A two-fold indicator of school performance and the cost of ignoring it

Citation for published version:
Troncoso, P 2019, 'A two-fold indicator of school performance and the cost of ignoring it', *International Journal of Educational Research*, vol. 95, pp. 153-167. https://doi.org/10.1016/j.ijer.2019.02.002

Digital Object Identifier (DOI):
10.1016/j.ijer.2019.02.002

Link:
Link to publication record in Heriot-Watt Research Portal

Document Version:
Publisher's PDF, also known as Version of record

Published In:
International Journal of Educational Research

Publisher Rights Statement:
© 2019 Elsevier Ltd.

A two-fold indicator of school performance and the cost of ignoring it

Patricio Troncoso

Department of Social Statistics, The University of Manchester, Manchester, United Kingdom

Correspondence address: Oxford Road, Manchester, United Kingdom, M13 9PL

E-mail address: patricio.troncoso@manchester.ac.uk

Patricio Troncoso currently works as a Lecturer in Social Statistics at the University of Manchester. Since 2011, he has been researching on school effectiveness. His research is focused on the methodological challenges of analysing school value added on Mathematics and Language attainment using complex multilevel models as well as substantively focusing on the socioeconomic inequalities of the education system. His research interests are mainly related to the broad areas of Statistics and Education with an emphasis on Multilevel Modelling and educational inequalities.

Word count: 8,142 (excluding abstract, references, acknowledgements, tables and appendices)

A two-fold indicator of school performance and the cost of ignoring it

Traditional contextualised value-added models are useful tools for research and policy-making, but they overlook key assumptions necessary to fully understand academic performance. These models ignore significant sources of variation and fail to acknowledge the relationship between subjects. Using data from the Chilean Ministry of Education, this study fits a bivariate 5-level cross-classified model to assess the contribution of schools to pupils' progress in Mathematics and Language. It is concluded that statistical models for school accountability need to fully account the variation between classrooms, primary schools and local authorities (in addition to variation between pupils and secondary schools), as well as the relationship between Mathematics and Language. This is to avoid undue lenience and/or harshness towards seemingly over-performing and under-performing schools.

Keywords: school value-added; multivariate multilevel modelling; Chilean education system

1. Introduction and background

1.1.Assumptions underlying traditional school effectiveness models and their implications

Traditionally, school effectiveness models have been employed to assess the variation in pupils' scores in standardised tests coming mainly from two sources: pupils and schools. This is done with the objective of distinguishing between the effects of pupils' abilities and characteristics from the effects of the schools' characteristics. Ultimately, one of the most important aims of such models is to ascertain the impact of school policy on pupils' progress. School policy, then, is assumed to be related to all those effects that arise as unexplained variance in a model, after controlling for relevant non-malleable school and pupils' characteristics. This is achieved by implementing a standard two-level model, where pupils are nested within schools.

Recent research (Goldstein et al., 2007; Leckie, 2009; Rasbash et al., 2010; Troncoso et al., 2015; Troncoso, 2015) has found that the somewhat standard approach to analyse school effectiveness is incomplete insofar as there are assumptions underlying these models that are simply untenable for statistical and, more importantly, substantive reasons. The results of this research constitute valuable information for implementing improvements to current practice in Chile, as it makes a move towards a more transparent analysis of the substantive implications of the most frequently unquestioned assumptions in school effectiveness research.

Firstly, schools may put streaming in practice, as is the case of Chile (Torche, 2005), which makes differences between classrooms (within schools) highly likely. In the case of Chile, Troncoso et al. (2015) and Troncoso (2015) have demonstrated the significance of the classroom level for analysing external school accountability. This has also been demonstrated by various authors in other national contexts (See for example: Cervini, 2009a, 2009b; Martínez, 2012; Murillo & Román, 2011).

Secondly, assuming that the last school a pupil has attended at the time of the standardised tests is the only one affecting their academic performance is untenable, because pupils often have attended other schools besides that one. Moreover, in the case of standardised tests taken during secondary schooling, it is highly relevant to differentiate pupils according to the primary schools they attended in order to estimate the long-term carry-over effects that might still be affecting subsequent attainment. This second point has been deeply explored and its relevance demonstrated by various authors (Goldstein et al., 2007; Hill & Goldstein, 1998; Rasbash et al., 2010).

Thirdly, the standard two-level model approach to school effectiveness also fails to discern geographical effects. When no other higher-level (above the school level) is specified, it is assumed that there is no other source of variation beyond the differences between schools. In doing so, the differences between areas are disregarded as trivial and geographical inequalities that might affect school effectiveness become obscure and confounded with the school-level effects. In other words, if pupils in a particular school are more likely to obtain similar results due to having shared a school environment throughout the years, then schools in a particular geographical area are also more likely to have similar levels of effectiveness due to sharing a geographical setting as compared to schools in other areas. This school geographical clustering has been shown to be statistically significant and substantively relevant in the specific case of progress in Mathematics and Language in Chile (Troncoso et al., 2015; Troncoso, 2015) and in other international studies (Cervini, 2009a, 2009b; Leckie, 2009; Plewis, 2011; Rasbash et al., 2010).

1.2.The practical implications of the traditional school value-added models

The previously mentioned adjustments of the traditional approach to school effectiveness are also of the utmost importance from the perspective of diverse stakeholders. Adjusting for classroom effects, from the perspective of the school administration, allows focusing school policy interventions on particular aspects that may be affecting classes differentially. While, from the school choice point of view, as discussed in Troncoso et al. (2015) and Troncoso (2015), classroom effects are a sort of combination of pupil and school effects. Hence, when classroom effects are unspecified, school effects are obscured, making school choice a less than optimum decision. This is because parents and guardians are generally unable to choose the class to which they would wish their pupils to be assigned. Hence, an overall measure of school effectiveness adjusted for classroom effects is more precise with regard to within-school differences.

Adjusting for effects from previously attended schools has implications for school administration and school choice as well (Goldstein et al., 2007; Leckie, 2009). Having information of the primary schools from which pupils come allows schools, for instance, to pursue remedial actions for disadvantaged pupils. From the guardians' perspective, this has two main implications: a) school choice is based on school accountability information that is less obscure and distinct from carry-over effects from other schools; and b) an estimation of the effect of pupil's mobility can be derived.

Moreover, as Troncoso et al. and Troncoso (2015) discuss, adjusting for geographical effects is useful for school accountability purposes, as these adjustments allow policy-makers and government officials to have geographically contextualised information about the schools, which is necessary to focus the use of resources in a more efficient way. On another front, this information is essential for policy interventions at the level of local governments, simply because they are limited to intervene on a circumscribed territory. Additionally, geographically contextualised information about school effectiveness is crucial for school choice, since guardians are most often limited to choose only from those schools located in the area in which they reside.

On a different note, Page et al. (2016) discuss these assumptions further, asserting that additional adjustments to value-added models would entail different definitions of the "reference school". The authors argue that such approach is still not as informative as it could be, as it does not provide a complete picture of the distribution of value-added within schools. The authors used quantile regression to analyse Chilean school data, which unveils differential school effects that would not have been apparent should another method have been used.

All these implications and assumptions refer to the case in which the models for school performance are univariate in nature. This is another overlooked assumption, which is explained and debunked in the next section.

1.3.School value-added as a multidimensional phenomenon

Traditional school value-added models assume that the subjects under assessment are unrelated to each other. Even though a school accountability system may include information on several subjects, as is the case of the performance tables published by the England and Wales Department for Education (Ray, 2006), the underlying statistical models do not take the relationships between them into account. Assuming no relationship between subjects would imply two basic assumptions: a) at the pupil level, pupils split up their learning process into different unrelated areas of knowledge; and b) at the school level, schools break down the teaching process into different dissociated subjects.

These are untenable assumptions for three main reasons. While it is true that the curriculum is divided into several components, connections between subjects are not prevented but encouraged; this is what Creemers (1994) denotes as the "consistency principle". Furthermore, even in the case in which the connections between subjects were neglected in a school or even discouraged, the plausible existence of shared teaching practices would still have common effects across subjects at the aggregated school level. Scheerens and Bosker (1997) indeed found some evidence of consistent school effects across subjects, although school value-added models showed only moderate levels of consistency. Additionally, pupils are taught and learn knowingly and/or unknowingly a number of transferrable skills throughout their educational trajectories, which makes the relationships between subjects even more plausible. In this respect, Mathematics and Language do have interconnections that previous studies have indeed found and some examples are discussed in the paragraphs to follow.

Several researchers have found that there is a significant link between language development and mathematical attainment. On the one hand, there is research on how linguistic impairments can affect mathematical skills development during childhood (Cowan et al., 2005; Donlan et al., 2007; Simmons & Singleton, 2008); while, on the other hand, there is research on how generic and specific language skills can affect the development of mathematical skills of pupils in general (Hecht et al., 2001; Vukovic & Lesaux, 2013).

