PEL Pensamiento Educativo. Revista de Investigación Educacional Latinoamericana
2014, 51(1), 103-122

Pontificia Universidad Católica de Chile
Facultad de Educación

Estimating Value-Added Models: Evidence on Teacher Effectiveness from São Paulo's Municipal Schools

Estimando modelos de valor agregado: evidencias sobre la eficacia de los maestros de las escuelas municipales de São Paulo

Gabriela Miranda Moriconi

Fundação Carlos Chagas, Brazil

Abstract

Given the importance of teachers to the educational process and availability of data for estimating a value-added model, this paper is part of the current efforts to estimate teacher effects and their variation, and to obtain evidence on what can explain it. This paper is intended to estimate individual value-added measures of a sample of 4th grade teachers from São Paulo's municipal schools, based on student achievement gains in tests on reading and math in 2010. The results indicated a variation from 0.062 to 0.45 standard deviation of teacher effects measured in terms of standard deviations of student achievement in reading and 0.059 to 0.43 in math. There were positive associations between the following variables and teacher effects: frequency of homework assignment and frequency of use of Support Workbooks. The analysis of reliability and stability of estimated teacher effect measurements has shown a limited capacity to support recommendations regarding personnel policies, but permitted the identification of 13% of the teachers with different effects from the estimated average. Due to this distinction, these teachers are the ideal target for future research on their teaching practices, especially those observed as being positively associated with the estimated teacher effects.

Keywords: effectiveness, teacher effects, value-added models

Post to:
Gabriela Miranda Moriconi
Department of Educational Research at Fundação Carlos Chagas, Brazil.
Rua Quitanduba, 363 – CEP: 05516-030 – São Paulo/SP, Brazil.
Email: gmoriconi@fcc.org.br

This paper is a result of the author's doctoral thesis in Public Administration and Government at Fundação Getulio Vargas (EAESP-FGV). The author thanks GVpesquisa and Coordination for Improvement of Higher Education Personnel (CAPES), for having funded the doctoral studies. She thanks the advisor, Professor Nelson Marconi, and Professors Reynaldo Fernandes, Paula Louzano, José Francisco Soares and André Portela, who in many ways contributed to the development of the thesis.

ISSN: 0719-0409 DDI: 203.262, Santiago, Chile
doi:10.7764/PEL.51.1.2014.9

Resumen

Dada la importancia de los docentes en el proceso educativo y la disponibilidad de datos para la estimación de modelos de valor agregado, este trabajo es parte de los esfuerzos para estimar los efectos docentes y su variación y obtener evidencias para explicarla. Este trabajo pretende estimar medidas de valor agregado individuales de una muestra de docentes de 4.º grado de las escuelas municipales de São Paulo, a partir del crecimiento en las puntuaciones de sus alumnos en pruebas de lectura y matemáticas en 2010. Se encontró una variación de 0.062 a 0.45 en la desviación estándar de los efectos docentes medidos en términos de desviaciones estándares de las notas de los alumnos en lectura y de 0.059 a 0.53 en matemáticas. Hubo asociaciones positivas entre las siguientes variables y los efectos docentes: la frecuencia de tarea y la frecuencia de uso de los libros de apoyo. El análisis de fiabilidad y estabilidad de los efectos estimados mostró una capacidad limitada para apoyar las decisiones de personal, pero permitió identificar un 13% de maestros con efectos diferentes del promedio. Así, son el público ideal para la investigación futura sobre sus prácticas docentes, sobre todo las positivamente asociadas con los efectos docentes estimados.

Palabras clave: eficacia, efectos docentes, modelos de valor agregado

Teachers have a central role in the educational process. They are the ones who interact everyday with the students and whose main priority is students' learning. Teachers are responsible for making many critical decisions such as: what methods they will use to present ideas and develop skills, what level and kind of participation they will demand from the students, what kind of procedures they will use to motivate their students, how they will assess their students' learning and so on. Therefore, it is easy to understand the amount of attention that policymakers and researchers have given to policies to improve teaching.

In recent decades, there have been many initiatives at both national and international level to assess students' learning through standardized tests, mainly on reading and math. Since students' skills on reading and math are considered very important for their lives in contemporary societies, and since teachers have a large responsibility to promote the development of these skills, it makes sense to try to develop methods to find out how much teachers have contributed to their students' achievement.

Recently, there has been an extensive debate on teacher effectiveness, that is, their capacity to improve their students' achievement in large-scale tests. There are two main approaches related to the use of estimates of teacher effectiveness. The first is to try to identify the attributes that can explain teachers' effectiveness and then use this information to subsidize policies to improve teachers' quality. Some attributes analyzed are their knowledge, practices and characteristics. The second is to consider estimates of teacher value-added to student achievement as good representations of individual teacher effectiveness and use them —isolated or in combination with other criteria— to make personnel decisions, such as tenure, promotion and compensation.

The emphasis of the American federal government on policies based on students' achievement and on measures of their teachers' contribution to this achievement provided impetus to the development of extensive literature on the methodological challenges and pitfalls of estimating value-added models to generate these measurements.

In Brazil, there is no information about any state or municipality that has implemented high-stakes policies based on measurements of teacher effects on students' achievement in tests. There are only a few experiences of testing the same students in many subsequent years, which may allow us to try to estimate individual measurements of teacher value-added.

While the concerns of American researchers are mainly related to the direct impacts of policies based on estimates of teacher value-added, in this study the motivation is to improve existing knowledge about teacher effectiveness in Brazil. Even though teaching is a complex profession and quality teaching is a concept that can embrace many other aspects, teachers' capacity to improve students' learning in reading and math is a very important theme to be explored.

This paper is intended to estimate individual value-added measurements of a sample of teachers from São Paulo's municipal schools, based on student achievement gains in *Prova São Paulo*, an assessment system that tests students on reading and math. The purpose is to analyze the potential of those measurements to be used for decisions on individual teachers and to provide evidence about characteristics and practices of effective teachers as a whole. This is a theme yet to be explored in Brazil, since there are only scarce and recent longitudinal data that allow these kinds of studies.

Method

Theoretical framework

Value-added models to estimate teacher and school effects are of considerable interest to both researchers and policymakers. McCaffrey, Lockwood, Koretz, and Hamilton (2003) point out that measurements of teacher effects are of interest as a means of answering at least two broad questions: (a) Do teachers have differential effects on student outcomes? (b) How effective is an individual teacher at producing growth in student achievement, and which teachers are most or least effective? From the answer to the second question we can ask another question of great interest: (c) What explains teacher effectiveness?

