, N_3, PE, PT_3, EI_3, IA; \mathbf{SC}) \qquad (2)$$

Thus those skills are determined by:

- Skills acquired by the end of time period 2 (S_2), which could be a vector
- Possible lingering effects of early childhood nutrition (N_1 and N_2)
- Nutrition during time period 3 (N_3)
- Parental education (PE)
- Parental time spent with child in time period 3 (PT_3)
- Educational inputs bought by households such as tutoring & children's books (EI_3)
- Innate ability (IA)
- A vector of school & teacher characteristics (**SC**, it is bold because it is a vector)

Note: The **SC** variables **include peer effects** as well as teacher & school characteristics.

As a **first approximation** to the process by which skills are formed, consider **linear specifications** of each of the above two equations:

$$S_2 = \alpha_0 + \alpha_1 N_1 + \alpha_2 N_2 + \alpha_3 PE + \alpha_4 PT_1 + \alpha_5 PT_2 + \alpha_6 PRE_2 + \alpha_7 IA + u_2 \quad (1')$$

$$S_3 = \beta_0 + \beta_1 S_2 + \beta_2 N_1 + \beta_3 N_2 + \beta_4 N_3 + \beta_5 PE + \beta_6 PT_3 + \beta_7 EI_3 + \beta_8 IA + \boldsymbol{\gamma' SC} + u_3 \quad (2')$$

where the **residual terms u_2 and u_3 represent two distinct phenomena**:

a) errors due to the linear approximation

b) measurement errors in S_2 in equation (1′) and in S_3 in equation (2′)

Both of these components of u_2 and u_1 are assumed to be uncorrelated with the explanatory variables in equations (1′) and (2′), respectively. Measurement error in the explanatory variables will be discussed below.

In general, as long as a nonlinear function is continuous, **a linear expression** of that function **can approximate the nonlinear function** very closely **by adding higher order** (e.g. quadratic) **terms** for each variable ***and*** by adding sufficient **interaction terms** between the variables in that function.

The **interaction terms** can be grouped into **three types**:

a) Those between child & household level variables (N vars, PE, PT vars, PRE_2, EI_3, IA)

b) Those between the school characteristic variables (**SC**)

c) Those between the child/ household variables and the school characteristic variables

One goal of this paper is to **investigate the extent to which countries' education systems reduce or reinforce gaps in learning outcomes**.

One mechanism by which reduction or reinforcement may occur is via the 3rd type of interaction, e.g. whether the impacts of school characteristics on child learning vary according to disadvantage in terms of wealth, initial learning and parental education.

One **problem** that plagues **estimates of education "production functions"** is **omitted variable bias**. There could be dozens, if not hundreds, of school characteristics (**SC**) that affect skill acquisition. Many that could be very important (teacher motivation, pedagogical practices used, teacher ability to diagnose students' learning difficulties) are very difficult to measure.

This paper avoids the problem of measuring all school characteristics that could affect students' acquisition of cognitive skills **by replacing γ′SC** in equation (2′) **with** a set of **school fixed effects**:

$$S_3 = \beta_0 + \beta_1 S_2 + \beta_2 N_1 + \beta_3 N_2 + \beta_4 N_3 + \beta_5 PE + \beta_6 PT_3 + \beta_7 EI_3 + \beta_8 IA + \sum_{s=1}^{S} \delta_s D_s + u_3 \quad (2'')$$

where D_s is a dummy variable indicating school s and δ_s indicates the total impact of that school's characteristics on student skill acquisition.

These **school fixed effects measure the impact of *all* school characteristics, *both observed and unobserved***, on student learning. They also incorporate all possible interaction terms between the various **SC** variables.

Interactions of the first type, those that are between school and household variables, are **easily accommodated** by using quadratic and interaction terms for those variables.

The **extent to which gaps in** students' acquisition of **skills are reduced or reinforced by the education system** can be divided into **two general phenomena**:

a) **Sorting**: The tendency for disadvantaged students to go to lower quality schools, while more advantaged students go to higher quality schools

b) The extent to which **disadvantaged students learn less** (or more) than advantaged students **within a given school**.

To distinguish between these 2 phenomena, and to allow for interaction terms of the third type (interactions of school characteristics & student/household characteristics), a **simple approach** is to **divide all students into a "disadvantaged" group and an "advantaged" group**. Assume that the impact of the school variables could differ across these two groups.

This implies that **equation (2′′)** could be **modified as follows**:

$$S_3 = \boldsymbol{\beta}'\mathbf{X} + \sum_{s=1}^{S} \delta_s D_s + \sum_{s=1}^{S} \theta_s D_s A + u_3 \quad (3)$$

where A is a dummy variable indicating that an "advantaged" student and $\boldsymbol{\beta}'\mathbf{X}$ denotes $\beta_0 + \beta_1 S_2 + \beta_2 N_1 + \beta_3 N_2 + \beta_4 N_3 + \beta_5 PE + \beta_6 PT_3 + \beta_7 EI_3 + \beta_8 IA$.

The **impact of a school s on a disadvantaged student** is captured by the term $\boldsymbol{\delta_s}$, while the impact of the same school on an **advantaged student** is estimated by $\boldsymbol{\delta_s + \theta_s}$.

If school s contributes equally to the learning of both advantaged and disadvantaged students, then $\theta_s = 0$.

