What does the National High School Exam (ENEM) tell Brazilian society?

Rodrigo Travitzki, Jorge Calero and Carlota Boto

ABSTRACT

This article assesses the limitations and potentials of the National High School Exam (ENEM) as an indicator of school effectiveness in Brazil, and considers the effects of introducing contextual variables. A multilevel regression analysis was performed on three levels (individual, school and state) using microdata on 17,359 schools from 2009 and 2010. Contextual factors made it possible to explain 79% of the difference between schools. The raw average and value-added (random effect at the school level) produced contrasting evaluations in 34% of cases; and the average was more stable (r = 0.8) than value-added (r = 0.5) in both years. Various shortcomings in the ENEM as an indicator of school effectiveness were identified. The results show that this league table reveals more about socioeconomic conditions than the schools' own merit, in other words the value-added they are supposedly providing to the students.

KEYWORDS High school education, examinations, schools, league tables, education quality, measurement, Brazil

JEL CLASSIFICATION I24, I28, C18

AUTHORS

Rodrigo Travitzki has a Ph.D. in Education from the University of São Paulo, Brazil. r.travitzki@gmail.com

Jorge Calero is Chair Professor of the Department of Public Economy, Political Economy and Spanish Economy of the University of Barcelona, Spain. jorge.calero@ub.edu

Carlota Boto is a Professor at the Faculty of Education at the University of São Paulo, Brazil. reisboto@usp.br

I
Introduction

In the 1990s, a number of countries published school league tables with the aims of improving the quality of the schools and making them more accountable to society, while also providing information to help parents choose a school for their children (Karsten, Visscher and De Jong, 2001). This type of strategy usually forms part of moderated accountability policies which aim to inform the government and families and identify good practices in the education system, but without linking the results to rewards or penalties, as is done in carrot-and-stick accountability policies (Martínez Arias, 2009). Although the way those policies are generally formulated is important, their effects depend mainly on the quality measurements used (Ladd and Walsh, 2002). Consequently, technical and methodological controversies arise on issues related to the production of the indicators used in the league tables, and their capacity to really promote quality in the schools. In practice, the countries adopt a variety of strategies.

In England, for example, school league tables have been published since 1992 and have been used to create incentive systems (West and Pennell, 2000). At the other extreme, education policies in Spain prohibit the publication of school league tables (Government of Spain, 2006, Art. 140). Brazil is currently in an intermediate situation, because the publication of league tables is not linked to incentives, although some states have accountability policies involving rewards and penalties. Some authors support the strengthening of those policies nationwide (Andrade, 2008), while others adopt more cautious and critical attitudes, stressing the trend towards greater inequality, for example (Franco and others, 2007).

The first controversial issue is the concept of quality itself, which is highly polysemic when applied in the education field (Murillo Torrecilla, 2005). In Brazil's recent history, the idea of quality has taken different forms. Firstly, it was related to the universalization of access, then to the flow and repetition rate, and then to the performance of students in large-scale examinations (Oliveira and Araujo, 2005). The capacity to achieve good results is normally referred to as "effectiveness"; but it is worth remembering that an effective school is not necessarily a quality school, since effectiveness is a necessary but insufficient condition (Murillo Torrecilla, 2005, p. 31). In other words, quality is a broader concept than efficiency, and it has different meanings.

Although school quality is a controversial subject, it has frequently been linked to the performance of students on standardized tests, not only because of the desire to find objective measures, but also for practical reasons such as cost and viability. Although fundamental, this study will not discuss that issue, but focus exclusively one aspect of school quality: effectiveness in preparing students to do standardized tests. This is a very narrow focus, but necessary bearing in mind that in Brazil, as in the rest of the world, this type of indicator plays an increasing role in education policies and people's imagination.

Other controversies surrounding school league tables relate to their overvaluation (Brandão, 2000), increased social exclusion (West and Pennell, 2000), the reproduction of class privileges (Apple, 2001), feedback that benefits the best and damages the worst (Ladd and Walsh, 2002) and a lack of attention to the tests themselves (Reckase, 2004). The use of aggregate individual indicators (rather than taking the average, for example) is also criticized as a means of evaluating schools (Meyer, 1997), as also is the type of information chosen for publication (Van Petegem and others, 2005). Lastly, few studies have considered how the schools can use that information to improve their students' learning process (Heck, 2000).

This study makes a critical analysis of this type of indicator, using quantitative data from the National High School Exam (ENEM), which serves as a selection criterion for Brazilian universities and could become the official indicator of school quality nationally (Passarinho, 2012).[1] Some of the conclusions of this study relate to the test itself, whereas others concern more general methodological issues concerning value-added models in education. The specific aim is to evaluate what type

☐ This study received support from the Brazilian research support agency, Coordination for the Improvement of Higher Education Personnel (CAPES).

[1] Index of Development of Basic Education (IDEB).

of information the ENEM league table provides to society and how the introduction of contextual variables interferes in the results by schools, both cross-sectionally and longitudinally.

The importance of context

Like many others, the ENEM league table is based on the publication of school averages. Nonetheless, as the schools have very heterogeneous starting conditions, particularly in developing countries, contextual factors must be taken into account to more accurately evaluate the effect obtained by each school (Heck, 2000) —in other words its merit—. According to Meyer (1997, p. 298), a school's average test score, which is one of the most widely used education indicators in the United States, is highly questionable as an indicator of school performance, and is a very weak or even counter-productive instrument of accountability.

In the specialized literature, many authors defend this position, particularly in terms of accountability (Willms, 2006); but others argue that both the "raw" average and the "net" average (in other words, the average obtained after controlling for the effect of contextual variables) can produce distorted results (Tekwe and others, 2004). Moreover, even multilevel value-added models, the "latest generation" of school-quality indicators, can produce a wide variety of results, depending on the contextual variables that are considered (Keeves, Hungi and Afrassa, 2005; Ladd and Walsh, 2002).

Despite these shortcomings, if the aim is to make fair comparisons between schools for accountability purposes, it is essential to take account of the context in which each one operates. As noted by Thomas (1998, p. 92), the publication of league tables based on raw averages assists neither the initially high-achieving nor the initially low-achieving school. "In the former, the need for improvement may not be appreciated; in the latter serious demoralization of staff may occur through no fault of their own".

The same article refers to a 1992 study published in the newspaper The Guardian, which reached the conclusion that 23% of schools were evaluated differently between the "raw" and "net" league tables. To what extent are these conclusions confirmed in the ENEM data?

Following the introduction, this article proceeds as follows: section II describes value-added models according to different concepts and authors; section III asks why it is necessary to evaluate schools; section IV presents the National High School Exam (ENEM) as a school quality indicator; section V presents the main results, and section VI draws relevant conclusions.

II
Value-added models

There are a variety of value-added concepts, some of which may even be mutually contradictory (Saunders, 1999). While some authors believe value-added should be based on longitudinal data (Martínez Arias, Gaviria Soto and Castro Morera, 2009),[2] this study uses a broader concept (Reckase, 2004), which includes, for example, the effect of the school. The general aim is to evaluate how much students improve thanks to the work of the school rather than to its prior conditions, in an attempt to eliminate the influence of factors that are outside the school's control (McCaffrey and others, 2004).