The first question requires estimates of the variability among teacher effects. If the data and statistical models can accurately describe the contributions of teachers to student achievement, the models can provide estimates of the variability among teacher effects and determine the proportion of variability in achievement or growth that is attributable to teachers (McCaffrey et al., 2003).

In order to do that, most recent studies have used a dummy variable to identify each teacher that taught the students in the sample and its estimated coefficient represents the teacher effect. The most common strategy is to calculate the standard deviation of estimated teacher effects, correct it for sampling errors and relate it to the students' scores, which are usually standardized to mean zero and variance one (Aaronson, Barrow, & Sander, 2007; Goldhaber & Hansen, 2010; Koedel & Betts, 2011; Rockoff, 2004; Rothstein, 2009).

This literature has developed mostly in the USA, since the No Child Left Behind Act from 2002 demanded that the states test every student from 3rd to 8th grades every year on reading and math. So, there are plenty of longitudinal data from American states to estimate value-added models. For example, in Hanushek and Rivkin's (2010) literature review, they found 11 studies with this approach, based on American states or cities' data, published from 2004 to 2010. These studies indicate variations from 0.08 to 0.36 standard deviation of teacher effects measured in terms of standard deviations of student achievement, with the average standard deviation of 0.11 for reading and 0.15 for math. As explained by the authors, the math results imply that having a teacher at the 25th percentile as compared to the 75th percentile of the quality distribution would mean a difference in learning gains of roughly 0.2 standard deviation in a single year. And this is a large impact for a single school factor —the teacher— in the educational production function. According to Wright, Horn, and Sanders (1997), some studies based on the Tennessee Value-Added Assessment System (TVAAS) data showed that teacher effectiveness was the most important determinant of student achievement, having even higher effects than student background.

As Ravitch (2010) pointed out, by concluding that the teacher is the most important factor affecting students' learning and that there is a great variation in teacher effectiveness, the immediate implication of these studies is seemingly that more can be done to improve education by improving the effectiveness of teachers than by any other single factor. The findings of those studies supported the creation of the American federal program called Race to the Top. The idea of this program is to generate a race for federal funds among the states, in which the states must adopt some reform policies. One of them is using measurements of teachers' value-added to student achievement to guide personnel decisions such as compensation, promotion and tenure (United States of America, Department of Education, 2009). Both No Child Left Behind and Race to the Top have led to an increase in initiatives that use teacher effect estimates to guide personnel decisions, but also to an increase in studies discussing the challenges and pitfalls of using the estimates for high-stakes proposals (Braun, 2005; Koedel & Betts, 2011; McCaffrey et al., 2003; Rothstein, 2009).

In fact, it is the possibility of using a kind of estimate to reward or sanction individual teachers that places interest on the second question pointed out by McCaffrey et al. (2003), which requires estimation of individual teacher effects.

In this case, the studies also estimate teacher effects by using dummy variables to identify each teacher. But they usually focus their attention on the analysis of the quality of these measurements, especially on issues like their precision and stability. The idea is that possible consequences tied to an estimated effect will determine the acceptable levels for the precision and stability of the estimate.

One of the many concerns related to the precision of teacher effectiveness estimates is the influence of sample errors (McCaffrey et al., 2003). A family problem is a good example of a factor that is unrelated to teacher quality and may affect a student's achievement in a specific test, which will lead to sample errors in teacher value-added estimates.

Like any other measurement, by increasing the sample size, the influence of errors on the estimated effects decreases. Since many teachers teach only moderately small numbers of students, the variability in sampling error in estimated effects is likely to be large relative to the true variance of teacher effects (McCaffrey et al., 2003).

Lockwood, Louis, and McCaffrey (2002) found that, unless the ratio of variance of sampling error to the variance of teacher effects is less than about 0.1, estimated rankings will not be sufficiently precise to permit differentiation between all but the most extreme teachers.

This seems to be a difficult target to achieve. In another study, McCaffrey, Lockwood, Koretz, Louis, and Hamilton (2004) found that the sampling errors in the estimated teacher effect equaled about 20% to 40% of the variance of teacher effects. Though the estimated rankings were not precise enough to distinguish between all teachers, the authors were able to identify between one-third and one-quarter of teachers as distinct from the mean. This information can be very useful for diagnostic purposes, like the one pointed out by McCaffrey et al. (2003): to identify teachers who might be low or high performing, being a starting point for administrators (such as principals or superintendents) to target teachers for more thorough review. From a researcher's point of view, these teachers with low or high teacher effects can also be the target for qualitative research in order to find patterns of effective teaching.

Since there are many policymakers interested in using estimates of teacher effects for high-stakes purposes, one major concern is also determining an acceptable standard of stability. An evaluation in which the estimates put the teacher in the 10th lowest percentile one year and in the 10th highest percentile the following year raises certain doubts about its quality. Meanwhile, full stability is not expected. If the correlation between the estimates of teacher effects in two subsequent years were perfect, it would mean that the most effective teacher one year would also be the most effective the following year (Goldhaber & Hansen, 2010). If this happened, it would be an indication that teacher effectiveness is something permanent with no variation from year to year.

Some studies also based on data from American states have been focused on the analysis of the stability of estimates of teacher effects. McCaffrey, Sass, Lockwood, and Mihaly (2009) have obtained correlations of estimated teacher effects from 0.22 to 0.67. According to them, these correlations imply rankings of moderate stability: about a third of teachers from the highest quintile in one year remain in the highest quintile the subsequent year while a tenth fall to the lowest quintile. Goldhaber and Hansen (2010) have worked with 10 years of estimates of teacher effects and obtained correlations from 0.32 to 0.59 for the pairs of years. According to them, the level of variation observed is consistent with measurements of workers from other occupations of high complexity like teaching, for whom the correlations were from 0.33 to 0.40. This evidence seems to confirm that there is a permanent component of teacher effects, but it does not represent the full teacher effects.

Goldhaber and Hansen (2010) indicate that, in addition to the persistent component and the sampling errors, there is a non-persistent component in teacher effects. This non-persistent component may capture the variation in teacher performance over the years, but may also capture "good chemistry" in a particular class or a teacher randomly being sick for a week during flu season, which will affect the estimated teacher effect in a single year.

McCaffrey et al. (2009) also began to analyze estimates of teacher effects as the sum of three independent variables: the persistent teacher effect, the non-persistent teacher effects and the sampling errors. Unlike the sampling errors and the persistent effects, both studies show that it is not possible to directly estimate the variability of non-persistent effects. Instead, it is necessary to indirectly estimate the variance of non-persistent teacher effects as the remaining portion of the variance of teacher effects estimates, after subtracting the portion due to persistent effects and sampling errors.