For further flexibility, the impact of child and household characteristics can also be allowed to vary over advantaged and disadvantaged students:

$$S_3 = \boldsymbol{\beta}_A'\mathbf{X}_A A + \boldsymbol{\beta}_{DA}'\mathbf{X}_{DA}(1 - A) + \sum_{s=1}^{S} \delta_s D_s + \sum_{s=1}^{S} \theta_s D_s A + u_3 \quad (3')$$

where the A subscript indicates advantaged children and the DA subscript indicates disadvantaged children.

Equation (3′) provides a convenient "Oaxaca-Blinder" framework for decomposing the (average) learning gap between the advantaged group and the disadvantaged group **into four components**:

a) **Differences child and household characteristics** (the **X** variables) that increase learning, such as parental education, **between the two groups**.

b) **Differences** across the two groups **in the impacts of the child and household characteristics** on learning.

c) **Sorting** of advantaged children into better schools.

d) **Differences in the impacts of school characteristics** on learning between advantaged and disadvantaged ***within*** **schools**. (Banerjee-Duflo hypothesis)

To **obtain this decomposition from equation (3′)**, note that the average learning of advantaged and disadvantaged children, denoted by $\overline{S}_{3,A}$ and $\overline{S}_{3,DA}$, are:

$$\overline{S}_{3,A} = \boldsymbol{\beta}_A' \overline{\mathbf{X}}_A + \sum_{s=1}^{S} (\delta_s + \theta_s) \overline{D}_{s,A} \qquad (4)$$

$$\overline{S}_{3,DA} = \boldsymbol{\beta}_{DA}' \overline{\mathbf{X}}_{DA} + \sum_{s=1}^{S} \delta_s \overline{D}_{s,DA} \quad (5)$$

These equations can be used in **two ways to decompose the gap in average test scores** between the advantaged and disadvantaged groups.

The **first decomposition** method divides this gap into four components mentioned above, plus two "interaction" terms:

$$\overline{S}_{3,A} - \overline{S}_{3,DA} = \boldsymbol{\beta}_A' \overline{\mathbf{X}}_A - \boldsymbol{\beta}_{DA}' \overline{\mathbf{X}}_{DA} + \sum_{s=1}^{S} (\delta_s + \theta_s) \overline{D}_{s,A} - \sum_{s=1}^{S} \delta_s \overline{D}_{s,DA} \qquad (6)$$

$$= \boldsymbol{\beta}_{DA}' (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA}) + (\boldsymbol{\beta}_A - \boldsymbol{\beta}_{DA})' \overline{\mathbf{X}}_{DA} + (\boldsymbol{\beta}_A - \boldsymbol{\beta}_{DA})' (\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA})$$

$$+ \sum_{s=1}^{S} \delta_s(\overline{D}_{s,A} - \overline{D}_{s,DA}) + \sum_{s=1}^{S} \theta_s \overline{D}_{s,DA} + \sum_{s=1}^{S} \theta_s(\overline{D}_{s,A} - \overline{D}_{s,DA})$$

The decomposition in equation (6) is **from the perspective of a disadvantaged student**.

The **first term**, $\boldsymbol{\beta}_{DA}'(\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA})$, indicates **how much of the gap is due to differences in average child and family characteristics** between advantaged and disadvantaged students, applied to the "productivity" of a disadvantaged student ($\boldsymbol{\beta}_{DA}$).

The **second term**, $(\boldsymbol{\beta}_A - \boldsymbol{\beta}_{DA})'\overline{\mathbf{X}}_{DA}$, is the **part of the test score gap due to differences** across advantaged & disadvantaged students **in "productivity" child & family characteristics**, applied to the average characteristics of a disadvantaged student ($\overline{\mathbf{X}}_{DA}$).

The **third term** allows for an **interaction effect between the first two terms**.

The **fourth term**, $\sum_{s=1}^{S} \delta_s(\overline{D}_{s,A} - \overline{D}_{s,DA})$, is the part of the gap due to the **2 types of students attending different schools**, measured by school's impacts on disadvantaged students (δ_s).

The **fifth term**, $\sum_{s=1}^{S} \theta_s \overline{D}_{s,DA}$, measures how much of the gap is due to **advantaged students learning more** than disadvantaged students **in the same school**.

The **last term**, $\sum_{s=1}^{S} \theta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$, allows for **interaction** effects **between the 4th & 5th terms**. A **visual interpretation** of this decomposition is given in **Figure 1**.

