

School effects on Chilean children's achievement growth in language and mathematics: an accelerated growth curve model

Lorena Ortega^{a*}

Lars-Erik Malmberg^a

Pam Sammons^a

^a *Department of Education, University of Oxford, Oxford, UK.*

(Original manuscript received 13 March 2017)

(First revision received 16 October 2017)

(Second revision received 2 December 2017)

(Third revision received 28 January 2018)

(Paper accepted on 16 February 2018)

Manuscript submitted to

School Effectiveness and School Improvement:

An International Journal of Research, Policy and Practice

Word count

8,205 words

Acknowledgements

The authors would like to thank the MIDE UC Assessment Centre at the Pontificia Universidad Católica de Chile and the Chilean Ministry of Education for granting access to the SEPA and SIMCE data, respectively. The authors are also grateful of the useful comments of the members of Educational Effectiveness Special Interest Group (EE-SIG) at the University of Tübingen on an earlier version of this article.

Funding

The completion of the article was supported by the National Commission of Scientific and Technological Research of Chile, through the project Centre of Advanced Studies on Educational Justice [CONICYT - PIA CIE 160007].

This work was supported by a PhD scholarship awarded to the first author by the Programme of Advance Human Capital, National Commission of Scientific and Technological Research of Chile [CONICYT - PFCHA 72100834].

* Corresponding author. Email: lcortega@uc.cl

School effects on Chilean children's achievement growth in language and mathematics: An accelerated growth curve model

Abstract

This study investigates school effects on primary school students' language and mathematics achievement trajectories in Chile, a context of particular interest given its large between-school variability in educational outcomes. The sample features an accelerated longitudinal design (three time-points, four cohorts) together spanning Grades 3 to 8 ($n = 19,704$ students in 156 schools). The magnitudes of school effects on students' growth trajectories were found to be sizeable (generally larger than school effects in Western industrialised countries) and moderately consistent across school subjects. School composition effects on student achievement status were found for both school subjects. However, there was no evidence of composition effects on student achievement growth. The study provides new evidence on the size and nature of school effects in a developing country context based on state-of-the-art methods (i.e., accelerated longitudinal and growth curve models).

Keywords

School effects, growth curve model, accelerated longitudinal design, composition effects, Chile

1. Introduction

The debate over the relative contribution of schools to educational outcomes gained momentum in the 1960's, when the influential Coleman Report underscored the little variance on student achievement explained by educational factors when compared to student background characteristics (Coleman et al., 1966). The field of Educational Effectiveness Research (EER) found its origins in studies that challenged this assertion by presenting evidence of significant and systematic differences between schools in their abilities to raise student achievement (i.e., Brookover, Beady, Flood, Schweitzer, & Wisenbaker, 1979 and Rutter, Maughan, Mortimore, & Ouston, 1979). Since then, the field has focused on disentangling and studying different aspects of the contribution of schools to students' outcomes, namely, school effects.

The term 'school effect' usually refers to the proportion of variance in student outcomes that is attributable to the school attended (Sammons, Davis, & Gray, 2016). Thus, school effects are an estimation of how much the school matters compared to other sources of variation on educational outcomes (such as student background characteristics). Small school effects indicate a weak contribution of school to variation on student outcomes and can be interpreted as homogeneity among schools (this could be due to, for example, similarities in terms of school resources and intake, shared minimum standards, similar educational processes in the educational system, among other factors).

The field Educational Effectiveness Research (EER) has used several different approaches to conceptualise and measure school effects (Sammons et al., 2016). When school effects on student cognitive outcomes have been estimated using "raw" student achievement adjusted for student background characteristics or gains in achievement from one time-point to the next, they have been found to be modest in size. For example, a review by Scheerens and Bosker (1997) estimated the between-school variance (the proportion of variance accounted for by differences between schools) to be around 10-15% when adjusted student achievement is used, although results are also known to vary according to the choice of outcome and control variables (Teddle & Reynolds, 2000). The context under examination is also an important aspect. When school effects in developing economies

are estimated using “raw” or adjusted achievement, the levels of between-school variance have been found to be higher (between 30-45%) than in Western industrialised countries (Hanushek, 1995; Riddell, 1997; Scheerens, 2001; Velez, Schiefelbein, & Valenzuela, 1993). The larger school effects found in developing countries have usually been interpreted as substantive differences across schools, in terms of their available resources and contexts, in those countries, which often feature more socially segregated and stratified educational systems.

EER has undergone significant methodological advances and developed new methods for estimating school effects, such as a multilevel longitudinal approach to estimate the effect of schools on students' achievement trajectories. This ‘dynamic perspective’ provides more reliable measures of students' progress and, consequently, increased power to detect educational effects (Guldemon & Bosker, 2009; Raudenbush, 1989). Indeed, substantive variation between schools in student achievement growth has been found when applying a three-level growth curve model where time-points are nested within students, which are in turn nested within schools. For example, using such a model, Bryk and Raudenbush (1989) found that 44-83% of the variance in reading and maths achievement growth was attributable to schools. This dynamic approach to the measurement of educational effects is an important component of the Dynamic Model of Educational Effectiveness, introduced by Creemers and Kyriakides (2008), a conceptual model that synthesises the vast and diverse research literature in the field, places the study of change (in contexts, processes, and student outcomes) at its heart, and highlights the multilevel nature of factors affecting student outcomes.

Based on the previously mentioned differences in the magnitude of school effects by context, one might argue that the effects of schools on student achievement growth in emerging economies are likely to be larger than those found in Western industrialised countries. However, most longitudinal research on school effects comes from industrialised economies, due to the availability of prospective data. In Latin America, very few studies have been conducted that allow for the estimation of school effects on student achievement growth. Minimum requirements of such studies are that student achievement is measured on at least three time points and that equated achievement scores are

calculated. As it has been stressed in reviews of the Latin-American literature, more longitudinal evidence is needed in order to understand the effect of schools on children's cognitive growth over time (Murillo, 2007).

The aim of this study is to analyse school effects on student achievement trajectories in Chile. Despite being among the highest-performing Latin-American countries in international assessments, Chile is also one of the systems with the highest within-country variability in outcomes in the region (Flotts et al., 2015). This situation warrants an in-depth investigation of the magnitude and sources of between-school variation in performance in the country.

This is the first study on the Chilean context that makes use of more than two points of annually collected longitudinal data on student achievement. To investigate the magnitude and consistency of school effects, as well as school composition effects, state-of-the-art analytical techniques are applied, namely, an accelerated longitudinal design and growth curve models.

2. Conceptual framework

2.1. From Value-Added Models (VAMs) to growth curve models

Early studies emphasised that schools (i.e., between-school differences) only accounted for a small proportion of variance over and above student background characteristics (e.g., Coleman et al., 1966). EER has since then focused on investigating systematic differences between schools and teachers in their abilities to raise student achievement (Creemers & Kyriakides, 2008). Several methodological approaches have been applied to estimate the effect of schools on student achievement.

Researchers in the field have often stressed that a school is not responsible for the absolute level of student achievement so much as for the progress made by pupils in its care and, therefore, a measure of the relative gain in achievement made by pupils is required in order to compare schools (Teddle & Reynolds, 2000). Thus, Value-Added Models (VAMs), based on student progress, are

often applied to measure the gain of being assigned to a given school (Raudenbush, 2004). These models require at least one lagged measure of the score representing a baseline, and the progress of students in each school is then compared jointly. As argued by Rowan, Correnti and Miller (2002) VAM approaches are likely to underestimate school effects on student achievement, due to the inclusion of test-level measurement error in the estimation of the student-level variance, which, in turn, inflates the total variance against which school-level variance proportions are calculated.

