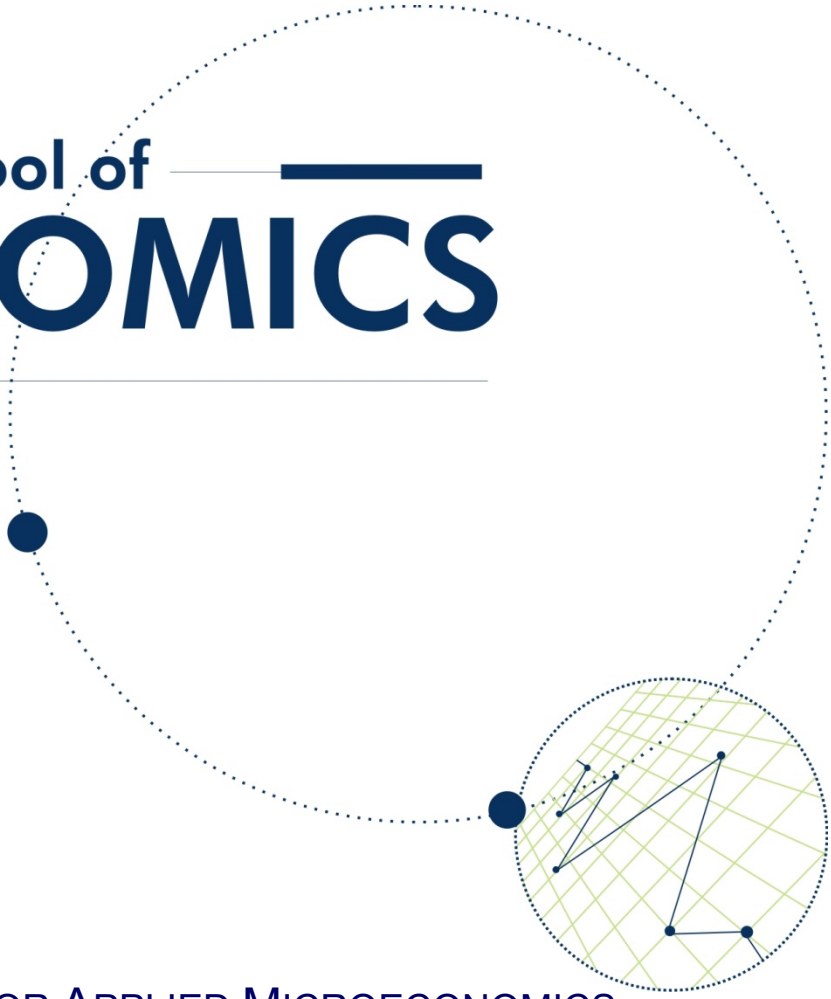
São Paulo School of
ECONOMICS

Working Paper
XXX

November 2011



CENTER FOR APPLIED MICROECONOMICS

17/2011

C-MICRO Series

Measuring Peer Effects in the Brazilian School System

SERGIO FIRPO
HUGO BORGES JALES

Measuring Peer Effects in the Brazilian School System*

Sergio Firpo† Hugo Borges Jales‡

November 2011

Abstract

This paper investigates the existence and magnitude of peer effects among pupils in Brazilian schools using a dataset on achievement of fifth-graders (around eleven years old) in Math, which is accompanied by detailed questionnaires completed by students, teachers and principals. We address the self-selection of group formation through a selection on observables hypothesis and use teacher's wage as a proxy for unobserved teacher's ability. A robustness check using as instrumental variable a question answered by the principal on how classrooms are formed at school was performed and similar results were found. Our results suggest that both contextual (exogenous) and endogenous peer effects are important determinants of students' achievement in Brazil. Furthermore, by being able to separate out contextual effects from endogenous peer effects, we show that even though boys perform slightly better than girls in Math, children in classes with larger proportions of boys tend to have worse results in Math.

JEL: I21 J24. KEYWORDS: Education, Brazilian School System, Peer Effects, Social Multiplier, Social Interactions.

*We have benefited from helpful comments from Cristine Pinto, Andre Portela Souza and Vladimir Ponczek and participants at seminar at EESP-FGV. Financial support from CNPq-Brazil is acknowledged. All errors remain our own.

†From São Paulo School of Economics, Getulio Vargas Foundation (Brazil), and IZA (Germany). Mailing address: R. Itapeva 474/1215, São Paulo - SP, Brazil, 01332-000. E-mail: sergio.firpo@fgv.br. Webpage: http://sites.google.com/site/sergiopfirpo

‡From The University of British Columbia, Department of Economics.

1 Introduction

Although most researchers in education would agree that peers' attributes and attitudes affect individual educational outputs as, for example, results in test scores, finding evidence of the magnitude of this effect is a complicated task. Social interactions, or peer effects, are not the only channel explaining the excess of variation in student outcomes between classrooms that are not explained by differences in the composition of the student body between classes. Indeed, the difficulty in identifying peer effects from other competing stories is not restricted to educational settings and the more general abstract problem has been raised by Manski (1993), who showed that in a very simple linear framework one would need to rely on strong assumptions to identify peer effects. Nonetheless, in the past twenty years, the literature in education faced a growing interest in more deeply understanding, both theoretically and empirically, within-class and within-school peer effects as reflected, for example, by the works of Arnott and Rowse (1987), Hoxby (2000), Hoxby and Weingarth (2006), Lavy et al. (2008), Sacerdote (2001), Ammermueller and Pischke (2006) among several others.

Despite the fact that there is a huge empirical literature on educational peer effects around the world, only recently those effects were estimated for the Brazilian school system. One fact that might explain that delay is that for Brazil availability of standardized exams is quite recent. We can find some works that marginally looked at that issue, example being the papers by Fletcher (1997), Albernaz et al. (2002) and Franco et al. (2007). However, estimation of peer effects clearly deserves more careful attention and except by the work of Pinto (2008) the problems of identification and estimation of peer effects on students' achievement in Brazil have not been addressed at all.

The first goal of this paper is to fill this gap. Using the identification strategy proposed by Graham and Hahn (2005), we estimated the magnitude of classmate effects in Math scores using Brazilian data from 2005. Our results show that both peer characteristics (exogenous peer effects) – like race, socioeconomic status and gender – and peer actions (endogenous peer effects) are important determinants of pupils' outcomes in the middle of elementary school (5th grade students, aged around 11 years old).