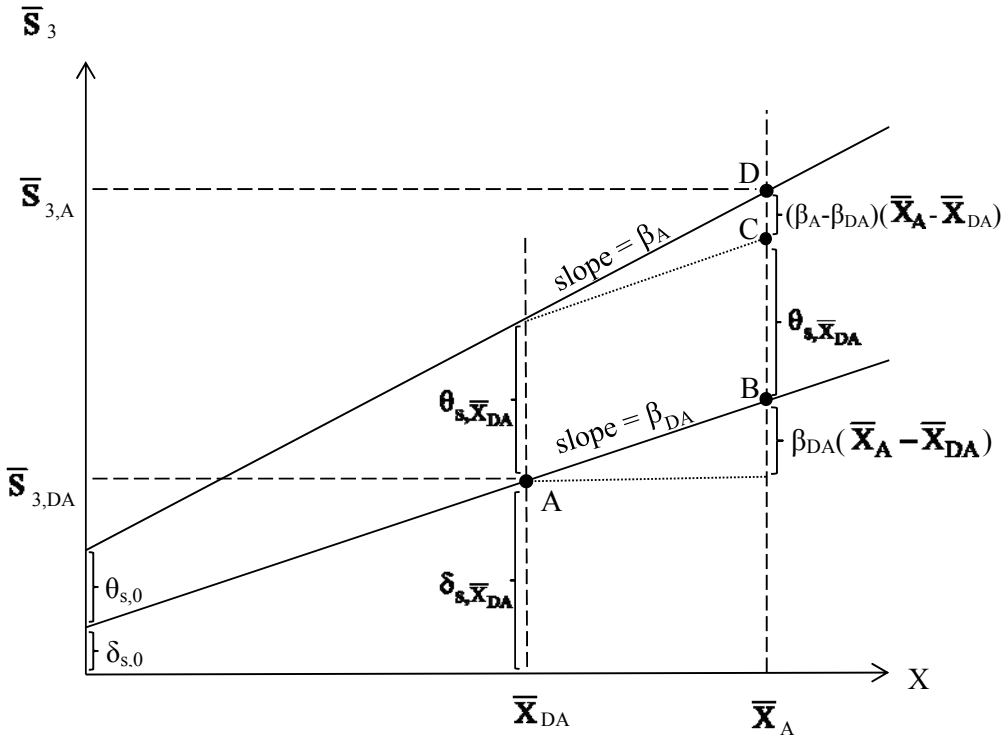

Figure 1: First Type of Decomposition (for a given school s)

The mean value of X for the **disadvantaged group** is given by $\overline{X}_{DA}$, which implies that the **mean value of test scores** for that group is given by $\overline{S}_{3,DA}$, and is also depicted on the graph by the **point A**.

Similarly, the mean value of X **for the advantaged group** is given by $\overline{X}_{A}$, and the mean test scores for that group is given by $\overline{s}_{3,A}$, and is depicted on the graph by the **point D**.

The **"standard" Oaxaca-Blinder decomposition** for equation (6) implicitly **takes as its "base group" a person with X = 0**, **and the associated school fixed effect terms** for that base group are $\boldsymbol{\delta_s}$ **and** $\boldsymbol{\theta_S}$, as shown in the lower left corner of Figure 1.

However, for the purpose of comparing the average disadvantaged student to the average advantaged student, it **makes more sense to set the base group as the average disadvantaged student, that is X =** $\overline{X}_{\mathbf{DA}}$, which implies that the relevant **school fixed effects are** $\delta_{s,\overline{X}_{DA}}$ **and** $\theta_{s,\overline{X}_{DA}}$**,** as shown in the center of Figure 1. **This choice for the base group will be used for all the analysis in this paper; it is easy to implement for all X variables by subtracting the mean of each X variable for the disadvantaged population from the values of that variable for all the children in the sample.**

The decomposition of the test score gap between advantaged and disadvantaged children in (6) is shown on the right side of Figure 1 for children in a given school; **the figure does not show the impact due to the fact that different children attend different schools.**

The **first term** in (6) shows the **"composition effect"**, what would happen to the test scores of an average disadvantaged child if he or she had the same mean characteristics as an advantaged child, but still the same "productivity" of a disadvantaged child. **In Figure 1**, this is shown as a **move from point A to point B**.

Once all X variables are normalized so that their means equal 0 for the disadvantaged group, the **second term in (6) disappears**. That term measures the extent to which the **gap between a child with all (pre-normalized) X variables equal to 0 and a child whose X variables are the means for the disadvantaged group** is due to differences in the "efficiency" of those variables, but that gap is **not the decomposition of interest**.

The **third term** in (6), $(\boldsymbol{\beta}_A - \boldsymbol{\beta}_{DA})'(\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA})$, accounts for the **interaction between the "increased productivity"** of the advantaged child **and the composition effect**; in Figure 1 it is represented by the **move from point C to point D**.

The **fourth** term in (6) measures the part of the gap, if any, due to **advantaged students attending better schools than the disadvantaged students**, but still measured in terms of school impacts on disadvantaged students; this **cannot be shown in Figure 1** since that figure is for a given school.

The **fifth term** in (6) is accounted for by $\theta_{s,\overline{X}_{DA}}$ in Figure 1; **advantaged students may learn more in the same school than disadvantaged students**, which moves them **from point B to point C** in that figure.

Finally, the **last term** in (6) accounts for the fact that **advantaged students may go** to better schools in the sense that they may go **to schools that have a higher "premium" for advantaged students**; this **cannot be shown in Figure 1** because it is due to differences across schools.

There is a **second type of decomposition**, which **generates "average" values** of the **β** coefficients and the school fixed effects over the two groups of children.