By being able to follow the same teacher for 10 years, Goldhaber and Hansen (2010) found evidence that the sampling errors have an important part to play in determining the stability of teacher effect estimates through time. They correspond to about a third of the variation of teacher effects. Of the remaining variance, between one and two thirds are due to the variation in the persistent effects, and between one and two thirds are due to variation in non-persistent teacher effects.

As already indicated, from the second question presented, it is possible to come up with another question of long-time interest in this literature: what explains the variation in teacher effectiveness? In other words, which teacher characteristics, practices, beliefs and attitudes can be associated with higher teacher effects?

There is a wide range of studies aimed at answering this question. In another literature review based on data from American states or cities by Hanushek and Rivkin (2004), a teacher's degree level was the variable with the weakest evidence of association with their student's achievement gains. This also appeared in other studies published after their literature review, like those from Rivkin, Hanushek, and Kain (2005); Clotfelter, Ladd, and Vigdor (2007); and Croninger, Rice, Rathbun, and Nishio (2007).

In these studies based on data from the United States, researchers usually compare teachers with a master's or PhD with teachers with only a bachelor's degree. This way, most studies have shown that having a master's or PhD is not associated with higher teacher effects.

On the other hand, there are many studies that present evidence that teaching experience has positive and significant effects on student gains (Clotfelter et al., 2007; Croninger et al., 2007; Goldhaber & Hansen, 2010; Rivkin et al., 2005; Rockoff, 2004). Some of the estimates by Rockoff (2004) and Rivkin et al. (2005) have shown that the increase in teaching experience only produces impacts in the first few years of teaching, but not after the third year of teaching.

However, there are many doubts about the evidence regarding the association between teaching experience and teacher effectiveness. The main concern is that more experienced teachers usually have preference in choosing the school and tend to choose the ones with higher performance students (Hanushek, 1986; Hanushek, Kain, O'Brien, & Rivkin, 2005). This generates the possibility of the opposite relationship: students with higher performance attract more experienced teachers, or the causality works in both directions. Those kinds of problems will only be reduced if more studies like the one done by Rockoff (2004) find positive effects of teaching experience when comparing teachers within schools or having the same teachers in different schools in a certain period of time.

Other characteristics, like teacher scores on achievement tests and the type of certification, have received considerable attention. Hanushek and Rivkin (2004) point out that teacher scores have more frequently been correlated with student outcomes than teacher experience and education, but the evidence is far from overwhelming. Meanwhile, the literature provides mixed evidence on the effects of certification on teacher quality (Ballou & Podgursky, 2000; Clotfelter et al., 2007; Croninger et al., 2007; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005; Hanushek et al., 2005; Kane, Rockoff, & Staiger, 2006; Rockoff, Jacob, Kane, & Staiger, 2008).

Although there is a smaller number, there are also studies based on Brazilian data focused on analyzing the relationship between teacher characteristics and student achievement. For example, those done by Barros et al. (2001), Albernaz et al. (2002), Soares, J. F. (2003), and Soares, T. M. (2005).

Except for Soares, T. M. (2003), those studies are particularly focused on teachers' degrees. Both Barros, Mendonça, Santos, and Quintaes (2001) and Albernaz, Ferreira, and Franco (2002) found empirical evidence of a relationship between primary teachers' degrees and student achievement. Soares,

J. F. (2005) obtained evidence that students taught by teachers with a bachelor's degree in mathematics had higher scores in this subject. In Brazil, studies usually test the differences between teachers with high school qualifications, bachelor's degree, master's degree (*especialização*), academic master's degree and PhD, as there are still teachers in the country with no higher education degree, especially in some poorer areas.

A very common qualification in Brazil, the *especialização* —which we will call a master's degree in order to distinguish it from an academic master's degree (a *mestrado*)— is a professional training course for graduates with a minimum class load of 360 hours. Unlike an academic master's degree, at the end of an *especialização* course, the students do not present a thesis, they present a monograph.

When using the educational production function to study the Brazilian case, the biggest challenge is the shortage of longitudinal data in the country. Only recently have some municipalities and states started to produce data that allow the estimation of value-added models. Due to this, none of the studies revised that were based on Brazilian data were able to estimate students' achievement gains, being limited to analyzing the relationship between teacher characteristics and students' scores at the end of the school year.

By being the first work to produce estimates of teacher effects based on data from a Brazilian municipality, this study has the aim of contributing to teacher effectiveness debates, offering evidence of the quality of such individual and group estimates and their usefulness for supporting public policies to improve teaching.

Model

To estimate individual teacher effects, the covariate adjustment model expressed in the following equation was applied:

$$A_{ijst} = \theta A_{it-1} + \beta\gamma_i + \alpha L_{it} + \tau_j T_{ijt} + \delta_s S_{ist} + \eta_i \quad (1)$$

It specifies the current score of student *i* taught by teacher *j* at school *s* in the year *t* as a function of:

A_{it-1}, which represents the score of student *i* in the year *t-1*;

γ_i, which represents a vector of fixed characteristics of student *i*, including gender and whether they attended pre-school;

L_{it}, which represents a vector of varying characteristics of student *i* in the year *t*, including family income, parental education level, proper age for the school grade, and participation in a program for children who were still considered illiterate in the previous evaluation (called PIC —*Programa Intensivo de Ciclo*);

T_{ijt}, which represents a vector of dummy variables identifying teacher *j* that taught student *i* in the year *t*;

S_{ist}, which represents a vector of dummy variables identifying the school *s* that student *i* went to in the year *t*;

η_i, which represents the random error.

The estimated effect for each individual teacher *j* in the year *t* is captured by the coefficient of this teacher's dummy variable in the model. All the students' scores are standardized to mean zero and variance one. This procedure allows the analysis of the estimated standard deviation of teacher effectiveness measured in terms of standard deviations of student achievement.

Teacher and school effects are specified as fixed effects. Teacher effect is then measured relative to the average of all teachers at a given school. As McCaffrey et al. (2003) pointed out, fixed and random effects tend to yield similar conclusions about the variability of teachers, but will provide different estimates of

individual teacher effects. Since the fixed-effects method only uses a teacher's students to estimate his or her effect, the estimates can be highly sensitive to sampling error because teachers tend to teach only small numbers of students. Therefore, procedures to estimate and adjust for sampling errors must be provided.