We also provide a detailed discussion of what parameters can be identified in the simplest case and what extra assumptions are necessary for identification. We invoke Manski's (1995) result that shows that excluding the effect of one classroom average covariate allows identification of structural parameters. However, we also define some parameters of interest that can be written as functions of structural parameters. We show that in some particular cases, identification is warranted even without that exclusion restriction.

Identification of structural peer effects parameters depend not only on the ability of expressing these parameters as an one-to-one mapping from reduced form ones: It crucially depends on the ability of expressing the reduced form parameters themselves as an one-to-one mapping from data. We would not be able to identify reduced-form parameters when there is a correlation between class-level observed characteristics such as class socioeconomic index and unobserved classroom components such as teacher ability. For the specific context of our study, we establish plausible conditions under which reduced form parameters can be expressed as functionals of data distribution. In particular, to control for potentially confounding unobserved characteristics at class level that might be correlated with peer composition, we use a rich data on school and teachers characteristics. Quite importantly, we use teachers' wage as a proxy for unobservable teacher's ability.

We show relatively large peer effects on the achievement of pupils in grade 5 at elementary schools in Brazil. Both exogenous (through peer characteristics) and endogenous (through peer actions) peer effects seem to be relevant in explaining individual performance. Peer characteristics, such as the socioeconomic index, race, gender and age present significant contextual effects. Our estimates of endogenous peer effects were about 0.4, which can be interpreted as evidence of a so-called "conformist" individual behavior, under which students face large costs to exert effort levels that are distant from what is believed to be the norm in the classroom.[1] Finally, those estimates of endogenous peer effects imply a social multiplier of about 1.7.[2] Thus, exogenous changes in student i's effort will propagate throughout

[1] See Akerlof (1997).

[2] See Scheinkman (2006), Glaeser and Scheinkman (2001) and Glaeser, Sacerdote and Scheinkman (2003) for a theoretical

the classroom having a large impact on the overall class effort.

Another important result that is in line with recent results on gender peer effects is the effect of increases in intra-class female composition. We obtain that an increase in the proportion of girls in the classroom have a positive effect on individual Math test scores, even though girls perform on average worse than boys in Math. Although our data set does not allow us to exploit the mechanisms behind these results, Lavy and Schlosser (2011) have recently shown from data on Israel that such findings could be explained by the fact that schools with higher proportions of female peers tend to benefit from higher levels of non-cognitive skills that are important to lower classroom disruption and violence.

The remainder of the paper is divided as follows. In section 2 we first present and discuss the assumptions and the methodology that allows us to identify and separate out endogenous from exogenous peer effects. In section 3 we present the data used in our study and present our main results. In section 4 we report some robustness checks while in section 5 we conclude.

2 Identification

2.1 A Simple Economic Model

Like most of the literature, we use the so-called "linear-in-means model" to estimate the size of peer effects on pupil's achievement. There are two main simplifications in that model: (*i*) the effort of a given student is affected by his/her peers only through the expectation of their attitudes and attributes; (*ii*) the relationship between individual effort and all covariates is linear.

Graham and Hahn (2005) follow the work of Akerlof (1997), who derived a simple static model of pupil's choices in the presence of social interactions that yields a linear-in-means model. In what follows we summarize the model's main features and predictions.

Let Y_{ig} be the effort of a student i at class g and assume that he/she maximizes his/her indirect utility function ν by optimizing the level of effort he/she will exert in a given academic year. The function ν is assumed to be quadratic on effort and to have the following parametric specification:

$$\nu(Y_{ig}|k_{ig}, m^e_{ig}) = -\frac{1-\beta_3}{2}Y^2_{ig} + k_{ig}Y_{ig} - \frac{\beta_3}{2}(Y_{ig} - m^e_{ig})^2 \tag{1}$$

where $k_{ig} = k(X_{ig}, Z_g, \epsilon_{ig}, \alpha_g)$. Note that student i takes into account his/her own characteristics X_{ig} and ϵ_{ig}, and group level characteristics Z_g and α_g in his/her optimization problem. Although ϵ_{ig} and α_g are visible for student i, we assume that all the econometrician observes is a random sample $\{Y_{ig}, X_{ig}\}$ of N individuals divided into $g = 1, 2, 3, .., G$ groups such that at each group the number of observations is N_g and $N = \sum_{g=1}^{G} N_g$. Finally, β_3 is unknown parameter; $k(\cdot)$ is unknown function; Z_g is some group level observable characteristics that might include, for example, the group level mean of X and teacher's observable characteristics such as teaching experience and salary; and m^e_{ig} is student's i expected effort level of group g. The dimension of X_{ig} is J and the dimension of Z_g is $M \geq J$.[3]

The first two terms of the previous sum are what Akerlof called intrinsic utility while the last term corresponds to the endogenous component of social interactions. Interestingly, student i will react to the expected effort of the group only if $\beta_3 \neq 0$. A test for $\beta_3 = 0$ is a test for the presence of endogenous peer effects. The quadratic form of the last term of the sum implies that students have disutility of being far away from what they believe is the norm inside the group, which generates the 'conformist' behavior discussed by Akerlof (1997). Thus, large values of β_3 indicate that students face large costs in exerting effort levels that are distant from what is believed to be the norm.

Considering that students choose optimal effort taking as given the behavior of the group, we can write the solution for the optimal effort level as

$$Y^*_{ig} = k_{ig} + \beta_3 m^e_{ig}, \tag{2}$$

discussion on social multipliers and Sacerdote (2001) for an application of the concept to students peer effects.

[3] Because the limitations on the data, we will assume that the student is influenced only by classmates, and equally by everyone in the classroom. Calvó-Armengol, Patacchini and Zenou (2009), using a dataset that includes information about the friendship network of the students, estimate the size of peer effects on the primary education on US using a model that flexibilizes both these hypotheses.

where Y^*_{ig} is a map from student's and group's characteristics (k_{ig}) and the expected behavior of the group (m^e_{ig}) into the optimal choice of the student.

As pointed by Brock and Durlauf (2001) and Blume and Durlauf (2005), this model only makes restrictions on the behavior of the students after we impose some conditions on how students form their beliefs about the behavior of the group. We impose a self-consistency condition, which restricts the expected behavior of the group m^e_{ig}, as seen by any individual i inside the group, to be equal to the within-group expectation of the optimal effort:[4]

$$m^e_{ig} = \mathbb{E}_g[Y^*], \forall i \in g, \tag{3}$$

where the notation we use is such that $\mathbb{E}_g[Y^*] = \mathbb{E}[Y^*|g]$.