This proceeds as follows, starting with the difference between equations (4) and (5):

$$\overline{S}_{3,A} - \overline{S}_{3,DA} = \boldsymbol{\beta}_A'\overline{\mathbf{X}}_A - \boldsymbol{\beta}_{DA}'\overline{\mathbf{X}}_{DA} + \sum_{s=1}^{S}(\delta_s + \theta_s)\overline{D}_{s,A} - \sum_{s=1}^{S}\delta_s\overline{D}_{s,DA} \qquad (7)$$

$$= \boldsymbol{\beta}_A'\overline{\mathbf{X}}_A + \boldsymbol{\beta}_{AVG}'\overline{\mathbf{X}}_A - \boldsymbol{\beta}_{AVG}'\overline{\mathbf{X}}_A - \boldsymbol{\beta}_{DA}'\overline{\mathbf{X}}_{DA} + \boldsymbol{\beta}_{AVG}'\overline{\mathbf{X}}_{DA} - \boldsymbol{\beta}_{AVG}'\overline{\mathbf{X}}_{DA}$$

$$+ \sum_{s=1}^{S}(\delta_s + \theta_s)\overline{D}_{s,A} - \sum_{s=1}^{S}\delta_{s,AVG}\overline{D}_{s,A} + \sum_{s=1}^{S}\delta_{s,AVG}\overline{D}_{s,A} - \sum_{s=1}^{S}\delta_s\overline{D}_{s,DA} + \sum_{s=1}^{S}\delta_{s,AVG}\overline{D}_{s,DA} - \sum_{s=1}^{S}\delta_{s,AVG}\overline{D}_{s,DA}$$

$$= \boldsymbol{\beta}_{AVG}'(\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA}) + (\boldsymbol{\beta}_A - \boldsymbol{\beta}_{AVG})'\overline{\mathbf{X}}_A + (\boldsymbol{\beta}_{AVG} - \boldsymbol{\beta}_{DA})'\overline{\mathbf{X}}_{DA}$$

$$+ \sum_{s=1}^{S}\delta_{s,AVG}(\overline{D}_{s,A} - \overline{D}_{s,DA}) + \sum_{s=1}^{S}((\delta_s + \theta_s) - \delta_{s,AVG})\overline{D}_{s,A} + \sum_{s=1}^{S}(\delta_{s,AVG} - \delta_{s,DA})\overline{D}_{s,DA}$$

One can interpret the six terms in the last expression as follows.

The **first term**, $\boldsymbol{\beta}_{AVG}'(\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA})$, is the portion of the gap due to differences in average child and household characteristics across advantaged and disadvantaged students.

The **second and third terms combined** show the impact of **differences in "productivity" of child and household characteristics** between advantaged and disadvantaged students, with the second term focusing on the difference between advantaged and average students (evaluated at mean values of **X** for advantaged students), and the third term adding the difference between average and disadvantaged students (evaluated at mean values of **X** for disadvantaged students).

The **fourth term** measures the part of the gap due to the fact that advantaged and disadvantaged students may attend different schools (**sorting**).

The **fifth and sixth terms together** account for the **differences in learning *within* schools** between advantaged and disadvantaged students.

This decomposition is shown in Figure 2, which again focuses on a single school and thus does not show the components of the decomposition that measure "sorting" of advantaged and disadvantaged students into different schools.

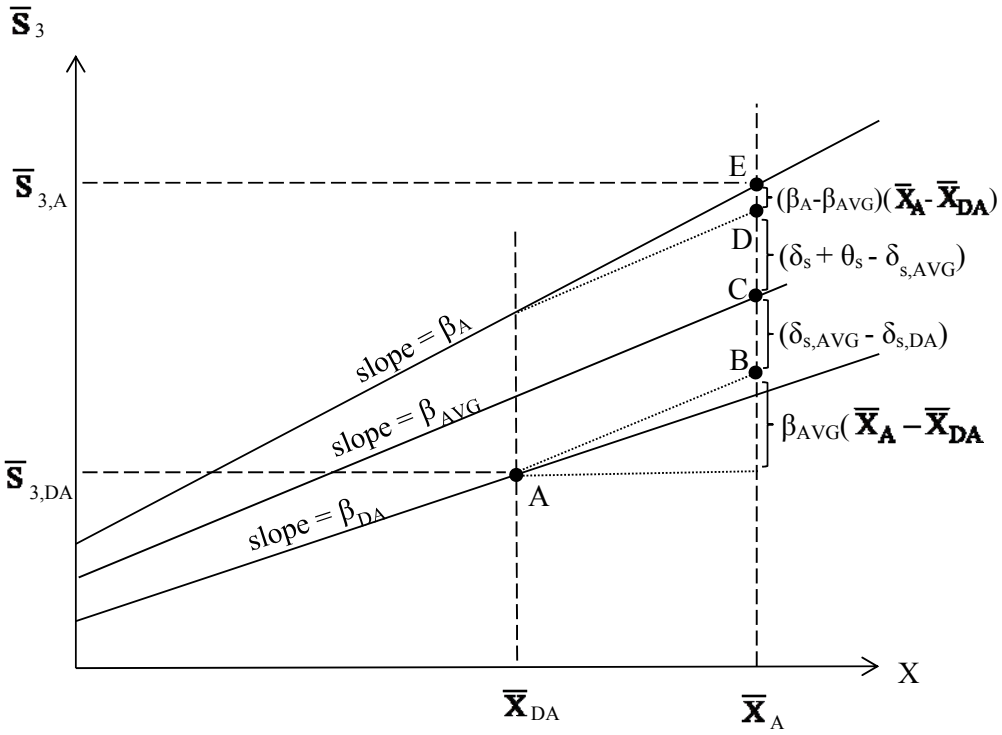

Figure 2: Second Type of Decomposition (for a given school s)

The gap between $\overline{S}_{3,A}$ and $\overline{S}_{3,DA}$ is in the movement from point A to point E, so the school fixed effects will again be measured at the point where $X = \bar{x}_{DA}$, not at X = 0.