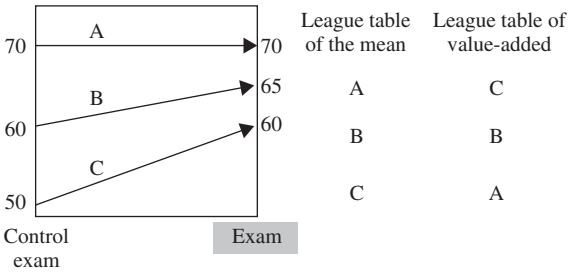
Figure 1 illustrates that general idea, which can be applied both to cross-section and longitudinal data. In the first case, which includes the school effect, there is only one test taken by the students, because the "control test" is estimated using contextual variables. In other words, the aim is to determine what the students' score (or that of the school) would have been if all had had the same contextual conditions at the outset. In the second case, the "control test" is real and, for example, was applied to students before they entered the school. In the second case, therefore, at least two different tests are applied to the same students. Each of the methodologies starts from its own assumptions and has its own limitations, so choosing one or the other depends largely on the data that are available.[3]

[2] The main argument is that previous performance condenses variables relating to socioeconomic level (Ferrão, 2009).

[3] Value-added with longitudinal data assumes, for example, that the measurement instruments used over the years have the same purpose, form, and degree of difficulty (in the case of the tests). In contrast, value-added with cross-section data not only only assumes the existence of contextual data, but that those data adequately represent the initial conditions of the student.

FIGURE 1

General concept of value-added



Source: prepared by the authors.

At the present time, value-added models are being used to guarantee accountability in Tennessee, Dallas, Chicago (United States) and in England (Martínez Arias, 2009). Models based on longitudinal data are recent, and they began to be used in England at the start of the new millennium, following the creation of an individual student identification number (Ray, Evans and McCormack, 2009). Studies on the school effect (value-added with cross-section and contextual data) have existed since the early 1980s, and they find that an estimated 5%-35% of the variance in scores obtained by students can be explained by the school in which they study (Martínez Arias, Gaviria Soto and Castro Morera, 2009).

The stability of the school effect over the years remains a controversial issue. There are data showing that few schools have results that are consistent (across different students) and stable (through time) (Thomas and others, 1997). Various studies have calculated the correlation coefficient of the school effect in different years, reporting coefficients ranging from zero (Linn and Haug, 2002) to around 0.6 (Mandeville, 1988); while (Luyten, 1994) finds correlation coefficients that are always between 0.35 and 0.65 for primary schools and between 0.70 and 0.95 in the case of high schools. Here again, it is worth asking to what extent the ENEM microdata confirm these results.

III

Why evaluate the schools?

In terms of the use made of the data produced by this type of indicator (based on standardized tests), there are at least two important questions that are very closely related to each other: what are the data used for and to whom are they directed? A minimally consistent approach to this topic requires its own research. For the purposes of this study, it is sufficient to identify two types of use: the accountability of public schools and the choice of a school by parents. The first aspect is related to state mechanisms, and the second to market mechanisms.

How is the aim related to the measurement instrument? It is possible to consider, for example, the size of an object. Given that "size" is an objective and consensual concept, the same rule could be used irrespective of the purpose of the measurement. In the worst of cases, the instrument is changed if the object is very large and the use of a common rule becomes impractical and inaccurate. Nonetheless, as school quality is not a consensual concept, it seems reasonable to use different indicators for different purposes. Research along these lines has sought to develop various forms for estimating the quality of schools, bearing in mind their usefulness for the family and the Government, for example (Meyer, 1997). In general, although raw averages are informative in terms of the schools' performance, they produce unfair comparisons for administrators, teachers and students (Willms, 2006). In view of this, two types of school effect were proposed, the first (Type A) related more to general performance, and the second (Type B) with the objective of isolating the factors over which the school has some control.

"The Type B effect is the effect school officials consider when evaluating the performance of those who work in schools. A school with an unfavourable context could produce a large Type B effect through the effort and talent of its staff. The school would rightly earn the respect of school evaluators even though parents shopping for a large Type A effect might not want to choose that school" (Raudenbush and Willms, 1995, p. 310).

The main ideas described above are summarized in figure 2, which also introduces another concept of value-added relative to school quality. Thus far, the value that the school adds to the student has been considered through cross-section or longitudinal data. Nonetheless, the school's value-added can be estimated over certain time period, whether by the school itself, or through educational policies, the community, or through economic and social changes, among other factors. This second concept of

value-added was used in Brazil to define the targets for each school proposed in the National Education Plan.[4]

It is clear, among other things, that the main source of information used to evaluate school quality consists of standardized individual tests (a fact that assumes a clear methodological shortcoming), and that different indicators could be used for different sectors of society. This plurality needs to be taken into account, because undue emphasis on parental choice tends to increase inequalities (Apple, 2001).

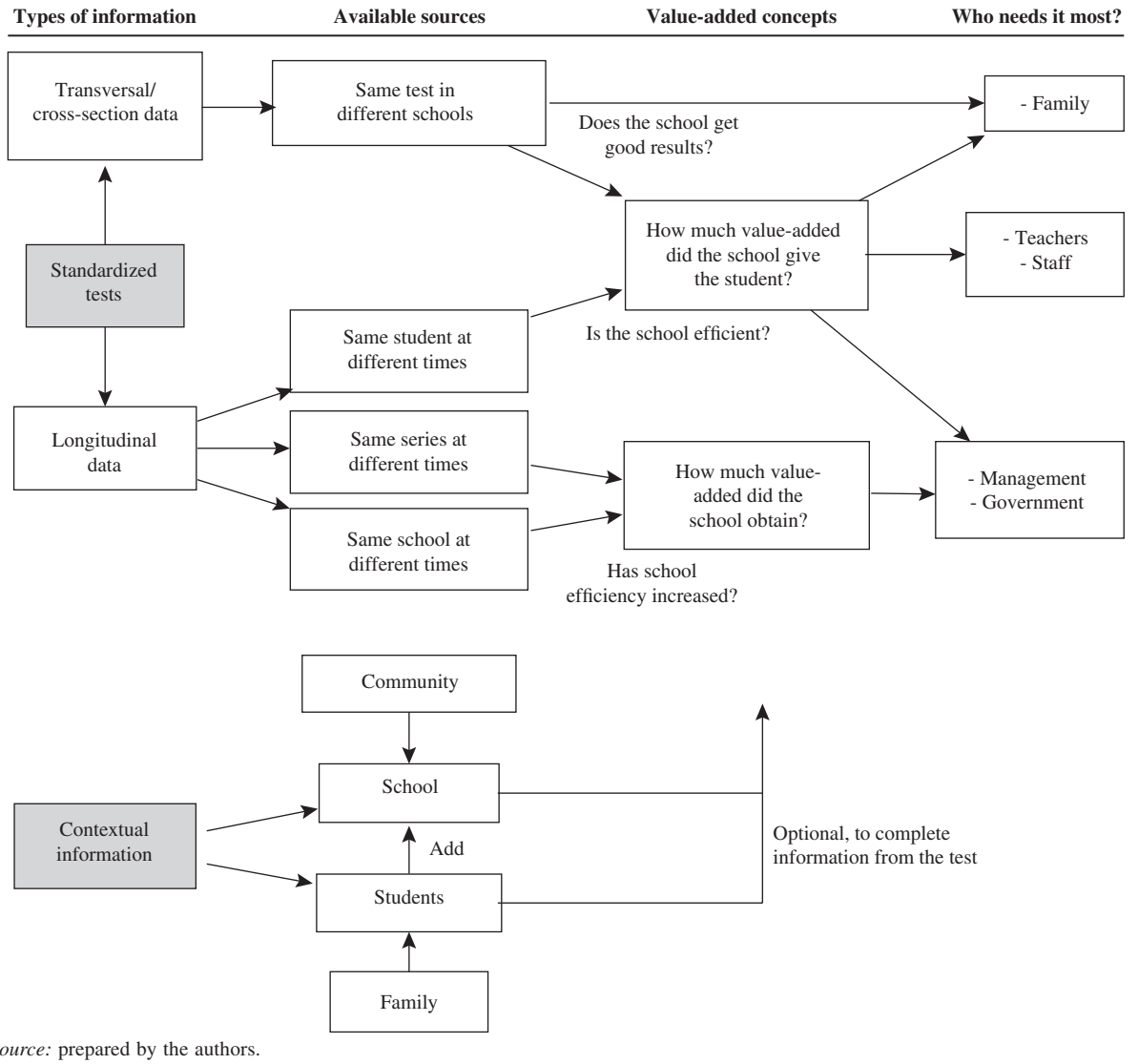
There seems to be an inherent tension in the evaluation field, which applies both to schools and to the students in each of them —namely the quality/equity dichotomy—. Dubet describes this tension in intra-school relations, by considering what a fair school system would be like, and demonstrating that there is no perfect solution but a combination of options and necessarily limited responses (Dubet, 2004, p. 540).