From the perspective of the linguistic impairments, Donlan et al. (2007) found that children with a specific language impairment performed significantly lower in a variety of mathematical tasks than the children in the control group. More specifically, these authors concluded that language impairment is an inhibitor factor when it comes to the development of mathematical skills, such as the acquisition of the number sequence, the development of calculation skills and the acquisition of the place-value principle; while this is not the case for learning arithmetic principles, which seems to be unaffected. Cowan et al. (2005) also found evidence that language comprehension affects mathematical skills, specifically tasks related to addition combinations and relative magnitude. Finally, Simmons and Singleton (2008), through meta-analysis, found that the difficulties endured by dyslexic pupils are not limited to language attainment, but are consistently related to mathematics performance as well.

From a wider perspective, other authors also found evidence of this underlying relationship between language abilities and mathematics performance of pupils in general. Hecht et al. (2001) reported that general verbal ability is persistently correlated with mathematics computation skills in pupils from Year 2 to Year 5 in the United

States. Along these lines, Vukovic and Lesaux (2013), using mediation analysis on a sample of Year 3 pupils, also found that verbal analogies had an indirect relationship with arithmetic knowledge via symbolic numeric skills, while phonological decoding was directly associated with arithmetic performance.

Sulis and Porcu (2014) provided further insight into the relationship between Mathematics and Language attainment at the level of schools and geographical areas. The authors found evidence that schools performing well in Mathematics also perform well in Language. One limitation of this study is that no control for prior attainment was specified, and hence, neither pupils' progress nor school effectiveness could be measured. This correlation between subjects at the school level seems to be pointing out that school performance is multifaceted.

Ortega et al. (2018) provided a brief account of the level of consistency of school effects across subjects in Chilean schools. The authors found a relatively high correlation between the predicted school effects (residuals) for Mathematics and Language, arising from univariate models. They conclude that schools that are effective in one subject are to some extent also effective in the other.

Educational researchers have found that language attainment is positively associated with Mathematics achievement (see for example: Tate, 1997; Wang & Goldschmidt, 1999) along with other socio-economic, educational and demographic characteristics at the level of schools and pupils, using models akin to the traditional 2-level approach. Even though the relationship between Mathematics and Language may seem obvious, the way in which this relationship is modelled is far from straightforward and a univariate multilevel model (or a set of them) cannot reliably account for this presumable correlation. Hypothetically speaking, if a model for performance in Mathematics is specified with a set of explanatory variables that include, for instance, socio-economic characteristics and language-related indicators, issues of multicollinearity and possibly endogeneity may arise. This is because such language proficiency indicators are also very likely to be associated with (or affected by) the socio-economic variables in the model.

As seen in other studies (Manzi et al., 2014; Mizala & Torche, 2012; Troncoso et al., 2015), progress in Mathematics and Spanish are both significantly associated with income, gender, year repetition, school institutional type and school SES, and hence in the event of specifying Spanish attainment as an additional predictor of Mathematics, spurious coefficients would be estimated, since attainment in the Spanish Language test could also be specified as a function of the other specified variables. A bivariate multilevel model is the best approach for taking into account this correlation without producing a model that suffers from unacceptable levels of multicollinearity or endogeneity.

1.4.School value-added models as a research tool for school accountability

One of the most important purposes of analysing the contribution of schools to the academic progress of pupils is school accountability. School accountability in its simplest form can be understood as the identification of the responsibility that lies with the schools with respect to their pupils' learning. The traditional approach to implement such notions conveys the definition of performance standards, school monitoring and inspection, pupils' achievement testing, as well as rewards and sanctions according to performance (Sahlberg, 2007, 2010). Furthermore, as in the case of England and Wales, the accountability system can also involve the publication of school performance tables for public scrutiny (Ray, 2006). This widely-known and traditional approach can be referred to as external school accountability (San Martín & Carrasco, 2013). Some authors (Au, 2007; Sahlberg, 2007, 2010) also refer to this approach as high-stakes testing, where the accent of the accountability system is on the consequences for schools.

Following Raudenbush's (2004) conceptualisation, as well as further extensions by Timmermans et al. (2011) and an application in the Chilean context by Troncoso et al. (2015), this paper implements a series of multilevel models to attempt to isolate the "true" school effects as reliably as possible by controlling for relevant non-malleable factors affecting pupils' progress in a realistically and sufficiently complex way, with the ultimate purpose of comparing schools fairly.

School accountability in Chile

Traditionally, government practice in Chile has employed school averages as indicators of effectiveness, although a new accountability system is being developed that takes into account some of the principles of value-added research (San Martín & Carrasco, 2012, 2013). This system developed by the Agencia de Calidad de la Educación (Agency for the Quality of Education), started a pilot stage in 2015; however, no results have yet been released to the public. This methodology for classification of the schools includes a series of indicators of school quality, with attainment being the most important amongst them (Agencia de Calidad de la Educación, 2014). This quality assurance system will hold schools accountable based on the achievement of their pupils in Years 2, 4, 6, 8 and 10 in standardised tests.

In this Quality Assurance System, schools will be classified into four categories: high, middle, lower middle and insufficient performance (Agencia de Calidad de la Educación, 2014). Two thirds of this classification will be dependent upon attainment scores averaged at the school level, which will be adjusted by a set of pupils' socio-economic and demographic characteristics to make allegedly fairer comparisons. Even though the specific details of the methodology have not been published yet, it has been announced that it will involve the implementation of multiple linear regression (MLR) models to make the necessary adjustments for the socio-economic and demographic characteristics of the pupils (Agencia de Calidad de la Educación, 2014). Although this is an improvement compared to current practice involving school averages only, the use of MLR models is indeed a major setback, given the overwhelming evidence in favour of the use of multilevel models. The advantages of using multilevel models over multiple linear regression models in the presence of clustering have been discussed by many authors (Goldstein, 2011; Snijders & Bosker, 2011). In this paper, the consequences of using a biased model to assess school performance are analysed in the results section.

The issues around the particular methods to be applied are not trivial. This quality assurance system can be said to have adopted a high stakes testing approach (Au, 2007; San Martín & Carrasco, 2013), where schools persistently judged to be performing insufficiently could face closure. It is, therefore, key to ensure a fair system, where schools are held accountable for what they can actually act upon.

As mentioned before, even though school accountability is highly relevant for many different reasons pertaining diverse stakeholders external to the schools, its purpose is also arguably applicable to the schools themselves. Sustained school improvement and/or maintenance of standards require permanent internal monitoring and assessment of pupils' attainment and progress. A comprehensive account as such should consider the way in which external and internal factors affect pupils' academic success, in order to distinguish between factors on which the schools can and cannot intervene. Nevertheless, an in-depth analysis of internal factors is beyond the scope of this paper.

2. Research questions and aims

This paper seeks to explore the relationship between progress made in the subjects of Mathematics and Spanish Language in pupils moving from primary to secondary schools in Chile, accounting for the influence of the wider shared environments in which pupils are inserted. As mentioned earlier, there are many plausible and known, but often ignored, contextual factors that affect pupils' progress, such as geographical factors, carry-over effects from previously attended educational settings, within-school streaming and the currently attended school. These environmental factors have been operationalised as the effects of local authorities, primary schools, classrooms and secondary schools.

School effectiveness research has traditionally treated academic subjects in a dissociated way. This is in spite of evidence supporting the idea of school effectiveness being a multidimensional phenomenon, as pupils neither learn subjects completely separated from each other, nor do schools teach subjects in such a way. Thus, neither in the learning nor in the teaching process do the connections between subjects need to be explicit or intentional. This is equivalent to assert that pupils who perform well in one subject can be expected (within a reasonable margin of error) to perform well in other subjects, after controlling for other relevant factors. Meanwhile, schools are reasonably expected to teach all subjects within the same standards (this is in line with the consistency principle, as mentioned in section 1.3); this implies that schools doing well in one subject would also be expected to be doing well in other subjects.

Given that traditional CVA models are allegedly inferior and that further adjustments are necessary for a more reliable estimation of school effects, it is worth investigating the usefulness of such extended (and more complex) models. School effectiveness models are recurrently used to inform the public about school performance for the main purposes of accountability and school choice via the construction of performance rankings. One of the aims of this paper is, therefore, to ascertain whether substantively relevant information can be extracted from the extension of the traditional models and whether this information is sufficiently different from what can be derived from simpler models. In sum, this paper attempts to address the following research questions:

(1) Is there a relationship between Mathematics and Spanish Language attainment in Chile at the disaggregated level of pupils and the aggregated levels of classrooms, primary schools, secondary schools and local authorities?

(2) Are there any differences between the effects of socio-economic and demographic characteristics of the pupils and the secondary schools on progress in Mathematics and Spanish Language in Chile?

(3) Are there any relevant (non-negligible) differences between school effectiveness accountability-oriented reports derived from univariate and multivariate models?

This paper builds up from existing research (Troncoso et al., 2015), to integrate the following elements in a single CVA model: a) assessing the effects of set of widely known influential socio-economic and demographic characteristics of the pupils and the schools; b) specifying the additional levels of classrooms and local authorities as significant sources of variation, in addition to the traditional 2-level models of pupils nested within schools; c) extending the CVA model further to assess the variation between primary schools; and d) analysing the correlation between different subjects, i.e. Mathematics and Language at various levels.