Growth curve models, an alternative approach facilitated by recent methodological advances and more readily available longitudinal data, permit capturing the dynamics of educational processes (Luyten & Sammons, 2010).). These models, that directly model the entire vector of student outcome scores jointly, are usually preferred for estimating educational effects, as they yield more precise estimates of achievement growth than VAMs (McCaffrey et al., 2003).

2.2. The magnitude of school effects

A meta-analysis of school effects showed that schools account for 19% of the variance in unadjusted student achievement levels, and 8% when controlling for initial differences between students (Scheerens & Bosker, 1997). Cross-nationally, the unadjusted between-school variance in mathematics achievement ranges from over 50% in countries with early school tracking to around 10% in Scandinavian countries (OECD, 2013).

The magnitude of school effects in longitudinal studies, where measurement error is controlled to a greater extent, tends to be larger than that of school effects estimated using cross-sectional or two time-points designs (Dumay, Coe, & Anumendem, 2014; Teddlie & Reynolds, 2000). Indeed, the application of three-level growth models has provided strong evidence of a substantive proportion of variation in student achievement growth that can be attributed to schools (ranging from 26 to 86% in outcomes related to language and mathematics) in a wide range of contexts (e.g., Guldmond & Bosker, 2009; Rowan et al., 2002; Zvoch & Stevens, 2006).

Previous reviews on the international evidence regarding school effects suggest that their

magnitude differs by academic subject, being larger in mathematics and science when compared to language (Reynolds et al., 2014). This is not surprising as the two former subjects are, to a greater extent, learned at school rather than in the family or the community.

Finally, it is important to mention that, when different school stages are compared, school effects tend to be larger in primary school than in secondary school" (Scheerens & Bosker, 1997).

2.3. The consistency of school effects

Another commonly investigated property of school effects is their consistency across academic subjects. A meta-analysis indicated that the correlations in unadjusted school effects across subjects are in the magnitude of 0.7, 0.8, or even 0.9, but tend to be considerably lower for school effects based on value-added models (Scheerens & Bosker, 1997). All in all, school effects based on value-added models show a moderate degree of consistency across subjects (Luyten, 1994; 1998; Marks, 2015; Sammons, Mortimore, & Thomas, 1996). For example, Luyten (2003) summarises previous studies on the consistency of school effects across language and mathematics, finding a median value of 0.43. This indicates that effective schools in one curricular area are not necessarily effective in other areas. However, longitudinal studies investigating the consistency of school effects on achievement growth across subjects are still scarce.

2.4. School composition effects

The impact of between-school differences in student intake on student achievement has been a popular research topic in EER. However, agreement has not been reached with regard to the nature and magnitude of school composition effects.

The school composition in terms of the socio-economic status (SES) of its student body has been shown to play a role in educational attainment in addition to that of student's individual SES (van Ewijk & Sleegers, 2010). Several studies have shown that low SES students perform worse in school than high SES students, both because they come from low SES families and also because they

tend to attend low SES schools (e.g., Caro & Lenkeit, 2012; Willms, 2010).

Other key compositional features examined in the literature are schools' aggregated levels of achievement and heterogeneity in terms of student performance. Most studies in this area have concluded that students benefit from attending a school with high average achievement levels (e.g., Opdenakker, van Damme, De Fraine, van Landeghem, & Onghena, 2002) while evidence on the effect of attending cognitively homogeneous schools is mixed and still inconclusive (Hanushek, Kain, Markman, & Rivkin, 2003; Rodríguez Menés & Donato, 2015).

The measurement of school composition poses significant methodological challenges, as the derived effects can be artifacts of poorly specified data at the individual level (Televantou et al., 2015). Furthermore, school composition research has often relied on cross-sectional data, which cannot disentangle the effects of school composition on achievement levels or status from the school composition effects related to achievement growth. The scarce longitudinal studies in this topic have shown that composition effects on achievement on a given measurement occasion differ from composition effects on learning growth, being considerably smaller for the latter (e.g., Belfi et al., 2014; Guldemon & Bosker, 2009).

2.5. School Effects in Chile

When comparing the evidence on school effects produced in different national contexts using “raw” or adjusted achievement, the levels of between-school variance in developing economies have been found to be considerably higher than in Western industrialised countries (Hanushek, 1995; Riddell, 1997; Scheerens, 2001; Velez, Schiefelbein, & Valenzuela, 1993; Willms & Somers, 2001). This phenomenon seems to reflect greater variability and influence of school resources, such as trained teachers, textbooks and materials in the former (Sammons et al., 2016).

It follows that the magnitude of school effects on student achievement growth in Latin American countries should also be larger than that found in Western industrialised economies. However, in this region the study of school effects has been hampered by the lack of suitable

longitudinal data. In Chile, most studies on school effectiveness have used cross-sectional student achievement data or, at most, two time-points data (e.g., Lara, Mizala, & Repetto, 2011; Troncoso, Pampaka, & Olsen, 2016; Willms & Somers, 2001). Furthermore, the existent research has mainly focused on the secondary school level and the later stages of primary education.

Previous studies on school effects in Chile using raw cross-sectional data have estimated a between-school variance ranging from 30 to 50% (e.g., Mizala, Romaguera, & Ostoic, 2004; OECD, 2013; Willms & Somers, 2001). When applying contextualised value-added models, school effects are found to be in the order of 6 to 20% (Muñoz-Chereau & Thomas, 2016, Troncoso et al., 2016). Also, in line with the international literature, Murillo and Roman (2011) found larger primary school effects in mathematics than in reading.

As mentioned above, Chile is among the highest-performing countries in Latin America and the Caribbean, but it is also one of the systems with the highest within-country variability in outcomes in the region (Flotts et al., 2015). The strength of the relationship between student performance and socio-economic status in the country is above the OECD average and one of the strongest in Latin America (OECD 2013; Treviño et al., 2015). In this context, where social background is a strong predictor of students' school destination and achievement status, it is relevant to investigate to what extent can schools ameliorate or exacerbate existing inequalities by affecting students' achievement growth.

This study analyses school effects on student achievement trajectories in Chile. The following research questions are addressed:

- (1) How large are the effects of schools on students' achievement status and growth over time?
- (2) How consistent are school effects across different academic subjects (i.e., language and mathematics)?
- (3) Are there school composition effects on students' achievement status and growth?

3. Sample and procedures

Several sets of data were linked to form a unique database of student and school records. The data used in these analyses derive from the SEPA¹ assessment programme developed by the MIDE UC Assessment Centre of the Pontificia Universidad Católica, as well as from the SIMCE² assessment system and the Student Enrolment Recording System (SERS), both maintained by the Chilean Ministry of Education. The SEPA assessment programme, the source of longitudinal student achievement data for this study, is a low-stakes assessment initiative designed to inform individual schools about their students' overall progress in comparison to that of students in other schools in the country. Thus, the SEPA project is not a school census nor is it a survey of randomly sampled schools. Instead, individual schools, or municipalities that administrate groups of public schools, voluntarily decide to participate in the project and have to pay for this service.

