Beyond traditional school value-added models: A multilevel analysis of complex school effects in Chile

Patricio Troncoso*, Maria Pampaka, Wendy Olsen

* The Cathie Marsh Institute for Social Research (CMIst), The University of Manchester
United Kingdom

* Corresponding author. Email: patricio.troncoso@manchester.ac.uk. Postal Address: Room G45, Humanities Bridgeford Street Building, The University of Manchester, Oxford Road, M13 9PL, Manchester, United Kingdom. Tel: +44 0161 275 4721

Patricio Troncoso is a PhD student in Social Statistics at the University of Manchester. Since 2011, he has been working on his research on school effectiveness in Chilean schools. His research work 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 socio-economic inequalities of the Chilean 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. Tel: +44 0161 275 4721. Email: patricio.troncoso@manchester.ac.uk

Maria Pampaka is currently holding a joint position, as a Lecturer and Research Fellow, at the Manchester Institute of Education and the Social Statistics group, at the University of Manchester. Her role involves leading an ESRC funded project in mathematics education (www.teleprism.com), and delivering courses in advanced quantitative methods. Her research expertise spreads along the spectrum of educational studies, focussing in teachers’ knowledge and practices and students’ attitudes and progress in their educational trajectories. She is particularly interested in the association between teaching practices and students’ learning outcomes. Methodologically, her expertise and interests lie within evaluation and measurement, and advanced quantitative methods, including complex survey design, longitudinal data analysis, and missing data and imputation techniques. Tel: +44 0161 275 7213. Email: maria.pampaka@manchester.ac.uk

Wendy Olsen works as a Reader in Socio-Economic Research at the University of Manchester. Her publications include Data Collection (Sage, 2012), Rural Indian Social Relations (1996) and The Politics of Money (2002) with F. Hutchinson and M. Mellor. Wendy Olsen currently holds an ESRC-DFID research grant on gender norms and labour supply in
Acknowledgements

We acknowledge the support of the Ministry of Education of Chile which has enabled this research by providing the national administrative dataset SIMCE for the years 2004 and 2006. We acknowledge the support of the CONICYT-Becas Chile Graduate Scholarship Programme which has funded this research. However, the research is conducted independently, and the views expressed here are held by the authors, not to be attributed to others. We would also like to acknowledge helpful comments in previous versions of this manuscript from Mark Tranmer, Natalie Shlomo and the anonymous reviewers of this journal.

Beyond traditional school value-added models: A multilevel analysis of complex school effects in Chile

School value-added studies have largely demonstrated the effects of socio-economic and demographic characteristics of the schools and the pupils on performance in standardised tests. Traditionally, these studies have assessed the variation coming only from the schools and the pupils. However, recent studies have shown that the analysis of academic performance could be significantly benefited from additional complexity in the model structure, incorporating non-hierarchical and unexplored levels of variation. Using data on secondary students from the Chilean National Pupil Database (2004-2006), this study shows how the traditional value-added models fall short in addressing the complex phenomenon of academic performance, because they largely overestimate school effects. A 4-level contextualised value-added model for progress in Mathematics was implemented and shown to avoid the masking of classroom and locality effects found in the traditional models. We also analyse the effects of important structural factors in Chile such as family income and school type.

Keywords: school value-added, multilevel modelling, Chilean education system

Introduction

Educational inequalities have been a great concern for practitioners, policy-
makers and researchers during the last decades; views vary between those authors who claim that 'schools do not matter' (see for example: Coleman et al., 1966; Jencks et al., 1972), to those who claim that in fact 'schools and teachers do make a difference' (or can add value to) in the educational trajectories of disadvantaged pupils (see for example: Reynolds, 1995; Scheerens & Bosker, 1997; Scheerens, 2000, 2004). The extent to which schools can contribute to overcome the gaps in students' attainment due to socio-economic inequalities has been one of the main research challenges in educational research; the question of how effective schools are, from the perspective of disadvantaged pupils, refers to how schools foster pupils' performance beyond what is expected given their initial attainment and the obstacles they have to overcome from a given socio-economic situation (Creemers, 1994; Goldstein, 1997; Reynolds, 1995; Scheerens & Bosker, 1997). In other words, the main “effectiveness question” from the perspective of this research field is ‘Which schools actually add 'value' to the educational trajectories of their pupils?’

The discussion about ‘value-added’ originates in the field of measuring school effectiveness, which simply refers to the amount of value added to any given student’s test outcome that is uniquely attributable to attending a particular school (Scheerens, 2000). School effectiveness itself is a controversial term. According to Goldstein (1997, 2001) value-added indicators are inherently uncertain and are relative to the performance of other schools. Thus the most accurate concept is "relative school effectiveness". An effective school is defined here as one that encourages the progress of all its pupils beyond what is expected given their initial levels of achievement and socio-economic (and cultural) background, ensuring in a sustainable manner, that all students get the highest possible outcomes and improvement in every aspect of their development and performance (Scheerens, 2000). Value-added is "an indication of the
extent to which any given school has fostered the progress of all students in a range of subjects during a particular time period in comparison to the effects of other schools in the same sample” (Sammons et al., 1997, p. 24).

The education systems of developing countries, such as Chile, have to deal with high levels of deprivation and multiple inequalities. Chile's inequality is reflected in the Gini coefficient$^1$ of 52.1 (The World Bank, 2013) and the loss of 19% in its Human Development Index (HDI)$^2$ due to inequality (United Nations Development Programme, 2013). Large HDI differences by region and between localities can also be found (United Nations Development Programme, 2004). This makes plausible the existence of significant differences in pupils' achievement according to the geographical location of their school. On another front, Chilean schools historically have put in place streaming practices, where pupils are assigned to different classrooms according to their ability level (Torche, 2005), which makes those schools likely to reveal differences in value-added between classrooms. Despite new evidence in favour of specifying value-added models that are more complex than the traditional 2-level model, i.e. pupils nested within schools, only two recent studies conducted in Chile (Manzi et al., 2013; Mizala & Torche, 2012) have addressed issues related to the effects of streaming, the geographical location of the schools or the carry-over effects from primary school.

We aim to contribute to the knowledge and analysis of school value-added in the specific context of Chilean schools, and thus potentially shed light on the topic in the broader context of developing countries. Ultimately, this paper aims to contribute to the development of a more precise school accountability system in Chile by implementing a more refined and complex approach that fits better with the complexity of the school effectiveness phenomenon. In section 2, we briefly outline the current state of the literature around measurement of school effectiveness and value added and conclude
with our particular research questions; in section 3, we describe the data and methods used in this paper; in section 4, we present the results from a series of multilevel models to analyse value-added in Mathematics; and finally in section 5, we discuss the results and draw some relevant conclusions and implications for educational policy and practice.

A brief overview of measuring school effectiveness and its implications

One indicator of school effectiveness has involved the use of averages in standardised tests. This has proved to be flawed given the risk of drawing seriously misleading conclusions as a result of the "ecological fallacy," since relationships between variables might well be different at the school level compared to the pupil level, which does not allow us to infer about differential effects when comparing groups of students within schools (Gibson & Asthana, 1998; Goldstein & Spiegelhalter, 1996; Goldstein, 1997, 2001). The problem is more pronounced when yearly comparisons of school averages are made, since these comparisons are performed between different cohorts and without controlling for pre-existing differences between pupils (Goldstein, 1997). For instance, a sudden rise in the average of a particular school could be due to an uncontrolled increase in the number of better-performing pupils who are admitted to the school instead of an actual improvement of the school's effectiveness; in other words, schools could appear to be more effective simply by recruiting better-performing pupils. Despite all these downsides, current government practice in Chile still employs school averages as indicators of effectiveness (San Martín & Carrasco, 2012).