3. Data and Methods

3.1.Data and selected variables

The data on pupils' performance come from the SIMCE (Measurement System for the Quality of Education, for its acronym in Spanish: *Sistema de Medición de la Calidad de la Educación*) database, which has been provided by the Chilean Ministry of Education. SIMCE is a series of standardised tests with the purpose of measuring the level of achievement of the goals defined in the National Curriculum. The main subjects evaluated in these tests are 'Mathematics' and 'Spanish Language and Communication'. The database also holds historical data on pupils and schools' socio-economic and demographic information.

Table 1: Attainment variables used to implement the bivariate CVA model for progress in Mathematics and Language in Chile

Attainment variables (pupil-level)	**Description**
Attainment in Mathematics (Outcome)	Mathematics test scores obtained by Year 10 pupils in 2006. Standardised scores range from -2.54 to 2.61†
Attainment in Spanish Language (Outcome)	Spanish Language test scores obtained by Year 10 pupils in 2006. Standardised scores range from -2.67 to 3.78†
Prior attainment in Mathematics (Lagged outcome)	Mathematics test scores obtained by Year 8 pupils in 2004. Standardised scores range from -2.96 to 2.93†
Prior attainment in Spanish Language (Lagged outcome)	Spanish Language test scores obtained by Year 8 pupils in 2004. Standardised scores range from -3.36 to 2.62†

† Original scores are estimated by the Ministry of Education using a three parameter logistic Item-Response Theory (3PL-IRT) model; scores range approximately from 100 to 400, with a mean of 250 and a standard deviation of 50.

At the time of this research, the SIMCE tests were administered every year to all students in the 4th grade and all students either in the 8th grade or in the 10th grade. The particular cohort to be analysed in this research comprises those pupils who sat the SIMCE tests in 2006 when they were in the 10th grade (2nd year of secondary school, age 15-16) and had also previously sat the tests in 2004, when they were in the 8th grade (final year of primary school, age 13-14). This cohort was the only one moving from the last year of primary school to secondary school for which data were available at the time of this research. The total number of observations in this database is 202,605 pupils who belong to 7,461 classrooms within 2,438 secondary schools and 320 local authorities; furthermore secondary schools are cross-classified with 5,537 primary schools. Tables 1, 2 and 3 provide further details of the variables used in this research.

Table 2: Pupil-level variables used to implement the bivariate CVA model for progress in Mathematics and Language in Chile

Pupil-level Variables	**Description and categories**	**Frequency (%)**
Gender (categorical)	Gender of pupils: Female (0)	103,496 (51.08%)
	Male (1)	95,916 (47.34%)
	Missing (.)	3,193 (1.58%)
Socio-economic level indicated by household income (categorical)	Average monthly household income in Chilean pesos (CLP) as reported by parents‡:	
	Low income (200,000 CLP or less) – reference category	99,939 (49.33%)
	Lower-middle income (200,001 CLP - 500,000 CLP)	54,086 (26.70%)
	Upper-middle income (500,001 CLP - 1,000,000 CLP)	18,147 (8.96%)
	High income (more than 1,000,001 CLP)	13,908 (6.86%)
	Missing (.)	16,525 (8.16%)
Year repetition (categorical)	It indicates whether the pupil has been made to repeat any year in primary school:	
	Not repeated (0)	171,014 (84.41%)
	Repeated (1)	18,956 (9.36%)
	Missing (.)	12,635 (6.24%)

‡ The parents' data were gathered by schools and sent as part of the required data for government auditing purposes. This paper links the student, school and parental data.

Table 3: School-level variables used to implement the bivariate CVA model for progress in Mathematics and Language in Chile

School-level Variables	Description and categories	Pupil Frequency (%)	School Frequency (%)
School institutional type (categorical)	State-funded schools (1) – reference category	83,127 (41.03%)	673 (27.55%)
	Subsidised independent schools (2)	103,819 (51.24%)	1,393 (57.02%)
	Independent schools (3)	15,659 (7.73%)	377 (15.43%)
	Missing (.)	0	0
School SES (categorical)	School Socio-economic status according to the classification of the Chilean Ministry of Education:		
	Low School SES (1) – reference category	36,035 (17.79%)	479 (19.61%)
	Lower-middle SES (2)	79,459 (39.22%)	665 (27.22%)
	Middle SES (3)	50,505 (24.93%)	584 (23.91%)
	Upper-middle SES (4)	21,942 (10.83%)	383 (15.68%)
	High SES (5)	14,664 (7.24%)	332 (13.59%)
	Missing (.)	0	0

3.2.Dealing with missing data

Needless to say, handling missing data appropriately is very important in statistical analysis; however, this in itself is a complicated matter. For the case of the models in this paper, the treatment of the missingness is listwise deletion. This is done for reasons of parsimony and convenience as in other previous studies in the topic (see for example: Leckie, 2009; Rasbash et al., 2010). A multilevel multiple imputation procedure was performed on a sample of the full dataset as a validity check for the complete case analysis, using the software package Realcom-Impute (Carpenter et al., 2011). More details of this procedure are available on request.

3.3.Model building

The modelling strategy is an adaptation of Hox's (2010) bottom-up approach, where the contributions to the overall fit of the specification of a multivariate model and additional levels of variation are included. All models were estimated with the software MLwiN (Rasbash et al., 2012) via the user-written Stata (StataCorp, 2011) module "runmlwin" (Leckie & Charlton, 2013). Models were fitted firstly by using the Iterative Generalised

Least Squares (IGLS) algorithm and the estimated coefficients were used as starting values for the Markov Chain Monte Carlo (MCMC) estimation.

In the model exploration stage, MCMC estimation was carried out with the Gibbs sampler using diffuse priors, a burn-in period of 500 iterations, a monitoring chain length of 5,000, storing all iterations (thinning equals to 1). Also, for the purpose of assessing the covariance between subjects, IGLS estimation was used for simplicity. For the full model in this paper, the monitoring chain length was increased to 215,000, storing all iterations. The overall model fit was assessed via the Deviance Information Criterion (DIC) (Spiegelhalter et al., 2002).

Following the general multilevel notation as described in Goldstein (2011) and the classification notation described in Browne et al. (2001) notation, the CVA model to be fitted in this study has the general form shown in the appendix section.

4. Results

4.1. The relationship between progress and school value-added in Mathematics and Language

In this section, the main aim is to ascertain whether there is an underlying relationship between Mathematics and Language. As mentioned before, although this may seem obvious, the way in which this relationship has been accounted for is not straightforward. Exploratory analyses show that the correlation between both subjects is sizable at the pupil and school level. This relationship could be considered evidence in favour of the multidimensionality of school value-added and pupils' progress, manifested through attainment in various subjects, in this case Mathematics and Language. To test the significance of the correlation between the two subjects, two bivariate multilevel models are fitted and compared: a) a baseline 2-level model with two outcomes whose correlations at both levels were constrained to be zero, and b) an unconstrained bivariate 2-level model with a full covariance matrix.

From a statistical point of view, another important advantage of using the multivariate specification is that the estimation of a joint model uses the information more efficiently (Goldstein, 2011). For instance, should there be pupils missing one test score but not the other, the covariance matrix could still be estimated efficiently using the available information on the other test score. In contrast, when specifying two separate models, missing information on one test score is simply disregarded. Furthermore, the multivariate specification allows testing hypotheses across outcomes; for instance, it is possible to test the significance of any explanatory variable on both outcomes jointly or separately, which can be of substantive interest for unveiling differential effects.

Table 4: Model comparison between constrained multilevel model with two uncorrelated outcomes and unconstrained bivariate multilevel model

	Parameters†	**Constrained model**	**Unconstrained model**
	Fixed Part	**Coef. (s.e.)**	**Coef. (s.e.)**
	Intercept Maths	0.018 (0.015)	0.018 (0.015)
	Intercept Language	0.027 (0.013)	0.024 (0.013)
Level	**Random Part**	**Coef.**	**Coef.**
Secondary Schools	Variance (Intercept Maths)	0.51	0.509
	Covariance (Language, Maths)	--	0.433
	Variance (Intercept Language)	0.392	0.392
Pupils	Variance (Intercept Maths)	0.569	0.569
	Covariance (Language, Maths)	--	0.347
	Variance (Intercept Language)	0.67	0.67
Model fit information	**-2*loglikelihood**	970233.877	888795.359
	AIC	970245.877	888811.359
	Number of parameters (k)	6	8
	Chi-square (X^2)	--	81438.518
	p(X^2)	--	<0.001
	N	202,605	202,605

† Parameters were obtained via IGLS estimation.

As observed in Table 4, the specification of an unconstrained bivariate multilevel model represents a significant improvement to the overall fit of the model

($p<0.001$), which implies that the Mathematics and Language test scores are significantly correlated. It is also appreciated that both sets of effects -fixed and random- remain mostly unchanged with this specification.