As Table A1 in the Appendix shows, the sample of the study is diverse, as it includes students living in urban and rural areas, attending public and private schools, and from a wide range of socio-economic backgrounds. However, students who are younger, come from more advantageous home environments and attend high-SES, private and large schools are slightly over-represented in the sample and those attending low-SES, public and private subsidised schools are slightly underrepresented.

The development of student achievement in mathematics and language is investigated in four cohorts. These cohorts were in Grade 3 (age 8), 4 (age 9), 5 (age 10) and 6 (age 11) of primary school in 2010. For each student there were three potential occasions of measurement (i.e., years 2010, 2011 and 2012). The measurements were taken at fixed regularly spaced occasions once every year towards the end of the school year³. Thus, the data feature a 3-year accelerated longitudinal design with four overlapping cohorts, permitting the study of grades 3 to 8. The grade levels in which these cohorts are each year are presented in Table 1, where cohorts are identified by Roman numerals.

[Insert Table 1 about here]

After performing multiple imputation, the language and mathematics samples became balanced, that is, the number of time-point observations is three for each of the students. The sample resulting from this linking strategy comprises 19,704 students in 156 schools assessed on three consecutive occasions. As shown in Table 2, the sample sizes vary by cohort, ranging from $n = 4,269$ for Cohort III to $n = 5,782$ for Cohort I. Descriptive statistics of the language and mathematics scores are summarised in the table.

[Insert Table 2 about here]

Figure 1 shows the mean at each grade for each cohort and its 95% confidence interval. These confidence intervals are rather small due, in part, to the large sample size for each cohort. Means are connected by lines to approximate the mean growth tendencies. This analysis provides a first indication of cohort differences. Adjacent cohort means are somewhat similar but, in most of the time points, confidence intervals of the cohorts do not overlap. Also, language and mathematics proficiency show non-linear upward growth. However, this evidence does not support an entirely developmental explanation because cohort and grade level are correlated and may interact. Thus, a more detailed analysis is needed.

[Insert Figure 1 about here]

3.1. Measures

The dependent variables of the study are the *language and mathematics test scores* obtained from the SEPA project. Both, the language and mathematics tests, consist of 35 multiple-choice items in Grade 3, 40 in Grades 4 to 7, and 50 in Grade 8. For each year and grade level considered, the language and mathematics achievement scales present satisfactory estimates of internal consistency (Cronbach's $\alpha > 0.85$). Scores were vertically and horizontally equated using Item Response Theory (IRT), which makes scores comparable across grade levels and cohorts.

The student-level control variables introduced into the models are described below. *Female* is a dichotomous variable that distinguishes boys (0) from girls (1). *Age* refers to student age, calculated

in years and months as in December of 2010 and cohort-mean centered. *SES* is a family socioeconomic status index obtained from a factor analysis of mother's education, father's education and family monthly income, and standardised to have a mean of zero and a standard deviation of unity. This index shows high internal consistency (Cronbach's $\alpha = 0.88$). Finally, *Number of books at home (books)*, a proxy variable for cultural capital and the value of scholarly culture, was reported by parents and categorised in five values (1 = None, 2 = Less than 10 books, 3 = between 10 and 50 books, 4= between 51 and 100 books and 5 = More than 100 books).

School-level predictors were included to depict composition effects. The school-level variables used were *Achievement Mean*, indicating school mean score on the SIMCE Assessment System test for the relevant subject, *Achievement SD*, referring to the within-school standard deviation in SIMCE test scores for the relevant subject, a measure of diversity in the levels of achievement of the student body, and *School SES*, a composite indicator created and calculated by the Chilean Ministry of Education⁴. Both, the student- and school-level variables were treated as time-invariant covariates. Descriptive statistics for these variables are shown in Table A1.

3.2. *Missing data*

The largest proportions of missing data were found in the student-level variables retrieved from the SIMCE Assessment System. *Family income* (19%), *mother's educational level* (19%) and *father's educational level* (22%), the three variables used for creating a *student socioeconomic status* (SES) indicator, had missing data, as it did the variable *number of books at home* (19%). The school-level variables, in turn, presented negligible proportions of missing data (below 1%).

Also, due to student and school attrition, as well as to the incorporation of new students and schools into the project each year, scores had a considerable proportion of missing values at each time point. For language test scores, the percentage of missing data was 31%, 44% and 45% in 2010, 2011 and 2012, respectively. Similarly, for mathematics test scores, the percentage of missing data was 33%, 42% and 42% in 2010, 2011 and 2012, respectively.

From the analysis of missing data mechanisms it was concluded that data were at least Missing At Random (MAR) (Little & Rubin, 2002). Thus, the results presented in this paper were obtained after performing Bayesian multiple imputation (MI) via Mplus (Muthén & Muthén, 2010). Based on recommendations by Rubin (1987) five imputed datasets were generated⁵. The missing data were imputed from an unrestricted two-level model and the hierarchical structure was accommodated by means of imputing data with test scores in wide format, students as level 1, and schools as level 2. The language and mathematics databases were linked together so the imputation of data in the language data set would benefit from information on mathematics test scores as auxiliary variables, and vice versa. Finally, all the results obtained from the five multiply imputed data sets were combined using Rubin's rules (1987).

It was decided to perform multiple imputation (MI) on all the variables in the models, including the dependent variables (i.e., language and mathematics test scores). The decision was based on well-established missing data treatment evidence that indicates that: (1) multiple imputation is an appropriate method under general MAR conditions that, compared to listwise deletion (LW) (i.e., complete-case analysis), makes better use of the observed information, increases robustness to non-ignorable missingness and improves estimation precision (Schafer & Graham, 2002); (2) the dependent variable should be included in the models used to impute independent variables, otherwise it would be tacitly assumed in imputation that there is no relationship between the independent and dependent variables and, when the imputed data are analysed, the estimated slope of the dependent variable on the independent variable would be biased toward zero (Allison, 2000; Von Hippel, 2007), and; (3) imputing outcome data is common practice and leads to correct inference when performed using multiple imputation (Little, 1992; Sullivan, Salter, Ryan, & Lee, 2015).

In order to check the reliability of the findings obtained using MI, the analyses were also run using two alternative approaches for dealing with missing data: (1) listwise deletion (LW) and (2) multiple imputation, then deletion (MID), an approach introduced by Von Hippel (2007) where all cases are used for imputation but, following imputation, imputed values on the dependent variables

are excluded from the analysis (this is, the dependent variables are used in the imputation model but kept as missing in the analyses). The results obtained under these two different approaches to missing data treatment for the most complex model fitted (Model 3) are presented in the Appendix (Table A2). The three approaches produce very similar estimates. The direction, magnitude and significance of the fixed effects are generally consistent across the different approaches, and, as shown by the overlap of the credible intervals for the variance components, the variances at Level 2 (student) and Level 3 (schools), which are used to calculate school effects, do not differ significantly from those obtained for Model 3 when using MI on all the variables (see Table 3)⁶.

3.3. *Analytical techniques*

Criteria for satisfactory inference in studies of educational effectiveness often include a longitudinal design with repeated measures on multiple cohorts (Goldstein, 1997). In this study, both aspects were combined through the use of accelerated growth curve models.

