Comunicaciones en Estadística
Statistics on the evaluation of education
June 2016, Vol. 9, No. 1, pp. 9–37

Value-added in primary and secondary education: Following cohorts over time [1]

Valor agregado en la educación primaria y secundaria: siguiendo cohortes en el tiempo

Jorge Duarte[a]
jduarte@contratista.icfes.gov.co

Silvana Godoy Mateus[b]
sgodoy@icfes.gov.co

Ximena Dueñas Herrera[c]
xduenas@icfes.gov.co

Abstract

Two Value-Added models were fitted using cohort comparisons on the Saber 3°; 5° and 9° test that is applied by the ICFES. The first one, which is only based on the results of the initial year to estimate growth in the final year, provides a measure in absolute terms of educational gains of students within a school. The second, controlling for several demographic, social and economic characteristics, cleans the result of conditions that the school cannot control, to give an estimate of the particular effect of the school on student learning. The analysis results show that economic conditions of students are strongly related to educational progress, but there are schools that, despite having underperforming students, have high-performance levels. These schools can maintain their high-achieving students at these levels but fail to leverage its low-performing students with other students.

***Keywords*:** value-added, education, standardized tests.

Resumen

Usando comparaciones de cohortes en las pruebas Saber 3°, 5° y 9° que realiza el Instituto Colombiano para la Evaluación de la Educación (Icfes) se estimaron dos modelos de valor agregado. El primero, que solo se basa en los resultados del año inicial para estimar el crecimiento en el año final, brinda una medida en términos absolutos de la ganancia educativa de los estudiantes dentro de un

[1] Duarte, J., Godoy, S., Dueñas, X. (2016) Value-added in primary and secondary education: Following cohorts over time. *Comunicaciones en Estadística*, **9**(1), 9-37.

[a] Economist, Subdivision of analysis and dissemination, Icfes, Colombia.
[b] Deputy Director, Subdivision of analysis and dissemination, Icfes, Colombia.
[c] General Director, Directorate General, Icfes, Colombia.

colegio. El segundo, que controla por varias características demográficas, sociales y económicas, limpia el resultado de condiciones que el colegio no puede controlar, para dar una estimación del efecto exclusivo del colegio sobre el aprendizaje de sus estudiantes. Los resultados del análisis muestran que las condiciones económicas de los estudiantes se relacionan fuertemente con el progreso educativo, pero que hay colegios que a pesar de tener bajo rendimiento tienen estudiantes en niveles de desempeño alto. Estos colegios logran mantener a sus estudiantes de desempeño alto en estos niveles, pero fallan en nivelar a sus estudiantes de bajo desempeño con el resto de sus estudiantes.

***Palabras clave*:** valor agregado, educación, pruebas estandarizadas.

1 Introduction

In the education field, the concept value-added (VA) refers to the students' achievement regarding their increase of knowledge, skills, capacities and other qualities obtained as an outcome of their experiences within an educational system over time (OECD 2008). It could also be defined as the effect magnitude of a certain school in the educational outcomes of its students, which go beyond their scores and affect their future opportunities (Saunders 1998).

From the concept VA comes the idea of VA models. These can on in a category of statistical models using data from the educational achievement of students over time, to measure learning gains (Doran & Lockwood 2006). Similarly, VA measures can be defined as the value added by a school to its students above what would be expected according to their background and previous achievements (Hill 1995).

There are important differences between the aim of VA models and the hypothesis of cross-sectional analysis with standardized tests data. Several studies use data during a period to compare students, schools, regions and other attachments. These studies reveal valuable information to know in absolute terms the student's performance level. Another kind of analysis uses data of a grade to make comparisons over time with the same grade. As long as learning standards and test design are maintained, these temporal analyses allow comparisons between different cohorts. On the contrary, VA models follow the same cohort at two points in time to make the change from the starting point to the arrival point.

The OECD identifies two main benefits of VA models. First, VA measures can be "fairer" than other tests because the result in VA takes into account the initial level of students (OECD 2008, Doran & Izumi 2004). It may be unfair to evaluate the contribution of a school to the achievement of its students by focusing only on the output conditions of these or the percentage of students who fulfill certain standards because the conditions of entry to schools are not equal (Reardon & Raudenbush 2009). In Colombia, the student performance varies widely depending on different socioeconomic conditions. VA measures, rather than showing the output performance of students, reveal the magnitude of changes over time, given the conditions of entry.

The second benefit of VA measurements, which are mentioned the OECD is that they can be 'more accurate' because they incorporate contextual characteristics of the schools and their students. Cross-sectional test scores provide crucial information for a variety of objectives of general interest, but they don't allow distinguishing the real contribution of schools with other features that also affect school performance. VA models can take into account social, economic and demographic conditions and other contextual factors of schools and students to generate a measure of the schools that recognize these characteristics.

Saber tests 3, 4, 5 and 9 offer the possibility of making estimates of VA. These tests evaluate every student in Colombia, with an annual periodicity since 2012. Among the results every school receives is a measure of central tendency of their performance and four measures of the distribution of its students regarding their performance. This distribution is based on four performance levels defined according to the educational standards of the Ministry of National Education (MEN) for each area and grade. To date it's possible to follow two cohorts:

> Students that were in third grade of primary school in 2012 and fifth grade of the same educational cycle in 2014.
>
> Students that were in fifth grade of primary school in 2009 and ninth grade of secondary school in 2013. For each of the two cohorts, it is possible to obtain results for the areas of mathematics and language.

The practical exercise of this document estimates two VA models. The first, which is only based on the results of the initial year to estimate growth in the final year, provides a measure in absolute terms of educational gains of students within a school. The second seeks to give an estimate of the particular effect of the school on student learning by controlling through several demographic, social and economic characteristics. We believe that both results are of interest to educational policies. The first model can be used to identify schools with low performance, with the aim of targeting aid or take plans of action. It is important to determine these schools, although they are not solely responsible for the results. The second model can help reward or recognize schools that, despite the characteristics of their students, manage to get a good performance.

After reviewing the literature, we can say that it could be the first approach to VA models for testing Saber 3, 5 and 9. Results should be interpreted with caution because they do not have student-level results; we cannot guarantee that students at the beginning of a period are the same as at the end (problems like dropouts, grade repetition, etc.)

However, we consider that the information of VA, accompanied by other information of schools can be a useful instrument of public policy. The document is divided into five sections including this introduction. In the next secti,n, we briefly describe the implications of the VA estimates. In the next two sections, we expose the methodology and show results. Finally, we conclude with a discussion and aspects to improve in future research.

2 Public policy implications

The school of school effectiveness is born by the need to identify the direct contribution of schools to its students' performance (OECD 2011, Downes & Vindurampulle 2007, Kim & Lalancette 2013). The first statistical analyses seeking this goal used the average of students from a school to make direct comparisons. Subsequent researches on the determinants of educational attainment showed that the performance of schools was related to the socioeconomic conditions of students and therefore, that the comparison of these gross measures did not recognize the contribution of schools in an accurate way (Hill 1995, Kim & Lalancette 2013). Raudenbush & Bryk (1986) make a pioneering study of VA with hierarchical linear models to analyze school effectiveness, taking into account not only the socioeconomic characteristics of students but also the previous achievements Haveman & Wolfe (1995) y Kim & Lalancette (2013).

Currently, VA models can be defined as tools describing the learning process. These models attempt to separate the contributions of the school in the development of students, as distinct from the contributions of other factors associated with performance (Willms 2009). Despite the many theoretical and practical definitions, all models of VA ensure that the evaluation of the effectiveness of schools takes into account every external influence, besides the socioeconomic conditions of students, their family history, and previous achievements (Wyatt 1996).

VA Models has been used for different purposes. In general, they have been used to evaluate, monitor and improve various aspects of an educational system (OECD 2008). These models allow identifying the schools that contribute most to the educational experience of its students, with the caution not to penalize or reward them for the conditions and background of their students (Dury & Doran 2003). VA models can be used as a tool for school improvement because they indicate how the academic performance of students is progressing regarding the performance of comparable students (OECD 2011, Downes & Vindurampulle 2007). Through VA models it is possible to identify schools with performance above and below projections, so that those schools with difficulties follow the practices of those who have relatively high-performance (OECD 2011, Downes & Vindurampulle 2007).