The comparison of these two models not only allows unveiling the statistical relevance of estimating a bivariate model, but also reveals that academic performance is not a unidimensional phenomenon, which is more important from a substantive point of view. This implies that underlying the observed test scores in Mathematics and Language; there exist other mechanisms and relationships that influence academic performance. In other words, subjects are not learned by pupils in an isolated way and subjects are not taught (intentionally or unintentionally) in an isolated way either. Furthermore, had the data been available, other relationships between different subjects could have been found.

Using the covariance and variance terms of the unconstrained model, it is possible to estimate the correlation of the Mathematics and Language scores at the level of secondary schools and pupils. At the secondary school level, it is found that this correlation is 0.969, i.e. 0.433/sqrt(0.509*0.392), which means for instance that a secondary school achieving high scores in Mathematics on average would also be expected to score high on average in the Language test. In contrast, at the pupil level this relationship is not so clear, but also fairly high (0.562). Having demonstrated that the specification of a bivariate multilevel model is indeed fruitful, this empty model is extended further with additional levels of variation to test whether the bivariate multilevel model benefits from increasing its complexity.

4.2.Decomposing the variation in Mathematics and Language test scores

The first step in analysing progress in Mathematics and Spanish Language jointly, is to implement the empty models to assess the main significant sources of variation in the data in the same way as it would be done for the univariate case. The values of the DIC of the models without cross-classification (2, 3 and 4 levels) were compared to the value of the DIC of the model with cross-classification (5 levels). In such comparison, a lower value of the DIC diagnostic for the more complex models in comparison to the less complex models, indicates a better fit (Browne, 2012). In Table 5, the results for the 2, 3, 4 and 5-level empty bivariate models are presented.

Table 5: Summary of the empty bivariate models for attainment in Mathematics and Language

	Parameter†	**2 levels**	**3 levels**	**4 levels**	**5 levels**
	Fixed Part	**Post. mean (s.d.)**	**Post. mean (s.d.)**	**Post. mean (s.d.)**	**Post. mean (s.d.)**
	Intercept Maths	0.012 (0.015)	0.002 (0.015)	-0.137 (0.023)	-0.189 (0.029)
	Intercept Language	0.018 (0.013)	0.007 (0.013)	-0.122 (0.021)	-0.172 (0.027)
Level	**Random Part**	**Post. mean**	**Post. mean**	**Post. mean**	**Post. mean**
Secondary Schools	Var. (Intercept Maths)	0.51	0.47	0.396	0.366
	Cov. (Language, Maths)	0.433	0.402	0.334	0.307
	Var (Intercept Language)	0.393	0.362	0.3	0.275
Pupils	Var. (Intercept Maths)	0.569	0.502	0.502	0.492
	Cov. (Language, Maths)	0.347	0.292	0.292	0.284
	Var. (Intercept Language)	0.67	0.615	0.616	0.606
Classes	Var. (Intercept Maths)	--	0.095	0.095	0.093
	Cov. (Language, Maths)	--	0.077	0.077	0.076
	Var. (Intercept Language)	--	0.076	0.077	0.075
Local Authorities	Var. (Intercept Maths)	--	--	0.069	0.073
	Cov. (Language, Maths)	--	--	0.063	0.067
	Var. (Intercept Language)	--	--	0.058	0.062
Primary Schools	Var. (Intercept Maths)	--	--	--	0.015
	Cov. (Language, Maths)	--	--	--	0.013
	Var. (Intercept Language)	--	--	--	0.015
Model fit	DIC	882,639.635	857,769.5	857,765.4	855,367.7
	pD	5,048.381	11,476.94	11,470.71	14,617.81
	N	202,605	202,605	202,605	202,605

† Obtained via MCMC, using IGLS estimates as starting values. Chain length: 5,000; burn-in: 500, storing all iterations.

Results from the empty bivariate models show that the addition of every level of variation is a significant contribution to the overall fit of the model. The DIC diagnostic measure is drastically reduced with the specification of the class level (difference between DIC's is 24,870.16), while this is not as pronounced for the case of the addition of the local authority level (difference is 4.062), but still a relevant improvement given that the effective number of parameters (pD) is also reduced because of this specification. These results are also consistent with what is found for the 2, 3 and 4-level models, where values of the AIC drop with every additional level specified. Finally, the addition of the level of primary schools, given its cross-classified structure, increases the overall complexity of the model with a larger number of effective parameters; however, the DIC indicates that this is a significant trade-off, with a large reduction of the deviance and hence a better fit.

The variance estimates at the school and pupil level differ quite considerably across the models, where the between-school variance in Spanish is much lower than for the case of Mathematics, which would be indicating that performance in Spanish is more associated with pupil characteristics and abilities than performance in Mathematics.

On another front, it can be easily noted that the school effects (school-level variance) are clearly overestimated in a basic 2-level model (pupils nested within secondary schools). In both subjects, the overestimation of the school effects ranges between 28% and 30%, when comparing the variance estimates of the 2-level model with the 5-level cross-classified model. In Table 6, the variance partitioning (VPC) is presented in detail.

According to the VPC, the variance at the higher levels is consistently larger across all the empty models for the case of the Mathematics test in comparison to the results in Spanish. Conversely, the variance due to the pupil level is higher in the Spanish test, which may be indicating that attainment in Language is more associated with pupil characteristics and abilities than attainment in Mathematics. It can also be observed that the carry-over effects from primary schools are comparatively small (although significant) with respect to the effects from secondary schools. However, amongst the four empty models, the most important (largest) source of variation is always the pupil level. These results are consistent with what has been found in previous Chilean-based research (Troncoso et al., 2015).

Since the specification of additional levels to the basic model structure of pupils nested within schools has proved to be fruitful, the upcoming bivariate analyses are fitted accordingly. On another front, these empty bivariate models cannot be considered as value-added models in the rightful sense, since they do not control for any measure of prior attainment. The next model to be fitted is the raw value-added bivariate model, where prior attainment in both subjects is specified as the only explanatory variable.

Table 6: Variance partition coefficients of the empty bivariate models for attainment in Mathematics and Language

Level	Test	2 levels	3 levels	4 levels	5 levels
Secondary Schools	Mathematics	47.27%	44.05%	37.29%	35.23%
	Language	36.97%	34.38%	28.54%	26.62%
Pupils	Mathematics	52.73%	47.05%	47.27%	47.35%
	Language	63.03%	58.40%	58.61%	58.66%
Classes	Mathematics	--	8.90%	8.95%	8.95%
	Language	--	7.22%	7.33%	7.26%
Local Authorities	Mathematics	--	--	6.50%	7.03%
	Language	--	--	5.52%	6.00%
Primary Schools	Mathematics	--	--	--	1.44%
	Language	--	--	--	1.45%
Total	Mathematics	**100%**	**100%**	**100%**	**100%**
	Language	**100%**	**100%**	**100%**	**100%**

4.3.A bivariate CVA model for progress in Mathematics and Language

After determining the relevance of the bivariate specification and the size of the variance components, the analytical steps included: a) specifying a type AA (Timmermans et al., 2011) or raw value-added model, in which only prior attainment in both subjects is specified as covariates; b) specifying a type A value-added model (Timmermans et al., 2011) with pupil-level covariates only; c) specifying a type B value-added model (Timmermans et al., 2011) with pupil-level and school-level covariates. Steps a), b) and c) are the foundations for the full model, which includes all the previous specifications, as well as random coefficients for prior attainment at the level of secondary schools and cross-level interaction effects between prior attainment and school-level characteristics. Results from these steps are available on request.

Detailed results of the fixed part of the bivariate CVA model are displayed in Table 7, while the random part is displayed in appendix B. Given that significant interaction effects are found in the fixed part, the main effects are not to be interpreted in their own right, but in combination with the variables with which they are interacted (Gelman & Hill, 2007). Some noticeable differences between progress in Mathematics and Language are evident from the fixed effects of the bivariate CVA model.

First of all, the gender gap manifests itself differently in both outcomes. Overall, female pupils are expected to make more progress than males in Language; however, this relationship is moderated by prior attainment and year repetition. With regard to year repetition, boys who have been made to repeat are better off than female pupils, but worse off when considering prior attainment in Language. In Mathematics, on the other hand, boys are expected to score higher overall, but this effect is moderated by income, prior attainment and year repetition. Boys are expected to make slightly less progress in Language than girls when considering their prior attainment and slightly less progress in Mathematics when they come from middle-income households. Nevertheless, male pupils who have been made to repeat at least one year in primary school are much better off than females, in terms of their progress in Mathematics. The harmful effect of year repetition is found in both subjects and is found to be larger for girls than boys; however, it is less pronounced for progress in Language.

Another noticeable difference between both subjects in the fixed part of the model is that the effect of gender on Language does not vary across income. This is clearly distinct from what is recorded in the case of Mathematics, where the effects of prior attainment and gender both vary significantly across income groups. Male pupils in non-low income households make significantly less progress in Mathematics than female pupils and pupils in low-income households. These findings are consistent with Radovic (2018), who carries out a comprehensive analysis of the gender gap in Mathematics attainment in Chile.