3.3.1. *An accelerated longitudinal design*

The study features an accelerated longitudinal design as it links adjacent segments of limited longitudinal data from different cohorts to create a common long-term developmental trend or growth curve. An advantage of this approach is that by studying several cohorts, rather than one, confidence in the generalizability of the results can be increased (Duncan & Duncan, 2012). However, this method has seldom been applied in EER.

An accelerated longitudinal design within a growth curve approach was implemented using Miyazaki and Raudenbush's convergence test (2000). Under this approach, the model fit of two models, namely, the *full* model, that includes the separate mean trajectories for all cohorts simultaneously, and the *reduced* model, that includes a common trajectory for all cohorts over grades in the study, are compared.

First, a cohort-based hierarchical model (the *full* model) was estimated, as shown in Equation 1. At the first level of the model (i), each person's observed development is conceived as a quadratic

function of grade level⁷ plus random error. At the second level of the model (i), the individual intercept and linear growth rate coefficients are assumed to vary as a function of cohort plus person-specific random effects. Thus, separate mean trajectories are estimated for each cohort. The specification for each individual i , and occasion t , is

$$Y_{ti} = \beta_{0i} + \beta_{1i}t_{ti} + \beta_{2i}t_{ti}^2 + e_{ti} \quad \text{Equation (1)}$$

$$\begin{aligned} \beta_{0i} &= \beta_0 + \sum_{c=1}^3 \beta_{0c}d_{ci} + u_{0i} \\ \beta_{1i} &= \beta_1 + \sum_{c=1}^3 \beta_{1c}d_{ci} + u_{1i} \\ \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} &\sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{pmatrix} \right] \\ e_{ti} &\sim N(0, \sigma_e^2), \end{aligned}$$

where Y_{ti} is the language/mathematics test score for individual i at occasion t . At the first level, student achievement is described by a linear and a quadratic function of time and the second level describes the variability in the individual growth curves. The fixed part of the model describes the average growth trajectory by means of three parameters: the average intercept (β_0), an average linear growth parameter (β_1) and an average quadratic growth parameter (β_2). The student specific residuals (u_{0i}, u_{1i}) are assumed to be generated by a bivariate normal distribution with average zero, variances σ_{u0}^2 and σ_{u1}^2 and covariance σ_{u0u1} . The error term at level 1 (e_{ti}) is the error at time t for the i th individual. These within person residuals are assumed to be mutually independent and normally distributed with mean zero and constant variance σ_e^2 .

The parameter β_{0i} denotes the achievement score of student i at time $t=0$. In this time-structured design, 'grade level' is the time metric chosen. The data are ordered by data collection occasion, with difference in age at the first measurement occasion introduced later as a student-level predictor. In the models, the time variable 'grade level' is scaled with the midpoint of the three grades

observed for each cohort coded as zero, this is, with the intercept as achievement status at the second measurement occasion. Also, $d_{ci} = 1$ if person i belongs to Cohort $(c+1)$ for $c = 1, 2, 3$; $d_{ci} = 0$ if not (Cohort I is the reference group).

Secondly, a common model for all cohorts (the *reduced* model) is estimated. The competing model does not include cohort effects. Instead, individual trajectories are viewed as varying randomly around a single mean-grade trajectory⁸. Finally, following Miyazaki and Raudenbush's convergence test (2000), the *full* and *reduced* models are compared. If the *full* model presents a better fit compared to the *reduced* model, then it cannot be assumed that all cohorts share the same mean trajectory.

3.3.2. Growth curve models

Growth curve models of language and mathematics achievement measurements (level 1) nested within students (level 2), nested, in turn, within schools (level 3), were estimated to characterise individual achievement growth trajectories and examine the effect of various socioeconomic, demographic and contextual variables on achievement growth. Equation 2 shows the specification of the baseline model that includes the cohort effects specified under the *full* model. The specification for each individual i , period t , and school j is

$$\begin{aligned}
 Y_{tij} &= \beta_{0ij} + \beta_{1ij}t_{tij} + \beta_2t_{tij}^2 + e_{tij} & \text{Equation (2)} \\
 \beta_{0ij} &= \beta_{00j} + u_{0ij} \\
 \beta_{1ij} &= \beta_{10j} + u_{1ij} \\
 \beta_{00j} &= \beta_{000} + u_{00j} \\
 \beta_{10j} &= \beta_{001} + u_{01j} \\
 \begin{pmatrix} u_{0ij} \\ u_{1ij} \end{pmatrix} &\sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u0}^2 & \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{pmatrix} \right] \\
 \begin{pmatrix} u_{00j} \\ u_{01j} \end{pmatrix} &\sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{u00}^2 & \\ \sigma_{u00u01} & \sigma_{u01}^2 \end{pmatrix} \right] \\
 e_{tij} &\sim N(0, \sigma_e^2).
 \end{aligned}$$

Student achievement is described using the same specification at Levels 1 and 2 as in Equation 1. However, at Level 3, the trajectory of each school is described by two parameters: a

school specific intercept (β_{00j}) and a school specific linear growth parameter (β_{10j}). The school residuals (u_{00j} , u_{01j}) are assumed to be generated by a bivariate normal distribution with average zero, variance σ_{u00}^2 , σ_{u01}^2 and covariance σ_{u00u01} .

Equation 3 includes differences in achievement status and growth related to student and school characteristic. The student variable *SES* at level 2 was grand-mean centered such that the school SES coefficient captures the composition effect directly (Enders & Tofighi, 2007). The structure of the random part of the model remains as in Equation 2. Equation (3)

$$\begin{aligned}
 Y_{tij} &= \beta_{0ij} + \beta_{1ij}t_{tij} + \beta_2t_{tij}^2 + e_{tij} \\
 \beta_{0ij} &= \beta_{00j} + \beta_{010}CohortII_{ij} + \beta_{020}CohortIII_{ij} + \beta_{030}CohortIV_{ij} + \beta_{040}Female_{ij} + \beta_{050}SES_{ij} \\
 &\quad + \beta_{060}Books_{ij} + \beta_{070}Age_{ij} + u_{0ij} \\
 \beta_{1ij} &= \beta_{10j} + \beta_{110}CohortII_{ij} + \beta_{120}CohortIII_{ij} + \beta_{130}CohortIV_{ij} + \beta_{140}Female_{ij} + \beta_{150}SES_{ij} \\
 &\quad + \beta_{160}Books_{ij} + \beta_{170}Age_{ij} + u_{1ij} \\
 \beta_{00j} &= \beta_{000} + \beta_{200}Achievement Mean_j + \beta_{300}Achievement SD_j + \beta_{400}School SES_j + u_{00j} \\
 \beta_{10j} &= \beta_{100} + \beta_{500}Achievement Mean_j + \beta_{600}Achievement SD_j + \beta_{700}School SES_j + u_{01j}
 \end{aligned}$$

The Intra-Class Correlation (ICC) is generally used as a measure of overall magnitude of school effects in EER. In this study, school effects are calculated as the proportion of variance in achievement scores (achievement status) and the proportion of variance in achievement growth that lies between schools, as recommended by Raudenbush and Bryk (2002). Thus, for Equation 3, the magnitude of school effects on student achievement status is defined as

$$\text{School ICC Achievement Status} = \frac{\sigma_{u00}^2}{\sigma_{u00}^2 + \sigma_{u0}^2},$$

and the magnitude of school effects on student achievement growth is defined as

$$\text{School ICC Achievement Growth} = \frac{\sigma_{u01}^2}{\sigma_{u01}^2 + \sigma_{u1}^2}.$$

This elicits the simultaneous comparison between school-to-school differences in achievement level and school-to-school differences in growth. Residuals from the model specified in

Equation 3 also permit the analysis of consistency of school effects by studying the correlation between school effects on language and mathematics.