In the United States information systems increasingly allow to identify and relate students to their teachers. This possibility has allowed VA models measure the specific contribution of teachers to the progress of their students and, consequently, that results can have important implications for them. This is the most controversial practice of the use of VA models (Koedel et al. n.d., Aaronson et al. 2007, Rivkin 2005). In Louisiana, New York and some counties from Tennessee, models are used to evaluate the effectiveness of the teacher preparation programs, to help with their improvement and study the teachers' distribution in poor schools (Armour-Garb 2009). However, the biggest criticisms come from Florida, Denver and other counties in Tennessee, where VA is used to define the permanence and payments of teachers. Teachers' unions in New York and California have banned any high-impact decision for teachers based on VA and particularly on standardized tests to students.

Groups that are in favor for high-impact decisions based on VA claim that these measures are less expensive than other mechanisms (such as observational evaluations of teacher performance) and, if the models are correctly specified, they can generate better results. Critics of these methods think that the estimates are not reliable enough to categorize teachers. We believe that practical exercises, such as this document, serve to publicize the advantages and disadvantages of these procedures and open discussion of what for and how the results can be used in the Colombian context.

Despite the many applications of VA models, the use of statistical models to measure the learning marginal gains generate several practical difficulties that, although theoretically understood, are difficult to answer empirically. To obtain accurate estimates appropriate specifications models should be used, given the properties of the data available and the public policy objectives (OECD 2008). In the next section, we present our methodology.

3 Methodology

For the empirical exercise, we used the ICFES Saber tests for the grades 3, 5 and 9, from the years 2009, 2012, 2013 and 2014, for the areas of mathematics and language. These tests evaluate all schools in the country. For each of the tests in these years, grades and areas we use the average score of school and the percentage of students in the four performance levels (insufficient, minimum, satisfactory and advanced).

Performance levels are defined according to educational standards of MEN and, therefore, provide a measure of distribution regarding the students' learning within the school. To reduce the number of response variables without losing information of distribution, we added the percentage of students at the minimum and insufficient performance levels to have a measure of the distribution's tail (it is, of the students with the most undesirable scores of the school). We excluded the satisfactory level and used the advanced level as the measure of the distribution's top. We made this decision because on average, distribution is skewed to the right. In this way we count on three measures of interest: the average score which gives a measure of the average performance of the school's students; the sum of the percentages of the students in insufficient and minimum levels, which shows the behavior of students in the tail of the distribution of school; and finally, the percentage of advanced, which shows the behavior at the top of the distribution.

The tests Saber 3, 5 and 9 do not give a particular score for the students. The reason is that every student presents only one part of the test for some of the evaluated areas to estimate a score for the school. This implies that no record identifies students in time. Therefore, the observation unit of our analysis is schools. This fact makes that, although we can follow a cohort specified in time, we cannot guarantee that students evaluated in two periods are exactly the same. This restriction can bias the estimates of the calculations in both directions. For example, if a school had excellent students in the fifth grade (primary school) and four years later, in the ninth grade (secondary school), had other not so good students, it may seem that there was no improvement in the school's performance. In the opposite case, the bias would be positive. Therefore the validity of our VA models with school level-observations depends on the strong assumption that the vast majority of students do not change their schools. This assumption cannot be verified because even knowing the proportion of students in the final year regarding the initial year, we do not know if they are the same students. The assumption could be unimportant if it is assumed that the students received by the school have similar conditions to the students it loses.

With the data we have, it's possible to follow two cohorts in time (see Table 1). The first cohort we studied corresponds to the students who were in the third grade of primary school in 2012 and two years later, in 2014, in fifth grade. We called this cohort Third 2012 to Fifth 2014. The second cohort corresponds to the students who were in the fifth grade of primary school in 2009 and four years later, in 2013, in ninth grade of secondary education. We called this cohort Fifth 2009 to Ninth 2013. The cohort Third 2012 to Fifth 2014 is of considerable interest because it is the one that guarantees the assumption of student stability in schools since both scores belong to the same school cycle.

Cohort Fifth 2009 to Ninth 2013 is the one that guarantees the least this assumption, but it's of interest because there is a change in the school cycle. Since 2012 Saber tests 3, 5 and 9 have been applied with an annual periodicity and since next year the test will be applied to seventh grade, so it will significantly expand the possibility of following cohorts over time.

Table 1: *Cohorts. Source: own elaboration.*

2009	2010	2011	2012	2013	2014
					Third
				Third	Fourth
			Third	Fourth	Fifth
		Third	Fourth	Fifth	Sixth
	Third	Fourth	Fifth	Sixth	Seventh
Third	Fourth	Fifth	Sixth	Seventh	Eight
Fourth	Fifth	Sixth	Seventh	Eight	Ninth
Fifth	Sixth	Seventh	Eight	Ninth	Tenth
Sixth	Seventh	Eight	Ninth	Tenth	Eleven
Seventh	Eight	Ninth	Tenth	Eleven	
Eight	Ninth	Tenth	Eleven		
Ninth	Tenth	Eleven			
Tenth	Eleven				
Eleven					

We used two different models for each of both cohorts, each of two areas and each of three scores. The first one, which we call growth model is a simple linear regression where the dependent variable is the final year's score and the independent the initial year's score. The second one, which we call VA model is a linear regression of fixed effects and interactions on two levels controlled by several contextual characteristics.

3.1 Growth Model

If the tests applied in the initial year and the final year were the same, the difference between both scores would be a measure of the change in the educational attainment, or growth, in the period existing between the two tests. As the Saber tests 3, 5 and 9 evaluate specific competencies in each school grade, tests between grades are not comparable. For example, if a school has excellent results in the initial year and the final year they are not so good, we can't affirm that it worsened because it possibly improved compared to the first test, but not enough to have a good performance in the second one. To know if, in fact, it worsened, the test of the initial year should to to students in the final year for the second time.

For this reason, the difference cannot be made between the result of the final year and the one of the second year, but between the results of the final year and an estimate of the final year's score based on the initial year's score. In other words, growth is the difference between the result reached by school and the expected result. We estimate using ordinary least squares (OLS) the following model:

$$F_i = \alpha_0 + \alpha_1 I_i + \epsilon_i \tag{1}$$

Where $i = 1, 2, \cdots, N_i$ indicates the number of schools. F_i shows the score of i in the final year, where the score is the average score, the sum of the percentage of students in the performance levels insufficient and minimum or the percentage of students at the advanced performance level. That is, three different models, one for each dependent variable are estimated. I_i Ii indicates the initial score of i. α_0 and α_1 are the regression coefficients and ϵ_i is the error term.

Prediction of F_i, $\hat{F_i}$, is the expected result of school for the final year, given the initial year and given the behavior of schools from all around the country. The difference $F_i - \hat{F_i}$ is the growth of the school i.

The growth model shows the marginal contribution of a school in time without being adjusted by other contextual variables, in other words, it shows an absolute measure of the change in the students' educational attainment of a school. We consider this model relevant because of two main reasons. The first is that its simplicity is an advantage in the light of public opinion; the second is: since the result is not affected by other different variables than those of the scores, comparisons between schools can be made without other considerations and based solely on the actual change.

3.2 Value Added Model

One of the VA advantages is that it's possible to include contextual information of the students and the school to make measures that take into account inequities or can separate the effects of schools of other characteristics from the students. Even controlling the students' initial level it is reasonable to think that changes in the educational attainment are affected by other variables outside the school. For example, parents who are the most committed to their children education, students with more food or more technology, the relevance of educational standards in some parts of the country, extracurricular student activities, minority groups, religious beliefs, disabilities and motivations of the students, among others. To this end and with the restriction of available data, the VA model includes other regressors with contextual characteristics of the school and its students.

We use the identification of Certified Local Authorities (ETC) as one of the fixed effects of the model. Law 715 of 2001 gives the category of ETC to all departments and municipalities that counted more than one hundred thousand inhabitants in 2002. Municipalities with fewer inhabitants who have the ability to administer the public education service can apply for certification if they meet certain technical, administrative and financial requirements. These requirements are that the municipal development plan is consistent with the national education policies, having state educational establishments organized to provide at least complete primary school, having teachers and executive staff defined according to national parameters and having institutional capacity to operate the information system of the education sector. Municipalities with more than one hundred thousand inhabitants must meet these requirements to follow certificates (decrees 2700, 2004 and 3940 2007). We consider that the ETC offers a better regional control than the municipality or department because it maintains the departmental identification, which is an important regional feature, but it differences within the department those municipalities with the above features. For this exercise, we used the 94 existing ETC until 2014. Since 2015 the municipality of Yumbo, Valle del Cauca, is certified.