With regard to the estimated fixed-effects of the school-level variables, it is first observed that subsidised independent schools have a significantly higher effect on pupils' performance in Mathematics and Language than State-funded schools (the reference category). Meanwhile, the effect of attending an independent school is not significantly different from zero, which implies that pupils in independent schools do not progress more than pupils in State-funded schools. This can be considered as evidence that most of the gap between low-achieving and high-achieving pupils is accounted for by socio-economic disparities. In other words, although pupils in Independent schools are high achievers, their progress in Mathematics and Language from primary to secondary schools is only as expected given their socio-economic advantages in comparison to their socio-economically disadvantaged peers attending State-funded schools.

On the other hand, as expected school socio-economic status has a large effect on Mathematics and Language scores. All non-low SES schools record, on average, significantly higher scores in both outcomes than low-SES schools. Although school SES is an ordinal variable, and hence, a linear effect cannot be estimated, an upwards trend is clearly appreciated. This trend shows that each higher school SES category has a larger difference with respect to the category of low-SES schools in both outcomes. Furthermore, none of the 95% credible intervals for the estimated coefficients of each school SES category overlaps with any other, which implies that there is a clear distinction between schools according to their average socio-economic status.

Nevertheless, both school-level variables have been found to interact significantly with prior attainment. Caution is, therefore, needed in interpreting the school-level main effects only. Since the random effects of prior attainment and gender proved to be significant improvements to the overall fit of the model in previous steps, it was deemed plausible that these effects also varied across school characteristics. Hence, cross-level interaction effects were also fitted. The full model includes the interaction between prior attainment and school SES, as well as school type. The cross-level interactions between gender and school characteristics did not yield statistically significant results.

Table 7: Fixed-effects parameters of the bivariate CVA model for progress in Mathematics and Language.

Fixed effects Mathematics†					**Fixed effects Language†**				
Pupil-level main effects‡	**Post. mean**	**s.d.**	**95% C.I.**		**Pupil-level main effects‡**	**Post. mean**	**s.d.**	**95% C.I.**	
Intercept	-0.342	0.016	-0.372	-0.312	Intercept	-0.234	0.012	-0.257	-0.210
Prior attainment	0.551	0.006	0.540	0.563	Prior attainment	0.615	0.006	0.604	0.626
Male	0.079	0.005	0.070	0.088	Male	-0.033	0.005	-0.042	-0.023
Low-mid income	0.016	0.004	0.008	0.025	Low-mid income	0.032	0.005	0.022	0.041
Up-mid income	0.032	0.007	0.018	0.046	Up-mid income	0.067	0.008	0.051	0.082
High income	0.044	0.010	0.025	0.062	High income	0.065	0.011	0.044	0.086
Held back	-0.269	0.008	-0.284	-0.254	Held back	-0.220	0.008	-0.236	-0.203
School-level main effects§	**Post. mean**	**s.d.**	**95% C.I.**		**School-level main effects§**	**Post. mean**	**s.d.**	**95% C.I.**	
Subs Indep. school	0.059	0.016	0.029	0.090	Subs Indep. school	0.049	0.012	0.026	0.072
Independent school	0.022	0.045	-0.067	0.110	Independent school	0.023	0.036	-0.048	0.094
Low-mid school SES	0.104	0.019	0.067	0.141	Low-mid school SES	0.071	0.015	0.043	0.100
Middle school SES	0.320	0.022	0.278	0.363	Middle school SES	0.254	0.017	0.222	0.287
Up-mid school SES	0.554	0.026	0.503	0.605	Up-mid school SES	0.436	0.020	0.396	0.475
Upper school SES	0.775	0.049	0.680	0.872	Upper school SES	0.622	0.039	0.544	0.700
Pupil-level interaction effects‡	**Post. mean**	**s.d.**	**95% C.I.**		**Pupil-level interaction effects‡**	**Post. mean**	**s.d.**	**95% C.I.**	
Prior att. & Male	-0.006	0.003	-0.012	0.0002	Prior att. & Male	-0.050	0.003	-0.056	-0.043
Male & Low-mid inc.	-0.013	0.006	-0.025	-0.001	Male & Low-mid inc.	-0.002	0.007	-0.016	0.012
Male & Up-mid inc.	-0.019	0.009	-0.038	-0.001	Male & Up-mid inc.	-0.014	0.011	-0.035	0.007
Male & High income	-0.002	0.012	-0.024	0.021	Male & High income	0.019	0.013	-0.006	0.044
Prior att. & Held back	-0.068	0.005	-0.078	-0.058	Prior att. & Held back	-0.069	0.005	-0.080	-0.058
Male & Held back	0.084	0.009	0.066	0.102	Male & Held back	0.029	0.010	0.009	0.049
Cross-level interaction effects‡§	**Post. mean**	**s.d.**	**95% C.I.**		**Cross-level interaction effects‡§**	**Post. mean**	**s.d.**	**95% C.I.**	
Prior att. & Subs. Indep.	0.012	0.005	0.002	0.022	Prior att. & Subs. Indep.	-0.001	0.005	-0.010	0.009
Prior att. & Indep. school	0.015	0.018	-0.020	0.050	Prior att. & Indep. school	0.033	0.019	-0.004	0.070
Prior att. & Low-mid SES	-0.010	0.006	-0.023	0.002	Prior att. & Low-mid SES	-0.023	0.006	-0.035	-0.011
Prior att. & Middle SES	0.022	0.007	0.008	0.036	Prior att. & Middle SES	-0.014	0.007	-0.028	-0.001
Prior att. & Up-Mid SES	0.023	0.009	0.005	0.040	Prior att. & Up-Mid SES	-0.010	0.009	-0.026	0.007
Prior att. & Upper SES	-0.016	0.019	-0.053	0.021	Prior att. & Upper SES	-0.052	0.020	-0.091	-0.013

† These parameters were obtained via MCMC estimation with Gibbs sampling. Monitoring chain length: 215,000; burn-in: 500; storing all iterations. All fixed-effects parameters have an effective sample size of at least 3,800. Posterior means are reported here. Prior distribution for the fixed-part parameters is: Normal(0, 10^8).

‡ Reference categories for pupil-level variables: Female; Low income and Not held back.

§ Reference categories for school-level variables: State-funded school and Low school SES.

From the cross-level interaction effects, some evidence is found that progress in Mathematics varies slightly according to schools' institutional type, where the interaction between prior attainment in Mathematics and subsidised independent school is significant. This implies that the slope of prior attainment in Mathematics is slightly steeper for pupils in subsidised independent schools in comparison to their peers in state-funded schools. On the other hand, there is no evidence that the slope of prior attainment is steeper for pupils in independent schools compared with pupils in State-funded schools. In the case of Language, the interaction between prior attainment and school institutional type did not turn out to be significant, which means that the relationship between prior attainment (covariate) and subsequent attainment (outcome) in Language does not vary according to institutional type.

In the case of average school SES, evidence is found that the slope of prior attainment in Mathematics varies across SES levels, although this slope is not significant for all levels. Only pupils in middle SES and upper-middle SES schools make slightly more progress, i.e. the linear relationship between prior and subsequent attainment is steeper, than that of their peers in low-SES schools (the reference category), while pupils attending upper-SES schools do not seem to make more progress than pupils in low-SES schools. Meanwhile, no such evidence has been found for progress in Language, where none of these interactions are significant.

4.4. The implications of the bivariate CVA model to school accountability measures

In a CVA model, a key assumption is that after controlling for all relevant factors that are beyond the control of schools, i.e. pupils' characteristics or compositional effects, as well as non-malleable school characteristics, the remainder of the variance at the level of schools can be used to estimate the amount of value that schools add to their pupils' educational trajectories.

Extending formulae from Goldstein (2011) to estimate school residuals with more than two levels of variation, CVA scores have been estimated for Mathematics and Language. From these CVA scores, about 9% of schools are below the national average in Mathematics, while about 9% are above average. These percentages in

Language are approximately 7%. Thus, it can be argued that secondary schools add more value in Mathematics than they add in Language to their pupils.

On another front, when comparing residuals from the univariate 4-level models for Mathematics and Language (details of this model are available on request), only a 4% (approx.) are below (and above) average in both subjects. In contrast, when comparing the residuals from the bivariate CVA model for both subjects, there is a higher rate of agreement between subjects. Approximately a 6% of schools are below (and above) the national average.

Based on whether the confidence interval of a school CVA estimate overlaps or not with the national average, i.e. it includes zero, a simple 3-level classification can be derived: 1) schools below the national average; 2) average schools; and 3) schools above the national average. In the bivariate case, though, there are two sets of school effects, i.e. one for Mathematics and another for Language. Therefore, schools can have diverse classifications according to subjects. When schools are classified as above average in one subject but not the other, they are classified as average. The reverse works in the same way: when schools are classified as below average in one subject but not the other, they are classified as average. It is worth noting that no schools are classified as above average in one subject and below average in another. Consequently, there are no huge discrepancies. In Table 8, classifications derived from the traditional (2-level) and the extended bivariate (5-level) CVA models are presented for comparison.