As already discussed, it is well known that the residual variance in students' scores due to measurement errors and sources of variability in effectiveness other than teachers, schools and other components of the model cause sampling errors in estimates of teacher effects. Because of that, a portion of the standard deviation of estimates of teacher effects is due to sampling errors.

The procedure used by Aaronson et al. (2007), Goldhaber and Hansen (2010) and Koedel and Betts (2011) was also adopted in this analysis to adjust for sampling errors. It assumes that the teacher effect estimated in the equation (1) is the sum of a real teacher effect and an error term, which is not correlated to the real teacher effect:

$$\hat{\tau}_j = \tau_j + \lambda_j \quad (2)$$

Then, the variance in estimated teacher effects can be decomposed into one portion due to the real variance of teacher effects and another one due to the sampling errors:

$$Var(\hat{\tau}) = Var(\tau) + Var(\lambda) \quad (3)$$

Like these authors did, the mean of the square of the standard error estimates of $\hat{\tau}_j$ was used as an estimate of the sampling error variance and this was subtracted from the observed variance of $\hat{\tau}_j$ to give an adjusted variance and, then, the adjusted standard deviation.

In order to better understand the variability of teacher effects, the estimated teacher effects and the variability of these estimates were decomposed considering the stability of these measures, as proposed by McCaffrey et al. (2009).

This way, the estimated teacher effects in a given year can be decomposed as:

$$\hat{\tau}_j = \tau_j pers + \tau_j npers + \lambda_j \quad (4)$$

The difference between equations (2) and (4) is that, in the second, the real teacher effect is divided into a persistent component and a non-persistent component. The persistent component refers to the portion of the effects that is common to all years of estimates, which can be understood as an intrinsic part of teacher effectiveness. The non-persistent component refers to the portion of the effects that is specific to that year.

This transient component not only captures variation in teachers' performance in different stages of their teaching career, but also other class factors in that particular year. In this paper's sample every teacher is only observed with one class per year. So, it is not possible to adopt procedures to separate the teacher effects from the class effects. Thus, the group of students that compose the class is probably responsible for a large portion of this transient component of the estimated teacher effects. Therefore the portion of the estimated effects that are not due to sample errors will not be called "real teacher effects" because it probably includes influences of factors other than teacher performance in that particular year.

The variability of the estimated teacher effects may be decomposed as:

$$Var(\hat{\tau}) = Var(\tau pers) + Var(\tau npers) + Var(\lambda) \quad (5)$$

Since data from 2010 are used for the main analysis, and the variance of the persistent teacher effects is available only for a restricted group of teachers that taught the 4th grade in both 2009 and 2010, the variance of estimated teacher effects in 2010 was decomposed for this smaller group.

From the results of estimating equation (1), the variance of estimated teacher effects was calculated for this group of teachers. Since the correlation between teacher effects for two subsequent years is considered to be a measurement of the coefficient of stability of teacher effectiveness, and this coefficient

can be seen as the ratio between the variance of persistent teacher effects and the variance of estimated teacher effects, as McCaffrey et al. (2009) point out, it is possible to obtain an estimate of the variance of persistent teacher effects.

The same procedure used to calculate the variance of sampling errors for the whole group of teachers can be applied here for this smaller group of teachers. By doing that, it is possible to decompose the variance of estimated teacher effects as described in equation (5).

One of the main reasons for estimating teacher effects using value-added models is to try to find out if there are any teacher characteristics and practices that are related to these effects. It can be very useful to find associations between some variables and the estimated effects to guide teacher quality policies.

Given that, the idea was to regress the estimated teacher effects for each individual teacher in the first stage $\hat{\tau}_j$ in a group of variables represented by Z:

$$\hat{\tau}_j = \phi Z_j + u_j \quad (6)$$

However, as Goldhaber and Hansen (2010) showed, ordinary least-squares estimation provides standard errors that are too small because it ignores the error in the dependent variable, which is the teacher effect, originating in the previous estimation. Instead, a generalized least-squares approach was adopted. It accounts for the uncertainty in the dependent variable by weighting observations in proportion to the reliability of each individual estimated teacher effect (Aaronson et al., 2007; Koedel & Betts, 2007; Goldhaber & Hansen, 2010).

By doing that, the next step was an analysis of the relationship between the estimated teacher effects and a group of variables based on the available administrative data and on a questionnaire that teachers answered as their students were being tested by *Prova São Paulo*.

Previous papers with similar analyses, all based on American states, commonly include variables such as teacher degree level and subject, experience, certification status and score, among others. In this database, there is only information on teachers' degree level and teaching experience, which were included in the model, as well as variables to control teachers' gender and race.

In the questionnaire for teachers there are almost 200 items, with questions about their socioeconomic characteristics, their opinion about the school environment, their classroom practices and other subjects.

To compose the model to analyze teacher effects, the questions chosen were those directly related to teachers and their work. The model only included the questions that could be considered objective, while excluding ones about teachers' opinions and perceptions.

Data

The data used for this study are derived from the records of the Assessment System of Students from São Paulo Municipal Schools, which is called *Prova São Paulo*. Other information is derived from administrative records and from questionnaires given to students and teachers concurrently with the tests.

Prova São Paulo uses standardized tests to measure student achievement from 3rd to 8th grades on reading and math. Unlike all the assessments found in the literature on teacher effects, *Prova São Paulo* is administered on a census basis to students from even grades and on a sample base to students from odd grades. This means that there are scores for almost every student in 4th, 6th and 8th grades, but only from 25% to 40% of these students were tested in the previous year. Given that, students' scores from the 4th grade were chosen for analysis because this is the tested grade with the largest amount of students with both the current score and the previous score: around 40% of all 4th grade students.

Unfortunately this data limitation left each teacher with only a few students with score gains: close to half of all teachers in the sample only have 10 students. In an attempt to guarantee a minimum number of students per teacher, the analyzed teacher sample contained only teachers with at least six students

with both scores. In order to allow the distinction between school effects and teacher effects, the teacher sample contained only teachers from schools with at least two teachers that taught the 4th grade in 2010.

The main estimates are based on 2010 data from students and teachers, with students' previous scores from 2009 data. For the stability analysis, we added 2009 data from students and teachers, and students' previous scores from 2008 data.