Researchers have advocated for a change in these practices, because of the indisputable amount of evidence in favour of more sophisticated statistical techniques for the analysis of school performance for public accountability (Goldstein & Thomas, 1996), such as multilevel modelling, which has become standard practice in educational
research. These models traditionally analyse variation in test scores at the level of pupils as well as schools, and have been proven to be fruitful.

There is a variety of multilevel models that have been used to analyse school effectiveness. For a start, "gross school effects" or "type 0" value-added models (Timmermans et al., 2011) specify a certain standard measure of attainment as the outcome and account only for the variance at the school and pupil level. They are empty models as described in the multilevel modelling literature (Goldstein, 2011). Nevertheless, it is not accurate to regard such models as value-added, because they do not measure progress. The simplest school value-added model that can be specified is the “raw” value-added (VA), which measures the progress made by the students controlling for their prior attainment. This model is said to measure the contribution of the school (school value-added) to the change in pupils' attainment (progress), but does not consider contextual information and thus it is deemed to be a "raw" measure of value-added. This definition corresponds to various names by various experts: “value-added” was previously used by the English Department for Education (Ray, 2006), “type AA” value-added (Timmermans et al., 2011) and “adjusted (school) comparisons” (Goldstein, 1997) are also amongst the ‘synonyms’.

From this model, it is possible to derive raw school value-added estimates, which correspond to the shrunken residual point estimates (also known as empirical Bayes residuals) of the multilevel model after controlling for prior attainment (Goldstein, 1997; Ray, 2006). The problem with this approach is that school prior attainment averages appear highly correlated with the value-added scores, which results from misspecification of the raw value-added models (Foley & Goldstein, 2012; Goldstein et al., 2007). This misspecification implies that the school effectiveness indicators are biased towards those schools serving the highest prior-attaining students.
Consequently, it is necessary to account for differential effects in the models (Foley & Goldstein, 2012). In other words, a (raw) VA model is misspecified insofar as it fails to control for relevant factors that influence academic performance while it also fails to meet relevant statistical assumptions.

Alternatively, contextualised value-added (CVA) models control for prior attainment in the subject of interest as well as variables related to pupil socio-economic and demographic characteristics and school context (Foley & Goldstein, 2012; many authors agree on this concept, see for example: Goldstein, 1997; Ray, 2006; Sammons et al., 1993). Others defined these models as “type A”, when controlling for student-level covariates only (Raudenbush, 2004; Timmermans et al., 2011), “type B”, when controlling for pupil-level covariates and compositional school-level variables (Raudenbush, 2004; Timmermans et al., 2011), and “type X”, when controlling for pupil-level covariates and compositional school-level variables and other non-malleable school factors (Timmermans et al., 2011). The "type X" value-added approach is an extension of the "type B" models that seeks to isolate the school effects on which the school has an influence (Timmermans et al., 2011). From this model, as with any multilevel model, it is possible to derive residual estimates which will be regarded as school CVA measures in this paper. Our focus in this research will be on this type of value-added since our primary aim is to develop models for more reliable school accountability systems.

Beyond the statistical superiority of this approach, as mentioned above, there have also been policy relevant arguments in favour of CVA scores estimation. In particular, educational research has demonstrated throughout the years that the main sources of differential effects in school performance are unsurprisingly associated with socio-economic and demographic characteristics of the pupils and the schools. With
respect to public policy, the use of these variables is relevant since using only school averages (without controlling for socio-economic variables) as indicators of effectiveness implicitly makes the schools accountable, for better or worse, for circumstances they cannot modify or for which they are not responsible (Ray, 2006).

However, CVA scores (and any other VA evaluation system) are not error-free, because they are products of statistical modelling processes, and hence, inherently uncertain, which is why confidence intervals should be and have been routinely estimated and published in performance tables in England (Foley & Goldstein, 2012; Ray, 2006). However, the use of CVA scores and their corresponding confidence intervals has some limitations related to high levels of uncertainty, especially when making predictions about future school performance (Leckie & Goldstein, 2009, 2011). Despite these limitations, Leckie and Goldstein (2011) point out that performance tables are still retrospectively informative and provide highly relevant information for holding schools accountable.

Another aspect of the measurement of school effectiveness is model complexity. Customarily, these value-added studies implement multilevel models where variation in standardised test scores comes from only two sources: schools and pupils. One of the main underlying assumptions of these models is that the school value-added is a latent trait that can be estimated from the analysis of the performance of the pupils within schools (Timmermans et al., 2012). Some other relevant and somewhat overlooked assumptions are: a) the last school a pupil has attended concentrates the whole amount of value-added throughout their educational trajectory; b) there are no differences between classrooms within schools; c) schools located in diverse geographical or administrative areas do not differ significantly. These assumptions are not only strong but rather unrealistic when analysing school performance.
More recent research has pointed out the necessity of increasing the complexity of the school value-added models by specifying additional levels of variation, which can be either completely nested or follow a cross-classification pattern (De Fraine et al., 2003; Goldstein et al., 2007; Leckie, 2009; Martínez, 2012; Rasbash et al., 2010; Timmermans et al., 2012). In these studies, the levels of neighbourhoods, classrooms, primary and secondary schools, as well as local education authorities were specified (although not all levels simultaneously in all studies) and they were all found to be relevant sources of variation in pupils' test scores. However, these studies are based in the United Kingdom, the United States and the Netherlands, which are all developed countries whose education systems are embedded in radically different socio-economic contexts compared to the Chilean education system.

Traditionally, as already mentioned, 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 new system is to be implemented by the Agencia de Calidad de la Educación (Agency for Quality of Education), from 2015 in a pilot stage. This methodology for classification of the schools will include a series of indicators of school quality, with attainment being the most important among them (Agencia de Calidad de la Educación, 2014). In this system, schools will be classified in four categories: a) high performance; b) middle performance; c) lower-middle performance; and d) 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 fairer comparisons. The main issue with this methodology is that it has been set out to implement multiple linear regression models.
to make these adjustments, which is clearly a major setback, given the overwhelming evidence in favour of the use of multilevel models.

Two recent multilevel studies in Chile actually address some of the issues around school effectiveness models that analyse the variation coming only from pupils and schools. Mizala and Torche (2012) implemented two univariate cross-sectional (without controlling for prior attainment) 2-level models for pupil attainment in Mathematics and Language in the 4th and 8th grade for the years of 2002 and 2004, respectively. Their analysis of the Chilean school system covered the effects of various pupil and school-level characteristics and emphasised the effects of within-school and between-school socio-economic stratification. The authors also attempted to control for within-school diversity by adding the standard deviation of an indicator of family SES, which they produced, as an explanatory variable at the school level. Regarding geographical location of the schools, the authors controlled for whether the schools were located in rural areas; however, this indicator variable does not account for other presumable regional differences, for instance between large cities and small towns. Another recent study by Manzi et al. (2013) proposes an alternative school value-added model to control for endogeneity present in the univariate value-added model for progress in Mathematics resulting from a 'traditional approach', i.e. a 2-level model of pupils nested within schools with random intercepts only, where prior attainment scores are correlated with the school value-added estimates, thus violating one of the assumptions of the multilevel modelling approach. The authors implemented a 2-level random coefficients contextualised value-added model controlling for parental qualifications, prior attainment, average school prior attainment and socio-economic status. They demonstrate how the inclusion of average school prior attainment controls for bias in the estimation of the school effects induced by school selectivity.
Nevertheless, these authors make no specific reference to within-school or between-localities effects.