2.2 Identification Cases

Taking the expectation inside the group for Equation (2) and substituting into (3), we have that for all values of $\beta_3 \neq 1$:

$$\mathbb{E}_g[Y^*] = (1+\lambda)\,\mathbb{E}_g[k]$$

where

$$\lambda = \beta_3 / (1-\beta_3)^{-1}.$$

In principle, the shape of function k can be very flexible. To simplify our analysis, we assume that k is an additively separable function of its arguments:

$$k(X_{ig}, Z_g, \epsilon_{ig}, \alpha_g) = \beta_0 + X^\top_{ig}\beta_1 + Z^\top_g\beta_2 + \alpha_g + \epsilon_{ig}, \tag{4}$$

thus group level unobserved heterogeneity α_g and observed Z_{mg} $(m = 1, ..., M)$ characteristics enter linearly in function k. We assume that Z_g can be broken down into two vector components: Z_{Ag} and Z_{Bg}. Moreover, we assume that $Z_{Ag} = \mathbb{E}_g[X]$. We split β_2 into two parts β_{2A} and β_{2B}, and call β_{2A} the J vector of contextual effects parameters, which capture the effect of the makeup of the class on individual effort, as pointed out by Hoxby (2002).

Using the specified functional form for k, the resulting equilibrium equation for effort takes the known 'linear in means' form:

$$\begin{aligned} Y^*_{ig} &= \beta_0 + X^\top_{ig}\beta_1 + Z^\top_g\beta_2 + \beta_3\mathbb{E}_g[Y^*] + \alpha_g + \epsilon_{ig} \\ &= \beta_0 + X^\top_{ig}\beta_1 + \mathbb{E}_g[X^\top]\beta_{2A} + Z^\top_{Bg}\beta_{2B} + \beta_3\mathbb{E}_g[Y^*] + \alpha_g + \epsilon_{ig}. \end{aligned} \tag{5}$$

But because $\mathbb{E}_g[Y^*] = (1+\lambda)\,\mathbb{E}_g[k]$, we have

$$\mathbb{E}_g[Y^*] = (1+\lambda)\left(\beta_0 + \mathbb{E}_g[X^\top](\beta_1 + \beta_{2A}) + Z^\top_{Bg}\beta_{2B} + \alpha_g + \mathbb{E}_g[\epsilon]\right). \tag{6}$$

We proceed following Graham and Hahn (2005) and rewrite the original individual and group-level equations using within and between regression equations. The latter is straightforward and we write:

$$\mathbb{E}_g[Y^*] = \pi_0 + \mathbb{E}_g[X^\top]\pi_2 + Z^\top_{Bg}\pi_3 + \eta_g \tag{7}$$

where

$$\begin{aligned} \pi_0 &= (1+\lambda)\,\beta_0 \\ \pi_2 &= (1+\lambda)\,(\beta_1 + \beta_{2A}) \\ \pi_3 &= (1+\lambda)\,\beta_{2B} \\ \eta_g &= (1+\lambda)\,(\alpha_g + \mathbb{E}_g[\epsilon]). \end{aligned}$$

[4] Another way to interpret that assumption has been pointed by some authors, like Manski (1995), Angrist, Grady and Imbens (2000) and Graham (2005), for whom it allows circumventing a missing data problem on beliefs that would be necessarty to identify parameters of this model.

Now, if we subtract group-level means from individual-level equations, we obtain the following individual class-demeaned variables

$$\ddot{Y}^*_{ig} = Y^*_{ig} - \mathbb{E}_g[Y^*], \quad \ddot{X}_{ig} = X_{ig} - \mathbb{E}_g[X], \text{ and } \quad \ddot{\epsilon}_{ig} = \epsilon_{ig} - \mathbb{E}_g[X].$$

Then, it is clear that the structural equation can be rewritten:

$$\begin{aligned}
\ddot{Y}^*_{ig} &= Y^*_{ig} - \mathbb{E}_g[Y^*] \\
&= \beta_0 + X_{ig}^\top \beta_1 + \mathbb{E}_g[X^\top]\beta_{2A} + Z_{Bg}^\top \beta_{2B} + \beta_3 \mathbb{E}_g[Y^*] + \alpha_g + \epsilon_{ig} \\
&\quad - \left(\beta_0 + \mathbb{E}_g[X^\top]\beta_1 + \mathbb{E}_g[X^\top]\beta_{2A} + Z_{Bg}^\top \beta_{2B} + \beta_3 \mathbb{E}_g[Y^*] + \alpha_g + \mathbb{E}_g[\epsilon]\right) \\
&= \ddot{X}_{ig}^\top \pi_1 + \ddot{\epsilon}_{ig}, \qquad (8)
\end{aligned}$$

where

$$\pi_1 = \beta_1$$

Somewhat interestingly, we can break down the estimation problem into two separate parts. In a first stage, we estimate β_1 (π_1) using within-regression (Equation 8). Then we estimate π_0, π_2 and π_3 using between-regression (Equation 7).

In what follows, we consider several different restrictions that allow us to identify structural parameters β. In all of them we assume that the reduced-form parameters π are identified, which is not an innocuous assumption. We discuss sufficient conditions for identification of π in the next subsection.

Case 1: $J \geq 1$, $M \geq J$. In this case, without imposing any other restriction, the problem of finding the structural parameters β has no unique solution as the number of equations is smaller than the number of structural parameters. In fact, there are $M + J + 1$ equations but $M + J + 2$ structural parameters. If $M = J$, that is, $\beta_{2B} = 0$, then there are $2J + 1$ equations but $2J + 2$ structural parameters.

In fact, because of non-identifiability of structural parameters, one usually imposes that some covariates have no contextual effect. Let us write $X_{ig}^\top = \left[W_{ig}^\top, R_{ig}^\top\right]^\top$, where R_{ig} is an L-vector random variable ($L \leq J$) of student's characteristics with no direct contextual effect, that is, if we write $\beta_{2A}^\top = \left[\beta_{2AW}^\top, \beta_{2AR}^\top\right]^\top$, then we have that $\beta_{2AR} = 0$. In what follows it is useful to write, for $j = 1, 2$, $\pi_j = \left[\pi_{jW}^\top, \pi_{jR}^\top\right]^\top$.