Table 8: Comparison of school classifications in the traditional 2-level CVA model and the bivariate 5-level CVA model

	Bivariate CVA model				
Traditional CVA model	**Below average**	**Average**	**Above average**	**Total**	**Percentage‡**
Below Average	130	106	0	236	9.68%
Average	16	1,880	8	1,904	78.10%
Above average	0	162	136	298	12.22%
Total	146	2,148	144	2,438	100%
Percentage†	5.99%	88.11%	5.91%	100%	

Note: The diagonal shows the agreement between the models. 2,146 (88.02%) schools remain in the same category in both models.

† Within bivariate CVA model classifications.

‡ Within traditional CVA model classifications. Classifications were derived from two univariate CVA models.

In Table 8, it is appreciated that the agreement between both models is quite high (diagonal elements in the table, i.e. 130+1880+136=2146; then 2146/2438*100%=88.02%). However, discrepancies can be deemed as highly relevant from a substantive point of view. In a traditional 2-level CVA model, 162 (6.64%) schools are wrongfully classified as “above average”, whereas they are “average” according to the 5-level bivariate model. In turn, in a traditional 2-level CVA model, 106 (4.35%) schools are unfairly classified as “below average”, whereas they are “average” according to the 5-level bivariate model. Nevertheless, this comparison between models can derive into a two-sided argument. On the one hand, one can argue that schools classified as above average in a traditional 2-level CVA model are unfairly downgraded to average in a 5-level bivariate CVA model, while schools classified as below average in the 2-level model are groundlessly upgraded to average in the 5-level model. On the other hand, one can make the same case, *mutatis mutandis*, to favour the 5-level model. To sort out these allegations, one should bear in mind the purpose of the CVA model. Given that the 5-level model effectively controls for statistically significant sources of variation external to the schools and the pupils that the traditional models do not take into account, one should presume that comparisons derived from it are fairer. It is argued here that, even though the traditional and the extended model represent two different concepts of value-added, the most useful specification from a public policy-making perspective is the one that contextualises the information on school performance according to socio-economic, demographic and geographic background.

These discrepancies are of the utmost importance when considering the repercussions of schools being classified in certain categories. As discussed previously, in a high-stakes accountability system, such as the one to be implemented in the Chilean education system, underperforming schools can be subject to extreme measures, such as closure. Naturally, the debate is still open with respect to how models for school accountability should be implemented. However, it is apparent from these analyses that the traditional approach is insufficient to deal with the complexity of the school performance phenomenon.

5. Conclusions

A 5-level cross-classified bivariate CVA model for progress in Mathematics and Language was implemented to analyse the intricacies of school performance and to draw relevant conclusions about the effectiveness of Chilean schools. The full model is a highly complex specification that attempts to give more insight into how much schools contribute to the educational trajectories of their pupils, by controlling for the most relevant factors that go beyond what any given school can regulate or intervene. This is done with the purpose of isolating the "true" school effects. Here, the conceptual meaning of “true” effects relates to the capacity of this model to isolate as far as possible the school effects from extraneous or diffuse influences from other sources to provide socio-economically and geographically contextualised school value-added estimates, which are useful for public policy-making.

From this analysis, it can be concluded, firstly, that progress in Mathematics and Language are undoubtedly related to each other and they need to be analysed accordingly with a sufficient level of complexity. This is also complementary to the idea that school effectiveness is not a unidimensional phenomenon, because schools neither teach curriculum subjects completely separated from each other nor do pupils learn in an isolated way.

As expected, this relationship varies in magnitude, but not in direction, across the levels of pupils, classrooms, primary schools, secondary schools and local authorities. Focusing on the pupils, the underlying relationship between Mathematics and Spanish Language is considerable when fitting an unconditional 2-level model, but only moderate after controlling for all relevant levels of variation and for socio-economic and demographic characteristics of the pupils themselves and the schools. This implies that a great deal of the relationship between Mathematics and Language arises as result of confounding. However, after confounders are removed, the remainder would seem to indicate that while pupils can be more inclined or able in one subject or another, they can still learn throughout their educational trajectories a set of skills (although undetermined in this research) that are applicable to both subjects. With regard to the schools, the underlying relationship between both subjects is quite sizable, which implies that secondary schools making significant contributions to their pupils' progress in one subject are also very likely to be doing the same in the other subject (and others as well). The opposite, of course, would also be true. Focusing on what school policy can intervene, this could be said to indicate that teaching practices shared within secondary schools are -on average- somewhat uniformly effective or ineffective irrespective of the subject.

Provided that the data are available, school value-added models should take into account all possible relationships between academic outcomes in different subjects, as well as non-academic educational outcomes. A thorough analysis of a model such as the latter, where academic and non-academic outcomes are analysed simultaneously is developed by Timmermans (2012). This is indeed a limitation of this research, as there are no data available to analyse neither further subjects nor non-academic outcomes, which is undoubtedly a shortcoming. However, this does not undermine the value of this analysis on its own right, because it brings together a more integrative vision of school effectiveness, along with the necessary technical sophistication.

The second main conclusion is that the bivariate CVA model is worthwhile in spite of its ever-increasing complexity. This is due to its power to discriminate better amongst underperforming, outperforming and average schools in a way that is still consistent with less complex models, such as the traditional 2-level CVA model. It is consistent because it successfully identifies broadly the same most and least effective schools as less complex models, while it discards others that were misclassified under those categories, when instead they were actually average-performing.

Thirdly, it has also been demonstrated that the analysis of pupils' educational outcomes needs to incorporate information about previously attended schools to estimate the contribution of the currently attended school more precisely. It has been demonstrated that carry-over effects from primary school into secondary school are relevant and their estimation can grant more reliability to school comparisons.

This extension to the traditional univariate CVA models embodies a richer and broader concept of school effectiveness. This, in turn, lays out a set of non-trivial and non-negligible adjustments of the utmost relevance when feeding back diverse stakeholders, namely: parents, head teachers, local education authorities, policy makers, etc. This bivariate CVA model is ultimately a valuable tool that takes a more thorough account of the complexity of the school performance phenomenon. However, the model although complex, has the limitation that it fitted homogeneous covariance structures at the higher levels. A further extension to the model presented in this paper is to fit a heterogeneous covariance structure at the level of secondary schools (or any other higher-level of interest), allowing variability at this level to depend not only on pupil-level characteristics (prior attainment), but also on school-level characteristics, as seen in Leckie et al. (2014) and Milla et al. (2016).

The overall contribution of this paper is related to three main aspects. From an international perspective, this study confirms once more that the school performance phenomenon is indeed complex, as shown by other studies in developed countries, such as the United Kingdom, the Netherlands and the United States. Furthermore, it shows that the challenges that this complexity conveys are even more marked in developing countries, such as Chile. As seen in this paper and other studies (Troncoso et al., 2015; Cervini, 2009a), the size of the school effects and the geographical effects are far greater in Chile than in developed countries, which makes the use of sophisticated methods even more relevant. In sum, assessing school performance in developing countries needs to be tackled with advanced modelling approaches to fully capture its complexity.

From a methodological point of view, the methods described and applied in this study are certainly not new and have been applied in other studies; however, none of the studies reviewed had applied all the different specifications presented here at once. For instance, as mentioned in section 2, a few researchers have pointed out the need for specifying further levels to the traditional 2-level structure of pupils nested within schools; some of those studies have incorporated the analysis of fully-nested structures, and some others have included the analysis of non-nested structures. On another front, a very small number of studies have specified multivariate multilevel models for analysing educational outcomes. Nevertheless, no other studies have specified further nested and non-nested levels (beyond the traditional 2-level structure) in a multivariate multilevel model to analyse educational outcomes. This study has demonstrated that this avenue, even though complex, is fruitful for analysing school performance and deriving fairer school accountability measures for policy-making purposes.

Finally, in the Chilean public policy context, this study contributes to the debate about how to assess school performance more fairly. As seen throughout this paper, the models presented are certainly an improvement with respect to the traditional approach of 2-level CVA models of pupils nested within schools, let alone with respect to the new policy of the Chilean Government. This study clearly shows that the choice of models is not trivial, especially when considering that an unfair/incomplete assessment can potentially have unduly harsh consequences on the schools and the pupils themselves.

References

Agencia de Calidad de la Educación. (2014). Minuta informe metodología de ordenación [Notes on report of ranking methodology]. Santiago, Chile. Retrieved from http://www.agenciaeducacion.cl/

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. Educational Researcher, 36(5), 258–267.

Browne, W. (2012). MCMC Estimation in MLwiN, v2.26. Centre for Multilevel Modelling, University of Bristol.

Browne, W., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. Statistical Modelling, 1(103-124).

Cervini, R. (2009a). Class, school, municipal, and state effects on mathematics achievement in Argentina: a multilevel analysis. School Effectiveness and School Improvement, 20(3), 319–340.