Estimation was performed using Bayesian estimation via Markov chain Monte Carlo (MCMC) methods as implemented in the software MLwiN (Browne, 2012). The MLwiN software was operated via the Stata command *runmlwin* (Leckie & Charlton, 2013). The means and standard deviations of the sampled parameters from the monitoring period were used as parameter estimates and standard errors while the 2.5th and 97.5th percentiles of the MCMC chain provided Bayesian 95% credible intervals, analogous to 95% confidence intervals.

Comparative model fit was assessed using the Deviance Information Criterion (DIC), a 'badness' of fit indicator developed specifically for use on complex hierarchical models and Bayesian estimators that allows the comparison of non-nested models and penalises for model complexity. Lower values reflect superior models and differences in DIC values of more than 5 units between two models are regarded as strong evidence in favour of the model with the smaller DIC (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2012).

4. Results

The results of the estimation of the *full* model, that includes the separate mean trajectories for all cohorts simultaneously, and the *reduced* model, that includes a common trajectory for all cohorts over grades in the study, are presented in the Appendix, Table A3. The magnitudes of the reduction in deviance associated with the *full* model were sizeable ($\Delta DIC = -7,285$ in language, and $\Delta DIC = -6,281$ in mathematics), suggesting that the additional cohort parameters are helpful in accounting for the data. Thus, all inferences regarding correlates of change were adjusted for cohort membership in subsequent models.

Findings are presented as follows. First, the magnitude of school effects on student achievement status and growth are analysed, followed by the analysis of the consistency of these effects across subject and the school composition effects.

4.1. *The magnitude of school effects*

The multilevel model specified in Equation 2, which introduces school-level variation on student achievement status and growth, is analysed in order to differentiate this source of variance from that arising from the student level. This multilevel model consists of three levels: schools, students within schools, and time points within students. By means of this baseline growth model, the raw unadjusted differences between schools (school variance) can be estimated. Results are shown in Table 3, Model 2.

[Insert Table 3 about here]

Comparing the models based on Equation 1 (*full* model in Table A3) and 2 (Table 3, Model 2), it is possible to see that the addition of school random effects significantly improves the model fit, as the DIC, in both language ($\Delta DIC = -2,888$) and mathematics ($\Delta DIC = -4,950$), decreases. This suggests that achievement growth trajectories differ significantly between schools.

Results from the model specified in Equation 3 (Table 3, Model 3) show a substantive variation on achievement status across schools. More interestingly, schools also contribute to differences in growth rates (see random effects). Indeed, the variance components on achievement status and growth rates, both at the student and school level, remain sizeable after student and school variables were introduced. Most of the variance in growth appears to be at the school level (i.e., 80.0% and 94.4% for the linear component in language and mathematics, respectively). In both language and maths, the school effects on the linear growth rate are larger than the school effects on students' achievement status (intercept), which are 6.9% and 7.9%, respectively. In addition, school effects on achievement growth are larger for mathematics than for language.

[Insert Table 3 about here]

The correlations between the random intercepts (the average proficiency in Grade 3 in a school) and random slopes at the school level are positive and small in language and mathematics (.12 and .17, respectively), indicating that students at schools with high overall achievement levels do not

tend to progress at a faster rate than those in schools with lower achievement levels. No evidence of ceiling effects on student achievement growth was found that could explain this trend (see Ortega, 2016).

To illustrate the magnitude of school effects, the 156 level-3 residuals are plotted in Figure 2, one for each school in the data set. These caterpillar plots graph each residual, obtained from the model specified in Equation 3, against their rank order, accompanied by error bars corresponding to confidence intervals.

[Insert Figure 2 about here]

It is possible to distinguish several outliers at both ends of each graph. In language, the confidence intervals of these residuals do not overlap zero for a group of 59 and 76 schools at the lower and upper end of the intercept and slope residual plots, respectively. In mathematics, the number of outlier schools is 60 for the intercept and 84 for the slope residuals. This means that more than 37% of the schools differ significantly from the average school effects at the 5% level.

To learn more about these outlier schools, highly effective and highly ineffective schools, identified in terms of their effect on student achievement growth (slope residuals based on Model 3), were compared on the basis of Rurality (whether they were located in a rural area), Sector (whether they were public, private subsidised or private non-subsidised), Size (number of students in primary school) and Student-teacher hours ratio (number of students per classroom teacher hours). Descriptive statistics for these variables can be found in Table A4. Chi-square tests and t-tests were performed to compare highly effective and highly ineffective schools on these variables and it was found that the two groups of schools did not differ significantly in any of these structural characteristics, except for Sector in mathematics, where highly ineffective schools were more likely to be private subsidised schools and highly effective schools more likely to be either public or private non-subsidised schools ($\chi^2(1) = 4.04, p < .05$).

4.2. The Consistency of School Effects across Subjects

In order to assess whether schools have consistent effects on students' achievement trajectories across academic subjects, correlation coefficients between school residuals in language and mathematics are calculated for the 156 schools in our sample, based on the model specified in Equation 3. Results are shown in shown in Table 4.

[Insert Table 4 about here]

High consistency ($r = 0.61$) was found between school effects on language and school effects on mathematics achievement levels. The more moderate correlation between slope residuals ($r = 0.43$) suggests a lower degree of coherence on school effectiveness across academic subjects. All in all, schools that are effective in language are, to some extent, also effective in mathematics. However, some schools might be better at promoting achievement in one of the two subjects, as this association is by no means perfect. In addition, a small positive correlation was found between the random intercepts and slopes in mathematics ($r = 0.26$), indicating that growth in achievement was not necessarily larger in schools attended by students with higher achievement levels.

4.3. School Composition Effects

The model specified in Equation 3, which includes student- and school-level variables, was used to address the third research question regarding composition effects (see Table 3, Model 3). School variables found to be associated with achievement status were *Achievement Mean*, *Achievement SD* and *School SES*, although the latter only in language. Importantly, none of the school-level variables tested were substantially related to achievement growth. This confirms that students in schools with high and homogeneous achievement levels do not necessarily progress faster than students in schools with low and heterogeneous achievement levels.

5. Discussion

The present study investigated school effects on the achievement growth trajectories of primary school students in Chile. This was facilitated by the use of three-level longitudinal growth curve models in combination with an accelerated longitudinal design.

In the field of EER, the development of three-level growth models has elicited a debate on whether schools have a larger effect on student status (intercept) than on student growth (slope) (Anumendem, De Fraine, Onghena, & Van Damme, 2013). Previous research has usually found that schools have a larger impact on their students' growth than on their students' outcomes at a certain point in time. The evidence presented is in line with this trend. In both language and maths, the school effects on the linear growth rate (slope) are larger than the school effects on students' achievement status (intercept).

Results indicate that the cross-cohort performance of schools differed depending on whether the mean achievement status or growth of students was considered. Across the four cohorts studied, the relationship between the school overall achievement levels and schools' subsequent achievement progress is weak, suggesting that schools with high achievement levels are not necessarily more likely than schools with low achievement levels to promote high mean achievement growth.