One could therefore ask what a fair school evaluation system that could take account of that tension and multiplicity would look like. Bearing in mind that Brazilian schools are currently only evaluated through individual exams, without taking account of context, it seems that much remains to be done in this regard.

[4] In September 2012, the Plan had not yet been approved by the National Congress.

FIGURE 2

Overview of value-added models



Source: prepared by the authors.

IV

The ENEM as an indicator of school quality

The National High School Exam (ENEM) was created in 1998 by the Anísio Teixeira National Institute for Educational Studies and Research (INEP), an organization attached to the Ministry of Education in Brazil, with the explicit aim of providing students with a self-evaluation at the end of basic education (MEC, 2002), although it is reasonable to assume that it also represents an initiative to create accountability policies in the country. This tool was initially not very ambitious, but it underwent various changes over the years and has become increasingly wide-ranging and deeply rooted in the educational culture. The ENEM is applied to roughly 5 million students each year, and is the second-largest high school exam in the world, after China's *goakao*, which covers 10 million (Zhang and Zha, 2010). This is explained by the fact that, despite being voluntary since its creation, the ENEM has come to be used as a selection test for admission to the public universities and for obtaining scholarships in private universities, among other purposes. Table 1 shows the rapid growth of the exam and the changes it has undergone over the years.

The ENEM league table covers 61% of Brazilian high schools and 7% of all schools, and is being used increasingly as a school-quality indicator.[5] Although the INEP itself does not publish a league table, the publication of the averages by schools enables the press to produce a league table. The first league table dates from 2006 and had at least two objectives: to galvanize society to improve teaching, and to help teachers, directors and administrators to identify shortcomings and good practices in the school environment (INEP/MEC, 2007). In 2011, along with the averages, the participation rate of each school (in four brackets) began to be published; and a recommendation was made to compare only schools with similar participation rates (INEP/MEC, 2011a). The purpose of this initiative was to minimize the sample bias caused by the voluntary nature of the exam; but, in any event, the lack of significance of a comparison of schools in which fewer than one quarter of their students participate in this indicator needs to be stressed.

[5] Based on 2009 data (see table A.1 of the annex).

TABLE 1

Brazil: historical summary of the National High School Exam (ENEM)

Year	No. of participants	Changes
1998	157 221	First ENEM: voluntary, for self-evaluation only.
2001	1 624 131	Introduction of the school identification number (ID), the same as the census but different each year.
2003	1 882 393	Major changes in the team responsible for the exam following the presidential elections of the previous year.
2005	3 004 491	The ENEM starts to be used as a vehicle for gaining access to higher education and as a criterion for receiving government scholarships under the "University for All" (ProUni) program in private institutions.
2006	3 742 827	Publication of the first ENEM schools league table.
2007	3 584 569	The school's ID number is kept the same over the years, making it possible to perform longitudinal analysis.
2009	4 148 721	Structural changes in the exam, including the skills matrix, the format of the test and its duration.
2010	4 626 094	The ENEM starts to be used as a certificate of completion of high school education for any citizen over 18 years of age. Most federal universities use the ENEM as a selection criterion.

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

1. Methodology

Given the hierarchical structure of the available data, this study performed multilevel regression analyses, the technique considered most appropriate for such cases (Raudenbush and Willms, 1995). The analysis encompassed three levels (individual, school and state) and used specific open-code software packages (Bliese, 2012). The microdata for the 2009 and 2010 ENEM and the 2009 School Census were obtained from the INEP portal.

The univariate and multivariate models were adjusted using a variety of variables pertaining to the schools and students (see table A.2 of the annex), including as a random effect not only the intercepts at the state and school levels, but also an indicator of the individuals' socioeconomic status (the SES component). In other words, the aim is to control for the effect of the states and isolate the effect of the school, allowing each school to have a different relation or gradient between the socioeconomic status and the student's score, in other words, its own socioeconomic gradient slope (GSE), as defined by the Organization for Economic Cooperation and Development (OECD, 2010).

The schools' value-added (obtained from cross-section data) was the random effect of this level in the complete models, which included all significant variables. The longitudinal value-added of the schools was obtained by simple subtraction from the results of 2010 and 2009. To calculate the explained variance of the multilevel models, their residuals were compared with the null models (Raudenbush and Bryk, 2002).

The methodology used suffers from a major shortcoming. Owing to the voluntary nature of the exam, there may be sampling bias, which would justify a two-stage regression analysis (Heckman, 1976). An analysis of that type could be undertaken in future studies, because there is not yet any version of the R program that performs multilevel regressions in two stages.

2. Calculation of the average scores by school

According to the official documentation (INEP/MEC, 2010), three averages were prepared for each school: one for writing, another for the objective test, and another general average calculated as the weighted-average of the first two. Nonetheless, the calculation of these averages only took account of students who:

(i) stated that they were in the final year of high school;
(ii) were attending regular high school or secondary education for young people and adults (consistent with the census data) either in traditional schools or in schools divided into cycles;[6] and
(iii) were present on the test days.

Given this sample, the documentation continues, the criteria for publishing the averages were as follows:

(i) do not publish if the participation rate was less than 2%;[7]
(ii) publish the average of the objective test if at least 10 students did the four tests;
(iii) publish the average of the writing test if at least 10 students did the writing; and
(iv) publish the overall average if at least 10 students did the four tests and the writing.

This study used the same criteria to clean up the data and calculate the averages, so as to reproduce the results posted on the Internet.[8] As part of that process, it was necessary to articulate the ENEM data with those of the census.

3. SES component

The socioeconomic status indicator (the SES component) was created using a methodology similar to that of the Programme of International Student Assessment (PISA) of 2009. A number of adaptations had to be made, particularly owing to the differences that exist in the data available in the questionnaire, not only between the ENEM and PISA program, but also, unexpectedly, between the Brazilian exams of 2009 and 2010.