Although there are no studies in Chile that explore within-school or between-locality effects, in neighbouring Argentina, Cervini (2009a, 2009b) demonstrates that classroom and geographical effects are indeed relevant. This author found that the between-school variation in Mathematics scores was greatly overestimated when using the traditional 2-level model in comparison with a 5-level model of pupils nested within classrooms, schools, municipalities and states. These studies are certainly relevant considering the similarities regarding the context in which the Chilean and the Argentinian education systems are embedded.

This study, thus, aims to fill in some of the limitations in previous work and contribute to school effectiveness research, around the issues of measuring value-added, including additional levels of variation, controlling for non-malleable pupil and school characteristics, differential progress measures and last but not least, the use of CVA measures for school accountability purposes. This will be accomplished by responding to the following research questions:

(1) Do the differences between classrooms and local authorities constitute relevant (non-negligible) sources of variation in pupils' progress in Mathematics in Chile, in addition to the effects at the level of pupils and schools?

(2) How do pupils' and schools' educational and socio-economic characteristics affect the progress made by pupils and the contribution of schools (school value-added) in Mathematics in Chile?

(3) Does pupils' academic progress in Mathematics significantly vary across classrooms, secondary schools and local authorities in Chile?
(4) Are there any relevant (non-negligible) differences between school value-added accountability measures derived from diverse statistical models?

Methodology: Data and Analytical Methods

Data and selected variables

The data on pupils’ performance in standardised academic tests come from the SIMCE (Measurement System for the Quality of Education, for its acronym in Spanish) 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. More details of the data collected by the SIMCE system can be found in Table 1 and in Agencia de Calidad de la Educación (2012).

The SIMCE tests have been administered every year, since 2006, to all students in the 4th grade and all students either in the 8th grade or in the 10th grade. Prior to 2006, pupils only sat the tests once in their whole educational trajectory, which made impossible to track pupils’ prior attainment. 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 is the only one moving from the last year of primary school to secondary school for which data are currently available. The total number of pupils for which data from the SIMCE database are available is 202,605 within 7,461 classrooms within
2,438 schools and 320 local authorities. Further details of the selected variables are given in Table 1.

The choice of explanatory variables at the pupil and school level is rooted in the literature, following the principle of controlling for relevant variables which schools cannot modify freely. Prior achievement in Mathematics is prerequisite for fitting any value-added model insofar it allows estimating the progress that pupils have made. With respect to the pupil-level variables gender and family income escape the schools' control and they have been found to impact on pupils' performance largely. Pupils' year retention is a variable over which schools have partial control since some can choose not to admit such pupils and therefore it may be a useful way for controlling, at least partially, for implicit selectivity school practices in the same way that prior achievement does (San Martín & Carrasco, 2013)⁴.

With respect to school-level variables, the socio-economic composition is a relevant influential factor identified in previous research and therefore School SES was included in the model. In addition, school type as Timmermans et al. (2011) point out is a non-malleable school characteristic that may be considered in type X value-added models. This is also a debatable choice; however, our stance is that schools cannot freely switch into a different institutional type, and hence a fair comparison should control for this, even at the expense of partially obscuring effects truly proceeding from diverse school practices associated with this. Chilean schools are classified in 3 main types: a) State-funded schools, which are (generally) comprehensive institutions managed by the Local Authorities and entirely funded by the State; b) Subsidised independent schools are institutions managed and partially funded by private parties, which receive sizable subsidies or "vouchers" from the State for the pupils they serve; c) Independent schools are institutions completely funded by private parties and parents
via fees and are only required to follow the National Curriculum loosely. Over the years, the coexistence of these three school types has produced a high level of segregation on the Chilean education system (Mizala & Torche, 2012; Torche, 2005), which is why it is relevant to ascertain the diverse effects that these schools have on pupils' performance.

(Table 1)

**Analytical Strategy: Modelling Value added within the multilevel framework**

The standard statistical techniques (e.g. multiple linear regression) are not robust in modelling CVA since they violate the assumption of independence of the observations (De Leeuw & Meijer, 2008; Hox, 2010; Snijders & Bosker, 2011). The most important consequence of ignoring the hierarchical structure of the data is the underestimation of standard errors, which produces spuriously significant regression coefficients (De Leeuw & Meijer, 2008; Hox, 2010; Snijders & Bosker, 2011). Multilevel modelling is an extension of the classical Multiple Linear Regression models (Hox, 2010) that allows to analyse the variation in an outcome at different levels of a hierarchical structure.

The analytical strategy followed in this study is adapted from Hox (2010) and includes the following steps:

1. specification of the empty models with 2 (pupils within schools), 3 (pupils within classrooms within schools) and 4 levels (pupils within classrooms within schools within local authorities) of variation and assess their significance. These are the variance components models that split the total variance into the different relevant sources. Variance partition coefficients (VPC) are also derived from the estimated variances as described by Snijders and Bosker (2011). VPC indicates
the percentage of the total variance that is due to a higher level in the data structure. This step corresponds to what Timmermans et al. (2011) denote as "type 0 value-added" or "gross school effects";

(2) specification of a baseline raw value-added model, where prior attainment is the only explanatory variable. This corresponds to a "type AA value-added" model (Timmermans et al., 2011);

(3) specification of a CVA model including pupil-level explanatory variables only. This corresponds to the "type A value-added" (Timmermans et al., 2011), which would be the most useful for school choice purposes from a substantive perspective. Methodologically, this model can be thought of as an intermediate model that allows to control for compositional effects, when the primary aim is to isolate the school effects;

(4) specification of a CVA model with pupil-level and school-level explanatory variables. This model and the forthcoming are all classified as "type X value-added" models (Timmermans et al., 2011), since school context and non-malleable factors have been included in this step onwards;

(5) specification of a CVA model where some of the pupil-level fixed effects are allowed to vary randomly at the school-level;

(6) specification of the full CVA model including cross-level interaction effects; and finally,

(7) fitting a 2-level model with the same specifications as the final 4-level CVA model for comparison. These models are comparable insofar as they are nested, i.e. the more complex model is a further specification of the simpler model.

All models were estimated with the software MLwiN (Rasbash et al., 2012), using the Iterative Generalised Least Squares (IGLS) algorithm. The significance of
individual variables was assessed via the Wald test and the overall fit of the models was assessed via the Likelihood ratio test (LR test) and the Akaike Information Criterion (AIC) as described by Snijders and Bosker (2011) and Goldstein (2011). The LR test was also used to assess the significance of the random components.

Following Goldstein's (2011) notation, the CVA model to be fitted in this study has the general form shown in Appendix 1. This notation allows the progress measure to lie in the slope coefficient on intake scores in Mathematics of each student, but also models the intercept term by positing random levels of 3 additional coefficients: one set of constant terms for classrooms, one for schools and one for local authorities. In addition, using the 4-level notation, it is possible to estimate random effects of each level’s units on progress. In other words, both the height and the slope of the curve can be different for each classroom, school, and local authority; and after controlling for these, the remaining explanatory slopes reflect contributions to (or correlates of) a student’s additional progress. Error terms take up further unexplained variation. This is a standard multilevel regression model of greater complexity than usual.

The model takes a confirmatory format with each variable reflecting an assertion rooted in the literature already cited. This model can measure the net effect of residing in a particular local authority, the effect of attending a particular school (usually called school effectiveness), the effect of being assigned to a particular classroom, and each pupil's heterogeneity. The effects of prior attainment in a particular local authority, at a particular school and a particular classroom, respectively, are also allowed for.