As a second level of fixed effects, we used the area (rural or urban) and the sector (official or non-official). Besides regional differences, the area and the sector factor capture defining characteristics for education. In Colombia, and generally in less developed countries, substantial differences between rural and urban areas are maintained, regarding socioeconomic status of families, the quality of education and dropout (Colbert 1999, Kazeem & Stokes 1995, Zhang 2006). Regarding the sector, non-official schools supply the demand that official reach not cover. This requirement comes from parents seeking better education for their children or parents who found no quota in public schools. Therefore, although on average non-official schools perform better than the officials or public schools; this feature is not a control for school performance (there are official schools with excellent performance and non-official with poor performance). The sector is important because there are marked differences in teacher incentives and curricular plans (Nunez et al. 2002, Tobón et al. 2008).

As the socioeconomic level approach to school, we use the index of the socioeconomic level (INSE) calculated by the ICFES. INSE is built with information from the socio-demographic survey that students respond at the time of the application of Saber tests. Calculation takes into account household's composition, infrastructure and possessions or belongings of housing, interaction with parents (talk about news, books or television shows) and a cultural component (in terms of its assistance to activities like theater, exhibitions, fairs, carnivals, parks, circuses, storytellers or puppets, libraries and cinema). This classification is particular to the educational field and different than other socioeconomic measures with other objectives; therefore the INSE is not comparable with poverty estimates of the National Administrative Department of Statistics (DANE).

INSE is calculated for each student, while schools are classified into four socioeconomic levels (NSE) based on the INSE. For our exercise, to have a continuous variable, we take the average per school of INSE as an explanatory variable. We use the four categories of NSE to present descriptive statistics in the following section. We include a measure of the stability of the cohort, defined as the difference between the number of students who took the test in the final year and the number of students who took the test in the initial year. This measure does not guarantee that students are the same. Finally, we use the average score's standard deviation of students of an individual school as a measure of the learning's dispersion within it. With these variables we estimate the following linear model of fixed effects and interactions at two levels:

$$F_{ijk} = \beta_0 + \beta_1 I_{ijk} + \beta_2 INSE_{ijk} + \beta_3 EST_{ijk} + \beta_4 DE_{ijk} + v_{1j} I_{ijk} ETC_j + \pi_{1k} I_{ijk} ZS_k + v_{2j} + \pi_{2k} + \mu_{ijk}$$

Where, $j = 1, 2, 3, N_j$ indicates the ETC number and k indicates zone and sector of i with k = rural official, urban official, rural non-official, urban non-official . $INSE_{ijk}$ is the $INSE$ of school i which belongs to j and to k, Est_{ijk} is the stability measure of the cohort of i and DE_{ijk} is the standard deviation of average score of i. ETC_j identifies each ETC and ZS_k zone and sector. $\beta_0, \beta_1, \beta_2, \beta_3$ and β_4 are the regression coefficients common to all schools. v_{1j} y v_{2j} are the slope and constant coefficients respectively common to the ETC and π_{1k} and π_{2k} are the coefficients common to every zone and sector. We used interactions because regression slopes must also be specific for the ETC levels and as well for the zone and sector levels. All regressions of the model of growth and the VA model are shown in Annex 1.

The model was estimated by fixed effects to solve the problem of endogeneity of I_{ijk}. Model does not include many contextual variables of students uncorrelated with I_{ijk} that are collected in μ_{ijk}. This means that β_1 is biased and the VA measure includes effects that are not attributable to school. By using fixed effects, we include in the model all the features of the ETC and area and sector that are constant along the time and are correlated with I_{ijk}, correcting, at least partially, the endogeneity problem. The model does not use random effects because these models are based on the assumption that $E[v_{2j} + \pi_{2k} + \mu_{ijk} \mid I_{ijk}, INSE_{ijk}, Est_{ijk}, DE_{ijk}] = 0$, , which is even harder to guarantee.

The assumption that should be guaranteed in order to obtain no biased estimates is that $E[\mu_{ijk} \mid I_{ijk}, INSE_{ijk}, Est_{ijk}, DE_{ijk}] = 0$. It is difficult to meet this assumption because almost the only way to ensure it is using data that come from a controlled experiment. That is, to know the effect of a school in the performance of their students, students with similar characteristics should be spread randomly in different schools and evaluated at the beginning and end of a period. Although we cannot guarantee that our model meets this assumption, we believe we have good information.

4 Results

Growth and VA estimates must be compared between schools or aggregates (municipalities, ETC and departments) with a similar performance in the initial year test, for two reasons. The first one is a mathematical reason. At least theoretically, schools with the worse performance in the initial year are the ones who have the bigger mathematical possibility to growth during the period. This fact may make it seems that, if schools are not separated by the result of the initial year, that a school with good performance in both periods adds less value to their students than a school that improved but had a weak performance in the initial year. As the mathematical possibility of improving is different than the actual probability of doing so, if a school with weak performance significantly improved in the period, it is appropriate to highlight this growth among schools with similar performance.

The second reason is that the learning difficulty can increase as more is learned.

For this reason, growth in learning should not be valued equally at different levels of this. For example, a low-performance school that increased one hundred points of the test does not generate the same value than a high-performance school that increased the same one hundred points. It was probably harder for the high-performance school to achieve this growth. Our models, especially the model of growth, do not take into account these marginal changes. Although the VA model bases its comparisons between similar schools, it is still a linear model. For both reasons, we divided the average score of the initial year of school performance in three categories: low, medium and high. We chose the categories so that in the early years of both cohorts and both subjects, on average, 40% of schools stay in poor performance, 40% in medium and 20% in high. This is equivalent to low-performing schools having scores between 100 and 290, medium performance schools between 291 and 350, and high-performance schools between 351 and 500. Estimates are comparable within these categories and not comparable between them.

Figure 1 shows an example of the six resulting quadrants of VA model for the average score of schools belonging to the cohorts Third 2012 to Fifth 2014 in the area of mathematics. Vertical comparisons can be made between schools within the same performance category. Besides, names could be given to the quadrants to give a qualitative measure of both dimensions. Categories of low performance and declining VA stand out because they contain the schools with the bigger learning deficiencies, and the category of high-performance schools and growing VA, because they are the schools with the best results.

Figure 2 shows the VA model example when the interest variable is the sum of the percentage of students in the insufficient and minimum performance levels. As the cut for the three categories of performance is based on the average score, there is not a perfect match between the court and the percentage of students in performance levels. For example, Figure 2 shows that there is a school ranked in high performance, despite having more than 40% of their students in minimum and insufficient. This school must have the rest of his students at a good performance level to have an average score greater than 350 and be ranked as a high performance school. Relevance of performance levels lies in the fact that if two schools with same average score have their students classified in different performance levels, the dispersion of student learning is different. In addition to analyzing the school's average behavior, it is important to know the dispersion of learning levels within it.

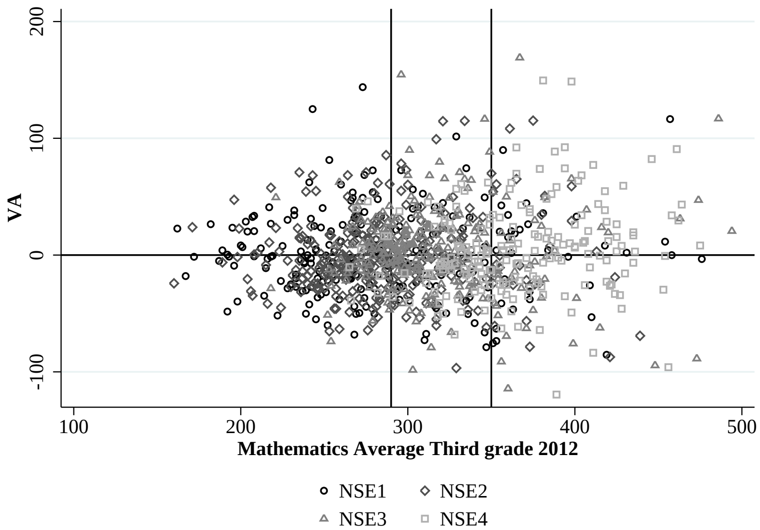

Figure 1: *Performance Quadrants. Cohort Third grade 2012 to Fifth grade 2014 Mathematics. Average. Source: own elaboration.*

Figure 3 continues with the example of VA model for the cohort Third Grade 2012 to Fifth Grade 2014 in the mathematics area, but the result of interest that it shows is the percentage of students at the advanced performance level; it is to say, analysis focus on the distribution top. In general, but not by definition, for a school to be ranked in high performance, it must have more than 40% of their students at the advanced performance level and schools in the lowest category does not have more than 20% of their students in advanced. Figures 1 to 3 show that as it advances in the performance categories, VA dispersion increases. The comparison group, formed by schools ranked in high performance for both the average score as for the ends of the distribution, varies more than the other comparison groups in the contribution concerning learning they provide to their students.