After analysing the correlation matrix with various variables, five variables were chosen and were grouped in trios to prepare four possible candidates for the SES component. These only differed in terms of the variables included (see table 2). Then the correlations between the four candidates and the school averages were calculated, along with other metrics, with the aim of comparing the candidates in terms of their power to explain the score obtained in the ENEM.

Among the four candidates, the SES 0 component corresponds to the trio of variables that is most similar to that included in the socioeconomic index of the 2009 PISA program: the sum of household possessions, highest education level of the parents, and highest professional

[6] Schools organized in cycles combine several year groups (series) in a single class, generally because they are smaller.
[7] The participation rate for schools organized by cycles is divided by three.
[8] Because the INEP does not provide microdata on ENEM schools, but only those of the students.

TABLE 2

Brazil: different ways of calculating the SES according to the 2009 PISA program model

Possible SES components	Variables included	Adjusted coefficient of determination[a] (*percentages*)	Missing data[b] (*percentages*)	Proportion of variance in the first component (*percentages*)	Likelihood (Akaike information criterion)[c]
SES 0	- Sum of possessions - Highest family education level - Highest family professional level	12.5	10.6	64	8 477 566
SES 1	- Family incomes - Highest family education level - Highest family professional level	12.0	10.8	60	8 481 460
SES 2	- Family incomes - Average family education level - Sum of possessions	13.4	10.6	67	8 470 428
SES 3	- Highest family professional level - Average family education level - Sum of possessions	13.2	11	66	8 472 236

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

[a] Simple regressions with variables as the individual level.

[b] The percentage of missing data is calculated in relation to the total number of records with questionnaires replied to and the general average as calculated (811,406).

[c] To be able to compare likelihoods, the missing data from all of those variables were removed, leaving 704,481 records with the average by school and the four SES components at the individual level. These were used to reconstruct the models and calculate the Akaike information criterion (AIC) for each one.

SES: socioeconomic status indicator.

PISA: Programme of International Student Assessment.

level of the parents.[9] Nonetheless, the indicator that best fit all the criteria was component SES 2. As the aim was to control for the contextual variables to the maximum, this component was chosen as the study's socioeconomic variable. Curiously, it consists of two economic variables, unlike the indicator used in the PISA programme. The significance of this result is controversial. It could be a matter of choice, a natural variance in the indices; or else the index of professional level, prepared using the results from more developed countries, may be unsuited to the context of Brazilian society, or simply inappropriate for some other reason. It is also possible that the economic background of the family interferes more in the students' scores than the parents' profession. This study has not investigated those hypotheses, but merely sought the model the best fitted the scores of the students in the 2009 ENEM, after filtering data according to the INEP criteria.

In essence, the SES component at the individual level is a principal components analysis with three variables:

(i) the sum of household possessions;
(ii) the parents' average educational level; and
(iii) family incomes.

Once calculated at the individual level, the index is aggregated, forming averages at the school and state levels (although the latter level was not significant and, therefore, does not form part of the analyses).

For the comparison between 2009 and 2010 in the longitudinal analysis, the SES component was calculated with a different trio of variables, namely the mother's education level, the father's education level, and family incomes. This was necessary because the questionnaire was truncated from one year to the next, and information on the professional level of the parents and household possessions in 2010 was eliminated. To test the stability of the longitudinal data, correlations between the two years were calculated (Pearson correlation, $p<0.01$), using the same methodology (in terms of calculating the SES component and the variables included in the complete model). Correlations were also calculated using different methodologies (the best model of 2009 and the common model of 2009 and 2010), to test the influence of the choice of variables on the stability of the longitudinal data.

[9] Respectively the HOMEPOS, PARED and HISEI, indices described in OECD (2010).

4. The available data

The ENEM microdata are highly heterogeneous. The basic cleaning process left about one quarter of the data from the two years (see table 3), largely owing to the absence of the school ID number in the individual records. Moreover, even among the identified schools, little over half of the data were left after the INEP validity criteria had been applied.

This is not problematic in itself, but can impose several limitations on the ENEM, in relation to the schools comparison, which depends on a number of partly arbitrary criteria, such as a minimum of 10 students, or a minimum 2% participation rate.

TABLE 3

Brazil: longitudinal data available on the schools in the National High School Exam (ENEM) 2009-2010

	ENEM 2009	ENEM 2010
Data with school identification	1 536 023	1 379 447
Data with school identification (*percentages*)	37,0	29,8
Number of schools identified	32 006	32 318
Number of schools participating in both years	28 010	
Number of schools valid in both years[a]	17 359	

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009 and 2010.

[a] According to INEP criteria for calculating and disseminating the averages by schools.

V
Results

Based on the foregoing observations, the first relevant question is whether the schools that participate in the ENEM constitute a representative sample of Brazilian schools. More specifically, it needs to be asked whether the lowest schools in the ENEM league table can be considered the country's worst in terms of enabling their students to enter higher education. The comparison with microdata from the 2009 school census (see table A.1 of the annex), shows that the schools that participate in the ENEM are usually better placed than the national average, so they do not constitute a particularly representative sample of all schools.[10]

In other words, the ENEM league table represents a distorted sample of Brazilian high school education overall, in which the best-placed schools are preselected. There seems to be a sample selection problem in relation to the schools, probably arising from the voluntary nature of the exam. This would be an argument against using the ENEM league table as an indicator of school effectiveness in Brazil in accountability policies. Even publication of the averages is a policy of this type (albeit moderated), so it would be desirable to investigate other arguments in greater depth, which oppose or support such publication.

A longitudinal comparison of the two years reveals a degree of stability in the results (see tables 4 and 5) despite the voluntary nature of the exam. This could be related to its consolidation in society and in the schools, and also to its use to gain access to higher education, and, in addition, to the item-response theory that was first used in this domain in 2009.[11] It is also possible that there is a relation between the stability of the averages and factors that are external to the school or exam, as will be discussed below.

[10] This result and some others presented here are also reported in Rodrigo Travitzki's Ph.D. thesis.

[11] Some studies find that item-response theory tends to produce more stable results over time, compared to the classical theory of tests or contrast, used until 2008 (Andrade, Tavares and Da Cunha Valle, 2000).

TABLE 4

Brazil: descriptive analysis of the longitudinal data on ENEM schools 2009-2010[a]

	2009		2010	
	Average	Standard deviation	Average	Standard deviation
Schools: general score	534	56	537	54
Schools: score on objective test	494	55	505	53
Schools: score on writing test	575	65	570	61
Schools: number of participants in the objective test	45	45	55	53
Schools: number of participants in the writing test	44	44	54	53

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009 and 2010.

[a] Only in the 17,359 schools that are valid in both years. The five differences between the years were statistically significant ($p<0.001$).

TABLE 5

Brazil: correlations between the averages scores obtained by the ENEM schools and the value added by the schools 2009-2010

Variable	Same method,[a] different year	Different method, different year	Different method, same year
Raw average	0.84	0.84	1
Value-added	0.46	0.43	0.96
Slope of the socioeconomic gradient (GSE)	0.18	0.17	0.83

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009 and 2010.

[a] In reality, one small part of the method: the trio of variables introduced in the principal components analysis to prepare the socioeconomic status (SES) component.

The results for 2010 were slightly better than those of 2009, thanks to the scores obtained on the objective test, since the scores on the writing test dropped. What does this difference mean: a value-added, a difference between generations, or a normal variance in indicators of this type?