Cervini, R. (2009b). Comparando la inequidad en los logros escolares de la educación primaria y secundaria en Argentina: Un estudio multinivel [Comparing inequity of educational achievement in primary and secondary education in Argentina: A multilevel study]. Revista Iberoamericana Sobre Calidad, Eficacia Y Cambio En Educación (REICE), 7(1), 5–21.

Cowan, R., Donlan, C., Newton, E., & Lloyd, D. (2005). Number Skills and Knowledge in Children With Specific Language Impairment. Journal of Educational Psychology, 97(4), 732–744.

Creemers, B. (1994). The effective classroom. London: Cassell.

Donlan, C., Cowan, R., Newton, E., & Lloyd, D. (2007). The role of language in mathematical development: evidence from children with specific language impairments. Cognition, 103(1), 23–33.

Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. New York: Cambridge University Press.

Goldstein, H. (2011). Multilevel statistical models (4th ed.). Chichester, UK: John Wiley and Sons, Ltd.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. Journal of the Royal Statistical Society: Series A (Statistics in Society), 170(4), 941–954.

Hecht, S., Torgesen, J., Wagner, R., & Rashotte, C. (2001). The relations between phonological processing abilities and emerging individual differences in mathematical computation skills: a longitudinal study from second to fifth grades. Journal of Experimental Child Psychology, 79(2), 192–227.

Hill, P., & Goldstein, H. (1998). Multilevel Modeling of Educational Data with Cross-Classification and Missing Identification for Units. Journal of Educational and Behavioural Statistics , 23(2), 117–128.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. Journal of the Royal Statistical Society: Series A (Statistics in Society), 172(3), 537–554.

Leckie, G., French, R., Charlton, C., & Browne, W. (2014). Modeling Heterogeneous Variance–Covariance Components in Two-Level Models. Journal of Educational and Behavioral Statistics, 39(5), 307-332.

Leckie, G., & Charlton, C. (2013). runmlwin : A Program to Run the MLwiN Multilevel Modeling Software from within Stata. Journal of Statistical Software, 52(11).

Manzi, J., San Martín, E., & Van Bellegem, S. (2014). School system evaluation by value added analysis under endogeneity. Psychometrika, 79(1), 130–153.

Martínez, J. F. (2012). Consequences of omitting the classroom in multilevel models of schooling: an illustration using opportunity to learn and reading achievement. School Effectiveness and School Improvement, 23(3), 305–326.

Milla, J., San Martin, E., Van Bellegem, S. (2016). Higher Education value added using multiple outcomes. Journal of Education Measurement, 53(3), 368-400.

Mizala, A., & Torche, F. (2012). Bringing the schools back in: the stratification of educational achievement in the Chilean voucher system. International Journal of Educational Development, 32(1), 132–144.

Murillo, F. J., & Román, M. (2011). School infrastructure and resources do matter: analysis of the incidence of school resources on the performance of Latin American students. School Effectiveness and School Improvement, 22(1), 29–50.

Ortega, L., Malmberg, L-E., & Sammons, P. (2018). School effects on Chilean children's achievement growth in language and mathematics: An accelerated growth curve model. School Effectiveness and School Improvement. 29(2), 308-337.

Page, G., San Martín, E., Orellana, J., & González, J. (2016). Exploring Complete School Effectiveness via Quantile Value-Added. Journal of the Royal Statistical Society, Series A (Statistics in Society), (in press).

Plewis, I. (2011). Contextual variations in ethnic group differences in educational attainments. Journal of the Royal Statistical Society: Series A (Statistics in Society), 174(2), 419–437.

Radovic, D. (2018). Gender differences in mathematics attainment in Chile. Revista Colombiana de Educacion. 74, 221-241.

Rasbash, J., Charlton, C., Browne, W., Healy, M., & Cameron, B. (2012). MLwiN version 2.26 [Computer software]. Bristol, UK: Centre for Multilevel Modelling, University of Bristol.

Rasbash, J., Leckie, G., & Pillinger, R. (2010). Children's educational progress: partitioning family, school and area effects. Journal of the Royal Statistical Society: Series A (Statistics in Society), 173(3), 657–682.

Raudenbush, S. (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.

Ray, A. (2006). School value added measures in England. London. Retrieved from https://education.gov.uk/publications/eOrderingDownload/RW85.pdf

Sahlberg, P. (2007). Education policies for raising student learning: the Finnish approach. Journal of Education Policy, 22(2), 147–171.

Sahlberg, P. (2010). Rethinking accountability in a knowledge society. Journal of Educational Change, 11, 45–61.

San Martín, E., & Carrasco, A. (2012). Clasificación de escuelas en la nueva institucionalidad educativa : contribución de modelos de valor agregado para una responsabilización justa [Classification of schools in the new educational institutional environment: the contribution of value-added mo. Temas de La Agenda Pública, 7(53), 1–18.

San Martín, E., & Carrasco, A. (2013). Criterios para evaluar la metodología oficial de clasificación de escuelas: ¿un asunto técnico o conceptual? [Criteria to assess the official methodology of classification of schools: a technical or conceptual matter?]. In I. Irarrázaval, M. Morandé, & M. Letelier (Eds.), Propuestas para Chile [Proposals for Chile] (pp. 85–114). Centro de Políticas Públicas, Pontificia

Universidad Católica de Chile. Retrieved from http://politicaspublicas.uc.cl/wp-content/uploads/2014/01/Libro-Propuestas-para-Chile_versión-web.pdf

Scheerens, J., & Bosker, R. (1997). The foundations of educational effectiveness. Oxford: Pergamon.

Simmons, F. R., & Singleton, C. (2008). Representations Impact on Review of Research into Arithmetic and Dyslexia. Dyslexia, 14, 77–94.

Snijders, T., & Bosker, R. (2011). Multilevel analysis: An introduction to basic and advanced multilevel modelling, 2nd edition. London, Thousand Oaks, New Delhi: SAGE Publications.

Spiegelhalter, D., Best, N., Carlin, B., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(4), 583–639.

StataCorp. (2011). Stata Statistical Software: Release 12. College Station, Texas.

Sulis, I., & Porcu, M. (2014). Assessing Divergences in Mathematics and Reading Achievement in Italian Primary Schools: A Proposal of Adjusted Indicators of School Effectiveness. Social Indicators Research, 1–28.

Tate, W. (1997). Race-Ethnicity , SES , Gender , and Language Proficiency Trends in Mathematics Achievement: An Update. Journal for Research in Mathematics Education, 28(6), 652–679.

Timmermans, A. (2012). Value added in educational accountability: Possible, fair and useful? University of Groningen, Groningen, The Netherlands.

Timmermans, A., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. School Effectiveness and School Improvement, 22(4), 393–413.

Torche, F. (2005). Privatization reform and inequality of educational opportunity: The case of Chile. Sociology of Education, 78, 316–343.

Troncoso, P. (2015). Beyond the traditional school value-added approach: Analysing complex multilevel models to inform external and internal school accountability in Chile. The University of Manchester.

Troncoso, P., Pampaka, M., & Olsen, W. (2015). Beyond traditional school value-added models: a multilevel analysis of complex school effects in Chile. School Effectiveness and School Improvement, 1–22.

Vukovic, R., & Lesaux, N. (2013). The relationship between linguistic skills and arithmetic knowledge. Learning and Individual Differences, 23, 87–91.

Wang, J., & Goldschmidt, P. (1999). Opportunity to Learn , Language Proficiency , and Immigrant Status Effects on Mathematics Achievement. The Journal of Educational Research, 93(2), 101–111.

Acknowledgements: This research has been funded with the support of the CONICYT-Becas Chile Graduate Scholarship Programme. Also, the Ministry of Education of Chile has enabled this research by providing the national administrative dataset SIMCE for the years 2004 and 2006. However, the research is conducted independently, and the views expressed here are held by the author, not to be attributed to others. I would also like to acknowledge helpful comments during this research and in previous versions of this manuscript from Prof. Mark Tranmer, Dr Maria Pampaka, Prof. Wendy Olsen and Dr Ana Morales-Gómez.