(Figure 1)

In Figure 1, a schematic diagram of the model implemented in this study (and more accurately presented in its algebraic form in Appendix 1) is depicted in its simplified form for clarity. In the diagram, the oval shapes represent estimated higher-
level effects that influence the observed variables which are represented by rectangular boxes, while the arrows indicate the direction of the relationship and the dashed lines delimit the boundaries between levels. It is appreciated that the outcome variable (with incoming arrows only) "Maths attainment" is measured at level 1 and is influenced not only by the variables at that level (1), i.e. "Maths prior attainment" and "Pupil characteristics", but also by the observed variables at school level (3) "School characteristics" as well as the estimated random effects at all the higher levels, i.e. "Classroom effects", "School effects" and "Local Authority effects", which also have an effect on Maths prior attainment. Prior attainment is also allowed to interact (arrows with incoming arrows in the figure) with pupil (level 1) and school level (3) variables.

Results

Variance components models: Decomposing the variation in Mathematics scores

As outlined in the analytical strategy, the first step in this analysis is to specify the most basic multilevel model, which is the empty model of pupils nested within schools, to build up sequentially towards more complex value-added models. Table 2 shows that the basic 2-level model of pupils nested within schools is significantly improved by the specification of the class level (3-level model), which is indicated by a highly significant chi-squared and a lower AIC. Likewise, the addition of yet another level, that of local authorities, produced a significant improvement to the overall fit of the model.

The addition of an intermediate level, i.e. the classroom level, has reduced the variance of both adjacent levels, i.e. the pupil level (from 0.57 to 0.502) and school level (from 0.51 to 0.47); this finding is consistent with what Martinez (2012) found for US school data, where the addition of the classroom level reduced the variance of both
the pupil and the school level. More generally, this is also consistent with Tranmer and Steel (2001) who found, using the UK census data, that the variance proceeding from an intermediate level was redistributed along the two adjacent levels when ignored. Meanwhile, the addition of the local authority level has further reduced the between-school variance (from 0.47 to 0.396). More importantly, it is clear from Table 2 that a 2-level model significantly overestimates the overall school effects.

(Table 2)

In Table 3, according to the Variance Partition Coefficient (VPC) of the 2-level model, it is estimated that 47.22% of the total variance is due to the variance between schools. An alternative interpretation is that the average correlation between two randomly selected pupils within the same school is expected to be 0.47. This result is yet another way of demonstrating how a single-level analysis would be severely misleading with a very high intra-class correlation; thus the implementation of a multilevel model is highly relevant. This result is also consistent with Munoz-Chereau (2013) who found a between-school variation of 47.3% in an empty 2-level model for Mathematics attainment in the same cohort of pupils.

(Table 3)

On another front, when specifying a 3-level model, the VPCs show that the between-school and between-pupil variation shrink to 44% and 47.1% of the total variation, whereas the variation exclusively explained by the class level is 8.9%. This results in an overestimation of the between-school variation by approximately 7% [i.e. (47.22%-44%)/47.22%*100] and approximately an 11% [i.e. (52.78%-47.1%)/52.78%*100] in the case of the between-pupil variation. In turn, in the 4-level model, the variance partition coefficients show that a relevant proportion of the variation that was believed to be due to between-school variability was actually due to
the variance between Local Authorities (6.5%); whereas the percentage of the total variation due exclusively to between-school variability is further reduced to 37.3%. This means that the school effects had been overestimated approximately by 21%, i.e. \((47.22\%-37.3\%)/47.22\% \times 100\).

These unconditional models are the starting point of the analysis of pupils' and schools' performance; however, they are not truly value-added models, since they do not measure pupils' progress (Goldstein, 1997). In the next section, a measure of prior attainment is included along with measures of socio-economic and demographic characteristics of the schools as well as the pupils in order to specify a contextualised value-added model.

A contextualised value-added model for progress in Mathematics

After specifying the raw value-added model which includes prior attainment as the only explanatory variable, it was found that school value-added estimates are highly correlated with the current attainment scores averaged at the school level\(^5\). As other authors have shown (Ferrão & Goldstein, 2009; Goldstein et al., 2007), this is indeed the result of a misspecification bias, which calls for the necessity of including variables to assess underlying differential school effects (Foley & Goldstein, 2012). Therefore, a somewhat 'true' school value-added model needs to incorporate socio-economic and demographic variables at the school and the pupil level.

Table 4 presents the estimated fixed-effects parameters of the traditional CVA model (pupils nested within schools) and the extended CVA model (pupils nested within classrooms within schools within local authorities). When comparing the parameters in Table 4, the main difference observed between the two models in both pupil and school-level main effects regards the size, while the direction of the relationships is maintained. In regards to the interaction effects, the significant
coefficients also varied only in size and not in direction; the only exception to this was the interaction between prior attainment and attending a subsidised independent school. This may indeed be a consequence of controlling for geographical differences with the addition of the local-authority level. Having noted the differences in the estimated fixed effects from both models, there is no need for separate interpretation of the coefficients of the 2-level model.

Given that significant effects were found for interactions involving all the explanatory variables, the main effects of the fixed part of the model are not to be interpreted in their own right, but in combination with the variables where the interactions have been found to be significant. For instance, the main effect of gender is significant although it varies by year repetition, income and prior attainment. Likewise, the main effect of year repetition is also significant but moderated by gender and prior attainment.

(Table 4)

Regarding the interaction effects, it is observed that the effect of prior attainment varies significantly according to gender, being held back, school SES and school type. In the case of gender, it can be appreciated that male pupils make slightly less progress than female pupils, or in other words, the relationship between prior and subsequent attainment in Mathematics has a steeper slope in the case of female pupils. This relationship can be more easily appreciated in Figure 2, where it can be observed that male pupils are more advantaged than female pupils, but the gap decreases as prior attainment increases. In the case of year repetition, being held back diminishes the effect of prior attainment and thus pupils who have been held back in primary school are expected to make less progress than pupils who have not been held back. Gender also significantly interacts with year repetition. When comparing males and females
who have been held back in one or more grades, it is observed that boys do significantly better than girls, so the effect of being held back is greater for girls.

(Figure 2)

With regards to the cross-level interaction effects (interactions between school-level and pupil-level variables), it is found that pupils in middle SES and upper-middle SES schools are expected to progress more than pupils in low-SES schools. In Figure 3, it is observed that pupils in middle-SES schools have a slightly steeper slope than pupils in lower-middle and low-SES schools, which increases as prior attainment increases. The same occurs for pupils in upper-middle SES schools who reduce the gap between themselves and pupils in Upper-SES schools as prior attainment increases. On another front, when comparing state-funded schools to subsidised independent schools, the effect of prior attainment is stronger on the latter. More interestingly, the effect of prior attainment does not vary between independent schools and state-funded schools.

(Figure 3)

While these results contribute in explaining pupils' progress in Mathematics and school value-added, the random part of the models can also reveal relevant differences in prior achievement at the higher levels of the structure. As other authors have pointed out (Goldstein et al., 2007; Leckie, 2009; Rasbash et al., 2010), the complexity of the network of effects on pupils' performance has not been traditionally taken into account in educational research and government policy, which can be misleading since school effects are usually overestimated in simpler models. A more realistically complex approach to analyse pupils' performance and school value-added includes specifying random coefficients for those covariates that are reasonably believed to vary randomly across the higher levels. In this study, random coefficients for prior attainment at the levels of classrooms, schools and local authorities have been estimated for the extended
CVA model. The random part of the extended CVA model is analysed and compared to the random part of the traditional CVA model in the next section.