The magnitudes of school effects on achievement growth in this study (80 to 94% of the variance attributed to schools) are on the higher end of the range found in studies using similar model specifications and outcomes, which goes from 26 to 86% (e.g., Bryk & Raudenbush, 1989; Guldmond & Bosker, 2009; Rowan et al., 2002; Zvoch & Stevens, 2006). This supports the evidence that school effects are larger in emerging economies when compared to Western industrialised countries (Scheerens, 2001; Willms & Somers, 2001).

Also, more substantial school effects were found in mathematics than in language. This result is in line with previous research suggesting that school effects are larger in subject areas that are

typically learned at school and where exposure is limited in the family and the community (Teddle & Reynolds, 2000; Thomas, Sammons, Mortimore, & Smees, 1997).

The findings of this study with regard to the moderate consistency of school effects across academic subjects are in agreement with previous research (Luyten, 1994; Marks, 2015; Sammons et al., 1996). This study also found that the association between school effects in language and school effects in mathematics is stronger when we consider school effects on achievement levels than when we look at school effects on achievement growth.

With regard to composition effects, school achievement levels and within-school variation in achievement were associated with achievement levels in both language and mathematics. In addition, it was found that school SES is positively related to the academic achievement status of students in language, even after controlling for family SES. However, these same factors were not statistically related to achievement growth rates. This accords with other longitudinal studies showing substantial composition effects on achievement at a given time point and no, or considerably smaller effects, on achievement growth (Belfi et al., 2014; Guldmond & Bosker, 2009).

This study presents two main limitations: the lack of representativeness of the sample for some important variables (see Table A1) and missing data. With regard to the former, the differences between the study's sample and the population are likely to be an artefact of both the way in which the SEPA project operates and Chile's highly socially stratified education system. As explained in Section 3, the SEPA project is not a school census nor is it a survey of randomly sampled schools. Instead, individual schools, or municipalities that administrate groups of public schools, voluntarily decide to participate in the project and have to pay for this service. This self-selection process may introduce bias and an equalising force across the sample, in the sense that those school and municipality administrators who are more confident about their schools' academic performance, have more sophisticated assessment practices in place, are less averse to external assessment and more motivated about improving their students' academic performance, and can fund the implementation of this assessment, are more likely to participate in the project. Students who come from higher SES

backgrounds and whose families show higher levels of cultural capital are, in turn, more likely to attend those schools.

Thus, students who are younger, come from more advantageous home environments and attend high-SES, private and large schools are slightly over-represented in the sample and those attending low-SES, public and private subsidised schools are slightly underrepresented. The implication of this for the study is a potential reduction in variance at the school level, which would mean that the estimates of school effects might appear lower than they actually are. Thus, the differences between the characteristics of the population and the composition of the study's sample demand caution with regard to the interpretation and generalisation of findings.

With regard to missing data, as explained in section 3.2, there were particularly large proportions of missing data on the dependent variables, due to attrition. This was the result of students changing schools as well as schools leaving (while others joined) the SEPA project during the period under study. As data were MAR, the issue of missing data was dealt with by implementing a suitable strategy for the data, namely, multilevel multiple imputation (MI) on all the variables in the models, including the dependent variables. Results obtained under this method were robust to those obtained with alternative approaches for dealing with missing data (i.e. LD and MID).

Despite these limitations, there are also noteworthy strengths in the study undertaken as it features a comprehensive sample, multiple measures of educational achievement, a longitudinal and multi-cohort design, a multilevel approach and appropriate missing data treatment.

The study addressed three themes in the measurement of school effects: magnitude, consistency and composition effects. Furthermore, the analysis of growth trajectories showed once again that school differences with respect to the growth of their students are bigger than usually demonstrated in studies with cross-sectional and two time points longitudinal designs. This study, thus, stresses the limitations of cross-sectional studies for estimating educational effects and highlights the need for more longitudinal studies in educational effectiveness research.

In general, both the student and school-level variables tested had significant effects on students' achievement status, but less so on their growth in achievement. The fact that a sizeable percentage of between-school variance remains after adjusting for both student background variables and school composition effects, indicates that malleable educational conditions (for example, school climate and teaching processes) rather than merely student selection factors are likely to account for this substantial between-school variance. As it has been suggested in the literature, when the analysis shifts from concern with students' achievement status to a concern with students' growth in achievement, social background and demographic characteristics, as well as school composition, become relatively insignificant predictors of academic development, and measures of instructional processes offer more explanatory power (Rowan et al., 2002).

More research on teacher effects and instructional processes at the classroom level could help to disentangle the large school effects observed in this context, as numerous international studies have found greater between-classroom than between-school variance in student achievement (Hill & Rowe, 1996; Luyten, 2003) and a large proportion of these classroom-level variance to be explained by what teachers do in the classroom (Muijs & Reynolds, 2011; Muijs et al., 2014). Furthermore, an extension of the present study was also able to confirm that teacher effects on students' language and mathematics achievement growth in Chile are large (54-66%) and exceed the remaining school effects (21% and 30%) (Ortega, Malmberg, & Sammons, 2017).

6. Conclusions

The large variation in educational achievement in Chile makes it a particularly interesting context for research on educational effectiveness. The present study investigated school effects in Chile with more advanced methods than in the past. Unlike this study that counted with three time points of data, previous research on the Chilean system has relied on cross-sectional or two time-points data that contain a very limited source of intra-individual variability and thus hamper the investigation of change in achievement. With three data points, the linear and curvilinear growth models can be evaluated and the precision of the parameter estimates improved when compared to

traditional VAMs. Furthermore, the linking of four overlapping cohorts of data in an accelerated longitudinal model adds robustness to the study of student achievement trajectories and school effects as grade and cohort effects can be disentangled.

Thus, this study addresses three important gaps in the literature as it, firstly, explores educational effects in the context of an emerging economy using a dynamic perspective, secondly, contributes further evidence on the properties of school effects on student achievement growth (i.e., magnitude, consistency and predictors) and, thirdly, advances the field methodologically by demonstrating the combined use of accelerated longitudinal designs and growth curve models.

The results presented provide a basis for future studies and comparisons on issues related to academic achievement in Chile and internationally. The evaluation of school effects based on the analysis of student attainment data is increasingly being used as a key strand of school inspection systems across the world. The present research has highlighted the importance of schools on student achievement growth, a finding consistent with previous international studies. In addition, this study has also confirmed that in Chile there is a strong relationship between student achievement levels, socio-economic background and school composition. Thus, assessment frameworks should include more sophisticated and dynamic approaches to evaluating school performance that can capture their contribution to achievement growth, net of the effects of socio-economic background, school composition and prior achievement levels. The results presented add to the on-going debate about fairer ways to judge the effectiveness of schools.