Case 2: $J \geq L = 1$, $M \geq J$. Under the exclusion restriction of no contextual effects for one covariate, that is, $\beta_{2AR} = 0$, we have that the solution for β is $\beta_0 = \pi_{2R}^{-1}\pi_0\pi_{1R}$, $\beta_1 = \pi_1$, $\beta_{2AR} = 0$, $\beta_{2AW} = \pi_{2R}^{-1}\left(\pi_{1R}\pi_{2W}^\top - \pi_{2R}\pi_{1W}^\top\right)$, and $\beta_{2B} = \pi_{2R}^{-1}\pi_{1R}\pi_3$, $\beta_3 = \pi_{2R}^{-1}\left(\pi_{2R} - \pi_{1R}\right)$. Note here that we have $M + J + 1$ equations and $M + J + 1$ parameters as we have set β_{2AR} to zero.

However, this nice exact result holds only when there is one exclusion restriction, allowing the model parameters to be exactly identified. It may be the case that there are several covariates with no contextual effect. In such case, we have more equations than parameters to be estimated

Case 3: $J \geq L > 1$, $M \geq J$. Under the exclusion restriction of no contextual effects for a vector of covariates, that is, if the L vector β_{2AR} is the null vector, we have an overidentified case or a lack of exact solutions for this problem. In that case, we have that λ has to satisfy the following condition: $\pi_{1R}(1 + \lambda) = \pi_{2R}$. One could exploit the identification conditions and propose an estimator for λ that uses all L entries and that minimizes $\|\widehat{\pi}_{1R}(1 + \lambda) - \widehat{\pi}_{2R}\|$ where $\|a\|$ is the norm of vector a. Once one has an estimator for λ, say $\widehat{\lambda}$, then all structural parameters can be consistently estimated by simply substituting $\widehat{\lambda}$ by λ in the following expressions: $\beta_0 = (1 + \lambda)^{-1}\pi_0$, $\beta_1 = \pi_1$, $\beta_{2AW} = (1 + \lambda)^{-1}\left(\pi_{2W}^\top - (1 + \lambda)\pi_{1W}^\top\right)$, $\beta_{2AR} = 0$, $\beta_{2B} = (1 + \lambda)^{-1}\pi_3$, $\beta_3 = (1 + \lambda)^{-1}\lambda$.

In many situations we are not only interested in identifying the structural parameters but rather meaningful combinations of these parameters. Consider, for example, the parameter $\delta_X = (1+\lambda)(\beta_1 + \beta_{2A})$. We can see that $\delta_X = \pi_2$. It is interesting to notice that this parameter has a meaningful interpretation as it is in fact the marginal effect of an increase in $\mathbb{E}_g[X]$ on expected effort in a given class g, that is,

$$\delta_X = \partial \frac{\mathbb{E}_g[Y]}{\partial \mathbb{E}_g[X]}.$$

We can define other interesting parameters. Consider, for example, $\delta_{XO} = (\beta_1 + \beta_{2A})\lambda + \beta_{2A}$, which is simply $\pi_2 - \pi_1$ and is hence identified. That parameter corresponds indeed to the effect in student i's effort of a change in $\mathbb{E}_g[X]$ netting out his own contribution to the class average of X. Below we state formally these results.

Case 4: $J \geq 1$, $M \geq J$. This is in fact case 1, but noting that in this setting, although we cannot point identify all structural parameters without imposing further restrictions, we can identify $\delta_X = (1+\lambda)(\beta_1 + \beta_2)$ by $\delta_X = \pi_2$ and $\delta_{XO} = (\beta_1 + \beta_2)\lambda + \beta_2$ by $\delta_{XO} = \pi_2 - \pi_1$.

Obviously, one need conditions for OLS estimators of β_1 to be consistent. These conditions are stated in a general level and then applied to our context in the next subsection.

2.3 Context-Specific Requirements for Identification

We would be in principle interested in identifying peer effects on student effort, which is in general a non observable variable. Therefore, we consider tests outcomes as our dependent variable instead.

We will assume that there is a scalar variable R, such that $\beta_{2AR} = 0$. In summary, we are in case 2 and as we are interested in identifying peer effects on students's tests outcomes, we will assume that a dummy variable indicating preschool attendance shall not present any contextual effect for fifth-graders (11 years old). That is, the proportion of students that have attended preschool in a class does not cause any direct effect on peers achievement.

Several authors suggest to use lagged variables to break the simultaneity present on the model. Lavy et. al. (2008), in a similar context, used lagged achievement to measure the effects of peers ability. The hypothesis that we make here is, in some sense, quite similar. We impose that the students that attended preschool get a shift in their own ability but, more importantly, the proportion of students that had that shift does not directly affect the achievement of the others.

Therefore writing preschool separately from the rest of the covariates vector and splitting the group level characteristics into observable and unobservable parts, we have that:

$$Y_{ig}^* = \beta_0 + W_{ig}^\top \beta_{1W} + R_{ig}\beta_{1R} + \mathbb{E}_g[W^\top]\beta_{2AW} + \beta_3 \mathbb{E}_g[Y^*] + \mathbb{E}_g[Z_B^\top]\beta_{2B} + \alpha_g + \epsilon_{ig}.$$

In our particular case, R_{ig} is an indicator that is equal to 1 when the student i at class g took preschool.

Two other hypothesis refer to the joint distribution of covariates and unobserved error terms. They are needed to consistently estimate the parameters of the model. Let $\Omega_{ig} = \left[X_{ig}^\top, Z_{ig}^\top\right]^\top$ and $\Omega_g = \left[\Omega_{1g}^\top, \Omega_{2g}^\top, ..., \Omega_{N_g g}^\top\right]$. In what follows, we assume that ϵ_{ig} is exogenous and that unobservable group-level terms are mean-independent of observable group-level components, that is, the impact of these unobservable on Y are simply random effects. More formally, we have:

$$E[\epsilon_{ig}|\Omega_g, \alpha_g] = 0 \tag{9}$$

$$E[\alpha_g|\Omega_g] = E[\alpha_g] = 0. \tag{10}$$

In fact, we have assumed, by imposing that $E[\alpha_g|\Omega_g] = E[\alpha_g]$, that we are able to separate the effects of α_g, the unobservable inputs shared by the members – like the teacher's ability – from exogenous peer effects, captured in the reduced form as the effect induced by changes in $\mathbb{E}_g[X^\top]$. Those two effects may not be distinguished from each other when the conditional independence assumption is violated, that is,

when the group formation process occurs taking simultaneously in consideration both α_g and $\mathbb{E}_g[X^\top]$, even after controlling for Z_{Bg}. Finally, as usual, the normalization $E[\alpha_g] = 0$ allows π_0 to be identified.