**Exploring heterogeneity in the relationship between prior and subsequent attainment across classrooms, schools and local authorities**

As part of the analysis, pupils' prior attainment in Mathematics (level 1 variable) was allowed to vary randomly at the higher levels, i.e. at the level of classrooms, schools and local authorities. This is performed in order to control for heterogeneity within and between schools and geographical differences in prior attainment. These random effects also allow unveiling differential school effects. The estimated random effects from this model are compared to those obtained when fitting the CVA model in the standard 2-level structure in table 5.

(Table 5)

In Table 5, neither the variance nor the covariance terms can be interpreted separately at any level, except for the pupil-level, which only has an intercept variance. Having noted that, it is best to give more attention to the correlations estimated for the variance terms at each level. At the level of the local authorities, the correlation between the intercept and the slope for prior attainment is very high (0.887), which indicates that the higher the intercept for a particular local authority, the steeper the slope of prior attainment. This implies that geographical inequalities in Maths attainment that are prevalent since primary school increase during secondary schooling.

From the correlation coefficients at the school level, it can be seen that the correlation between the intercept and the slope for prior attainment in the extended CVA model is rather weak (0.122); however it does indicate that schools with high intercepts tend to have slightly steeper slopes for prior attainment. Higher-achieving schools have a slightly better chance of making a greater contribution towards the
progress of their pupils than lower-achieving schools. It is also appreciated that the correlation between the slope of prior attainment and the school intercept is quite higher (0.266) in the case of the traditional CVA model. This implies that the diverse rates of progress that schools have are more related to classrooms differences and geographical differences than to differences between the schools themselves.

At the classroom level, the correlation between prior attainment and the school intercept is 0.317, which indicates that the higher the intercept of a classroom within a school, the steeper the slope for prior attainment. This indicates that classrooms with higher average prior attainment are more likely to progress more than other classrooms, although this differential rate of progress is not too pronounced.

At the pupil level, it is observed that the variance in the extended CVA model (0.28) is considerably reduced in comparison to the traditional CVA model (0.307). This indicates that the traditional CVA model erroneously attributes part of the variance in Mathematics progress to pupils' abilities and characteristics, while obscuring that this part of the total variance is more likely to be related to classrooms' characteristics as seen in the empty models in Table 2.

Finally, it is observed that the variance components of the CVA model have been largely reduced when comparing with the random part of the previous models. In the case of the school-level variance, this has been reduced from 0.396 as estimated in the empty 4-level model (table 2) to 0.057 in the extended CVA model, indicating that 85.6%, i.e. \((0.396-0.057)/0.396*100\%\), of the variance in scores at the school-level is accounted for by this model. This implies that the effectiveness of schools is heavily dependent upon prior attainment and non-malleable characteristics of the pupils and the schools. It is also important to note that the estimated variances for the intercepts at all levels do not change substantially when prior attainment is not allowed to vary
randomly across the higher levels (results not shown in this paper), which is why the choice between a random coefficients CVA model or a random intercepts CVA model is not critical.

The differential effects derived from the random coefficients CVA models (traditional or extended) displayed in Table 4, are of the utmost importance when attempting to ascertain how schools are more effective for certain groups of pupils than for others (Kyriakides, 2004). They also contribute to understanding that progress is not uniform across schools, classroom or local authorities, which is also further evidence of within and between schools as well as country-wide inequalities.

As mentioned before, from the random part of the models, it is possible to derive residual estimates that can be used to represent school value-added. Given the differences between the estimates from the random part of the traditional and the extended CVA models observed in Table 5, it is foreseeable that a school accountability system derived from one of these models will differ substantially from the other. In the next section, this is illustrated in more detail.

**A practical application of the extended CVA model to inform school performance**

After fitting the traditional and extended CVA models, the analysis turns to examine school value-added estimates more thoroughly to check for relevant differences in regards to their usefulness to inform school accountability. Residual point estimates and their corresponding standard errors are useful to make comparisons between the higher level units and the overall average. School residuals and their corresponding confidence intervals are estimated by using the shrunken residuals formulae described by Goldstein (2011). Since the residual point estimates at the school level represent the amount of value-added that schools have and considering that the
model assumes that the grand mean is equal to zero, when estimating confidence intervals using standard errors, one is also analysing whether the effect of a particular school on its pupils is significantly different from the national average.

The estimation of these CVA scores allows ranking schools based on their relative effectiveness levels. In Figures 4a and 4b, the X-axis represents the rank of the schools according to their residual point estimates, which are plotted in the Y-axis. Furthermore, the bars represent the 95% confidence intervals for each school's residual point estimate (CVA score) and the line across the 'y' axes at 0 represent the national average. The ranking of estimated school effects depicted in Figure 4a (left) is derived from a 2-level CVA model, while the ranking in Figure 4b (right) is derived from the 4-level CVA model.

(Figure 4)

In Figure 4, it can be appreciated that the estimated CVA scores for the Chilean schools analysed in this paper vary greatly according to the model that is fitted. In the case of the traditional CVA model (Figure 4a, left), it is observed that there is a larger number of schools performing significantly below or above the national average. In contrast, with the extended CVA model (Figure 4b, right), the number of average-performing schools is remarkably larger. These differences are explored in more depth in Table 6.

Based on whether the confidence interval of a school CVA estimate overlaps or not with the national average, 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 Table 6, the classifications derived from the traditional CVA model (2-level model) and the extended CVA model (4-level model) are presented for comparison.

(Table 6)
With the traditional model, 51.25% of schools are classified as average compared to 80.19% classified as average when using the extended CVA model. In the 2-level model, 25.83% of schools are found to outperform the national average, while 22.92% perform below the national average. In a 4-level model, these results are 9.95% and 9.86%, respectively. On another front, both models agree on 70.65% of the school classifications as noted in Table 6.

This shows that there are considerable differences in the way in which schools are classified between the traditional CVA model and the extended CVA model. One could argue that the traditional model misclassifies a large proportion of schools at both ends of the distribution and hence it is misleading, but, naturally, there are two sides to this argument. On the one hand, at the lower end of the distribution, many schools are unfairly classified as below average when using a 2-level model, whereas they might be considered as average when using a 4-level model. On the other hand, at the higher end of the distribution, many schools that could be considered as above average in a 2-level model, are classified as average in a 4-level model.

As mentioned earlier in this paper, the main concern of a school accountability system should be fairness to all schools (See for example: Fitz-Gibbon, 1997; OECD, 2008), which implies that schools should ideally be assessed based only on the circumstances over which they have control. Since there is strong evidence of geographical differences (as seen in Table 2), holding everything else in the models constant, one should be inclined to regard the classifications derived from the extended CVA model as fairer than the classifications derived from the traditional CVA model. In sum, these results show that the way in which a CVA model is specified is a sensitive matter, which can have significant policy implications when reporting back to diverse stakeholders.
**Discussion and Conclusions**

In order to analyse progress in Mathematics made by pupils in Chilean secondary schools in 2006, a 4-level random coefficients model has been implemented. Since the implemented model for attainment in Mathematics is conditional upon prior attainment (it measures pupils' progress) and controls for socio-economic variables at the level of pupils and schools, it can be regarded as a contextualised school value-added model (Ray, 2006). With respect to the specific research questions of this paper, it is found that: a) school-level effects are not the only relevant source of variation in Mathematics scores, since variation between classrooms and local authorities is not negligible; b) there are a number of non-malleable educational, socio-economic and demographic characteristics at the level of pupils and schools that significantly affect pupils' progress in Mathematics that need to be controlled for in order to implement a fair school accountability system in Chile; c) pupils' progress in Mathematics significantly varies across classrooms, schools and local authorities; and d) differences between school performance rankings derived from diverse statistical models are not negligible.