References

- Allison, P. D. (2000). Multiple Imputation for Missing Data: A Cautionary Tale. *Sociological Methods and Research*, 28, 301–309.
- Anumendem, N., De Fraine, B., Onghena, P., & Van Damme, J. (2013). The impact of coding time on the estimation of school effects. *Quality & Quantity*, 47(2), 1021–1040.
- Belfi, B., Goos, M., Pinxten, M., Verhaeghe, J. P., Gielen, S., De Fraine, B., & Van Damme, J. (2014). Inequality in language achievement growth? An investigation into the impact of pupil socio-ethnic background and school socio-ethnic composition. *British Educational Research Journal*, 40(5), 820–846.
- Brookover, W. B., Beady, C., Flood, P., Schweitzer, J. and Wisenbaker, J. (1979). *School social systems and student achievement: Schools can make a difference*. New York: Praeger.
- Browne, W. J. (2012). *MCMC estimation in MLWin*, v2.26. Bristol: Centre for Multilevel Modelling, University of Bristol.
- Bryk, A. S., & Raudenbush, S. W. (1989). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 159–204). San Diego, CA: Academic Press.
- Caro, D., & Lenkeit, J. (2012). An analytical approach to study educational inequalities: 10 hypothesis tests in PIRLS 2006. *International Journal of Research & Method in Education*, 35(1), 3–30.
- Coleman, J. S., Campbell, E. Q., Hobson, C. J., Mcpartland, J., Mood, A., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Government Printing Office.
- Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London: Routledge.
- Dumay, X., Coe, R., & Anumendem, D. N. (2014). Stability over time of different methods of estimating school performance. *School Effectiveness and School Improvement*, 25(1), 64–82.
- Duncan, S. C., & Duncan, T. E. (2012). Accelerated longitudinal designs. In B. Laursen, T. D. Little, & N. A. Card (Eds.) *Handbook of developmental research methods* (pp. 31–45). New York, NY: Guilford Publications.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological methods*, 12(2), 121–138.
- Flotts, P., Manzi, J., Jiménez, D., Abarzúa, A., Cayuman, C., García, M. J. (2015). Informe de resultados *Tercer Estudio Regional Comparativo y Explicativo: Logros de aprendizaje* [Results of the Third Regional Comparative and Explicative Study: Learning achievement]. Paris: UNESCO.
- Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement*, 8(4), 369–395.
- Guldmond, H., & Bosker, R. J. (2009). School effects on students' progress – a dynamic perspective. *School Effectiveness and School Improvement*, 20(2), 255–268.
- Hanushek, E. A. (1995). Interpreting recent research on schooling in developing countries. *World Bank Research Observer*, 10(2), 227–246.

- Hanushek, E. A., Kain, J. F., Markman, J. M., & Rivkin, S. G. (2003). Does peer ability affect student achievement? *Journal of Applied Econometrics*, 18(5), 527–544.
- Hill, P. W. and Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School Effectiveness and School Improvement*, 7(1), 1–34.
- Lara, B., Mizala, A., & Repetto, A. (2011). The effectiveness of private voucher education: Evidence from structural school switches. *Educational Evaluation and Policy Analysis*, 33(2), 119–137.
- Leckie, G., & Charlton, C. (2013). Runmlwin: A program to run the MLWin multilevel modelling software from within Stata. *Journal of statistical software*, 52(11), 1–40.
- Little, R. J. A. (1992). Regression with Missing X's: A Review. *Journal of the American Statistical Association*, 87(420):1227–1237.
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley.
- Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2012). *The BUGS book: A practical introduction to Bayesian analysis*. Boca Raton, FL: Chapman and Hall-CRC.
- Luyten, H. (1994). Stability of school effects in Dutch secondary education: The impact of variance across subjects and years. *International Journal of Educational Research*, 21(2), 197–216.
- Luyten, H. (1998). School effectiveness and student achievement, consistent across subjects? Evidence from Dutch elementary and secondary education. *Educational Research and Evaluation*, 4(4), 281–306.
- Luyten, H. (2003). The size of school effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement*, 14(1), 31–51.
- Luyten, H. & Sammons, P. (2010). Multilevel modelling. In B. P. M. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 246–276). London: Routledge.
- Marks, G. N. (2015). The size, stability, and consistency of school effects: Evidence from Victoria. *School Effectiveness and School Improvement*, 26(3), 397–414.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.
- Miyazaki, Y., & Raudenbush, S. W. (2000). Tests for linkage of multiple cohorts in an accelerated longitudinal design. *Psychological methods*, 5(1), 44–63.
- Mizala, A., Romaguera, P., & Ostoic, C. (2004). *A hierarchical model for studying equity and achievement in the Chilean school choice system*. Santiago: Universidad de Chile.
- Muijs, D., Kyriakides, L., van der Werf, G., Creemers, B., Timperley, H. and Earl, L. (2014). State of the art: Teacher effectiveness and professional learning. *School Effectiveness and School Improvement*, 25(2), 231–256.
- Muijs, D. & Reynolds, D. (2011). *Effective Teaching: Evidence and Practice* (3rd ed.). London: Sage.
- Muñoz-Chereau, B. & Thomas, S. M. (2016). Educational effectiveness in Chilean secondary education: Comparing different 'value added' approaches to evaluate schools. *Assessment in Education: Principles, Policy & Practice*, 23(1), 26–52.

- Murillo, F.J. (2007). School effectiveness research in Latin America. In T. Townsend (Ed.), *International handbook of school effectiveness and improvement* (pp. 75–92). New York, NY: Springer.
- Murillo, F., & Román, M. (2011). The school or the cradle? Evidence of their contribution to students' performance in Latin America. Multilevel study about the estimate of school effects. *Profesorado*, 15, 27–50.
- Muthén, L. K., & Muthén, B. O. (2010). *MPLUS user's guide*. Los Angeles, CA: Muthén & Muthén.
- OECD (2013). *PISA 2012 results: Excellence through equity: Giving every student the chance to succeed (Volume II)*. Paris: OECD Publishing.
- Opdenakker, M., van Damme, J., De Fraine, B., van Landeghem, G., & Onghena, P. (2002). The effect of schools and classes on mathematics achievement. *School Effectiveness and School Improvement*, 13(4), 399–427.
- Ortega, L. (2016). *Educational Effectiveness and Inequalities in Chile: A Multilevel Accelerated Longitudinal Study of Primary School Children's Achievement Trajectories*. (Unpublished doctoral dissertation). University of Oxford, Oxford, United Kingdom.
- Ortega, L., Malmberg, L., & Sammons, P. (2017). *Teacher effects on Chilean children's achievement growth: A cross-classified multiple membership accelerated growth curve model*. Manuscript submitted for publication.
- Raudenbush, S. W. (1989). The analysis of longitudinal, multilevel data. *International Journal of Educational Research*, 13, 721–740.
- Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121–129.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.
- Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): a state-of-the-art review, *School Effectiveness and School Improvement*, 25(2), 197–230.
- Riddell, A. (1997). Assessing designs for school effectiveness research and school improvement in developing countries. *Comparative Education Review*, 41, 178–204.
- Rodríguez Menés, J., & Donato, L. (2015). Social capital, social cohesion and cognitive attainment. In Y. Li (Ed.), *Handbook of research methods and applications in social capital* (pp. 324–343). Cheltenham: EdwardElgar.
- Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools. *Teachers College Record*, 104, 1525–1567.
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.
- Rutter, M., Maughan, B., Mortimore, P., & Ouston, J. (with Smith, A.). (1979). *Fifteen thousand hours: Secondary schools and their effects on children*. London: Open Books.
- Sammons, P., Davis, S., & Gray, J. (2016). The scientific properties of school effects. In C. Chapman, D. Muijs, D. Reynolds, P. Sammons, & C. Teddlie (Eds.), *The Routledge international*