One example of violation of the conditional mean independence assumption occurs when there is a positive correlation between observed characteristics, such as class socioeconomic index and the unobserved determinants of achievement at the class level. This could occur if more fortunate parents register their kids at schools where the teacher's ability is higher. Since this ability is hardly observed by the econometrician, this process would lead us to overestimate the contextual effects of the socioeconomic index.

We try to overcome this problem using a vector of controls, Z_{Bg}, that is broad enough to be credible to expect that the unobservable determinants of achievement at group level are not correlated with the socioeconomic index after conditioning on the information included in Z_{Bg}. Specifically to our problem we have used control demographic variables (dummy for school location in state capital and dummy for rural area), school administration (federal, state, municipal or private), principal and teacher's characteristics (race, age, gender, schooling and experience), along with indicators of inputs, physical installations and violence episodes at school.

We think that is reasonable to assume that the information available in this specification is nearly as good as the information available to the parents when they register their children at school. If so, variations in achievement between classes, once we control for this information, should not reflect any self-selection problem.[5]

Finally, we have also included the teacher's wage as a proxy for her unobserved ability. Assuming that ability is observed by employer and that there exists a teacher market, teacher's wage should reflect, at least partially, her productivity. Notice also that if we expect that there is a correlation between unobserved teacher ability and class socioeconomic index, this correlation should disappear when we condition on the teacher's wage. So, we conclude that the teacher's wage possesses the characteristics needed to function as a proxy for her ability.

Under our assumptions we can therefore identify endogenous peer effects, the parameter β_3, using a linear in means model. In the next section, we focus on results of standardized Math test scores and present estimates for the structural parameters for the population of fifth graders (11 years old) at elementary Brazilian schools. We first describe the data and then present our results.

3 Results

3.1 Data and Descriptive Statistics

We used the data set known as SAEB, acronym for *Sistema de Avaliação da Educação Básica*, or "Basic Education Evaluation System". It consists of students, teachers and principal's questionnaires and students' test scores. SAEB is run by the Ministry of Education and was created in 1990. Since 1995 it uses Item Response Theory (IRT) and is based on a representative at state level sample of schools that have at least 10 students enrolled in a tested grade. Since 2001 it evaluates only students' Math and Language proficiency at grades 5, 9 and 12 (11, 15 and 18 years old).

In this work we use information regarding the 2005 Math test scores of students at grade 5. Tables 1 and 2 display the main features of the data.

[**insert Table 1 around here**]
[**insert Table 2 around here**]

By inspection of these tables we can notice that boys perform slightly better than girls in Math. The same occurs with the students that reported their races as white, when comparing to other races. The variables parental education, socioeconomic index and the dummy for private school were the ones that explained most of the students variation in achievement. It is also important to notice that this table

[5] We are aware that, as pointed out by Heckman (1996, cited by Blume and Durlauf, 2005) *"persons making decisions have more information about the outcomes than the statisticians studying them"*, but we tried to make the bias associated with this "fundamental ignorance" to negligible levels.

shows that the school quality and school violence indexes present coefficients with the expected sign. In the appendix we explain how we constructed these index variables.

3.2 Within and Between Regressions

Table 3 presents parameter estimates of the within-regression. After controlling for several characteristics, girls had an average score 5.79 points smaller than boys. In the same fashion, students that reported race as white had an average score higher than the others in 1.39 points. All other variables present the expected sign, and the point estimates are quite similar to those found in the literature.[6]

[**insert Table 3 around here**]

Table 4 shows estimates of parameters obtained through a between-regression using class averages as units of observation. All three specifications, using teachers' and principals' characteristics yielded very similar results. Basically, we can say that for all variables associated to the student body, except for the proportion of female students and average age we found coefficients statistically different from zero. The violence and school quality indexes also presented significant coefficients and expected sign, as most of the demographic, teachers' and principals' characteristics. Interestingly, teacher's wage seems to be a relevant control and in the three models reported in Table 4 it appears with a positive sign.

[**insert Table 4 around here**]

Once we have reduced-form parameters from previous within and between regressions, we can proceed and measure peer effects using the formulae presented in section 2.2, case 2. Table 5 reports the estimates for the structural parameters. Standard errors were obtained through bootstrap replications.

[**insert Table 5 around here**]

As we can see, all three models indicate the presence of endogenous peer effects, measured by parameter β_3, which is statistically different from zero. Regarding contextual effects (β_{2AW}), it is interesting to notice that the sign of the contextual effect of female variable is the opposite of the effect of this variable at the individual level. In other words, despite the fact that girls had on average lower test scores, an increase in the proportion of girls in the classroom induces a positive effect on the achievement in that class. Thus, when looking at the overall effect of an increase in the proportion of girls in a classroom, as measured by $\delta_X = \partial \mathbb{E}_g[Y]/\partial \mathbb{E}_g[X]$, we will be summing two components with opposite signs yielding a non-significant effect.

Notice that the results in Table 4, whose coefficients for X can be interpreted as δ_X, are valid even when no credible exclusion restriction is available. Thus, even if we were unsure about the validity of the exclusion restriction that preschool did not present any contextual effect, results for δ_X would still be valid.

Furthermore it is clear that in all specifications we found a strong component of endogenous peer effects. The coefficient β_3 around 0.4 implies a social multiplier of 1.67 ($= (1 - 0.4)^{-1}$), which should be interpreted as the impact of exogenous changes in individual student i's achievement, propagating throughout the classroom, on average achievement in the class.

Finally, we conclude that peer effects seem to be an important part of the determination of the achievement of fifth graders in Brazil, both in the contextual form – through the characteristics of their peers like race, age and socioeconomic status – and in the endogenous form, through the achievement of their classmates.

[6] For example, Pinto (2008) using SAEB 2003 found similar results as those presented in Table 3.

4 Robustness Checks - Instrumental Variables

In this section we use alternative conditions with the purpose to verify the robustness of our results.

As discussed before, the hypothesis that is specially restrictive is $E[\alpha_g|\Omega_g] = E[\alpha_g]$, which imposes the absence of correlation between the average socioeconomic characteristics in the group and the unobserved component α_g. Being education a normal good one could imagine that wealthier families would buy better education that could be represented by Z_{Bg}.