Firstly, one might interpret this as genuine progress in the school results, and that the disparity in the scores should not be attributed to a different degree of difficulty in the tests, because the questions are analysed in advance, and the scores are calculated (in reality, they are estimated) using the three-parameter logistic function of item-response theory developed by Birbaum in 1968 (INEP/MEC, 2011b). Nonetheless, the ability scales can only be adequately equalized if there are common items in the two tests (Andrade, Tavares and Da Cunha Valle, 2000), something which is unviable in a standardized and printed exam that is used as an admission test into good-quality and free universities.

Secondly, one can start from the principle that it is difficult for the school really to improve in just one year. In this regard, it would be crucial for the exam-based school effectiveness measures to be multiple, in other words that they covered more than one year. In longitudinal models of value-added (relative to how much the student improves over time), "most authors recommend using at least three measures" (Martínez Arias, 2009, p. 225).

1. Multilevel analysis of the 2009 ENEM

The 2009 ENEM microdata consists of the records of a total of 4,148,721 students, of whom 2,218,191 answered the socioeconomic questionnaire, and just 1,339,445 were in the final grade of high school, according to their own declarations. Applying the INEP validity criteria left 811,406 individual records for the multilevel analysis of the school league tables.

The concordance at level 1 (states) was 0.77, and at level 2 (schools) it was 0.82, which shows that there is more consistency between the scores obtained by students from the same schools than between students from the same state, as would be expected. These numbers also suggest that both levels are significant in the analysis, which was confirmed by comparing the likelihood of models with and without these variables.

The intra-class correlation coefficient, for each of the levels separately, was 0.25 for the schools and

0.03 for the states. Nonetheless, when the intra-class correlation coefficient is calculated in the three-level model, the contribution of the state remained at 3%, while that of the school dropped to 22%, which still left 75% of the total variance for the individual level (see table A.4 of the annex). This means that 3 percentage points of the 25% initially attributed to differences between the schools can in fact be attributed to the difference between states. In a multilevel regression study using results from another Brazilian exam, the Basic Education Assessment System (SAEB), it was estimated that the proportion of the variance in individual results that could be explained by the school was 39%, which is considerably more than the proportion normally found in developed countries (around 20%), and could be due to the large differences between schools in Brazil (Franco and others, 2007). In that regard, the differences between the schools included in the ENEM league table are more like those found in developed countries than between Brazilian schools generally. This is probably due to the different characteristics of the two tests, because the SAEB is done by sampling and aims to represent all Brazilian schools, whereas the ENEM is voluntary and serves as a higher education admission exam.

In order to investigate the influence of the characteristics of the students and schools on ENEM scores, the various multilevel models were adjusted using the SES component and other variables available in the microdata from the test and the questionnaire (see table A.2 of the annex).

Table A.3 of the annex reports the coefficients of those models, making it possible to determine the extent to which context alters the characteristics of the students' scores. The effect of most of the variables declines when the two SES context variables are introduced, but it is hardly altered by the introduction of the others, which suggests that the variable constructed with the 2009 PISA program methodology has a high level of explanatory power.

The variables that change most at the school level include the administrative dependency of the school. In an initial analysis, private schools seem much better than state schools, for example, but the difference between them declines significantly in the second column. Federal schools, meanwhile, remain well ahead of the others once context is taken into account, which shows that they are highly efficient institutions and capable of producing good results even in unfavourable circumstances.

Something similar happens with the proportion of individuals of white race, which is five times less important after taking account of contextual factors. Nonetheless, if one considers that skin colour is one of the variables introduced at the individual level, the fact that the proportion of white students in the school remains significant in the complete models suggests that this characteristic has a major influence on the results generated by Brazilian education. This conclusion is corroborated by the relative stability of the influence of skin colour at the individual level, as can be seen in table A.3 of the annex.

Unlike administrative dependency, the type of school seems to have an influence that is less related to context. Schools for adults, for example, achieve inferior results to all of the others in the three columns, with small variances. This suggests that the differences between the types of school are structural, and that the degree of comparison is small, a fact which should be taken into account in the way ENEM results by schools are published.

Table A.3 of the annex shows that context has a greater influence with respect to the school than to the individual. To verify this hypothesis, a model was fit with the two standardized variables (z-score), and a coefficient of 39 was obtained for the school SES and one of 10 for the individual SES.

Consequently, it can be said that for a family in less favourable circumstances, it would be worthwhile taking their children to a school attended by students from more favourable backgrounds. The fact that this analysis is based on the results of a selection test for admission to free universities (and to gain access to scholarships in the case of private universities), reinforces the conclusion.

Given the objective of investigating the effect of contextual variables in the ENEM schools league table, the magnitude of the explanatory power of context in relation to the students score, or, to be more specific, the percentage variance explained by the models, is particularly relevant. In this regard, the results of this study are significant both for the discussion of value-added methodologies, and for the ENEM schools league table as such.

On the methodological front, corroborating the foregoing conclusions, it can be seen that the SES component, inspired in the PISA program, has high explanatory power in terms of the schools' results (75%) and that introduction of the other variables increased the explained variance only slightly to 79% (see table A.4 of the annex). Accordingly, it seems reasonable to only use the SES component to control for context if the aim is to make a more practical analysis with a minimum of missing data. Nonetheless, for the purposes of this study, it is more appropriate to use the

complete model, which fit the data better according to the likelihood analysis.

In terms of using ENEM as an indicator of school quality, table A.4 of the annex highlights a number of significant limitations. Although the result with respect to individuals continues to be sufficiently explained by the variables contained in the model, the same is not true in relation to the schools and the states. That means that, at most, 21% of the variance in scores obtained in these institutions can be attributed to the school's effort and merit, since that is the percentage that is not determined by factors outside its control. If the same analysis is done separately for the two "sub-averages", this number drops to 13% in the scores on the objective test and rises to 38% in the scores on writing (see table A.5 of the annex).

Although this result (unprecedented with ENEM data) is no novelty in the international research scenario, it is still worrying, because this exam has consolidated its status as a relevant indicator of school quality in Brazil year by year.