Results from this CVA model reveal the extent of the inequalities of the Chilean education system; they show that progress in Mathematics is positively associated with living in a non-low income household, being male, attending a subsidised independent, non-low SES school. On the other hand, progress in Mathematics is negatively associated with living in a low-income household, being held back at least once during primary school, being a female, attending a low-SES school and attending a state-funded school.

The greatest differences are found when comparing the progress of pupils according to the SES of the school they attend to and the level of income of the household they live in. Thus, the largest effect sizes are found in upper SES schools and
middle SES independent schools, where pupils are expected to score higher than pupils in lower SES schools and middle SES state-funded schools.

It is also found in the CVA models that male pupils progress more than female pupils in general, a tendency that is only remedied when (average) prior attainment is high in a particular school. Furthermore, it is found counter-intuitively that in general pupils in independent schools do not progress more than those in state-funded schools. However, progress must not be confused with attainment. Pupils in independent schools have not attained more than expected, given their socio-economic background and their prior attainment; in other words, they have not progressed more than pupils in state-funded schools. That is independent schools do not seem to add value to their pupils' educational trajectories.

Nevertheless, a word of caution is needed on these results. Differences between schools' estimated value-added scores can partially be the result of a selection process in the school system. This selection process can be reflected upon a correlation between prior attainment and the CVA measures due to school choice based on school averages, which produces bias in the estimated random effects (Ebbes et al., 2004). This is an issue that it is not directly addressed in this paper, but it certainly requires more investigation. In this paper, school selectivity is at least partially controlled for by specifying prior attainment and year repetition as mentioned before.

From a more general perspective, it is worth noticing that previous research in the United Kingdom has found that schools only account for between 5% and 20% of the total variation (Rasbash et al., 2010); however a much larger percentage is found in the Chilean education system, where schools account for near half the variation in Mathematics tests scores; this is a comparison between 2-level empty models.
Variation between local authorities reveals that Chilean schools have different levels of performance according to their geographical location. This, most likely reflects upon a highly unequal distribution of wealth and resources amongst different areas in Chile. Furthermore, this is certainly an outstanding difference with respect to what has been found in UK-based research, where Local Authorities only account for between 1% and 3% of the total variation (Rasbash et al., 2010). These are signs of the massive differences that can be found in a highly unfair and unequal education system such as the Chilean system.

Leckie and Goldstein (2009) have argued that estimated CVA scores have important limitations to inform parents' school choice. According to them standard CVA estimates do not account for the uncertainty of predicting future school performance and hence they are useless for informing parents. They show that when this uncertainty is accounted for by computing standard errors in a non-standard way, the confidence intervals of the estimated future CVA scores are so wide that hardly any school can be distinguished from the national average or even from any other school. This is certainly a limitation from which the models presented here are not free, although as Leckie and Goldstein (2011) point out, performance tables derived from these models are still retrospectively informative and provide highly relevant information for holding schools accountable.

Some authors (Foley & Goldstein, 2012; Leckie & Goldstein, 2011) have argued that the meaning and limitations of these CVA scores and their confidence intervals have not been properly disseminated to the public, which prevents them from properly interpreting and using this information. In order to minimise the downsides of using CVA scores, Leckie and Goldstein (2011) propose a new methodology to make multiple tailor-made school comparisons to inform parents' choice. This methodology involves
simulating the probabilities of pairs of (or several) schools to be ranked higher than the other(s), thus making school performance information more accessible to a non-statistical audience, because of the familiarity with the concept of probability in the wider public. Obviously, this debate is open because no consensus has been reached about how school performance should be presented to the public. However, it is clear from this and other research that the standard approach in educational research fails to disentangle the rather vast complexity of analysing pupils and schools' performance.

In sum, the CVA model presented in this paper extends and improves the traditional 2-level school value-added models insofar as it explicitly assesses the variation between pupils, classrooms, secondary schools and local authorities, which have been shown to be highly significant and crucial to estimate more precise and informative school effects for accountability purposes mainly. More specifically, adjusting for the effect of local authorities allows contextualising the information about school performance within a given territory to which most often parents are limited to choose from and to which local government administrations are limited to intervene. On the other hand, adjusting for classroom effects allows unveiling differences within schools that traditional models obscure; such differences might be relevant for policy interventions and for parents who could make choices based on overall school effects rather than a mixture of effects proceeding from the schools and the classrooms. Parents cannot choose from the latter and local as well school authorities ought to identify within-school differences for intervention, and hence the relevance of distinguishing between the two. Although we have focused on school accountability, these extensions to the traditional CVA models, mainly specifying further levels of variation, constitute non-trivial and non-negligible adjustments when feeding back diverse stakeholders,
because they take a more thorough account of the complexity of the school performance phenomenon.

In light of these apparent differences in pupils' progress due to socio-economic inequalities and within and between-school differences and geographical effects, it is evident that a school accountability system that does not thoroughly consider them will undoubtedly be unfair, especially when considering what the stakes are. The new Chilean Education Quality Assurance System (Sistema Nacional de Aseguramiento de la Calidad de la Educación Escolar, SAC) foresees the utilisation of the school classifications for supporting schools with insufficient performance; however, persistent insufficient performance may result in closure (Agencia de Calidad de la Educación, 2014; San Martín & Carrasco, 2013). The choice of methods is therefore a key player in this high stakes game as seen in this paper and should not be taken as a trivial matter.

The extended CVA model presented in this paper was set up to analyse Mathematics attainment only, but it can be easily implemented on other outcome variables, either separately or simultaneously in a multivariate multilevel model. Other theoretically possible extensions of this CVA model include assessing additional contextual random effects such as neighbourhood and family as well the carry-over effects from primary schools. However, these specifications go beyond the scope of this paper because they are not straightforward extensions to the standard multilevel model and data are not available to analyse all these effects. An upcoming research paper by the authors is in preparation where this CVA model is extended further to assess the random effects of primary schools and all the higher levels specified in this paper on Mathematics and Language attainment using a multivariate cross-classified multilevel model.
Notes

1 The GINI index is a measure of inequality that ranges from 0 to 100, where 0 indicates complete equality and 100 indicates complete inequality. For a more thorough explanation, see: The World Bank (2011)

2 The Human Development index (HDI) is a summary measure of achievements in key areas of Human Development, namely: Health, Education and Income. For a more thorough explanation, see: (Anand & Sen, 1994)

3 The ecological fallacy (Robinson, 1950) is a mistaken conclusion about the individuals of a group drawn from the sole analysis of group (aggregated) data.

4 This is debatable and other variables should be included, such as number of students who have applied to a certain school and have been rejected or number of students who have sat selectivity tests; however, there are no reliable measures of explicit school selectivity practices in the SIMCE database.