Table 6 presents the estimates of peer effects using a slightly modified version of the original model. Because our previous estimates could be biased through some self selection process that would not be controlled by our "selection on observables" specification, we use an instrumental variable for the socioeconomic index, considered here our endogenous variable. We use the allocation rule of the students into classes as our instrumental variables, exactly as proposed by Pinto (2008), who describe in detail a microeconomic model of the principal's decision and shows the conditions in which the allocation rule of the students could work as an instrument.

There is a question in the principal's questionnaire on how classrooms are formed. Principals have to choose one out of five options; classes are formed favoring: (*i*) age homogeneity, (*ii*) past achievement homogeneity, (*iii*) age heterogeneity, (*iv*) past achievement heterogeneity, (*v*) no criteria or random allocation. We constructed five dummies based on these answers and their means can be seen in Table 1. We used the first four as instrumental variables and excluded the last one to avoid multicollinearity.

[**insert table 6 here**]

By inspection of Table 6, it is clear that the change in the estimation procedure kept results almost unchanged from the previous model and we could find some small reductions on the coefficients associated with race and previous retention. However and importantly, the coefficient associated with endogenous peer effects, was not statistically different from zero, except at 10% on the specification with the full vector of controls.

Finally, it is interesting to notice that the point estimate of the effect of the socioeconomic indicator was higher in the IV model than in the OLS specification. Under some assumptions, such result could serve as indirect evidence that the vector of controls that we used in the OLS specification is in fact removing selection bias, since one would expect a positive rather than negative bias. In part, that may be explained by the fact that we were able to include as a control variable the teacher's salary. Therefore we are, even though this might not be perfect, controlling for teacher's ability, which removes in principal a substantial part of potential biases.

5 Conclusion

Despite the effort to warrant universal access to basic and elementary education in Brazil, its educational system still shows serious deficiencies. According to the assessment realized by the Unesco, Brazil figures only at the 88th position, with worse achievement than other developing countries like Paraguay (72th), Bolivia (79th) and Colombia (75th) that have lower per capita income levels. The results from the assessment made by the PISA program point also at the same direction. Out of the 57 countries evaluated at the year of 2006, Brazil was in the bottom 7 in proficiency in Math and language. A better understanding of the relationship between educational inputs and student achievement is of great importance to Brazilian educational policy.

Although this paper focused only on Math test results for the student population at a single grade, our results might be helpful to inform policy as it shows that, by finding evidence of positive endogenous peer effects, there is a non-trivial impact of exogenously increasing individual student achievement on classmates. Fifth graders (11 years old) do seem to follow the norm in a given classroom, and therefore, policies that are not necessary universal but rather target students in a given class might end up helping a larger fraction of students through existing social interactions. This is good news for policy-makers as fewer resources than expected might be necessary to improve quality of education in Brazilian schools.

This paper also contributes to shedding some light on gender differences in school achievement. In Math at age 11, Brazilian girls tend perform poorer than boys, after we net out any social interactions. However, classes with larger proportions of girls have typically no worse grades in Math than in classes with larger proportions of boys. This surprising result shows the importance of being capable of separating out endogenous from exogenous peer effects. In this case, an increase in the proportion of girls have a positive contextual effect (exogenous peer effect) on individual Math scores, which implies that gender diversity at class level might be one important way to increase student achievement. Although we do not have a richer data set that could allow us to exploit further the mechanisms behind this result, that is in line with the existing literature on gender peer effects, which attribute these positive effects to better non-cognitive skills of female pupils.

6 References

Akerlof, G. A. (1997). "Social distance and social decisions". *Econometrica*, 65:1005–1027.

Albernaz, A., F. Ferreira and C. Franco. (2002). "Qualidade e equidade no ensino fundamental brasileiro". *Pesquisa e Planejamento Econômico*, 32:453–476.

Ammermueller, Andreas; Pischke, J.-S. (2006). "Peer effects in european primary schools: Evidence from PIRLS". IZA Discussion Paper.

Angrist, Joshua D.; K. Graddy, and G. imbens. (2000). "The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish". *Review of Economics Studies*, 67 (3):499–527.

Arnott, Richard; Rowse, J. (1987). "Peer group effects and educational attainment". *Journal of Public Economics*, 32:287–305.

Blume, L. E. and S. Durlauf. (2005). "Identifying social interactions: a review". Working paper.

Brock, William A. ; S. Durlauf. (2001). *"Interaction-based Models", Hand-book of Econometrics"*, chapter 5, pages 3297–3380. J. Heckman e & E. Leamer.

Calvo-Armengol, Antoni; Patacchini, and Y. Zenou. (2009). "Peer effects and social networks in education". Mimeo.

Fletcher, P. (1997). "À procura do ensino eficaz". Technical report, Ministério da Educação e Cultura - Avaliação da Educação Básica.

Franco, Creso, I. Ortigão, A. Albernaz, A. Bonamino, G. Aguiar, F. Alves, N. Sátyro (2007). "Qualidade e equidade em educação: Re-considerando o significado dos fatores intra-escolares". *Ensaio: aval. pol. públ. Educ.*, Rio de Janeiro, v.15, n.55, p. 277-298, abr./jun. 2007.

Glaeser, Edward L., Bruce I. Sacerdote, and Jose A. Scheinkman, (2003). "The Social Multiplier," *Journal of the European Economic Association*, MIT Press, vol. 1(2-3), pages 345-353, 04/05.

Glaeser, E. and J. Scheinkman (2001), "Measuring Social Interactions," in Social Dynamics, ed. by S. Durlauf and P. Young. Cambridge: MIT Press.

Graham, B. S. (2005). "Identifying social interactions through excess variance contrasts". Working Paper.

Graham, Bryan S.; Hahn, J. (2005). "Identification and estimation of the linear-in-means model of social interactions". *Economic Letters*, 88:1–6.

Hoxby, C. (2000). "Peer effects in the classroom: Learning from gender and race variation". NBER Working Paper.

Hoxby, C. (2002). "The power of peers: How does the makeup of a classroom influence achievement?". *Education Next*, 2:57–63.

Hoxby, Caroline ; Weingarth, G. (2006). "Taking race out of the equation: School reassingnment and the structure of peer effects". Working Paper.

Lavy, Victor, D. Paserman and A. Schlosser. (2008). "Inside the black box of ability peer effects: Evidence from the variation in high and low achievers in the classroom". .NBER Working Papers 14415.