Tables 1 and 2 contain descriptive statistics of 4th grade students and teachers from São Paulo municipal schools in 2010. The first two columns in these tables show the characteristics of the population, that is, all students and teachers from the 4th grade in 2010. The third and fourth columns of these tables show the characteristics of the students and teachers of the sample of this analysis, that is, the students from the 4th grade in 2010 that had been selected for testing the previous year while they were attending 3rd grade—and their respective teachers. Since students from PIC (*Programa Intensivo de Ciclo*) are tested on a census basis in 3rd grade and the other students are tested on a sample basis, the PIC students are overrepresented in the sample compared with the population. Beyond that fact, there were no other significant differences between students from the sample and the population, and also between teachers from the sample and the population. This is evidence that reinforces the idea that the sample does not represent a group of students and teachers that is distinct from the original population. Even though there seem to be no significant differences between the sample and the population, the inferences provided by this study are restricted to the sample analyzed.

Table 1
Descriptive statistics for 2010 4th grade students

	Population		Sample	
	% of students	% of answers	% of students	% of answers
Female	47%	100%	44%	100%
Age				
Under 10 years old	1%	100%	1%	100%
10 years old	68%	100%	57%	100%
11 years old	22%	100%	29%	100%
12 years old	7%	100%	10%	100%
Over 12 years old	2%	100%	3%	100%
Participation in PIC	7%	100%	14%	100%
Attended pre-school	57%	69%	56%	72%
Father's level of education				
Lower than primary	22%	71%	24%	72%
Primary	28%	71%	29%	72%
Lower secondary	21%	71%	21%	72%
Upper secondary	26%	71%	23%	72%
Higher education	3%	71%	3%	72%
Mother's level of education				
Lower than primary	16%	75%	18.40%	77%
Primary	29%	75%	30.20%	77%
Lower secondary	22%	75%	21.50%	77%
Upper secondary	30%	75%	27.70%	77%
Higher education	3%	75%	2.30%	77%
Family income				
Under R$ 850.00	43%	56%	45%	57%
R$ 851.00 to R$ 1275.00	31%	56%	30%	57%
R$ 1276.00 to R$ 2125.00	18%	56%	18%	57%
R$ 2126.00 to R$ 4250.00	7%	56%	6%	57%
Over R$ 4250.00	1%	56%	1%	57%
Number of students		70,367		25,777
Number of teachers		535		519

Note: Based on 2009 and 2010 data from *Prova São Paulo*.

Table 2
Descriptive statistics for 2010 4th grade teachers

	Population		Sample	
	% of teachers	% of answers	% of teachers	% of answers
Female	96%	100.00%	96%	100.00%
White or Asian[a]	70%	87%	70%	87%
Age				
Under 26 years old	2%	100%	2%	100%
26 to 35 years old	17%	100%	18%	100%
36 to 45 years old	39%	100%	39%	100%
46 to 55 years old	29%	100%	28%	100%
Over 55 years old	13%	100%	13%	100%
Higher degree obtained				
Higher Education (*Licenciatura*)	76%	88%	75%	88%
Master's (*Especialização*)	23%	88%	24%	88%
Academic Master's (*Mestrado*)	1%	88%	1%	88%
PhD	0%	88%	0%	88%
Teaching experience				
Up to 5 years	7%	89%	7%	89%
6 to 10 years	12%	89%	12%	89%
11 to 15 years	17%	89%	17%	89%
16 to 20 years	22%	89%	22%	89%
Over 20 years	42%	89%	42%	89%
Family income				
R$ 1000.00 to R$ 2000.00	2%	86%	2%	86%
R$ 2001.00 to R$ 3000.00	16%	86%	15%	86%
R$ 3001.00 to R$ 4000.00	28%	86%	29%	86%
R$ 4001.00 to R$ 5000.00	25%	86%	25%	86%
More than R$ 5000.00	29%	86%	29%	86%
Father's level of education				
Lower than primary	30%	87%	30%	87%
Primary	37%	87%	37%	87%
Lower secondary	13%	87%	13%	87%
Upper secondary	13%	87%	13%	87%
Tertiary	7%	87%	7%	87%
Number of teachers		2,209		2,055

Note: Based on 2009 and 2010 data from Prova São Paulo.

Results

Table 3 contains the results from the estimation of the covariate adjustment model expressed in the equation (1). As previously discussed, the dummy variables included in the equation permitted the estimation of each individual teacher effect, estimated by the coefficients of $\hat{\tau}_j$, and also analysis of how much these effects vary between the teachers in the sample. This also allowed the estimation of the impacts of changes in teacher effectiveness on students' scores by analyzing the estimated standard deviation of teacher effectiveness measured in terms of standard deviations of student achievement.

[a] In São Paulo, there's a considerable population of Japanese, Chinese and Korean descendants, which we will call Asian.

Table 3
Results from estimation of equation (1)

	Reading	Mathematics
Previous score	0.645***	0.640***
	[0.009]	[0.008]
Female	0.100***	-0.016
	[0.015]	[0.015]
Proper age for grade	0.082***	0.115***
	[0.027]	[0.026]
Participation in PIC	-0.044	0.228
	[0.288]	[0.256]
Attended pre-School	-0.006	-0.010
	[0.015]	[0.015]
Mother's level of education		
Primary	0.016	0.052**
	[0.026]	[0.025]
Lower secondary	0.044	0.039
	[0.028]	[0.027]
Upper secondary	0.071***	0.117***
	[0.028]	[0.027]
Higher education	0.158***	0.132***
	[0.053]	[0.051]
Father's level of education		
Primary	0.038*	0.037*
	[0.023]	[0.022]
Lower secondary	0.002	0.074***
	[0.025]	[0.024]
Upper secondary	0.049*	0.084***
	[0.026]	[0.025]
Higher education	0.126**	0.168***
	[0.050]	[0.048]
Family income		
R$ 851.00 to R$ 1275.00	0.050***	0.038**
	[0.018]	[0.018]
R$ 1276.00 to R$ 2125.00	0.085***	0.067***
	[0.022]	[0.022]
R$ 2126.00 to R$ 4250.00	0.146***	0.119***
	[0.034]	[0.033]
More than R$ 4250.00	0.067	0.058
	[0.067]	[0.064]
Standard deviation		
Of the estimates of teacher effects	0.622	0.621
Of the estimates of school effects	0.500	0.483
R^2	0.665	0.677
Number of students	9,425	9,567
Number of teachers	2,055	2,055

Note. *** denotes 1% significance level; ** denotes 5% significance level; * denotes 10% significance level.