The **first term** in (7) is the **main composition effect**. It shows how test scores would increase if the mean of the X variable among disadvantaged children were changed to the mean for advantaged children; this mean is valued at the average value of β (β_{AVG}). This is the increase in $\overline{S}_3$ is shown as the **move from point A to point B** in Figure 2.

The **second term boosts the effect of the first term further**, by valuing the change in the mean of the X variable at β_A, which is slightly higher than β_{AVG}. This **"increased productivity effect"**, $(\beta_A - \beta_{AVG})'\overline{X}_A$, is the move **from point D to point E** in Figure 2 (since X is normalized to $\overline{X}_{DA} = 0$ this term can be written as $(\beta_A - \beta_{AVG})'(\overline{X}_A - \overline{X}_{DA})$).

The **third term** equals zero since X is normalized so that $\overline{X}_{DA} = 0$. In effect it captures the **movement from X = 0 to X = $\overline{X}_{DA}$,** which is **not of interest** for this decomposition.

The **fourth term** reflects variation in the distrib. of students to schools; it **can't be shown.**

The **fifth term** shows the increase in skills going from an "average student" FE to an "advantaged student" FE; the **sixth** does the same for going from a "disadvantaged student" FE to an "average student" FE; these the **moves from C to D, and from B to C**.

B. Estimation Issues.

Several complications arise concerning these estimations. This section presents several estimation problems and the methods used to address them.

1. Omitted variable bias. While the Young Lives data contain a wealth of information dating back to when the students were 1-2 years old, they contain little information on education inputs (EI_3). The approach taken here is to include household wealth, which is used to purchase these items. Technically, this implies that equation (2″) is no longer a "pure" production function because other variables in it, such as parental education and child ability, also affect educational inputs. But this should have little or no effect on estimates of the school fixed effects.

Another important omitted variable is lack data on innate ability (IA). This could also lead to bias in estimates of the structural impacts of observed variables since some of those variables may be correlated with IA, such as S_2 and PE. The approach taken here is to assume that S_2 accounts for much of the impact of innate ability on S_3, so the omitted variable bias due to lack of data on that variable should be minimal.

2. Measurement error in explanatory variables. Examples are S_2, N and PT variables. Fortunately there are previous or multiple measures of most of these, which should be useful IVs to deal with this problem. **[Have not done yet.]**

3. Sample selection. Some of the schools from both the Peru and Vietnam samples do not have both advantaged and disadvantaged children in them. **For schools without advantaged children, it is not possible to estimate θ_s; for schools with no disadvantaged children it is impossible to estimate δ_s,** although one can still estimate $\delta_s + \theta_s$.

Thus the main estimates focus only on schools that have both types of children, which could suffer from sample selection bias. It is probably relatively small for **Vietnam**, where **82.3% of the sample** (931 out of 1131) **attend schools that have both types** of children when the advantaged group is defined as children in the top two quintiles of the wealth distribution and the disadvantaged group is defined as the bottom three quintiles of that distribution.

However, in **Peru only 55.3% of the sample** (302 out of 546) **are in schools that contain both** advantaged and disadvantaged children similarly defined.

V. Results

The following tables present the two types of decompositions for mathematics scores in Peru and Vietnam. **Table 3 presents simple descriptive statistics** for those test scores as well as for the control variables used in the regressions and associated decompositions. The test scores were normalized to have means of 500 and standard deviations of 100 for the whole sample, but for both countries children who attend schools that have all advantaged children or all disadvantaged children are dropped from the sample, so that the means are slightly above 500 and the standard deviations are slightly below 100.

The results for the **decompositions** are shown in **Tables 4 and 5**. In both tables, the advantaged group is defined as children in the top two quintiles of the wealth distribution, and the disadvantaged group is defined as those in in the bottom three quintiles of the wealth distribution.

Table 3A: Descriptive Statistics for Peru

Variables	Peru (all)	Peru (non-poor)	Peru (poor)	difference in means
Maths IRT score	539.99	563.41	507.73	55.68***
	(86.30)	(80.74)	(83.5)	
Male	0.45	0.48	0.41	0.08
Age in months	64.47	65.27	63.37	1.9***
	(4.57)	(4.68)	(4.18)	
Height for age z-score (age 5)	-1.28	-1.09	-1.54	0.45***
	(1.04)	(0.91)	(1.14)	
Rasch score PPVT (age 5)	313.34	315.29	310.66	4.63
	(42.39)	(41.99)	(42.96)	
Rasch score on CDA test (age 5)	306.60	313.61	296.94	16.66***
	(43.13)	(40.39)	(45.04)	
Number of siblings (age 5)	1.75	1.53	2.06	-0.54***
	(1.67)	(1.44)	(1.9)	
Mother's educ,: secondary plus	0.23	0.33	0.11	0.22***
Father's educ.: secondary plus	0.23	0.30	0.13	0.16***
Wealth index (age 5)	0.58	0.71	0.40	0.30***
	(0.18)	(0.09)	(0.13)	
Parent's educ, aspiration for child	0.85	0.90	0.79	0.11***
Caregiver: Cantril well-being ladde	5.07	5.27	4.78	0.48**
	(1.72)	(1.70)	(1.73)	
Total Observations	302	175	127	