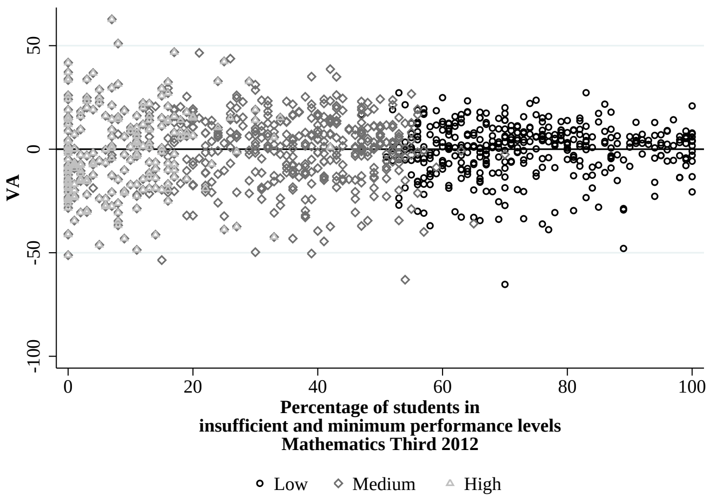

Figure 2: *Cohort Third 2012 to Fifth 2014 mathematics. Insufficient and Minimum. Source: own elaboration.*

Here we show the results of both aggregated cohorts nationwide. The national aggregate is the average of the country's schools weighted by the number of students who took the test at both time points. For the cohort Third 2012 to Fifth 2014 we introduce the results of the growth model and the VA model only for the area of language and for the cohort Fifth 2009 to Ninth 2013 we introduce the results of both models only for the area of mathematics. We present only one area for each cohort because results among areas are very similar nationwide. Correlation of Saber tests in school level for the national aggregate between areas of the same degree is never less than 0.7. This trend does not hold for other kinds of comparisons.

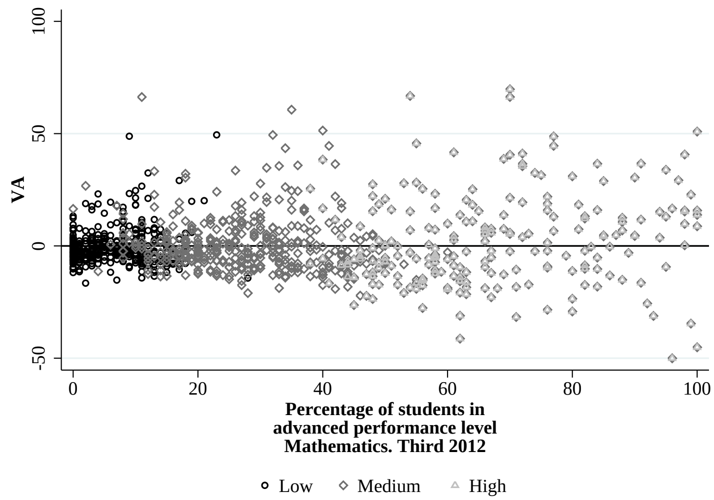

Figure 3: *Cohort Third 2012 to Fifth 2014 mathematics. Advanced. Source: own elaboration.*

National results are only a part of the findings of the analysis because, with certain precautions and for different purposes, it is possible to make comparisons between schools, municipalities, ETC or departments. The database with results for these aggregates, for both cohorts and areas and the results of the three independent variables, may be requested to authors.

4.1 National results of cohort Third 2012 to Fifth 2014

Figure 4shows the national aggregate for growth and VA models in the language area, for the average score of schools by performance category and NSE. Results between categories are not comparable, but inside of each category, it is possible to compare between NSE. In the growth model, you can clearly see a positive correlation marked between the change in learning and the socioeconomic level of students. Growth is negative in all three performance categories for the two lowest socioeconomic levels. This shows that, regardless of the school's performance, on average, schools of students with more economic needs do not generate any value at the end of primary cycle; compared with the learning they had in third grade. Similarly, regardless of performance, schools with students with a better economic situation are the ones that more value generate at the end of primary cycle.

The right side of Figure 4 shows the results of VA model. These results are not on the same scale of the growth model because the results of VA "shrink" compared to the previous model's results. The reason is that in VA model the difference between the observed and expected school's result is, on average, lower than the one of the growth model because the expected value is calculated compared to similar schools.

On the contrary, the expected value in the growth model is calculated for all schools of the country. VA model removes the effect of several contextual characteristics (variables that could be affecting the academic performance of schools) to give a "cleaner" estimate. Results show that correlation between valued added of schools and socio-economic levels gets lost.

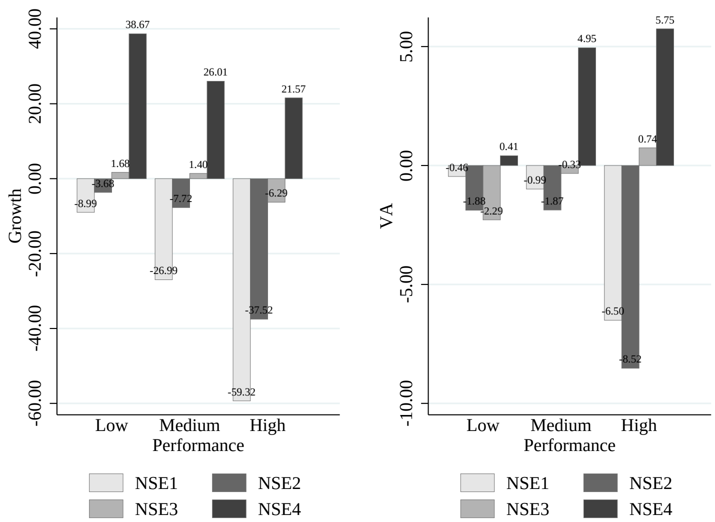

Figure 4: *Growth and value added for average score by performance and socioeconomic level category. Cohort Third 2012 to Fifth 2014. Language area. Source: own elaboration.*

It is important to note that both models provide relevant information to public policy objectives. VA model removes from school's effect characteristics of students it cannot control and, therefore, the result is an estimate of the direct and exclusive effect of school. This result has equity effects when comparing schools. On the contrary, growth model provides an absolute measure of how students are improving in time, independently of their demographic and socioeconomic characteristics. The difference between the growth model and VA model can be interpreted as the difference in the increase of student's learning that was not caused directly by the uncontrolled variables in the model, among them the school.

The most striking results are the ones in which there is not sign's change between both models. In estimates of 4it happens to schools of NSE 3. For example, for low-performance schools, average increase, in absolute terms, is an effect of contextual characteristics, and for high-performance schools, the VA model shows that these schools, on average, are contributing value to their students. It is noteworthy that high-performance schools in VA model, categorized in the most desirable NSE, are the ones that on average generate a bigger added value in learning. This shows that beyond the contextual features, schools serving this kind of students are those who are doing the job better.

Figure 5 and 6 show the result for the ends of school's distribution. Figure 5 shows the progress regarding insufficient and minimum performance levels and must be interpreted in such a way that a decrease is positively valued. This is because learning makes itself evident in reducing the percentage of students at these levels of performance. Results show that correlation between performance and NSE is maintained. Growth models reveal that high-performing schools that serve students from less desirable socioeconomic status are those that generate lower profits in learning.