Lavy, Victor, and A. Schlosser. (2011), "Mechanisms and Impacts of Gender Peer Effects at School," *American Economic Journal: Applied Economics*, American Economic Association, vol. 3(2), pages 1-33,

Manski, C. F. (1993). "Identification of endogenous social effects: The reflection problem". *The Review of Economic Studies*, 60:531–542.

Manski, C. F. (1995). "Identification problems in the social sciences". Cambridge, MA: Harvard University Press.

Pinto, C. (2008). "Semiparametric estimation os peer effect in classrooms: Evidence for brazilian schools in 2003". Mimeo.

Sacerdote, B. (2001). "Peer effects with random assignment results for Dartmouth roommates". *The Quarterly Journal of Economics*, 05:681–704.

Scheinkman, J. A. (2006). "Social Interactions". Princeton University and NBER.

Smith, L. I. (2002). "A tutorial on principal component analysis". Mimeo.

UNESCO (2010). "Education for all global monitoring report". Paris: UNESCO.

Appendix – Index Calculation

We constructed three synthetic indexes in this paper: (*i*) a socioeconomic index; (*ii*) a school quality index; and (*iii*) a school violence index. The three indexes were constructed using principal component analysis. For a revision of that procedure see, for example, Smith(2002). Below we have summarized the variables that we used in order to construct each one of those indexes. Table A.1 reports summary statistics of these variables.

[**insert table A.1 around here**]

For the socioeconomic index, we used the information regarding durable goods inventory present at student's household and housing characteristics. We used number of: cars, bathrooms, rooms per person, TV sets and radio sets. We also used dummies for presence of DVDs, refrigerator, washing machine, vacuum cleaner and finally the weekly frequency that the family used the services of a maid.

SAEB data includes a huge list of variables from principal's questionnaire that can be used to construct the school quality and school violence indexes. However, due to the high incidence of non-response to all these variables, we used a somewhat smaller list of variables for index construction. Note that most of variation in indexes using the full list of variables comes from the restricted list. In fact, for the restricted sample of schools with no missing data on these variables, the correlation between our indexes and the one that uses all information was larger than 0.8 for both quality and violence indexes.

Thus, for the violence index we used the information regarding: presence of drug dealing at the school (from a insider agent), presence of weapons (melee or firearms), action of gangs around the school and, finally, if there was any murder attempt against any of the school workers.

For the school quality index we used information regarding the physical structure of the school, like material used for roof, walls, doors, windows and floor. We also used information regarding cleanliness of building entrance and of the doors.

Table 1: Summary Statistics

Variable	Obs	Mean	Standard Deviation	Minimum	Maximum
Student Characteristics					
Math test score	41783	188.90	48.94	65.43	373.44
Socioeconomic Index	27123	0.00	2.12	-4.23	6.07
Female	40964	0.50	0.50	0.00	1.00
White	39751	0.38	0.48	0.00	1.00
Parental Education	31103	6.04	2.64	1.00	9.00
Age	41170	10.66	1.20	8.00	15.00
Ever Failed	40036	0.28	0.45	0.00	1.00
Preschool	39432	0.81	0.39	0.00	1.00
School Characteristics					
Violence Index	38074	0.00	1.36	-0.66	9.50
School-Quality Index	41887	0.00	1.94	-7.30	1.50
Urban	43823	0.94	0.23	0.00	1.00
State Capital	43823	0.48	0.50	0.00	1.00
Federal School	43823	0.01	0.08	0.00	1.00
State School	43823	0.35	0.48	0.00	1.00
Municipal School	43823	0.35	0.48	0.00	1.00
Private School	43823	0.30	0.46	0.00	1.00
Allocation Rules					
Age Homogeneity	39358	0.43	0.49	0.00	1.00
Past Achievement Homogeneity	39358	0.07	0.26	0.00	1.00
Age Heterogeneity	39358	0.10	0.30	0.00	1.00
Past Achievement Heterogeneity	39358	0.13	0.33	0.00	1.00
No criteria / Random Allocation	39358	0.27	0.44	0.00	1.00
Teacher Characteristics					
Wage (Monthly Salary in BRL$)	36953	905.35	628.38	0.00	3100.00
Female	38209	0.93	0.26	0.00	1.00
Age	38079	39.11	9.76	20.00	60.00
White	37878	0.46	0.50	0.00	1.00
Years of Experience	37759	13.48	6.76	0.00	22.00
Schooling	36789	14.00	1.74	8.00	15.00
Homework Assignment*	41779	2.96	0.23	0.00	3.00
Principal Characteristics					
Female	40555	0.81	0.39	0.00	1.00
Age	40706	44.56	8.86	22.00	60.00
White	40613	0.49	0.50	0.00	1.00
Schooling	38627	13.75	0.84	8.00	14.00
Years of Experience	40309	6.96	6.43	1.00	20.00

* Homework Assignment is a discrete variable assuming 4 values: 3 if homework is assigned and graded every week; 2 often but not every week; 1 never or rarely;) no homework is assigned.

Table 2: Univariate Regressions of Math Test Score on Class-level Characteristics

	Coefficient/Std. Error	Constant	N	R^2
Students' Characteristics				
Female	-3.387***	191.336***	40,964	0.001
	(0.482)	(0.340)		
White	16.142***	183.601***	39,751	0.026
	(0.500)	(0.307)		
Parental Education	7.104***	147.633***	31,103	0.143
	(0.098)	(0.649)		
Age	-12.267***	319.997***	41,170	0.091
	(0.191)	(2.048)		
Socioeconomic Index	11.285***	193.011***	27,123	0.239
	(0.122)	(0.260)		
Ever Failed	-38.148***	200.514***	40,036	0.122
	(0.511)	(0.270)		
Preschool	32.612***	163.370***	39,432	0.069
	(0.605)	(0.544)		
School Characteristics				
Violence Index	-3.325***	188.831***	38,074	0.009
	(0.183)	(0.250)		
School-Quality Index	6.482***	189.047***	39,956	0.066
	(0.122)	(0.237)		
Teachers' Monthly Salary	0.014***	177.680***	36,953	0.030
	(0.000)	(0.442)		
Urban	21.499***	168.695***	41,783	0.011
	(1.001)	(0.971)		
State Capital	10.783***	183.658***	41,783	0.012
	(0.476)	(0.332)		
Federal School	55.510***	188.550***	41,783	0.008
	(3.010)	(0.239)		
State School	-18.885***	195.354***	41,783	0.033
	(0.496)	(0.290)		
Municipal School	-27.413***	198.593***	41,783	0.072
	(0.483)	(0.287)		
Private School	48.558***	174.412***	41,783	0.206
	(0.466)	(0.255)		