Total Schools	36	36	36

Table 3B: Descriptive Statistics for Vietnam

Variables	Vietnam (all)	Vietnam (non-poor)	Vietnam (poor)	difference in means
Maths IRT score	504.09	523.68	487.88	35.81***
	(83.75)	(97.51)	(91.23)	
Male	0.51	0.51	0.51	0
Age in months	64.52	64.66	64.41	0.26
	(2.78)			
Height for age z-score (age 5)	-1.26	-1.06	-1.42	0.36***
	(0.96)	(0.97)	(0.91)	
Rasch score PPVT (age 5)	304.77	315.91	295.54	20.37***
	(45.12)	(46.77)	(41.54)	
Rasch score on CDA test (age 5)	307.10	312.70	302.47	10.22***
	(46.14)	(43.4)	(47.83)	
Number of siblings (age 5)	1.12	1.10	1.13	-0.03
	(0.90)	(0.89)	(0.90)	
Mother's education: secondary plus	0.16	0.28	0.06	0.22***
Father's education: secondary plus	0.23	0.37	0.12	0.26***
Wealth index (age 5)	0.55	0.69	0.44	0.25***
	(0.16)	(0.08)	(0.12)	
Parent's educ. aspiration for child	0.81	0.92	0.71	0.21***
Caregiver Cantril well-being ladder	4.31	4.94	3.78	1.16***
	(1.49)	(1.38)	(1.36)	
Total Observations	932	422	510	

Total Schools	48	48	48

Table 4: First Decomposition

Component	Detail	Peru	Vietnam
$(\bar{S}_{3,A} - \bar{S}_{3,DA})$	Difference	55.68*** (9.59)	35.59*** (6.24)
$\beta_{DA}'(\bar{X}_A - \bar{X}_{DA})$	Background composition effect	14.580 (20.74)	22.441** (11.602)
$(\beta_A - \beta_{DA})'\bar{X}_{DA}$			
$(\beta_A - \beta_{DA})'(\bar{X}_A - \bar{X}_{DA})$	"Increased productivity" of advantaged person	-27.393 (32.847)	11.838 (20.641)
$\sum_{s=1}^{S} \delta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	School composition effect	10.22 (13.19)	30.528*** (11.130)
$\sum_{s=1}^{S} \theta_s \bar{D}_{s,DA}$	School "coeff" effect – advantaged students learn more	59.812** (25.921)	-4.957 (18.438)
$\sum_{s=1}^{S} \theta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	Advantaged students go to better schools AND learn more	-1.535 (19.666)	-24.258* (12.601)

Details of First Decomposition: Peru

VARIABLES*	endowments	interaction
Male	1.893	-1.515
	(1.868)	(1.914)
Age in months	11.20**	-7.520
	(5.111)	(5.632)
Height for age z-score (age 5)	8.770**	-3.365
	(3.976)	(4.549)
Wealth index (age 5)	-21.18	-22.38
	(22.64)	(32.31)
Rasch score PPVT (age 5)	10.22	-15.99
	(14.13)	(20.83)
Rasch score PPVT sq (age 5)	-9.401	16.29
	(13.61)	(21.23)
Rasch score on CDA cognitive test (age 5)	5.353	-46.32
	(22.51)	(36.11)
Rasch score on CDA cognitive test sq (age 5)	-5.636	48.64
	(20.50)	(34.67)
Mother's education: secondary plus	3.858	5.314
	(6.084)	(7.004)
Father's education: secondary plus	1.858	-2.272
	(4.392)	(5.254)
Number of siblings (age 5)	10.37	-23.66*
	(7.904)	(13.05)

Number of siblings (sq) (age 5)	-5.921	22.48*
	(5.693)	(13.58)
Caregiver's Cantril's subjective well-being ladder (child age 5)	-0.582	0.880
	(2.217)	(2.843)
Caregiver's educ. aspiration for child is university (child age 5)	-1.528	3.991
	(2.368)	(3.685)
Total background effect	14.580 (20.74)	-27.393
		(32.847)

Details of First Decomposition: Vietnam

VARIABLES*	endowments	interaction
Male	0.0598	0.0227
	(0.367)	(0.153)
Age in months	0.0526	-0.0395
	(0.378)	(0.592)
Height for age z-score (age 5)	0.195	2.339
	(1.570)	(2.498)
Rasch score on PPVT (age 5)	4.085*	6.971*
	(2.293)	(3.725)
Rasch score on CDA cognitive test (age 5)	3.466**	-2.384
	(1.435)	(1.859)
Number of siblings (age 5)	0.181	-0.438
	(0.387)	(0.905)
Mother's education: secondary plus	-0.183	5.504
	(3.689)	(4.557)
Father's education: secondary plus	2.175	0.843
	(3.140)	(4.148)
Parent's education aspiration for child (when child was 5): university	0.765	7.171*
	(1.874)	(4.310)
Caregiver's Cantril's subjective well-being ladder (child age 5)	3.367	-7.028
	(3.490)	(5.469)
Wealth index (age 5)	7.226	0.327
	(11.69)	(21.50)
Total Background effect	22.44** (11.60)	11.838 (20.641)