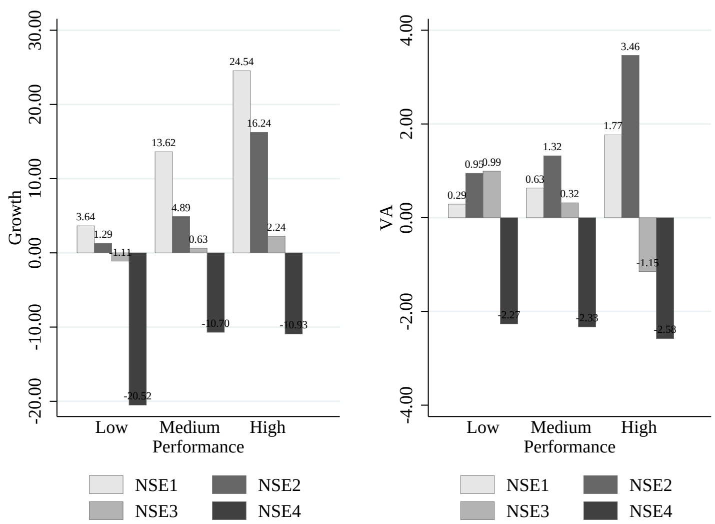

Figure 5: *Growth and value added for insufficient and minimum performance levels by categories of performance and socioeconomic level. Cohort Third 2012 to Fifth 2014. Language area. Source: own elaboration.*

When comparing the panels of each figure, it is evident that the decrease in absolute terms on the tips of the distribution of schools in the less desirable socioeconomic level changes drastically when adjusted for contextual features. Schools of first socioeconomic level manage to include students in the advanced level but fail to reduce the percentage in insufficient and minimum.

Comparison between the results of the average school and the tips of the distribution (figures 4, 5, 6) shows that schools that added value did it mainly by moving the center of the distribution to students with poorer outcomes, and not necessarily by moving them to the right. Many schools grew on average, despite decreasing the percentage of students at the advanced performance level. Similarly, schools that did not add value on average did it mainly because they did not take their students out of the two less desirable performance levels. This fact suggests that, regardless of performance or socioeconomic status, public policy efforts should focus on leveling students who are below the average for their respective schools.

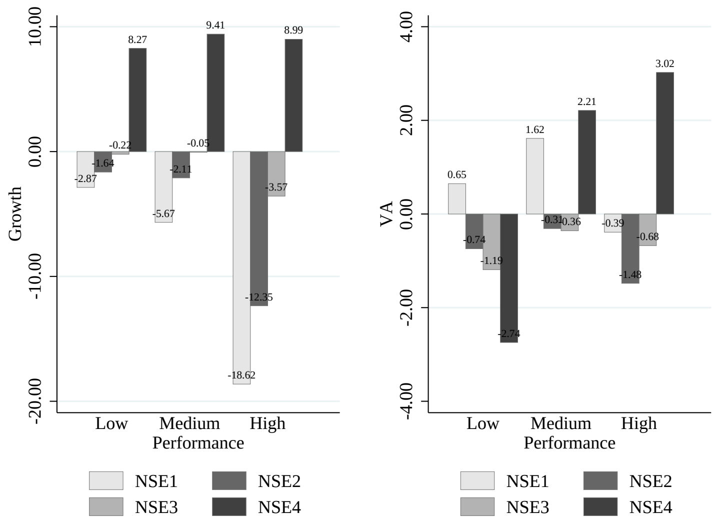

Figure 6: *Growth and added value for advance performance level by categories of performance and socioeconomic level. Cohort Third 2012 to Fifth 2014. Language area. Source: own elaboration.*

Figure 7 shows results for average in the area of language by categories of performance and zone and sector. Both in the growth model as in the VA schools that generate more value to their students are non-official and rural. These schools of the countryside are generally bilingual, with totally different characteristics to the official rural schools. This result confirms, once again, that education received by students of a high socioeconomic level is significantly higher than that received by other students.

There are two important conclusions for official schools in the results by zone. After removing the effect of contextual characteristics, rural schools still do not generate value to their students, but urban schools show better behavior, especially high-performance schools are those that create more value from all comparison groups.

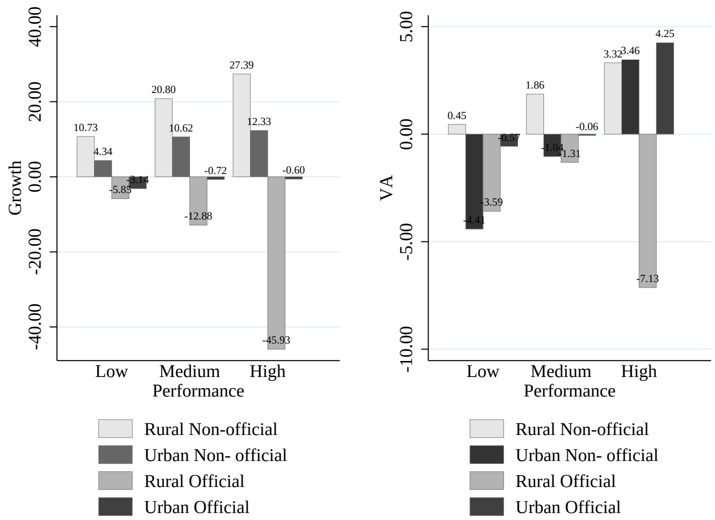

Figure 7: *Growth and value added for average score by categories of performance, zone and sector. Cohort Third 2012 to Fifth 2014. Language area. Source: own elaboration.*

4.2 National Results of the cohort Fifth 2009 to Ninth 2013

There are four years of difference in the period of study of cohort Fifth 2009 to Ninth 2013. Besides, fifth grade is the last grade of primary cycle and ninth the last of secondary. The magnitude of change in learning of this cohort in this period is determinant for the near future of students. Figure 8 shows that, as for the cohort Third 2012 to Fifth 2014, high-performance schools of NSE 4 are the schools that more value add to their students. Besides, in VA model the effect of the school is double than in the previous cohort. Therefore, it is emphasized that such schools, even after removing the effect of socioeconomic characteristics, continue reporting the best results. In the same way, it is evident that the education of students of high socioeconomic status is categorically better than the education received by less fortunate students.

Figure 10 shows analysis by zone and sector. Like in cohort Third 2012 to Fifth 2014, schools that generate more value to its students are the rural non-official and those that less value generate are the rural officials. Unlike the previous cohort, there is no evidence of good results for official urban schools compared to rural official schools.

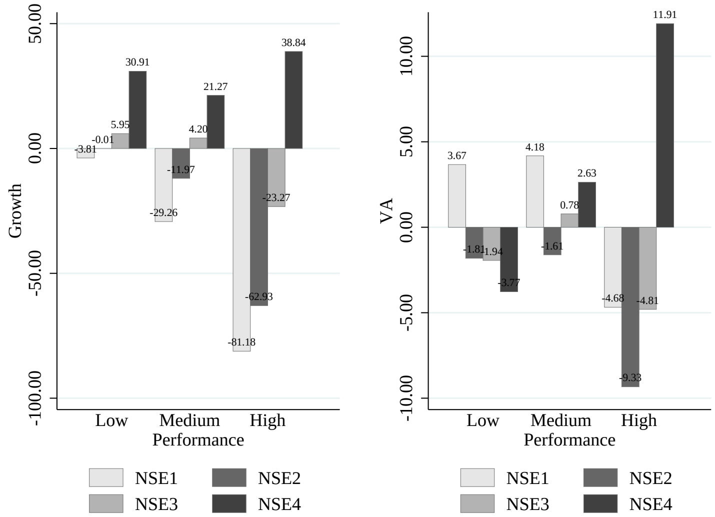

Figure 8: *Growth and value added for average score by categories of performance and socioeconomic level. Cohort Fifth 2009 to Ninth 2013. Mathematics area. Source: own elaboration.*

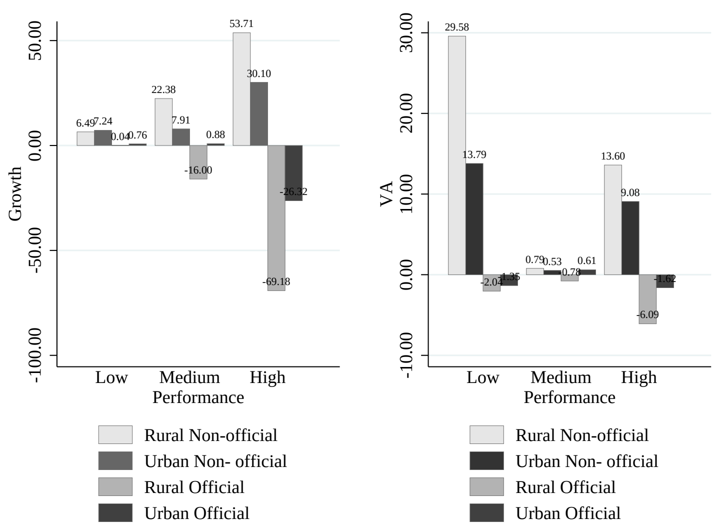

Figure 9: *Crecimiento y valor agregado para el puntaje promedio por categorías de desempeño, zona y sector. Cohorte Quinto 2009 a Noveno 2013. Área de matemáticas. Fuente: elaboración propia.*

In figure 10, growth model shows an evident but modest increase in the percentage of students in advanced level in low-performance schools of all socioeconomic levels. It indicates that, although most students do not achieve the expected goals for the ninth grade, schools do get the most academically advanced students to reach them.