- handbook of educational effectiveness and improvement. Research, policy, and practice* (pp. 25–76). New York: Routledge.
- Sammons, P., Mortimore, P., & Thomas, S. (1996). Do schools perform consistently across outcomes and areas? In D. Gray, D. Reynolds, C. Fitz-Gibbon & D. Jesson (Eds.), *Merging traditions: The future of research on school effectiveness and school improvement* (pp. 3–29). London: Cassell.
- Schafer, J.L. & Graham, J.W. (2002). Missing Data: Our View of the State of the Art. *Psychological Methods*, 7, 147–177.
- Scheerens, J. (2001). Monitoring school effectiveness in developing countries. *School Effectiveness and School Improvement*, 12(4), 359–384.
- Scheerens, J., & Bosker, R.J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon Press.
- Sullivan, T. R., Salter, A. B., Ryan, P. & Lee, K.J. (2015). Bias and precision of the "multiple imputation, then deletion" method for dealing with missing outcome data. *American Journal of Epidemiology*, 182(6), 528–34.
- Teddlie, C., & Reynolds, D. (Eds.) (2000). *The international handbook of school effectiveness research*, London: Falmer.
- Telavantou, I., Marsh, H., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement*, 26(1), 75–101.
- Thomas, S., Sammons, P., Mortimore, P., & Smees, R. (1997). Differential secondary school effectiveness: Comparing the performance of different pupil groups. *British Educational Research Journal*, 23(4), 451–469.
- Treviño, E., Fraser, P., Meyer, A., Morawietz, L., Inostroza, P., & Naranjo, E. (2015). *Informe de resultados Tercer Estudio Regional Comparativo y Explicativo: Factores asociados* [Results of the Third Regional Comparative and Explicative Study: Associated factors]. Paris: UNESCO.
- Troncoso, P., Pampaka, M., & Olsen, W. (2016). Beyond traditional school value-added models: A multilevel analysis of complex school effects in Chile. *School Effectiveness and School Improvement*, 27(3), 293–314.
- van Ewijk, R., & Sleegers, P. (2010). The effect of peer socioeconomic status on achievement: A meta-analysis. *Educational Research Review*, 5(2), 134–150.
- Velez, E., Schiefelbein, E. & Valenzuela, J. (1993). *Factors affecting achievement in primary education*. Washington: The World Bank.
- Von Hippel, P. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, 37, 83–117.
- Willms, J. D. (2010). School composition and contextual effects on student outcomes. *Teachers College Record*, 112(4), 1008–1037.
- Willms, J.D., & Somers, M.A. (2001). Family, classroom, and school effects on children's educational outcomes in Latin America, *School Effectiveness and School Improvement*, 12(4), 409–445.

Zvoch, K., & Stevens, J. J. (2006). Longitudinal effects of school context and practice on middle school mathematics achievement, *The Journal of Educational Research*, 99(6), 347–357.

Appendix

[Insert Appendix about here]

[Insert Table A1 about here]

[Insert Table A2 about here]

[Insert Table A3 about here]

[Insert Table A4 about here]

Table captions

Table 1: Cohorts in the sample using a three-year data window

Table 2: Descriptive statistics of language and mathematics achievement scores by grade level and cohort

Table 3: Models based on Equation 2 (Model 2) and Equation 3 (Model 3)

Table 4: Correlation between school residuals (from model based on Equation 3)

Figure captions

Figure 1: Language and mathematics test scores by grade level and cohort

Figure 2: School residuals for language and mathematics based on Model 3

Appendix captions

Table A1: Descriptive statistics for the sample and the population

Table A2: Models based on Equation 3 (Model 3) using listwise deletion (LD) and multiple imputation, then deletion of imputed missing values of the dependent variables (MID)

Table A3: Full model based on Equation 1 and Reduced model

Table A4: Comparison of outlier schools based on Model 3 school-level residuals on achievement growth (slope)

Notes

¹ Spanish acronym for 'Learning Progress Evaluation System'.

² Spanish acronym for 'Educational Quality Measuring System'.

³ Students who changed school during the period of the data collection were not followed to their new schools. Students who were retained once or more times during the period analysed were not included in the analyses. Analyses were carried out considering only those schools with 22 or more students in each of the years assessed.

⁴ The variables considered in this indicator are: Mother's level of education, Father's level of education, Average monthly household income and the Social risk index of the school. The final variable was categorised in five ordered levels, defined from 1 to 5, with level 1 being the lowest socio-economic level.

⁵ Rubin (1987) suggests that the efficiency of estimation based on a finite number of imputations, say μ , relative to one based on an infinite number is $(1 + \lambda/\mu)^{-1}$, where λ is the rate of missing information. Replacing μ with 5 and λ with .13 (a value that approximates the rate of unit-missing data in the sample) gives an estimate of .97 – high enough for the purposes of this analysis.

⁶ With regard to estimates for the fixed effects in Model 3, comparison of the results obtained with MI of all variables in Table 3 and those obtained with LW and MID in Table A2, show several important similarities and very few significant differences, at a significance level of .05. The direction of these fixed effects, their magnitude and significance are generally consistent across the three approaches. Some effects that were borderline significant under MI became borderline non-significant under LW. These were student SES effects on language and mathematics achievement status and school achievement SD on language achievement status. Other effects were borderline non-significant under MI and became borderline significant under MID and/or LW. These were the student gender effect on language achievement growth, student age and SES effects on mathematics achievement growth, and school SES on mathematics achievement status. The only coefficient for which the direction of the effect changed significantly was the estimate of the difference between the rates of growth of Cohorts I and II (i.e., Time x Cohort II) in mathematics, which changed direction and become significantly negative under MID and LW. In relation to the random effects, the biggest differences across the different missing data treatment approaches can be found on the variances at Level 1 (occasion), which tend to be smaller under MID and LW. However, this source of variance does not affect the calculation of school effects, the focus of this article. The credible intervals for the variances at Level 2 (student) and Level 3 (school), which are used to calculate school effects, do overlap under the three approaches.

⁷ Preliminary inspections of the data confirmed that the mean trajectory had a quadratic shape.

⁸ The *reduced* model is written as: $Y_{ti} = \beta_{0i} + \beta_{1i}t_{ti} + \beta_{2i}t_{ti}^2 + \beta_{3i}t_{cohort_{ti}} + \beta_{4i}t_{cohort_{ti}}^2 + e_{ti}$, where $t_{cohort_{ti}}$ is a linear component of the function of grade, centered (i.e., coding the zero-point in time) around the second observed grade for each cohort (i.e., 4 for Cohort I, 5 for Cohort II, 6 for Cohort III and 7 for Cohort IV) and $t_{cohort_{ti}}^2$ is a quadratic component of the function, grade centered around the second observed grade for each cohort. The inclusion of t_{ti} , $t_{cohort_{ti}}$ and their squares in the same model serves the purpose of ensuring that both the *full* and the *reduced* models have the same covariance structure. Collinearity is avoided by using the following specification at level 2 of the model: $\beta_{0i} = \beta_0 + u_{0i}$, $\beta_{1i} = \beta_1$, $\beta_{3i} = u_{3i}$, where u_{3i} is the random effect of person i on the rate of increase at the second observed point of the cohort. The random effect associated with t is restricted to zero. In addition, fixed effects for t_{ti} and t_{ti}^2 but not for $t_{cohort_{ti}}$ and $t_{cohort_{ti}}^2$ are estimated. For more details on the specification of the *reduced* model see Ortega (2016).