Table 4.5: Further Checks on Peru, First Decomposition

Component	Detail	School Sample (302 children)	Household Sample (593 children)	
		Wealth Split	Wealth Split	Ability Split
$(\bar{S}_{3,A} - \bar{S}_{3,DA})$	Difference	55.68*** (9.59)	7.13*** (1.18)	7.93 (1.12)***
$\beta_{DA}'(\bar{X}_A - \bar{X}_{DA})$	Background composition effect	14.58 (20.74)	4.15*** (1.15)	7.14 (2.11)***
$(\beta_A - \beta_{DA})'\bar{X}_{DA}$				
$(\beta_A - \beta_{DA})'(\bar{X}_A - \bar{X}_{DA})$	"Increased productivity" of advantaged person	-27.39 (32.85)	0.45 (1.49)	3.23 (2.8)
$\sum_{s=1}^{S} \delta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	School composition effect	10.22 (13.19)	1.59 (1.32)	-1.33 (1.4)
$\sum_{s=1}^{S} \theta_s \bar{D}_{s,DA}$	School "coeff" effect: advantaged students learn more	59.81** (25.92)	2.68* (1.54)	0.71 (2.5)
$\sum_{s=1}^{S} \theta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	Advantaged students go to better schools AND learn more	-1.54 (19.67)	-1.73 (1.65)	-1.83 (2.15)

Table 4.6: Further Checks on Vietnam, First Decomposition

Component	Detail	School Sample (932 children)	
		Wealth Split	**Ability Split**
$(\bar{S}_{3,A} - \bar{S}_{3,DA})$	Difference	35.59*** (6.24)	32.13*** (5.91)
$\beta_{DA}'(\bar{X}_A - \bar{X}_{DA})$	Background composition effect	22.44** (11.60)	14.29 (12.82)
$(\beta_A - \beta_{DA})'\bar{X}_{DA}$			
$(\beta_A - \beta_{DA})'(\bar{X}_A - \bar{X}_{DA})$	"Increased productivity" of advantaged person	11.84 (20.641)	20.87 (17.78)
$\sum_{s=1}^{S} \delta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	School composition effect	30.53 (11.13)***	-8.87 (7.39)
$\sum_{s=1}^{S} \theta_s \bar{D}_{s,DA}$	School "coeff" effect: advantaged students learn more	-4.96 (18.44)	-5.50 (14.05)
$\sum_{s=1}^{S} \theta_s(\bar{D}_{s,A} - \bar{D}_{s,DA})$	Advantaged students go to better schools AND learn more	-24.26* (12.60)	11.41 (9.70)

Table 5: Second Decomposition

Component	Detail	Peru	Vietnam
$(\overline{S}_{3,A} - \overline{S}_{3,DA})$	Difference	55.68*** (9.59)	35.59*** (6.24)
$\boldsymbol{\beta}_{AVG}'(\overline{\mathbf{X}}_A - \overline{\mathbf{X}}_{DA})$	Background composition	18.67 (13.27)	19.74*** (7.76)
$(\boldsymbol{\beta}_A - \boldsymbol{\beta}_{AVG})'\overline{\mathbf{X}}_A + (\boldsymbol{\beta}_{AVG} - \boldsymbol{\beta}_{DA})'\overline{\mathbf{X}}_{DA}$	Background coeff	-15.61** (7.87)	9.18 (7.21)
$\sum_{s=1}^{S} \delta_{s,AVG}(\overline{D}_{s,A} - \overline{D}_{s,DA})$	School composition	3.96 (6.92)	9.21** (4.44)
$\sum_{s=1}^{S} ((\delta_s + \theta_s) - \delta_{s,AVG})\overline{D}_{s,A} + \sum_{s=1}^{S} (\delta_{s,AVG} - \delta_{s,DA})\overline{D}_{s,DA}$	School coeff	48.66*** (13.39)	-2.54 (11.80)

Details of Second Decomposition: Peru

VARIABLES*	endowment	coefficient
Male	0.720	0.536
	(0.793)	(0.793)
Age in months	7.642***	0.393
	(2.826)	(1.688)
Height for age z-score (age 5)	6.138**	1.217
	(2.432)	(1.138)
Wealth index (age 5)	-11.23	-19.35**
	(12.87)	(8.383)
Rasch score PPVT (age 5)	-1.782	5.281
	(4.955)	(10.78)
Rasch score PPVT sq (age 5)	2.933	-5.485
	(5.535)	(11.07)
Rasch score on CDA cognitive test (age 5)	-8.883	-5.245
	(16.45)	(10.89)
Rasch score on CDA cognitive test sq (age 5)	9.425	5.395
	(15.34)	(11.19)
Mother's education: secondary plus	8.447***	-2.355
	(3.203)	(2.161)
Father's education: secondary plus	0.240	0.663
	(1.836)	(1.464)
Number of siblings (age 5)	0.00988	0.411
	(3.326)	(5.107)