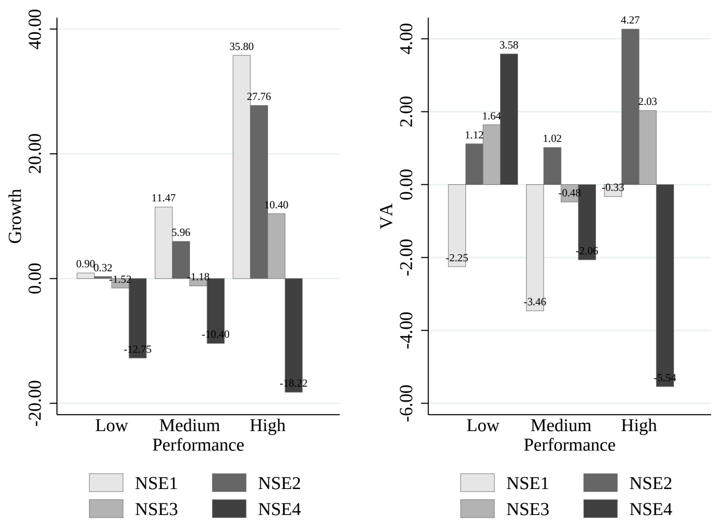

Figure 10: *Growth and value added for average scores by categories of performance, zone and sector. Cohort Fifth 2009 to Ninth 2013. Mathematics area. Source: own elaboration.*

In figure 11, growth model shows an evident but modest increase in the percentage of students in advanced level in low-performance schools of all socioeconomic levels. It demonstrates that, although most students do not achieve the expected goals for the ninth grade, schools do get the most academically advanced students to reach them.

The aspect that stands out in Figures 11 and 9 is that, unlike cohort Third 2012 to Fifth 2014, low NSE schools in the VA model managed to reduce the percentage of students in insufficient and minimum level and increase the proportion of advanced. Annex 2 shows the national results of the mathematics area for cohort Third 2012 to Fifth 2014 and language for the cohort Fifth 2009 to Ninth 2012. Conclusions in the analysis by NSE are very similar for both areas, but there is an important difference in the analysis by zone and sector. VA model shows good results in both cohorts for urban official (public) schools in language area, but not for the mathematics area.

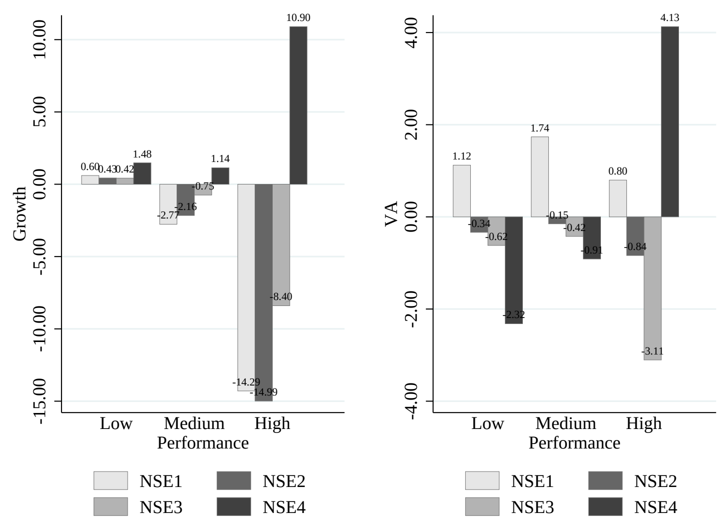

Figure 11: *Growth and value added for insufficient and minimum performance levels by categories of performance and socioeconomic level. Cohort Fifth 2009 to Ninth 2013. Mathematics area. Source: own elaboration.*

5 Conclusions

To the best of our knowledge, this article is the first approach to VA models for Saber tests 3, 5 and 9 in Colombia. This methodology allows us to follow two cohorts over time. The first cohort is of students who were in Third grade of primary school in 2012 and Fifth grade in 2014. The ICFES will implement Saber test 7, in 2016 it will be possible to follow this same group of students again. The second cohort we followed is of students who were in Fifth grade of primary school in 2009 and four years later, in 2013, in ninth grade of secondary cycle. Next year, when the ICFES scores and publishes the test results of Saber 11 of this year, this cohort could be followed again. It implies that we could study this group of students in the three cycles that complete primary, secondary and middle education.

Estimates of VA are relevant because they allow us to follow the same cohort over time, with the purpose of estimating the change in the students learning. This possibility allows knowing the result of a school in an endpoint discounted by what their students already knew at an initial point. Cross-sectional studies, on the contrary, allow comparisons at one point of time or temporal comparisons between different cohorts only. VA models are an excellent complement to the cross-sectional analyzes because knowing how much students have improved is as important as knowing how they are at the end of a period.

Our empirical exercise is based on two models: one without contextual settings and other that controls by different demographic, social and economic characteristics. Model without adjustment gives a measure of the change in learning experimented by students who belong to a school during a period. Model with adjustment separates the exogenous effects from the direct effect of school. Both models have practical implications for public policy. The first model can be used to identify schools with negative results and focalize attention toward them and the second model to reward or recognize the efforts of the best ones.

Results of the national analysis show that economic conditions of students are strongly related to educational progress. Students belonging to higher socioeconomic levels who go to expensive schools are the ones who improve their skills further. This result is evident in the model that separates the students' characteristics from the direct effect of school: is not only the favorable conditions of these students what allow them a better performance, but also the fact that schools provide better training. Our analysis, besides of seeing the average behavior of schools, focuses on the distribution of students within it. Results show that there are schools that despite having a low performance have students in high-performance levels. These schools can maintain their high-achieving students at these levels. The problem is that they fail to level their underperforming students with the rest of their students.

The interpretation of the results of our VA models must take into account the initial performance of schools, because increases in learning have different meanings depending on the level they start. For this reason, we present our estimates in three performance categories based on the results of the initial year. Estimates of schools or other aggregates are comparable only within each category of performance and not comparable between these categories. There are other ways to show the VA results. One of them is showing the results with measures of relative contribution. These measures do not show the accurate estimate of VA, but a comparison of each school with respect to others with certain similar conditions. Results of the study of the relative contribution of all universities in the country are published on the ICFES website. Research's projects management and evaluation's management 2015 uses Saber tests 11and Saber Pro to estimate the VA of the reference groups of universities. Results introduce reference group for each institution, as the relative contribution with respect to institutions whose students have more similar results in Saber 11.

Another way to show VA results could base on a single measure of the behavior of a group of similar schools. In a research of (de Gestión de Proyectos de Investigación y Dirección de Evaluación. 2015) comparison groups that are not based on cognitive outcomes, but contextual characteristics are being built. The objective of that analysis is to know how is each school compared to schools with similar socioeconomic conditions. By combining this exercise with the VA's valuable information could be obtained.

A shortcoming of our work is that, because Saber 3, 5 and 9 tests do not offer a particular score for students, observation unit for analysis are schools. This fact implies that estimates are valid under the assumption that there is student stability in schools. This assumption could be verified if there were student-level scores for Saber 3, 5 and 9 tests. A less expensive option could be given if it was possible to know from other sources the students' proportion from a cohort that continues in each school. In this way would be more confidence for estimates of schools with verified student stability.