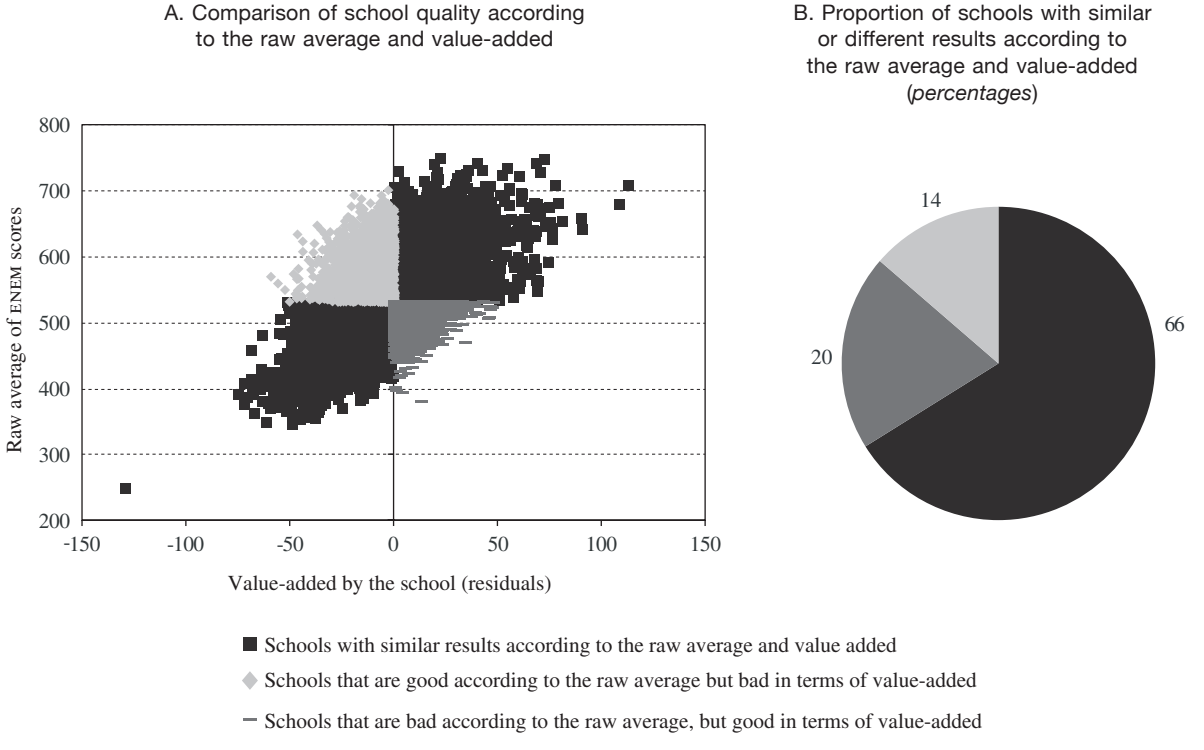
In other words, the analysis of the variance of the residuals at the different levels of the hierarchical models suggests that this individual exam could evaluate the students' merits, but contains little information on the merit of the schools and the states, when contextual conditions are taken into account.

Lastly, it is worth asking to what extent the "raw results" per school (the averages published annually) are different from the "net results", in other words the value-added after taking account of contextual conditions. There is some correlation between the two scores ($r = 0.51$ $p<0.001$) as can be seen in figure 3, where each point represents a school located in the plane defined by two quality measures.

Classifying the schools simply as "good" or "bad" (above and below the average, respectively), in 66% of cases the two indicators produced similar results. Nonetheless, some 7,000 schools were evaluated contrastingly by the two criteria, which is very significant number. More specifically, 14% of the schools were rated "good" according to the raw average, but "bad" in

FIGURE 3

Brazil: comparison of the raw average of scores in schools participating in the ENEM, and the value-added by their schools, 2009



Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

terms of value-added, while in 20% of cases the opposite occurred (see figure 3).

This finding is deeply problematic, because a fifth of the schools were evaluated as bad in the ENEM league table, but obtained better than expected results when their contextual conditions were take in to account. This means they were schools of merit, probably operating in an unfavourable context; yet they could be being undervalued by the annual publication of averages by schools presented in the form of a league table by the mass media. This effect is probably intensifying owing to the differences found between the schools included in the INEM and the average of Brazilian schools generally. Thus, in its current form, this policy to promote the quality of school teaching may in many cases be having the opposite effect.

2. Longitudinal analysis: 2009 and 2010

Thus far, this study has considered the value that the schools have supposedly added to the students (in terms of the results obtained on the Brazilian ENEM), using cross-section data. To conclude, the other concept of value-added (see figure 2) will be applied to the ENEM microdata using longitudinal data, in other words how much value has been added to the schools over a certain period. Given the shortcomings in current data (see table 1), this part will probably be more useful for the methodological analysis then for actually evaluating Brazilian schools.

How does the introduction of contextual variables affect the stability of the longitudinal data? Correlations were calculated between the raw and "net" averages (value-added) of the schools in 2009 and 2010, using two different methods in 2009. The results reveal considerable stability in the raw averages, but this is lost when contextual conditions are controlled for (see table 5). This can be taken as an argument in favour of publishing the raw averages as is currently done, because they would constitute a reasonably stable and robust measure of school effectiveness. Nonetheless, bearing in mind that 79% of the variance in the results by schools is explained by context (see table A.4 of the annex) and that the correlation between the SES components of the schools in 2009 and 2010 is 0.95, it seems plausible to conclude that the stability of the ENEM averages is more reflective of contextual conditions than the schools' own merit.

Table 5 also shows that the use of other contextual variables does not produce very different results in the schools, unlike what is reported in other studies. One possible explanation of this stability of value-added based on different variables would be the inclusion of many variables in the models (see table A.3 of the annex), which could cause a group effect that is reasonably resistant to change in any of its constituent parts. Moreover, with respect to the states, the use of different methods reduces the stability of value-added (see table 6), which demonstrates the complexity of this type of methodology.

Table 5 also shows that the slope of the GSE, in other words the magnitude of the change in scores based on the SES component, varies greatly from year to year. This raises a number of questions about the reliability of this type of measure at the school level. Among the states, however, both the slope and the value-added are highly stable between the two years when the same method is used (see table 6).

Another relevant question concerning the longitudinal stability of the ENEM league table is whether that stability is homogeneous across the different strata, in other words between schools considered good, medium, or bad. When the three strata are analysed separately, the position in the league table varies more among schools

TABLE 6

Brazil: correlations between the averages of scores corresponding to the states in the ENEM and values added by the states, 2009-2010

Variable	Same method,[a] different year	Different method, different year	Different method, same year
Raw average	1	1	1
Value-added	0.93	0.60	0.66
Slope of the socioeconomic gradient (GSE)	0.91	0.85	0.95

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009 and 2010.

[a] In reality, one small part of the method: the trio of variables introduced in the principal components analysis to prepare the socioeconomic status (SES) component.

rated bad than among those rated good (see table A.6 of the appendix), which could be taken as a second piece of evidence[12] that the ENEM league table is not a very reliable source of information for comparing the worst schools. Nonetheless, it may be a good reference for comparing the better schools, although more in relation to effectiveness than school merit as such.

In the case middle-ranked schools particularly, small differences in their average scores cause large differences in their position in the league table, which supports the idea that the averages are being interpreted too precisely.

[12] The first is obtained from table A.1 of the annex.

VI
Conclusions

The results obtained here can be divided into two groups, one relating to value-added methodologies, and the other to the exam as such. The first case investigated the effects caused by the introduction of contextual variables. The second aimed to identify the type information provided by the ENEM league table to Brazilian society.

In methodological terms, the multilevel analysis of the 2009 ENEM showed that, on its own, the contextual variable based on the PISA programme has a high level explanatory power for the students' score, particularly in relation to the schools (75%), whereas the introduction of the other variables only increased the proportion of the variance explained to 79%. Context is found to be four times more important for schools than for individuals, whereas between states it was not significant compared to the other two. When comparing the performance of the schools in terms of raw average and value-added, 34% of the results were contradictory; in other words, the merit of one third of the institutions was not adequately evaluated when the different contexts were taken into account.

The longitudinal analysis showed that there was reasonable stability in the raw average between the two years (r = 0.8), which decreased when the contextual variables were introduced (r = 0.5). The slope of the GSE, in turn, was highly unstable (r = 0.2), which raises a number of questions about the reliability of this variable, particularly when a single measurement is made. At the state-level, all indicators behaved stably between 2009 and 2010.

In contrast, the performance differences between public and private schools were substantially reduced when contextual conditions were considered, which did not happen when normal schools were compared with schools for adults. Performance differences between the different "races" (skin colour) were also maintained after considering the effect of socioeconomic context, although further studies are needed on this.