As Table 3 shows, the calculated standard deviation of the estimated teacher effects was around 0.62 in both reading and math. But this measure is expected to be inflated since the error term is particularly problematic when the estimates of teacher effects are based on small samples like this one. In these cases, the sample variations may overestimate or underestimate teacher effects, because few observations with results close to the extreme results of the distribution may influence the estimated teacher effects (Aaronson et al., 2007).

After correcting for sampling errors, the adjusted standard deviation of estimated teacher effects was 0.53 in both reading and math.

This can still be considered a very large variation in teacher effects, since 10 studies with similar analyses revised by Hanushek and Rivkin (2010) obtained adjusted standard deviation of teacher effects between 0.08 and 0.26 in reading and between 0.11 and 0.36 in math.

A reliability coefficient of these estimates was then produced, being calculated as the ratio of the variance in the adjusted teacher effects to total variance in the estimates, following McCaffrey et al. (2009).

The calculated reliability coefficient was 0.73, a similar result to those obtained by McCaffrey et al. (2009), which were from 0.6 to 0.8. All these empirical results provide estimated rankings that are not precise enough to distinguish effectiveness between two teachers. According to Lockwood et al. (2002), in order to do that, the ratio of the variability of sampling errors to the variability in the estimates should be less than 0.1, which means a reliability coefficient of at least 0.9.

As McCaffrey et al. (2004) did, the estimated coefficients for each teacher effect were tested to verify if they could be considered different from zero, which means that these teachers' effectiveness could be considered to be significantly different from the mean.

While these authors' analyses demonstrated that between a third and a quarter of teachers in their sample could be considered significantly different from the mean, the results obtained from *Prova São Paulo* indicate that only 12% of teachers had estimated effects significantly distinct from the mean in reading and 15% of teachers had estimated effects distinct from the mean in math.

To analyze the stability of individual teacher effects, the estimated effects for the same teacher in two subsequent years were correlated (McCaffrey et al., 2009). This is the simplest way of calculating the stability coefficient and was the only available procedure given the limitation of our data. If there were 12 years of data on a sample of teachers as in Goldhaber and Hansen's (2010) analysis, it would have been possible to conduct a more sophisticated analysis.

Teacher effects were estimated only for teachers who taught in 4th grade both in 2010 and 2009 in order to have two measurements for each teacher. Of the 2055 teachers that taught in the 4th grade in 2010, only 714 had also taught in 4th grade in 2009. Other teachers were either in their first year of teaching or, more likely, were teaching in other grades.

Since the new sample was different from the original one, it was worth verifying whether the standard deviation of the estimated teacher effects for this group in 2010 was not substantially different from the initial estimation. The standard deviation of estimated teacher effects for this group was 0.64 in reading and 0.61 in math —close to the 0.62 obtained for the original group in both areas. The adjusted standard deviation of estimated effects was 0.54 in reading and 0.51 in math— also close to the 0.53 obtained for the original group in both areas. This supports the idea that the variability in teacher effects of this new sample may represent the variability of the original sample of teachers.

The correlation between the two years of estimated teacher effects was 0.097 in reading and 0.081 in math. These correlations demonstrate a much smaller stability of teacher effects than those obtained by McCaffrey et al. (2009) —from 0.22 to 0.67— and Goldhaber and Hansen (2010) —from 0.32 to 0.59— when these authors also worked with correlations between the estimates of pairs of subsequent years.

In an attempt to better understand the variability in the estimated teacher effects in this paper, a decomposition of the variance has been provided for the group of teachers that taught in 4th grade in both 2009 and 2010. This decomposition is shown in Table 4.

Table 4
Decomposition of the variance of estimated teacher effects

	Reading		Math	
	Variance	%	Variance	%
$Var(\hat{\tau})$	0.403	100%	0.368	100%
Decomposition:				
$Var(\tau pers)$	0.039	9.7%	0.030	8.1%
$Var(\tau mpers)$	0.250	61.9%	0.230	62.7%
$Var(\lambda)$	0.114	28.4%	0.108	29.2%

Note: Based on 2008, 2009 and 2010 data from Prova São Paulo.

Only 9.7% of the variance of estimated teacher effects in reading and 8.1% of the variance of estimated teacher effects in math are due to persistent effects. This means that it is only possible to assure that around 9% of the variance of estimated teacher effects is due to the portion of teacher effects that do not change over the years. These values can be considered lower limits for the variation of the estimated effects that are really due to individual teachers and can be seen as real teacher effects.

This empirical evidence means that moving the distribution of teacher effects one standard deviation up would raise the distribution of students' scores by at least 0.062 standard deviation in reading and 0.059 standard deviation in math.

These data do not allow the part of the variance of non-persistent teacher effects that is due to changes in teachers' performances over time to be distinguished from the part that is due to the quality of that specific class, and from the part that is due to the interaction —or the "chemistry"— between the teacher and the class. It is possible that all of the non-persistent effects are due to teacher effectiveness. Therefore, the higher limit for the variance of real teacher effects is 71.6% in reading and 70.8% in math. This means that an increase of one standard deviation in the distribution of teacher effects would raise the distribution of students' scores by at most 0.45 standard deviation in reading and 0.43 standard deviation in math.

Even though it is possible, it is unlikely that all of the variation in non-persistent teacher effects is only due to teacher effectiveness. So, the higher limit of the variation of real teacher effects is probably being overestimated.

In an attempt to simulate a more likely distribution of the variation of non-persistent teacher effect, it can be assumed that the non-persistent teacher effects are equally due to: (a) teachers, (b) classes, and (c) the interaction between teachers and classes. Then, each of these three factors would respond for a third of the variation due to non-persistent components. In this simulation, the portion of the variance of real teacher effects would be 30.3% in reading and 29% in math. In this hypothetical situation, an increase of one standard deviation in teacher effects would mean an increase of 0.19 standard deviation in students' scores in reading and 0.18 standard deviation in math. This seems to be a more realistic estimate of the differential effects teachers have on student achievement.

The next attempt was to estimate the relationship between teacher characteristics and practices and the estimated teacher effects from the first stage. The results are shown in Table 5.