Our measures of VA can provide evidence of student learning in schools to those responsible for educational policies. These evidences can have accountability purposes; can be used internally by the institutions for their improvement plans or identify strengths and weaknesses of schools (Steedle 2010). However, we believe that the complexity of the educational environment requires that interpretations of VA estimates include several warnings for a fair and correct interpretation (OECD 2008).

Even correcting the problems of our empirical exercise, several difficulties remain when estimating the effect of schools on the educational progress of their students. Everything you learn in a school can be translated into accumulation of knowledge; skills, tools, customs and ethical values, in addition to the effect on the way students think, feel and act (Bennett 2001), (Harvey & Green 1993). VA models use data of standardized tests that can only measure a part of the whole education given in a school.

Received: February 12, 2016
Accepted: April 14, 2016

References

Aaronson, D., Barrow, L. & Sander, W. (2007), 'Teachers and student achievement in the chicago public high schools.', *Journal of Labor Economics.* **25**(1), 95–135.

Armour-Garb, A. (2009), 'Should "value-added" models be used to evaluate teachers?', *Journal of Policy Analysis and Management.* **28**(4), 692–712.

Bennett, D. C. (2001), 'Assessing quality in higher education.', *Liberal Education.* **87**(2).

Colbert, V. (1999), 'Mejorando el acceso y la calidad de la educación para el sector rural pobre: El caso de la escuela nueva en colombia.', *Revista Iberoamericana de Educación.* .

de Gestión de Proyectos de Investigación y Dirección de Evaluación., O. A. (2015), Reporte de resultados en saber pro, medidas de aporte relativo y otros indicadores de calidad de las instituciones de educación superior 2013-2014., Technical report, Instituto Colombiano para la Evaluación de la Educación. ICFES.

Doran, H. C. & Izumi, L. T. (2004), 'Putting education to the test: A value-added model for california.', *Pacific Research Institute.* .

Doran, H. C. & Lockwood, J. R. (2006), 'Fitting value-added models in r.', *Journal of Educational and Behavioral Statistics.* **31**(2), 205–230.

Downes, D. & Vindurampulle, O. (2007), 'Value added measures for school improvement.', *Office for Education Policy and Innovation.* (13).

Dury, D. & Doran, H. (2003), 'The value of value-added analysis. policy research brief.', *National School Boards Association.* **3**(1), 1–4.

Harvey, L. & Green, D. (1993), 'Defining quality.', *Assessment and Evaluation in Higher Education.* **18**(1), 9–34.

Haveman, R. & Wolfe, B. (1995), 'The determinants of childrens attainments: A review of methods and findings.', *Journal of Economic Literature.* **33**(4), 1829–1878.

Hill, P. (1995), 'Value-added measures of achievement.', *Incorporated Association of Registered Teachers of Victoria Seminar Series.* **44**.

Kazeem, A. andJensen, L. & Stokes, S. (1995), 'School attendance in nigeria: Understanding the impact and intersection of gender, urbana rural residence, and socioeconomic status.', *Comparative Education Review* **54**(2), 295–319.

Kim, H. & Lalancette, D. (2013), Literature review on the value added measurement in education., Technical report, OECD.

Koedel, C., Mihaly, K. & Rockoff, J. (n.d.), Value added modelling: A review. Unpublished article. (2015).

Nunez, J., Steiner, R., Cadena, X. & Pardo, R. (2002), '¿Cuáles colegios ofrecen mejor educación en Colombia? ', *CEDE, Universidad de los Andes* .

OECD (2008), Measuring improvements in learning outcomes: Best practices to assess the value-added of schools., Technical report, OECD.

OECD (2011), La medición del aprendizaje de los alumnos: Mejores prácticas para evaluar el valor agregado de las escuelas., Technical report, OECD.

Raudenbush, S. W. & Bryk, A. S. (1986), 'A hierarchical model for studying school effects.', *Sociology of Education.* **59**, 1–17.

Reardon, S. F. & Raudenbush, S. W. (2009), 'Assumptions of value added models for estimating school effects.', *Education Finance and Policy.* **4**(4), 492–519.

Rivkin, S. (2005), 'Teachers, schools and academic achievement.', *Econometrica.* **73**(2), 17–58.

Saunders, L. (1998), '"Value Added" measurement of school effectiveness: An overview.', *National Foundation of Economic Research* .

Steedle, J. T. (2010), 'Improving the reliability and interpretability of value-added scores for post-secondary institutional assessment programs.', *Paper presented at the Annual Meeting of the American Educational Research Association.* .

Tobón, D., Valencia, G., Ríos, P. & Bedoya, J. (2008), 'Organización jerárquica y logro escolar en Medellín. Lecturas de Economía.', *Universidad de Antioquia.* **68**, 145–173.

Willms, D. (2009), 'Value added models in education. policy brief.', *Canadian Research Institute for Social Policy* .

Wyatt, T. (1996), 'School effectiveness research: dead end, damp squib or smouldering fuse?', *Issues in Educational Research.* **6**(1), 79–112.

Zhang, Y. (2006), 'Urban-rural literacy gaps in sub-saharan africa: The roles of socioeconomic status and school quality.', *Comparative Education Review.* **50**(4).

Appendix

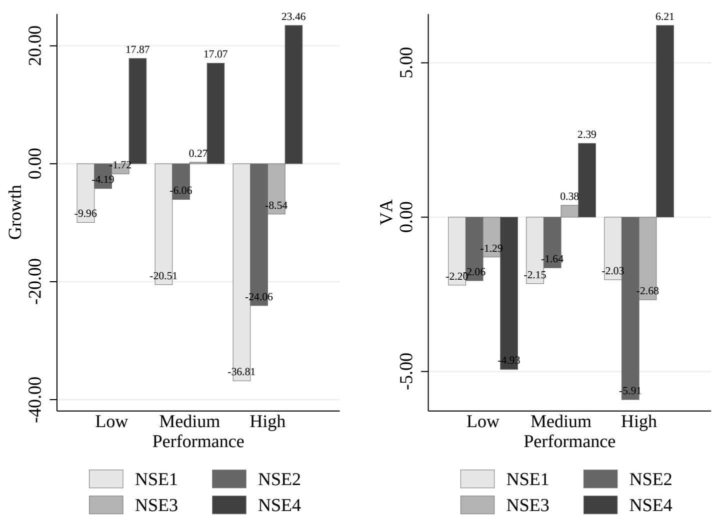

Figure 12: *Growth and value added for average score by categories of performance and socioeconomic level. Cohort Third 2012 to Fifth 2014. Mathematics area. Source: Own elaboration.*

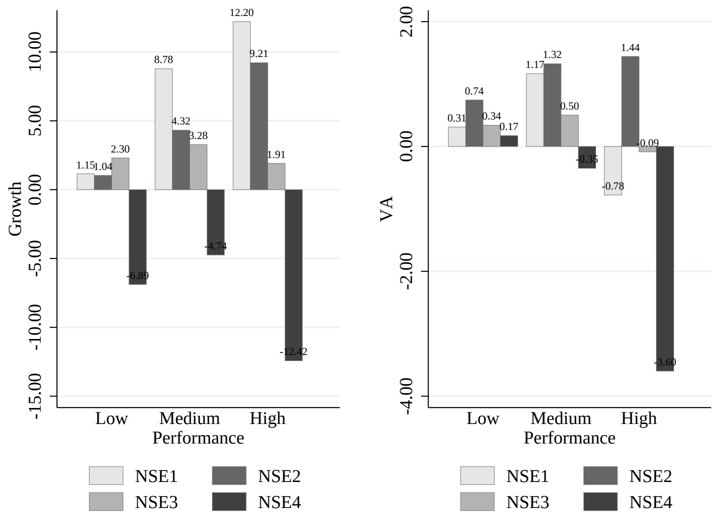

Figure 13: *Crecimiento y valor agregado para niveles de desempeño Insuficiente y Mínimo por categorías de desempeño y nivel socioeconómico. Cohorte Tercero 2012 a Quinto 2014.Área de matemáticas. Fuente: elaboración propia.*