The 2009 ENEM league table covers 35% of schools providing secondary education in Brazil that have infrastructure conditions above the national average, owing to the purpose and voluntary nature of the exam. The multilevel analysis revealed that 3% of the variance in scores can be attributed to the state, 22% to the schools, and 75% to the students. The scores on the objective test were more influenced by context (87%) than were the scores on writing (62%), which could mean that writing is fairer (in terms of merit) or less reliable than the former. Further studies need to be done to investigate these two hypotheses. When comparing the two years, it can be seen that the averages are more stable in the "better" schools, and that the "middle-ranked schools" display large variations in terms of their position in the league table, but small variations in terms of average score.

These results show that the ENEM league table suffers from major shortcomings as an indicator of school quality at the national level. The lowest schools in this league table should not be considered the worst in Brazil, and the difference in averages between the "middle-ranked" schools is very small. The best schools, on the other hand, display some stability in terms of raw average. This could mean that the exam is more informative for the higher strata of the ability scale, which would be understandable given its use as a higher education selection criterion. Other studies need to be done to verify this hypothesis.

Ultimately, what type of information does the ENEM school league table provide to Brazilian society? The results of this study show that, when confined to raw performance, the league table is more representative of

the socioeconomic conditions of the schools than their merit, bearing in mind the contextual differences. This is due to the fact that: (i) context can explain four-fifths of the variance in scores between schools; (ii) the raw averages are stable, and value-added is unstable; and (iii) the contextual conditions of the schools are even more stable (r = 0.95) in the two years analysed.

Consequently, this indicator of school quality could appropriately be used by families wishing to choose a university for their children and who enjoy good economic circumstances. Nonetheless, for less favoured schools and for the state (in relation to responsibility and accountability policies), the ENEM league table provides little information and can even be misleading. Giving it undue importance could aggravate inequalities between schools, because it would under-rate institutions that do a good job in precarious conditions, while favouring those that cater to the upper socioeconomic strata of Brazilian society. The fact that these conclusions confirm other results reported in international literature points to the need to create other indicators of school quality in democratic countries.

ANNEX

TABLE A.1

Brazil: comparison of schools present in the ENEM with total schools, 2009
(Number of schools and percentages)

Characteristic	Schools in the ENEM league table	All high schools[a]	All schools
Number of schools[b]	18 605	30 554	255 445
Urban[c]	97.2	92.1	53.0
Private[c]	24.6	30.9	19.8
Public water supply network[c]	93.5	90.3	64.1
Public sewerage network[c]	69.0	66.2	40.7
Computer on[c]	90.3	81.8	23.2
Science laboratory[c]	53.2	43.4	7.2
Sports field[c]	80.1	68.6	21.5
Library[c]	75.1	71.3	25.3
Photocopier[c]	67.6	63.5	30.7
Broadband Internet[c]	76.5	71.4	32.2

Source: Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) and School Census of 2009.

[a] Only regular education and the education of young people and adults.
[b] Number of schools.
[c] Percentages.

TABLE A.2

Brazil: variables included in the complete model using ENEM 2009 data

Variable	Level[a]	Type
Average	0	Numerical
Skin colour	0	Categorical
Sex	0	Categorical
Religion	0	Categorical
Individual socioeconomic status component (SES)	0	Numerical
School socioeconomic status component (SES)	2	Numerical
Administrative dependency	2	Categorical
Modality	2	Categorical
Proportion of white students	2	Numerical
Proportion of students who completed the preparatory course	2	Numerical

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

[a] The levels are numbered from the most general to the most specific (10 = state; 2 = school), and level 0 corresponds to the individual.

TABLE A.3

Brazil: effects of the introduction of contextual variables using ENEM 2009 data

	Univariate models	Models with socioeconomic status components (SES) in the two levels	Complete model
Socioeconomic status component (SES)			
SES (individual)	10	8	8
SES (school average)	41	36	22
Individual level			
Sex [female]	0	0	0
Sex [male]	-15	-17	-17
Colour [white]	0	0	0
Colour [brown]	-9	-7	-6
Colour [black]	-16	-13	-11
Colour [yellow]	-11	-10	-10
Colour [indigenous]	-34	-31	-29
Religion [Catholic]	0	0	0
Religion [Protestant/Evangelical]	7	7	9
Religion [Spiritism]	13	9	10
Religion [Umbanda/Candomblé]	-9	-13	-8
Religion [other]	13	11	14
Religion [no religion]	17	14	18
School level			
Administrative dependency [federal]	0	0	0
Administrative dependency [state]	-108	-65	-66
Administrative dependency [municipal]	-97	-57	-57
Administrative dependency [private]	-17	-39	-41
Type [regular]	0	0	0
Type [young people and adults]	-48	-42	-42
Proportion of white students	162	41	30
Proportion of students completing the preparatory course	109	8	8

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

Note: in reality, the zeros in these categories are references of the categorical variables, whereas the coefficients in the other categories relate to the first.
Coefficients of various models ($p<0.01$), with the students' score as the response variable. In the first column, the explanatory variables are shown alone; in the second they are accompanied by the SES component with respect to the individuals and schools; and the third column shows the model consisting of or joint variables. As the variables are not standardized, comparisons must be made horizontally. The only vertical comparisons that make sense are those between factors of the same categorical variables (identified by square brackets).

TABLE A.4

Brazil: variance of the residuals and explained variance of the ENEM results of 2009

	Level 1 intercept (state)	Level 2 intercept (school)	Individual residuals
Variance in model 0	356	2 507	8 482
Variance in model 1 (only individual level variables)	197	1 658	8 191
Variance in model 2 (SES in the two levels)	121	637	8 305
Variance in model 3 (complete)	74	529	8 129
Variance within model 0	0.03	0.22	0.75
Explained variance of model 1	0.45	0.34	0.03
Explained variance of model 2	0.66	0.75	0.02
Explained variance of model 3	0.79	0.79	0.04

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

TABLE A.5

Brazil: variance of the residuals of the ENEM objective and writing test of 2009
(Absolute values and percentages)

	Level 1 intercept (state)	Level 2 intercept (school)	Individual residuals
Score on the objective test (null model)	453	2 500	4 133
Score on the writing test (null model)	348	3 064	23 497
Score on the objective test (complete model)	88	312	3 929
Score on the writing test (complete model)	240	1 160	22 413
Explained variance of the objective test (*percentages*)	80.6	87.5	4.9
Explained variance of the writing test *(percentages)*	31.0	62.1	4.6

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009.

TABLE A.6

Brazil: standard deviations of the differences in results, by strata, 2009-2010

	Standard deviation of the difference between raw averages	Standard deviation of the difference in positions in the league table
Best 2 000 in 2009	26.6	1 055
Middle 2 000 in 2009	25.8	3 300
Worst 2 000 in 2009	32.4	2 803

Source: National Institute for Educational Studies and Research (INEP), microdata from the National High School Exam (ENEM) of 2009 and 2010.

Bibliography

Andrade, E.C. (2008), "'*School accountability*' no Brasil: experiências e dificuldades", *Revista de Economia Política*, vol. 28, No. 3, São Paulo.

Andrade, D.F., H.R. Tavares and R. da Cunha Valle (2000), *Teoria da resposta ao item: conceitos e aplicações*, São Paulo, Brazilian Statistical Association.