5 Detailed results of this step and all the intermediate models described in the analytical strategy are available on request. They are not presented in this paper, because they are not essential to the central argument.
References


Appendix 1. Contextualised value-added model for progress in Mathematics

\[ y_{t(ijk)} = \beta_{0ijk} + \beta_{1ijk} y_{t-1(ijk)} + (X\beta)_{ijkl} \]

\[ \beta_{0ijk} = \beta_0 + f_{0l} + v_{0kl} + u_{0jkl} + e_{0ijkl} \]

\[ \beta_{1ijk} = \beta_1 + f_{1l} + v_{1kl} + u_{1jkl} \]

where:

\[
\begin{pmatrix}
 f_{0l} \\
 f_{1l}
\end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{f0} & \sigma_{01f} \\ \sigma_{01f} & \sigma^2_{f1} \end{pmatrix} \right)
\]

\[
\begin{pmatrix}
 v_{0kl} \\
 v_{1kl}
\end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{v0} & \sigma_{01v} \\ \sigma_{01v} & \sigma^2_{v1} \end{pmatrix} \right)
\]

\[
\begin{pmatrix}
 u_{0jkl} \\
 u_{1jkl}
\end{pmatrix} \sim N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2_{u0} & \sigma_{01u} \\ \sigma_{01u} & \sigma^2_{iu} \end{pmatrix} \right)
\]

\[ e_{0ijkl} \sim N(0, \sigma^2_{e}) \]

Here $y_{t(ijk)}$ corresponds to the standardised Mathematics test scores obtained by 10th grade pupils in Chilean schools in 2006; subscripts "t", "t-1", "i", "j", "k", "l", correspond to occasion 2, occasion 1, pupil, classroom, secondary school and local authority, respectively; $\beta_{0ijk}$ corresponds to the sum of the national average denoted by $\beta_0$ and all the random intercepts (see explanation below); $\beta_{1ijk}$ is the expected increase in the outcome given prior attainment denoted by $y_{t-1(ijk)}$, which is allowed to vary randomly at all higher levels (see explanation below); and $(X\beta)_{ijkl}$ corresponds to a vector of explanatory variables either at the pupils or the schools' level. The random intercepts in this model are the following: $f_{0l}$ corresponds to the effect of residing in a particular local authority; $v_{0kl}$ corresponds to the effect of attending a particular school; $u_{0jkl}$ corresponds to the effect of being assigned to a particular classroom; and $e_{ijkl}$ corresponds to the pupils' heterogeneity or residual variance of the individuals. With respect to the random coefficients in this model: $f_{1l}$ is the random effect of prior attainment for local authority "l"; $v_{1kl}$ is the random effect of prior attainment for
secondary school "k" in local authority "l"; and $u_{1jkl}$ is the random effect of prior attainment for classroom "j" in a secondary school "k" in a local authority "l".

The random intercepts are normally distributed with a mean of zero and a variance to be estimated and when random coefficients are specified at the higher levels, the coefficients (or slopes) are correlated with the random intercepts by a covariance term to be estimated as well.
<table>
<thead>
<tr>
<th>Variable</th>
<th>Type of data</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
<td>Continuous (standardised)</td>
<td>Mathematics test scores obtained by 10th grade pupils in 2006. Scores range from -2.54 to 2.61. 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 250 and a standard deviation of 50.</td>
</tr>
<tr>
<td><strong>Pupil-level variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prior attainment</td>
<td>Continuous (standardised)</td>
<td>Mathematics test scores obtained by 8th grade pupils in 2004. Scores range from -2.96 to 2.93. 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 250 and a standard deviation of 50.</td>
</tr>
<tr>
<td>Gender</td>
<td>Binary variable</td>
<td>Gender of pupils. Coded 0 for female pupils and 1 for male pupils.</td>
</tr>
<tr>
<td>Socio-economic level indicated by household income</td>
<td>Indicator variable (set of dummy variables)</td>
<td>Average monthly household income as reported by parents†. Reference category: Low income (200,000 CLP or less). Other categories: Lower-middle income (between 200,001 CLP and 500,000 CLP); Upper-middle income (between 500,001 CLP and 1,000,000 CLP); and High income (more than 1,000,000 CLP). Income is expressed in Chilean pesos (CLP).</td>
</tr>
<tr>
<td>Retention</td>
<td>Binary variable</td>
<td>Pupils' retention. It indicates whether the pupil has been held back once or more at any grade before the first occasion of the SIMCE tests. Coded 0 for non-held-back pupils and 1 for held-back pupils.</td>
</tr>
<tr>
<td><strong>School-level variables</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>School SES</td>
<td>Indicator variable (set of dummy variables)</td>
<td>School Socio-economic status according to the classification of the Chilean Ministry of Education. Reference category: Low School SES. Other categories: Lower-middle, Middle, Upper-middle and High.</td>
</tr>
</tbody>
</table>

† The parents’ data were gathered by schools and sent as part of the required data for government auditing purposes. This research links the student, school and parental data.
Table 2: Summary of the variance components models

<table>
<thead>
<tr>
<th>Parameter</th>
<th>2-level model</th>
<th>3-level model</th>
<th>4-level model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept</td>
<td>0.018 (0.015)</td>
<td>-0.0003 (0.015)</td>
<td>-0.138 (0.024)</td>
</tr>
<tr>
<td>Random part</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pupil</td>
<td>0.57 (0.002)</td>
<td>0.502 (0.002)</td>
<td>0.502 (0.002)</td>
</tr>
<tr>
<td>Classroom</td>
<td>--</td>
<td>0.095 (0.002)</td>
<td>0.095 (0.002)</td>
</tr>
<tr>
<td>School</td>
<td>0.51 (0.015)</td>
<td>0.47 (0.015)</td>
<td>0.396 (0.014)</td>
</tr>
<tr>
<td>Local Authority</td>
<td>--</td>
<td>--</td>
<td>0.069 (0.013)</td>
</tr>
<tr>
<td>Deviance</td>
<td>469085.2</td>
<td>452933.5</td>
<td>452758</td>
</tr>
<tr>
<td>AIC</td>
<td>469091.2</td>
<td>452941.5</td>
<td>452768</td>
</tr>
<tr>
<td>Number of parameters</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>--</td>
<td>16151.7</td>
<td>175.5</td>
</tr>
<tr>
<td>N</td>
<td>202,094</td>
<td>202,094</td>
<td>202,094</td>
</tr>
</tbody>
</table>
Table 3: Variance partition coefficients of the empty models for Mathematics in Chile

<table>
<thead>
<tr>
<th>Level</th>
<th>2-level model</th>
<th>3-level model</th>
<th>4-level model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pupil</td>
<td>52.78%</td>
<td>47.10%</td>
<td>47.30%</td>
</tr>
<tr>
<td>Classroom</td>
<td>--</td>
<td>8.90%</td>
<td>8.90%</td>
</tr>
<tr>
<td>Secondary School</td>
<td>47.22%</td>
<td>44.00%</td>
<td>37.30%</td>
</tr>
<tr>
<td>Local Authority</td>
<td>--</td>
<td>--</td>
<td>6.50%</td>
</tr>
<tr>
<td>Total</td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
<td><strong>100%</strong></td>
</tr>
</tbody>
</table>
Table 4: Fixed-effects parameters of the traditional CVA model (2 levels) and the extended CVA model (4 levels)