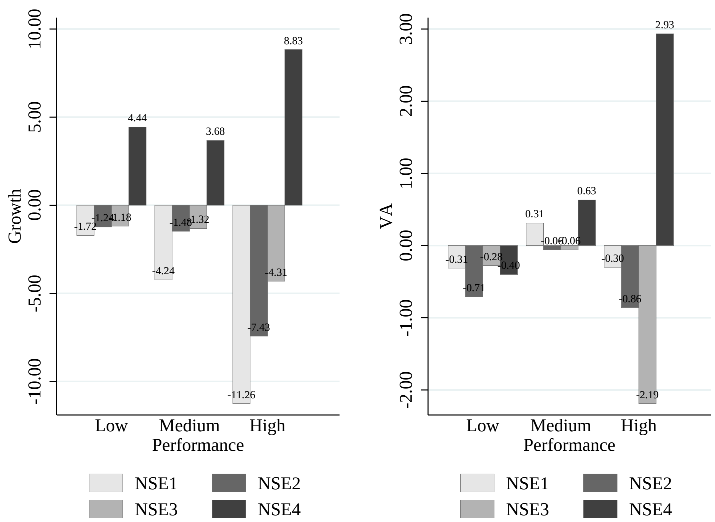

Figure 14: *Growth and value added for insufficient and minimum performance levels by categories of performance and socioeconomic level. Cohort Third 2012 to Fifth 2014. Mathematics area. Source: own elaboration.*

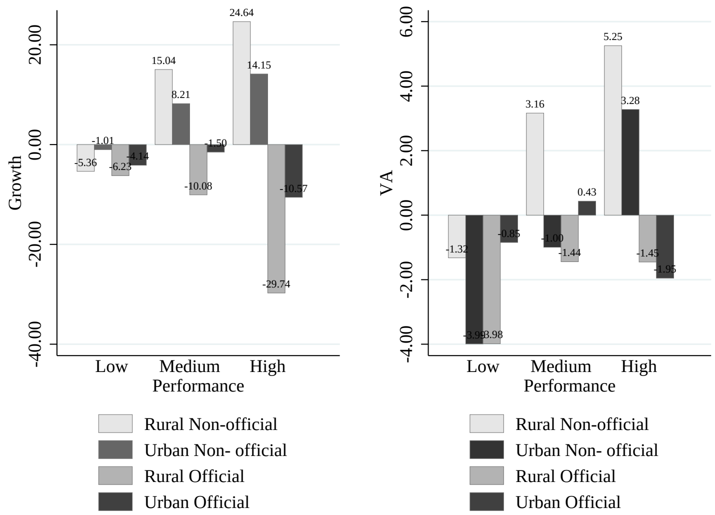

Figure 15: *Growth and value added for advanced performance level by categories of performance and socioeconomic level. Cohort Third 2012 to Fifth 2014. Mathematics area. Source: own elaboration.*

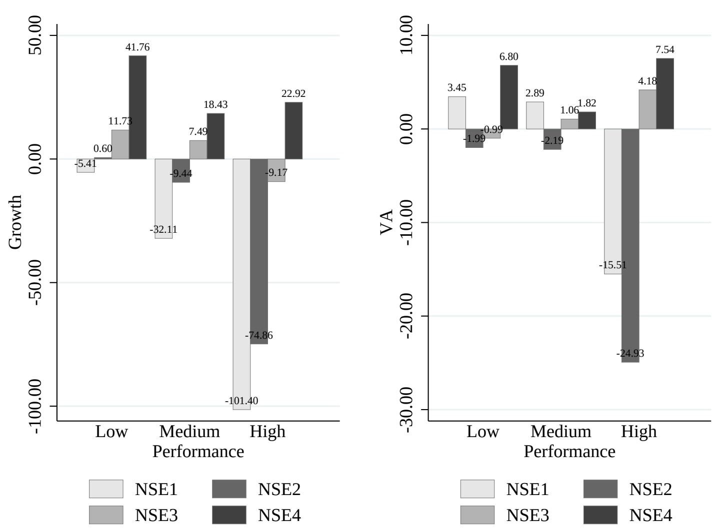

Figure 16: *Growth and value-added for average score by categories of performance, zone and sector. Cohort Third 2012 to Fifth 2014. Mathematics area. Source: own elaboration.*

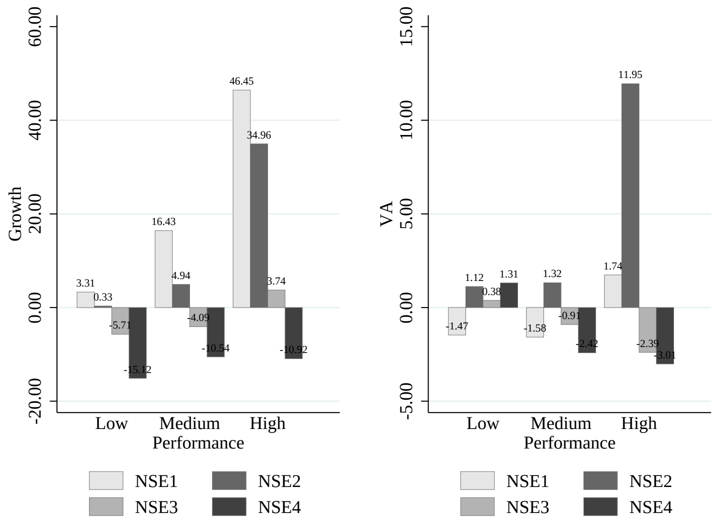

Figure 17: *Growth and value for average score and socioeconomic level. Cohort Fifth 2009 to Ninth 2013. Language area. Source: own elaboration.*

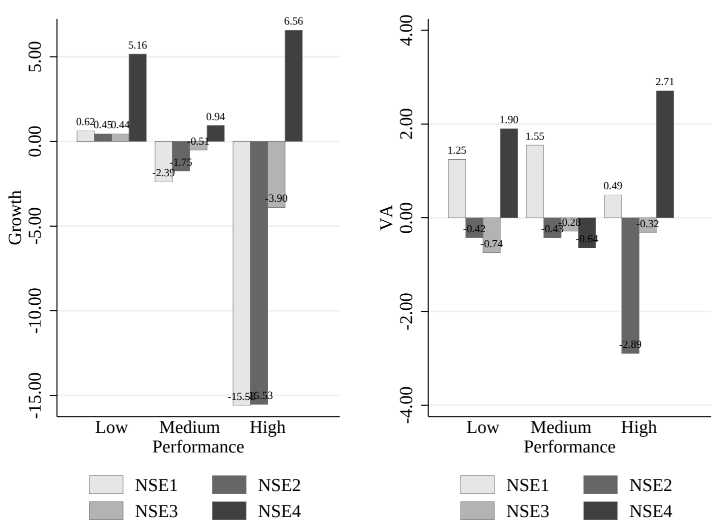

Figure 18: *Growth and value added for insufficient and minimum performance levels by categories of performance and socioeconomic levels. Cohort Fifth 2009 to Ninth 2013. Language area.Source: own elaboration.*

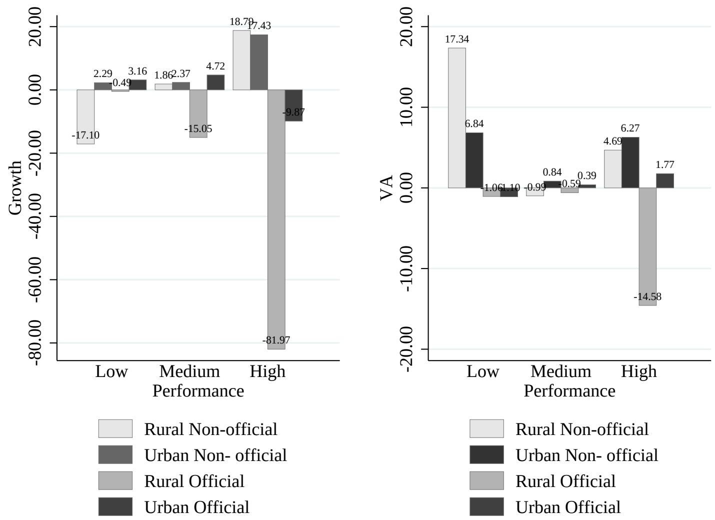

Figure 19: *Growth and value added for advanced performance level. Cohort Fifth 2009 to Ninth 2013. Language area. Source: own elaboration.*