Apple, M.W. (2001), "Comparing neo-liberal projects and inequality in education", *Comparative Education*, vol. 37, No. 4, Taylor & Francis.

Bliese, P. (2012), "Multilevel: Multilevel Functions. R package version 2.4" [online] http://cran.r-project.org/package=multilevel.

Brandão, Z. (2000), "Fluxos escolares e efeitos agregados pelas escolas", *Em Aberto*, vol. 17, No. 71, Brasilia, National Institute for Educational Studies and and Research (INEP).

Dubet, F. (2004), "O que é uma escola justa?", *Cadernos de Pesquisa*, vol. 34, No. 123.

Ferrão, M.E. (2009), "Sensibilidad de las especificaciones de los modelos de valor añadido: Midiendo el estatus socioeconómico", *Revista de Educación*, No. 348, Madrid, Ministry of Education, Culture and Sport.

Franco, C. and others (2007), "Qualidade e eqüidade em educação: reconsiderando o significado de 'fatores intra-escolares'", *Ensaio*, vol. 15, No. 55, Rio de Janeiro.

Government of Spain (2006), "Ley orgánica 2/2006, de 3 de mayo, de educación" [online] http://www.boe.es/aeboe/consultas/bases_datos/doc.php?id=BOE-A-2006-7899.

Heck, R.H. (2000), "Examining the impact of school quality on school outcomes and improvement: a value-added approach", *Educational Administration Quarterly*, vol. 36, No. 4, SAGE.

Heckman, J.J. (1976), "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models", *Annals of Economic and Social Measurement*, vol. 5, No. 4, Cambridge, Massachusetts, National Bureau of Economic Research (NBER).

INEP/MEC (National Institute for Educational Studies and Research/Ministry of Education) (2011a), "Nota técnica. Médias do ENEM 2010 por escola" [online] http://inep.gov.br/.

_____ (2011b), "Nota técnica. Procedimento de cálculo das notas do Enem" [online] http://inep.gov.br/.

_____ (2010), "Nota técnica. Médias do ENEM 2009 por escola" [online] http://inep.gov.br/.

_____ (2007), "Nota técnica. Notas médias do Enem 2006 por município e por escola dos alunos concluintes do ensino médio em 2006" [online] http://inep.gov.br/.

Karsten, S., A. Visscher and T. de Jong (2001), "Another side to the coin: the unintended effects of the publication of school performance data in England and France", *Comparative Education,* vol. 37, No. 2, Taylor & Francis.

Keeves, J.P., N. Hungi and T. Afrassa (2005), "Measuring value-added effects across schools: should schools be compared in performance?", *Studies in Educational Evaluation*, vol. 31, No. 2-3, Amsterdam, Elsevier.

Ladd, H.F. and R.P. Walsh (2002), "Implementing value-added measures of school effectiveness: getting the incentives right", *Economics of Education Review*, vol. 21, No. 1, Amsterdam, Elsevier.

Linn, R.L. and C. Haug (2002), "Stability of school-building accountability scores and gains", *Educational Evaluation and Policy Analysis*, vol. 24, No. 1, Washington, D.C., American Educational Research Association.

Luyten, H. (1994), "Stability of school effects in Dutch secondary education: the impact of variance across subjects and years", *International Journal of Educational Research*, vol. 21, No. 2.

Mandeville, G.K. (1988), "School effectiveness indices revisited: cross-year stability", *Journal of Educational Measurement*, vol. 25, No. 4, Wiley.

Martínez Arias, R. (2009), "Usos, aplicaciones y problemas de los modelos de valor añadido en educación", *Revista de Educación*, No. 348, Madrid, Ministry of Education, Culture and Sport.

Martínez Arias, R., J.L. Gaviria Soto and M. Castro Morera (2009), "Concepto y evolución de los modelos de valor añadido en educación", *Revista de Educación*, No. 348, Madrid, Ministry of Education, Culture and Sport.

McCaffrey, D.F. and others (2004), "Models for value-added modeling of teacher effects", *Journal of Educational and Behavioral Statistics*, vol. 29, No. 1, SAGE.

MEC (Ministry of Education) (2002), *Examen Nacional de Ensimo Médio (ENEM): documento básico 2002*, Brasilia.

Meyer, R.H. (1997), "Value-added indicators of school performance: a primer", *Economics of Education Review*, vol. 16, No. 3, Amsterdam, Elsevier.

Murillo Torrecilla, F.J. (2005), *La investigación sobre eficacia escolar*, Barcelona, Octaedro.

OECD (Organization for Economic Cooperation and Development) (2010), *PISA 2009 Results. Overcoming Social Background: Equity in Learning Opportunities and Outcomes (Volume II)*, Paris, OECD Publishing.

Oliveira, R.P. de and G.C. Araujo (2005), "Qualidade do ensino: uma nova dimensão da luta pelo direito à educação", *Revista Brasileira de Educação*, No. 28, Rio de Janeiro.

Passarinho, N. (2012), "MEC vai substituir a Prova Brasil pelo Enem em cálculo do Ideb, diz ministro" [online] http://g1.globo.com/vestibular-e-educacao/noticia/2012/08/mec-vai-substituir-prova-brasil-pelo-enem-em-calculo-do-ideb-diz-ministro.html.

Raudenbush, S.W. and A.S. Bryk (2002), *Hierarchical Linear Models: Applications and Data Analysis Methods*, Thousand Oaks, SAGE Publications.

Raudenbush, S.W. and J.D. Willms (1995), "The estimation of school effects", *Journal of Educational and Behavioral Statistics*, vol. 20, No. 4, Washington, D.C., American Educational Research Association.

Ray, A., H. Evans and T. McCormack (2009), "El uso de los modelos nacionales de valor añadido para la mejora de las escuelas británicas", *Revista de Educación*, No. 348, Madrid, Ministry of Education, Culture and Sport.

Reckase, M.D. (2004), "The real world is more complicated than we would like", *Journal of Educational and Behavioral Statistics*, vol. 29, No. 1, Washington, D.C., American Educational Research Association.

Saunders, L. (1999), "A brief history of educational 'value-added': how did we get to where we are?", *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, vol. 10, No. 2, Taylor & Francis.

Tekwe, C.D. and others (2004), "An empirical comparison of statistical models for value-added assessment of school performance", *Journal of Educational and Behavioral Statistics*, vol. 29, No. 1, Washington, D.C., American Educational Research Association.

Thomas, S. (1998), "Value-added measures of school effectiveness in the United Kingdom", *Prospects*, vol. 28, No. 1, Springer.

Thomas, S. and others (1997), "Stability and consistency in secondary schools' effects on students' GCSE outcomes over three years", *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, vol. 8, No. 2, Taylor & Francis.

Van Petegem, P. and others (2005), "Publishing information on individual schools?", *Educational Research and Evaluation: An International Journal on Theory and Practice*, vol. 11, No. 1, Taylor & Francis.

West, A. and H. Pennell (2000), "Publishing school examination results in England: incentives and consequences", *Educational Studies*, vol. 26, No. 4, Taylor & Francis.

Willms, J.D. (2006), *Learning Divides: Ten Policy Questions About the Performance and Equity of Schools and Schooling Systems*, Montreal, UNESCO Institute for Statistics.

Zhang, X. and Y. Zha (2010), "On 'abandoning examination" phenomenon of Gaokao" [online] http://en.cnki.com.cn/Article_en/CJFDTOTAL-HJKS201012007.htm.