Table 5
Results of the estimation of the relationship between teacher characteristics and practices and teacher effects

	Reading	Math
Female	0.196**	-0.064
	[0.088]	[0.083]
White or Asian	-0.075*	0.026
	[0.039]	[0.039]
Master's (*Especialização*)	-0.103***	-0.072*
	[0.040]	[0.041]
Academic Master's (*Mestrado*)	-0.254	-0.030
	[0.186]	[0.178]
From 1 to 5 years of teaching experience	0.219	0.067
	[0.220]	[0.224]
From 6 to 10 years of teaching experience	0.201	0.154
	[0.216]	[0.221]
From 11 to 15 years of teaching experience	0.153	0.097
	[0.216]	[0.220]
From 16 to 20 years of teaching experience	0.190	0.154
	[0.217]	[0.221]
More than 20 years of teaching experience	0.146	0.118
	[0.216]	[0.220]
From 1 to 5 years teaching at this school	-0.019	0.040
	[0.051]	[0.051]
From 6 to 10 years teaching at this school	0.039	0.053
	[0.061]	[0.060]
From 11 to 15 years teaching at this school	-0.000	0.002
	[0.066]	[0.065]
From 16 to 20 years teaching at this school	0.074	-0.005
	[0.088]	[0.086]
More than 20 years teaching at this school	-0.026	-0.026
	[0.106]	[0.100]
Teaches at more than one school	-0.031	-0.008
	[0.037]	[0.037]
Has a professional activity other than teaching in basic education	0.045	-0.109
	[0.110]	[0.105]
Dedicates from 4 to 8 hours per week to pedagogical work outside the school	0.055	0.025
	[0.047]	[0.046]
Dedicates more than 8 hours per week to pedagogical work outside the school	0.071	-0.025
	[0.058]	[0.057]
Uses computer and internet in classes	-0.013	0.020
	[0.038]	[0.038]
Invites parents to talk about their children	0.101	0.145
	[0.144]	[0.144]

High frequency of homework assignment	0.039	0.106**
	[0.046]	[0.045]
Uses Support Workbooks once or twice a week	0.219*	-0.027
	[0.125]	[0.122]
Uses Support Workbooks three or four times a week	0.176	-0.016
	[0.126]	[0.122]
Uses Support Workbooks every day	0.284**	0.016
	[0.140]	[0.137]
Covered from 50 to 70% of the curriculum	0.119	0.245
	[0.247]	[0.215]
Covered more than 70% of the curriculum	0.120	0.255
	[0.246]	[0.212]
Spends from 20 to 30 minutes on students' organization	0.060	-0.109
	[0.081]	[0.081]
Spends from 10 to 20 minutes on students' organization	0.049	-0.086
	[0.070]	[0.072]
Spends less than 10 minutes on students' organization	0.109	-0.040
	[0.075]	[0.077]
Constant	-0.784**	-0.436
	[0.375]	[0.368]
R^2	0.031	0.024
Number of Teachers	1,161	1,155

Note. *** denotes 1% significance level; ** denotes 5% significance level; * denotes 10% significance level.

The results of these estimations must be analyzed with some caution. Firstly, the number of teachers in the sample has been reduced to almost half between the first and the second stages due to the lack of teachers' answers to the questionnaire. Secondly, the analysis is based on cross-sectional data in which teacher effects from a year are regressed for teacher characteristics and practices captured in the same year. This means that the results provide evidence of the association between the dependent and independent variables, but they do not allow the establishment of causal relationships between them.

Moreover, the highest value of the R^2 of the estimations was 0.11, which demonstrates that the largest portion of the variation in teacher effects is not explained by the available teacher characteristics and practices. This evidence is similar to that found by Aaronson et al. (2007) and Koedel and Betts (2007). However, as these authors explain, the R^2 is an underestimation of the explanatory power of this relationship, since a considerable portion —maybe a third— of the variation in $\hat{\tau}_j$ is due to sampling errors.

As Aaronson et al. (2007) did, by multiplying the total sum of the squares by a 50% rate to account for the sampling errors —which is considered conservative by the authors— the R^2 value would be double. Even so, it would be at most 0.22, which can still be considered a low value for the explanatory power of a regression.

The results did not indicate any relationship between teacher experience and teacher effectiveness, one of the traditional variables used in studies on this theme. On the other hand, a negative and statistically significant association was found between having a master's (*especialização*) degree and teacher effects in both reading and math.

A positive relationship or no relationship was expected between these variables, but not a negative and statistically significant one. This result shows that teachers with a master's (*especialização*) degree produce, on average, lower effects on student achievement. Teachers with academic master's degrees, on the other hand, have similar teacher effects as teachers with only higher education degrees (*licenciatura*).

Since teachers can participate in many kinds of master's (*especialização*) degree courses with different themes and approaches, it is hard to try to find explanations for these results that are valid for all courses. Since it is the strongest association found in the analysis, it will be worth trying to gather information that may help explain these results. This effort should include collecting information on common characteristics of teachers that participated in master's (*especialização*) courses. For example, it could be started by analyzing the criteria to select teachers for the courses that were offered or provided by the government. These criteria may be responsible for the selection of a group of less effective teachers to participate in the courses, and these courses may not be able to improve the effectiveness of teachers.

Teachers who assign homework activities more often than others have higher teacher effects in math. Also, teachers who use Support Workbooks more frequently have higher teacher effects in reading. The Support Workbooks are special material to guide teacher practices on topics in which students showed greater needs in previous tests. Given these results, it would also be interesting to produce further and deeper studies into the use of Support Workbooks in reading classes in order to understand what makes teachers who use them more frequently more effective. There is a range of possible explanations, such as, for example, the quality of this pedagogical resource compared with other resources that are used by teachers, teacher participation in training courses for using these Support Workbooks, or the fact that this resource has a structuring nature which helps teachers organize their classes, providing better use of time to teach students to read.

Discussion

This paper aimed to use a value-added model to estimate and analyze the effectiveness of a sample of 4th grade teachers from São Paulo's municipal schools.

From these estimations, some evidence was obtained to contribute to the understanding of teacher effects on students' achievement gains in *Prova São Paulo* and the challenges of using those measurements.

By estimating individual teacher effects, the calculated standard deviation of the estimated teacher effects was around 0.62 in both reading and math. And even after correcting for sampling errors, the adjusted standard deviation of estimated teacher effects was 0.53 in both reading and math.

This can still be considered a very large variation in teacher effects compared with similar analyses revised in this paper. It represents impacts which are too big for a single school factor, even if we are talking about teachers. Given that, there was an effort to decompose the variance of estimated effects, in order to try to separate the portion that is, in fact, due to individual teachers and not due to other factors that can influence the estimates.

By decomposing the variance of teacher effects, around 29% of the variance of estimated teacher effects was observed to be due to sampling errors. And only around 9% of the variance of estimated teacher effects is due to the portion of persistent teacher effects, and is really due to individual teachers. On the other hand, around 62% of the variance of estimated teacher effects is due to the portion of non-persistent teacher effects. As these data did not allow the part of the variance of non-persistent teacher effects that are due to teachers to be distinguished from the part that is due to the class and to the interaction between teachers and classes, the variance of real teacher effects is between 9% and 62% of the variance of estimated teacher effect. This means that an increase of one standard deviation in teacher effects would mean an increase of between 0.062 and 0.45 standard deviation in students' scores in reading and between 0.059 and 0.43 standard deviation in math.