<table>
<thead>
<tr>
<th>Fixed-effects parameters</th>
<th>Traditional CVA (2-level model)</th>
<th>Extended CVA (4-level model)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Pupil-level†</strong></td>
<td>Coef. (s.e.)</td>
<td>Coef. (s.e.)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.276 (0.015)**</td>
<td>-0.318 (0.015)**</td>
</tr>
<tr>
<td>Prior attainment</td>
<td>0.621 (0.006)**</td>
<td>0.582 (0.006)**</td>
</tr>
<tr>
<td>Male</td>
<td>0.056 (0.003)**</td>
<td>0.068 (0.003)**</td>
</tr>
<tr>
<td>Low-mid income</td>
<td>0.017 (0.003)**</td>
<td>0.012 (0.003)**</td>
</tr>
<tr>
<td>Up-mid income</td>
<td>0.031 (0.005)**</td>
<td>0.024 (0.005)**</td>
</tr>
<tr>
<td>High income</td>
<td>0.049 (0.008)**</td>
<td>0.041 (0.008)**</td>
</tr>
<tr>
<td>Held back</td>
<td>-0.289 (0.008)**</td>
<td>-0.257 (0.008)**</td>
</tr>
<tr>
<td><strong>School-level†</strong></td>
<td>Coef. (s.e.)</td>
<td>Coef. (s.e.)</td>
</tr>
<tr>
<td>Lower-middle SES</td>
<td>0.084 (0.018)**</td>
<td>0.115 (0.018)**</td>
</tr>
<tr>
<td>Middle SES</td>
<td>0.263 (0.02)**</td>
<td>0.329 (0.021)**</td>
</tr>
<tr>
<td>Upper-middle SES</td>
<td>0.47 (0.024)**</td>
<td>0.549 (0.025)**</td>
</tr>
<tr>
<td>Upper SES</td>
<td>0.664 (0.045)**</td>
<td>0.746 (0.047)**</td>
</tr>
<tr>
<td>Subsidised independent</td>
<td>0.036 (0.016)*</td>
<td>0.061 (0.015)**</td>
</tr>
<tr>
<td>Independent</td>
<td>0.011 (0.042)</td>
<td>0.026 (0.043)</td>
</tr>
<tr>
<td><strong>Interaction effects†</strong></td>
<td>Coef. (s.e.)</td>
<td>Coef. (s.e.)</td>
</tr>
<tr>
<td>Prior attainment&amp;Male</td>
<td>-0.008 (0.003)**</td>
<td>-0.01 (0.003)**</td>
</tr>
<tr>
<td>Prior attainment&amp;Held back</td>
<td>-0.082 (0.005)**</td>
<td>-0.067 (0.005)**</td>
</tr>
<tr>
<td>Male&amp;Held back</td>
<td>0.083 (0.009)**</td>
<td>0.078 (0.009)**</td>
</tr>
<tr>
<td>Prior attainment&amp;Low-mid SES</td>
<td>-0.003 (0.007)</td>
<td>0.0001 (0.007)</td>
</tr>
<tr>
<td>Prior attainment&amp;Middle SES</td>
<td>0.026 (0.008)**</td>
<td>0.044 (0.008)**</td>
</tr>
<tr>
<td>Prior attainment&amp;Up-mid SES</td>
<td>0.021 (0.01)*</td>
<td>0.049 (0.01)**</td>
</tr>
<tr>
<td>Prior attainment&amp;Upper SES</td>
<td>-0.018 (0.021)</td>
<td>0.017 (0.02)</td>
</tr>
<tr>
<td>Prior attainment&amp;Subs. independent</td>
<td>-0.018 (0.006)**</td>
<td>0.013 (0.005)**</td>
</tr>
<tr>
<td>Prior attainment&amp;Independent</td>
<td>-0.006 (0.02)</td>
<td>0.021 (0.019)</td>
</tr>
<tr>
<td><strong>Model fit information</strong></td>
<td>Traditional CVA</td>
<td>Extended CVA</td>
</tr>
<tr>
<td>Deviance</td>
<td>308,971</td>
<td>300,478</td>
</tr>
<tr>
<td>AIC</td>
<td>309,029</td>
<td>300,560</td>
</tr>
<tr>
<td>Number of parameters‡</td>
<td>26</td>
<td>32</td>
</tr>
<tr>
<td>Chi-squared</td>
<td>--</td>
<td>8,493</td>
</tr>
<tr>
<td>N</td>
<td>181,867</td>
<td>181,867</td>
</tr>
</tbody>
</table>

* p<0.05    ** p<0.01    *** p<0.001
† Reference categories: Female; Low income; Not held back; State-funded school and Low school SES.
‡ Includes random-effects parameters.
Table 5: Comparison between the random-effects parameters of the traditional CVA model (2 levels) and the extended CVA model (4 levels)

<table>
<thead>
<tr>
<th>Parameters†</th>
<th>Traditional CVA</th>
<th>Extended CVA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coef. (95% C.I.)</td>
<td>Coef. (95% C.I.)</td>
</tr>
<tr>
<td>Level 4: Local Authority</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Var(Intercept)</td>
<td>--</td>
<td>0.007 (0.004; 0.01)</td>
</tr>
<tr>
<td>Cov(Prior att./Intercept)</td>
<td>--</td>
<td>0.002 (0.001; 0.003)</td>
</tr>
<tr>
<td>Var(Prior attainment)</td>
<td>--</td>
<td>0.001 (0.001; 0.001)</td>
</tr>
<tr>
<td>Level 3: School</td>
<td>Coef. (95% C.I.)</td>
<td>Correlation</td>
</tr>
<tr>
<td>Var(Intercept)</td>
<td>0.076 (0.072; 0.081)</td>
<td>--</td>
</tr>
<tr>
<td>Cov(Prior att./Intercept)</td>
<td>0.006 (0.004; 0.007)</td>
<td>0.266</td>
</tr>
<tr>
<td>Var(Prior attainment)</td>
<td>0.006 (0.005; 0.006)</td>
<td>--</td>
</tr>
<tr>
<td>Level 2: Classroom</td>
<td>Coef. (95% C.I.)</td>
<td>Correlation</td>
</tr>
<tr>
<td>Var(Intercept)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Cov(Prior att./Intercept)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Var(Prior attainment)</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Level 1: Pupil</td>
<td>Coef. (95% C.I.)</td>
<td>Correlation</td>
</tr>
<tr>
<td>Var(Intercept)</td>
<td>0.307 (0.305; 0.309)</td>
<td>--</td>
</tr>
</tbody>
</table>

† 95% confidence intervals are reported for informational purposes only. Significance has been assessed via LR test as described in the methods section.
Table 1: Comparison of school classifications in traditional (2 levels) and extended (4 levels) CVA models

<table>
<thead>
<tr>
<th>Traditional CVA model</th>
<th>Below average</th>
<th>Average</th>
<th>Above average</th>
<th>Total</th>
<th>Percentage‡</th>
</tr>
</thead>
<tbody>
<tr>
<td>Below Average</td>
<td>237</td>
<td>323</td>
<td>0</td>
<td>560</td>
<td>22.92%</td>
</tr>
<tr>
<td>Average</td>
<td>4</td>
<td>1,247</td>
<td>1</td>
<td>1,252</td>
<td>51.25%</td>
</tr>
<tr>
<td>Above average</td>
<td>0</td>
<td>389</td>
<td>242</td>
<td>631</td>
<td>25.83%</td>
</tr>
<tr>
<td>Total</td>
<td>241</td>
<td>1,959</td>
<td>243</td>
<td>2,443</td>
<td>100%</td>
</tr>
</tbody>
</table>

| Percentage†           | 9.86%         | 80.19%  | 9.95%         | 100%  |

Note: The diagonal shows the agreement between the models. 1,726 (70.65%) schools remain in the same category.
† Within traditional CVA model classifications
‡ Within extended CVA model classifications
Figure 1: Diagram of the CVA model implemented in this study
Figure 2: Interaction effect between gender of pupils and prior attainment on subsequent attainment in Mathematics
Figure 3: Cross-level interaction effect between prior attainment and School SES on Mathematics scores
Figure 4: Comparison of school rankings based on CVA scores derived from a 2-level model and a 4-level model

Note: The reference line at 0 in the Y-axis of both figures represents the national average. A school is different from the mean when its 95% CI does not overlap with this reference line.