Note: *** p<0.01, ** p<0.05, * p<0.1

Table 3: Within-Regression Results

	Coefficient/Std. Error
Socioeconomic_Index	0.377
	(0.233)
White	1.387**
	(0.665)
Female	-5.785***
	(0.599)
Parental Education	0.732***
	(0.152)
Age	-2.248***
	(0.355)
Ever Falied	-16.013***
	(0.871)
Preschool	9.096***
	(0.883)
Constant	215.027***
	(4.074)
R^2	0.229
N	18,438
F	135.670
σ^2_{α}	32.340
σ^2_{ε}	36.153
ρ (Fraction of the variance due to α_g)	0.445

Note: *** p<0.01, ** p<0.05, * p<0.1

Table 4: Between-Regression Results

	I	II	III
	Coefficient/Std. Error	Coefficient/Std. Error	Coefficient/Std. Error
Students' Characteristics			
Socioeconomic Index	6.231***	6.108***	6.004***
	(0.502)	(0.520)	(0.555)
White	7.331***	7.496***	8.658***
	(1.721)	(1.779)	(1.855)
Female	-2.305	-2.325	-3.326*
	(1.737)	(1.778)	(1.862)
Parental Education	0.951**	0.795**	0.901**
	(0.390)	(0.402)	(0.423)
Age	0.345	0.299	0.446
	(0.724)	(0.739)	(0.770)
Ever Failed	-19.685***	-20.160***	-20.789***
	(2.223)	(2.267)	(2.347)
Preschool	15.294***	15.426***	15.672***
	(2.192)	(2.238)	(2.289)
School Characteristics			
Violence Index	-0.805**	-0.751**	-0.942**
	(0.356)	(0.365)	(0.398)
School-Quality Index	0.511*	0.606**	0.706**
	(0.260)	(0.265)	(0.276)
Teacher's Monthly Salary	0.004***	0.004***	0.004***
	(0.001)	(0.001)	(0.001)
Urban	-1.103	-1.630	-1.165
	(1.785)	(1.832)	(1.918)
State Capital	2.119**	2.523**	2.998***
	(1.006)	(1.049)	(1.105)
Federal School	31.840***	19.209**	18.093**
	(6.781)	(7.539)	(7.900)
State School	3.249***	3.767***	3.631***
	(1.109)	(1.167)	(1.221)
Private School	20.329***	20.741***	18.798***
	(1.666)	(1.709)	(1.821)
Controls:			
State Dummies	Yes	Yes	Yes
Teachers' Characteristics	No	Yes	Yes
Principals' Characteristics	No	No	Yes
Intercept	158.957***	144.351***	148.115***
	(9.083)	(11.894)	(14.529)
R-squared	0.62	0.63	0.63
Number of Observations	14,400	13,517	12,266
Number of Classes	2,761	2,590	2,355

Note: In all regressions, dependent variable is Math scores.
Clustered standard errors at classroom level. *** p<0.01, ** p<0.05, * p<0.1

Table 5: Endogenous and Contextual Peer Effects

	I	II	III
	Coefficient/Std. Error	Coefficient/Std. Error	Coefficient/Std. Error
Endogenous Peer Effects	0.405***	0.410***	0.420***
	(0.121)	(0.120)	(0.120)
Contextual Peer Effects			
Socioeconomic Index	3.329***	3.225***	3.108***
	(0.788)	(0.791)	(0.793)
White	2.972**	3.033**	3.637**
	(1.417)	(1.409)	(1.477)
Female	4.414***	4.414***	3.855***
	(1.355)	(1.401)	(1.385)
Parental Education	-0.167	-0.263	-0.209
	(0.314)	(0.311)	(0.323)
Age	2.453***	2.425***	2.507***
	(0.552)	(0.554)	(0.568)
Ever Failed	4.306	4.126	3.947
	(2.983)	(3.059)	(3.092)
Controls:			
State Dummies	Yes	Yes	Yes
Teachers' Characteristics	No	Yes	Yes
Principals' Characteristics	No	No	Yes

Note: *** p<0.01, ** p<0.05, * p<0.1

Table 6: IV Estimates of Endogenous and Contextual Peer Effects

	I	II	III
	Coefficient/Std. Error	Coefficient/Std. Error	Coefficient/Std. Error
Endogenous Peer Effects	0.323	0.362	0.391*
	(0.272)	(0.222)	(0.230)
Contextual Peer Effects			
Socioeconomic Index	7.433	5.835	4.610
	(8.981)	(7.004)	(6.772)
White	1.106	1.870	2.846
	(3.215)	(2.577)	(2.284)
Female	4.478**	4.390***	3.759**
	(1.750)	(1.639)	(1.659)
Parental Education	-1.415	-1.124	-0.673
	(2.112)	(1.650)	(1.549)
Age	2.323***	2.294***	2.442***
	(0.802)	(0.730)	(0.750)
Ever Failed	3.283	3.392	3.580
	(4.762)	(4.206)	(4.073)
Controls:			
State Dummies	Yes	Yes	Yes
Teachers' Characteristics	No	Yes	Yes
Principals' Characteristics	No	No	Yes
First Stage F- Statistic	3.600***	3.640***	3.470***
Number of Observations	13796	12934	11755

note: Student allocation rules as instrumental variables. *** p<0.01, ** p<0.05, * p<0.1

Table A.1: Summary Statistics of Variables Used in Indexes Construction

Variable	Obs	Mean	Standard Deviation	Minimum	Maximum
Violence Index					
Fire Arms	40215	0.03	0.16	0	1
White Arms	40210	0.19	0.39	0	1
Gangs (External to the School)	40114	0.19	0.40	0	1
Gangs (Inside the School)	40093	0.03	0.17	0	1
Murder Atempt	39236	0.02	0.13	0	1
School-Quality Index					
Conservation of:					
Roof	43143	1.39	0.63	1	4
Walls	43105	1.36	0.58	1	4
Floor	43091	1.41	0.64	1	4
Doors	42972	1.45	0.67	1	4
Windows	42944	1.44	0.71	1	4
Cleanliness of:					
Entrance	43118	0.09	0.29	0	1
Doors	42992	0.15	0.36	0	1

WORKING PAPER 17/2011 • NOVEMBER, 2011 •