Based on the papers of Lockwood et al. (2002), McCaffrey et al. (2009) and Goldhaber and Hansen (2010), it is possible to say that these data have limited capacity to provide recommendations for personnel policies. According to Lockwood et al. (2002), the ratio between the variance of sampling errors and the variance of estimated teacher effects should be less than 0.1, so the estimated ranking could be precise enough to permit differentiation in teacher effectiveness. And this does not happen in this study, since this ratio was 0.27.

Despite these limitations, the analysis showed that 12% of the teachers in the sample had effects that could be considered to be distinct from the mean in reading and 14% of the teachers in the sample had effects that could be considered to be distinct from the mean in math.

When analyzing what factors can be associated with teacher effects, having a master's (*especialização*) degree was found to be negative and was significantly associated with teacher effects in both reading and math. On the other hand, the frequency of homework assignments and the frequency of use of Support Workbooks showed positive and significant relationships with teacher effects.

Given these results, a deeper investigation based on teachers who produced effects that were distinct from the mean can be recommended. A starting point should be their teaching practices, with special attention to homework assignments and the use of pedagogical resources.

This is an interesting example of using quantitative research to raise certain hypotheses that can generate ideas for qualitative research, leading to a better understanding of teacher effectiveness. Together, both kinds of research have great potential to inform policies to improve teacher effectiveness.

Due to the potential of teacher value-added models to contribute to the development of teacher policies, it would be recommendable to make additional efforts to obtain larger samples of students, teachers and schools in order to improve the estimation of teacher effects and their use to inform teacher policies.

The original article was received on May 2nd, 2013
The revised article was received on August 18th, 2013
The article was accepted on November 22nd, 2013

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teacher and student achievement in Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135. Retrieved from http://www.press.uchicago.edu/ucp/journals/journal/jole.html

Albernaz, A., Ferreira, F., & Franco, C. (2002). Qualidade e equidade no ensino fundamental brasileiro. *Pesquisa e Planejamento Econômico*, *32*(3), 453-476. Retrieved from http://www.memoria.nemesis.org.br/index.php/ppe

Ballou, D., & Podgursky, M. (2000). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record, 102*(1), 5-27. Retrieved from http://www.tcrecord.org/

Barros, R. P., Mendonça, R., Santos, D. D., & Quintaes, G. (2001). Determinantes do desempenho educacional no Brasil. *Pesquisa e Planejamento Econômico*, *31*(1), 1-42. Retrieved from http://www.memoria.nemesis.org.br/index.php/ppe

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models. A Policy Information Center's Report.* Princeton, NJ: Educational Testing Service (ETS).

Clotfelter, C., Ladd, H., & Vigdor, J. (2007). Teacher credentials and student achievement: Longitudinal analysis with student fixed effects. *Economics of Education Review*, *26*(6), 673-682. doi:10.1016/j.econedurev.2007.10.002

Croninger, R., Rice, J. K., Rathbun, A., & Nishio, M. (2007). Teacher qualifications and early learning: effects of certification, degree, and experience on first-grade student achievement. *Economics of Education Review*, *26*(3), 312-324. doi:10.1016/j.econedurev.2005.05.008

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives*, *13*(42), 1-51. Retrieved from http://epaa.asu.edu/ojs/

Goldhaber, D. D., & Hansen, M. (2010). *Is it just a bad class? Assessing the stability of measured teacher performance*. (CEDR Working Paper 2010-3). Seattle, WA: University of Washington.

Hanushek, E. (1986). The economics of schooling: Production and efficiency in public schools. *Journal of Economic Literature*, *24*(3), 1141-1177. Retrieved from http://www.aeaweb.org/jel/

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). *The market for teacher quality*. (Working Paper 11154). Massachusetts: NBER.

Hanushek, E. A., & Rivkin, S. G. (2004). How to improve the supply of high-quality teachers. *Brookings Papers on Education Policy*, *7*, 7-25. doi: 10.1353/pep.2004.0001

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review*, *100*(2), 267-271. doi:10.1257/aer.100.2.267

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). *What does certification tell us about teacher effectiveness? Evidence from New York City*. (Working Paper 12155). Massachusetts: NBER.

Koedel, C., & Betts, J. (2007). *Re-examining the role of teacher quality in the educational production function.* (Working Paper). San Diego, CA: University of Missouri.

Koedel, C., & Betts, J. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*(1), 18-42. doi: 10.1162/EDFP_a_00027

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics*, *27*(3), 255-270. doi: 10.3102/10769986027003255

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., Louis, T. A., & Hamilton, L. S. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*(1), 67-101. doi: 10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The inter-temporal variability of teacher effects estimates. *Education Finance and Policy*, *4*(4), 572-606. doi: 10.1162/edfp.2009.4.4.572

Ravitch, D. (2010). *The death and life of the great American school system: How testing and choice are undermining education*. New York: Basic Books.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458. Retrieved from http://www.econometricsociety.org/aims.asp

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review*, *94*(2), 247-252. Retrieved from http://www.aeaweb.org/aer/index.php

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2008). *Can you recognize an effective teacher when you recruit one?* (Working Paper 14485). Massachusetts: NBER.

Rothstein, J. (2009). Student sorting bias in value added estimation: Selection on observables and unobservables. *Education Finance and Policy*, *4*(4), 537-571. doi:10.1162/edfp.2009.4.4.537

Soares, J. F. (2005). Qualidade e equidade na educação básica brasileira: Fatos e possibilidades. In C. Brock, & S. Schwartzman (Eds.), *Os desafios da educação no Brasil* (pp. 91-117). Rio de Janeiro, Brasil: Nova Fronteira.

Soares, T. M. (2003). Influência do professor e do ambiente em sala de aula sobre a proficiência alcançada pelos alunos avaliados pelo Simave 2002. *Estudos em Avaliação Educacional*, *28*, 103-123. Retrieved from http://www.fcc.org.br/institucional/2012/08/31/estudos-em-avaliacao-educacional-4/

United States of America, Department of Education (2009). *Race to the Top Program: Executive Summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, *11*, 57-67. Retrieved from http://old.library.georgetown.edu/newjour/j/msg03680.html