Number of siblings (sq) (age 5)	1.613	1.920
	(3.372)	(4.604)
Caregiver's Cantril's subjective well-being ladder (child age 5)	-0.586	0.374
	(1.126)	(0.894)
Caregiver's educ. aspiration for child is university (child age 5)	-0.193	0.343
	(1.314)	(1.053)
Total background effect	18.672	-15.611**
	(13.272)	(7.865)

Details of Second Decomposition: Vietnam

VARIABLES*	endowment	coefficient
Male	0.0782	-0.00368
	(0.477)	(0.0792)
Age in months	0.0819	-0.0510
	(0.274)	(0.138)
Height for age z-score (age 5)	1.323	0.153
	(1.145)	(0.518)
Rasch score on PPVT (age 5)	6.968***	0.944
	(1.970)	(0.996)
Rasch score on CDA cognitive test (age 5)	2.170**	-0.0224
	(1.087)	(0.523)
Number of siblings (age 5)	0.0122	-0.0535
	(0.0966)	(0.462)
Mother's education: secondary plus	3.856*	-1.047
	(2.003)	(1.137)
Father's education: secondary plus	3.336*	-0.702
	(1.806)	(0.788)
Parent's education aspiration for child (when child was 5): university	2.066	2.620*
	(1.662)	(1.459)
Caregiver's Cantril's subjective well-being ladder (child age 5)	0.411	-0.889
	(2.761)	(1.081)
Wealth index (age 5)	-1.118	8.523
	(7.904)	(7.731)
Total background effect	19.7***(7.8)	9.18 (7.21)

Table 5.5: Further Checks on Peru, Second Decomposition

Component	Detail	School Sample (302 children)	Household Sample (593 children)	
		Wealth Split	Wealth Split	Ability Split
$(\bar{S}_{3,A} - \bar{S}_{3,DA})$	Difference	55.68*** (9.59)	7.13*** (1.1)	7.90*** (1.02)
$\boldsymbol{\beta}_{AVG}'(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_{DA})$	Background composition	18.67 (13.27)	4.09*** (0.81)	8.60*** (1.32)
$(\boldsymbol{\beta}_A - \boldsymbol{\beta}_{AVG})'\bar{\mathbf{X}}_A + (\boldsymbol{\beta}_{AVG} - \boldsymbol{\beta}_{DA})'\bar{\mathbf{X}}_{DA}$	Background coefficients	-15.61** (7.87)	0.28 (0.36)	0.20 (0.68)
$\sum_{s=1}^{S} \delta_{s,AVG}(\bar{D}_{s,A} - \bar{D}_{s,DA})$	Sorting (school composition)	3.96 (6.92)	0.16 (0.81)	-0.98 (0.85)
$\sum_{s=1}^{S} ((\delta_s + \theta_s) - \delta_{s,AVG})\bar{D}_{s,A} + \sum_{s=1}^{S} (\delta_{s,AVG} - \delta_{s,DA})\bar{D}_{s,DA}$	Differential learning within schools (school coeffs.)	48.66*** (13.39)	2.60 (0.96)***	0.10 (1.46)

Table 5.6: Further Checks on Vietnam, Second Decomposition

Component	Detail	School Sample (932 children)	
		Wealth Split	Ability Split
$(\bar{S}_{3,A} - \bar{S}_{3,DA})$	Difference	35.59*** (6.24)	32.13*** (5.9)
$\boldsymbol{\beta}_{AVG}'(\bar{\mathbf{X}}_A - \bar{\mathbf{X}}_{DA})$	Background composition	19.74*** (7.76)	28.96*** (7.01)
$(\boldsymbol{\beta}_A - \boldsymbol{\beta}_{AVG})'\bar{\mathbf{X}}_A + (\boldsymbol{\beta}_{AVG} - \boldsymbol{\beta}_{DA})'\bar{\mathbf{X}}_{DA}$	Background coefficients	9.18 (7.21)	-3.59 (4.02)
$\sum_{s=1}^{S} \delta_{s,AVG}(\bar{D}_{s,A} - \bar{D}_{s,DA})$	Sorting (school composition)	9.21** (4.44)	-5.38 (4.28)
$\sum_{s=1}^{S} ((\delta_s + \theta_s) - \delta_{s,AVG})\bar{D}_{s,A} + \sum_{s=1}^{S} (\delta_{s,AVG} - \delta_{s,DA})\bar{D}_{s,DA}$	Differential learning within schools (school coeffs.)	-2.54 (11.80)	12.14 (9.01)^

VI. Tentative Conclusions

1. Both decompositions using **school data** suggest that **almost all of the test score gap in Peru can be explained by differential learning within schools**, but in **Vietnam** we find **no such effect**.

2. Both decompositions using **household data** suggest that **35-40% of the test score gap in Peru is due to differential learning within schools**, while the rest is due to composition effects.

3. The above differential learning effect results are based on splitting the sample by **wealth**. When the sample is split by **predicted ability at age 5** then there are **no differential learning effects for Peru**.

4. In **Vietnam most of the gap** seems to be explained by the fact that advantaged Vietnamese children have better "endowments" of child and household characteristics (**composition effect**). The most robust endowment effect is that advantaged children start out with higher cognitive skills at age 5, even before they start school.

Questions for the Audience

1. Setting aside estimation issues, is this decomposition analysis useful?

2. Has it been done before?

3. What estimation issues are most worrisome?