Appendix A: Bivariate CVA model for progress in Mathematics and Language

Equation 1 follows the general multilevel notation as described in Goldstein (2011) and the classification notation described in Browne et al. (2001).

$$y_{t(i)} = \beta_{01}z_{1i} + \beta_{02}z_{2i} + \beta_{11}z_{1i}y_{t-1(i)} + \beta_{12}z_{2i}y_{t-1(i)} + (\mathrm{X}\beta)_{1i}z_{1i} + (\mathrm{X}\beta)_{2i}z_{2i}$$

$$\beta_{01} = \beta_{01i} + u^{(1)}_{1,local(i)} + u^{(2)}_{1,secondary(i)} + u^{(3)}_{1,primary(i)} + u^{(4)}_{1,classroom(i)} + e_{1i}$$

$$\beta_{02} = \beta_{02i} + u^{(1)}_{2,local(i)} + u^{(2)}_{2,secondary(i)} + u^{(3)}_{2,primary(i)} + u^{(4)}_{2,classroom(i)} + e_{2i}$$

$$\beta_{11} = \beta_{11i} + u^{(2)}_{3,secondary(i)}$$

$$\beta_{12} = \beta_{12i} + u^{(2)}_{4,secondary(i)}$$

$$where$$

$$z_{1i} = \begin{Bmatrix} 1\ if\ Mathematics \\ 0\ if\ Language \end{Bmatrix}\ and\ z_{2i} = \begin{Bmatrix} 1\ if\ Language \\ 0\ if\ Mathematics \end{Bmatrix} \qquad (1)$$

$$\begin{bmatrix} u^{(1)}_{1,local(i)} \\ u^{(1)}_{2,local(i)} \end{bmatrix} \sim MVN\left(0, \Omega^{(1)}_u\right): \Omega^{(1)}_u = \begin{bmatrix} \sigma^2_{u11} & \sigma_{u11,12} \\ \sigma_{u11,12} & \sigma^2_{u12} \end{bmatrix}$$

$$\begin{bmatrix} u^{(2)}_{1,secondary(i)} \\ u^{(2)}_{2,secondary(i)} \\ u^{(2)}_{3,secondary(i)} \\ u^{(2)}_{4,secondary(i)} \end{bmatrix} \sim MVN\left(0, \Omega^{(2)}_u\right): \Omega^{(2)}_u = \begin{bmatrix} \sigma^2_{u21} & \sigma_{u21,22} & \sigma_{u21,23} & \sigma_{u21,24} \\ \sigma_{u21,22} & \sigma^2_{u22} & \sigma_{u22,23} & \sigma_{u22,24} \\ \sigma_{u21,23} & \sigma_{u22,23} & \sigma^2_{u23} & \sigma_{u23,24} \\ \sigma_{u21,24} & \sigma_{u22,24} & \sigma_{u23,24} & \sigma^2_{u24} \end{bmatrix}$$

$$\begin{bmatrix} u^{(3)}_{1,primary(i)} \\ u^{(3)}_{2,primary(i)} \end{bmatrix} \sim MVN\left(0, \Omega^{(3)}_u\right): \Omega^{(3)}_u = \begin{bmatrix} \sigma^2_{u31} & \sigma_{u31,32} \\ \sigma_{u31,32} & \sigma^2_{u32} \end{bmatrix}$$

$$\begin{bmatrix} u^{(4)}_{1,classroom(i)} \\ u^{(4)}_{2,classroom(i)} \end{bmatrix} \sim MVN\left(0, \Omega^{(4)}_u\right): \Omega^{(4)}_u = \begin{bmatrix} \sigma^2_{u41} & \sigma_{u41,42} \\ \sigma_{u41,42} & \sigma^2_{u42} \end{bmatrix}$$

$$\begin{bmatrix} e_{1i} \\ e_{2i} \end{bmatrix} \sim MVN(0, \Omega_{\mathrm{e}}): \Omega_{\mathrm{e}} = \begin{bmatrix} \sigma^2_{e1} & \sigma_{e12} \\ \sigma_{e12} & \sigma^2_{e2} \end{bmatrix}$$

$y_{t(i)}$ is a twofold set of outcome variables defined by the dummy variables z_{1i} (Mathematics) and z_{2i} (Language). The data set has a long format with two observations per case, which adds an artificial level to fit two equations simultaneously. β_{01} and β_{02} correspond to the intercepts of each subject, which are allowed to vary randomly at the levels of primary and secondary schools, classrooms and local authorities. β_{11} corresponds to the effect of prior attainment score in Mathematics, denoted by $z_{1i}y_{t-1(i)}$; while β_{12} corresponds to the effect of the prior attainment score in Language, denoted by $z_{2i}y_{t-1(i)}$. β_{11} and β_{12} are allowed to vary randomly at the secondary school level, which makes covariance matrix $\Omega_u^{(2)}$ depend on pupil-level prior attainment. $(\mathrm{X}\beta)_{1i}z_{1i}$ and $(\mathrm{X}\beta)_{2i}z_{2i}$ are the coefficients of the set of covariates (specified in Tables 2 and 3), which are allowed to have different effects on each outcome. The u terms correspond to the random effects at the higher levels indicated in the subscript (local, primary, secondary or classroom) on the outcome indicated by the number subscript (1 or 2); they have a multivariate normal distribution with mean 0 and covariance matrices Ω_u which are superscripted with the number of the corresponding higher level. Each u term has a pair of variances σ_u^2 for each outcome and a covariance σ_u to allow for the non-vanishing relationship between Mathematics and Language. Finally, the e terms correspond to the unobserved heterogeneity of pupils (error term); they have a multivariate normal distribution with mean 0 and covariance matrix Ω_e which contains variances σ_e^2 for each outcome and covariance σ_e for the relationship between both outcomes.

Appendix B: Random-effects parameters of the bivariate CVA model for progress in Mathematics and Language

Levels	Parameters†	Post. mean	s.d.	Correlation	95% C.I.	
Secondary Schools	Var. (Intercept Maths)	0.059	0.003	--	0.054	0.065
	Var. (Intercept Language)	0.028	0.002	--	0.025	0.031
	Cov. (Int. Maths, Int. Language)	**0.035**	**0.002**	**0.873**	**0.032**	**0.039**
	Cov. (Prior att. Maths, Int. Maths)	0.003	0.001	0.204	0.002	0.004
	Cov. (Prior att. Maths, Int. Language)	0.002	0.0005	0.214	0.001	0.003
	Var. (Prior att. Maths)	0.004	0.0003	--	0.003	0.005
	Cov. (Prior att. Language, Int. Maths)	0.003	0.001	0.288	0.002	0.005
	Cov. (Prior att. Language, Int. Language)	0.003	0.0005	0.344	0.002	0.004
	Cov. (Prior att. Language, Prior att.	**0.003**	**0.0002**	**0.836**	**0.002**	**0.003**
	Var. (Prior att. Language)	0.002	0.0003	--	0.002	0.003
	Cov. (Male Maths, Int. Maths)	0.002	0.001	0.111	-0.0004	0.004
	Cov. (Male Maths, Int. Language)	0.002	0.001	0.149	0.0001	0.003
	Cov. (Male Maths, Prior att. Maths)	0.00004	0.0003	0.010	-0.001	0.001
	Cov. (Male Maths, Prior att. Language)	-0.0001	0.0003	-0.044	-0.001	0.0005
	Var. (Male Maths)	0.004	0.001	--	0.003	0.006
	Cov. (Male Language, Int. Maths)	0.004	0.001	0.292	0.002	0.006
	Cov. (Male Language, Int. Language)	0.004	0.001	0.422	0.002	0.005
	Cov. (Male Language, Prior att. Maths)	0.0002	0.0003	0.054	-0.0005	0.001
	Cov. (Male Language, Prior att. Language)	0.0005	0.0003	0.182	-0.0001	0.001
	Cov. (Male Language, Male Maths)	**0.002**	**0.001**	**0.647**	**0.001**	**0.003**
	Var. (Male Language)	0.003	0.001	--	0.002	0.004
Classes	Var. (Intercept Maths)	0.039	0.001	--	0.037	0.041
	Cov. (Int. Maths, Int. Language)	**0.025**	**0.001**	**0.784**	**0.023**	**0.026**
	Var. (Intercept Language)	0.025	0.001	--	0.023	0.026
Pupils	Var. (Intercept Maths)	0.276	0.001	--	0.274	0.278
	Cov. (Int. Maths, Int. Language)	**0.079**	**0.001**	**0.248**	**0.077**	**0.081**
	Var. (Intercept Language)	0.368	0.001	--	0.365	0.370
Local Authorities	Var. (Intercept Maths)	0.004	0.001	--	0.002	0.006
	Cov. (Int. Maths, Int. Language)	**0.003**	**0.001**	**0.858**	**0.001**	**0.004**
	Var. (Intercept Language)	0.002	0.001	--	0.001	0.004
Primary Schools	Var. (Intercept Maths)	0.007	0.0004	--	0.007	0.008
	Cov. (Int. Maths, Int. Language)	**0.004**	**0.0003**	**0.765**	**0.003**	**0.005**
	Var. (Intercept Language)	0.004	0.0003	--	0.003	0.004

† These parameters were obtained via MCMC estimation with Gibbs sampling. Chain length: 215,000; burn-in: 500; thinning: 1. All random-effects parameters have an effective sample size of at least 1,000. DIC=613,323.3. Prior distribution for the random-part is: $Inverse\ Wishart_6(\widehat{\Omega}, 6)$ for the secondary school level and $Inverse\ Wishart_2(\widehat{\Omega}, 2)$ for the rest of the levels, where $\widehat{\Omega}$ is each level's covariance matrix estimated via IGLS. Trajectories mix well with approximately normally-distributed posteriors; however, it is not presented here for it exceeds the scope of this paper. Full details are available on request.