Inequality of opportunity in educational achievements in Spain

C. Lasso de la Vega[1], *A. Lekuona*[2] *and S. Orbe*[3]

1. Introduction

The aim of this work is to analyze to what extent the inequalities in the educational achievements in Spain are due to students' characteristics or socio-economic factors and how much they are due to effort related factors. The theory of inequality of opportunity (Roemer (1998)) is concerned with the answer to these questions. According to this theory two kinds of determinants can be distinguished in the results obtained by the individuals. On the one hand, there are "circumstances" beyond an individual's control, such as genetic characteristics and socio-economic background, and on the other hand, there are "efforts" related to individual responsibility. Differences emerging from the former factors are deemed ethically illegitimate, while differences due to the latter ones are considered tolerable. Inequality of opportunity should reflect how much of the total outcome inequality is due to circumstances.

This paper evaluates the inequalities of opportunities in educational achievements in Spain taking into account different circumstance variables. In addition, we identify different sub-populations according to gender and geographical location. We take data from the PISA 2012 database. Following the non-parametric methodology proposed by Checchi & Peragine (2010), we construct counterfactual distributions in which all the differences due to effort have been removed and the only differences that remain are due to circumstances. We then evaluate the inequality of these counterfactuals using the variance. We analyze the results according to geographic characteristics and gender differences.

2. Data set

Data is taken from the cross-country student assessment PISA conducted in 2012. PISA is a three-yearly international program managed by OECD and its main objective is to provide comparable insights into 15-year-old students on their cognitive abilities in reading, scientific and mathematical literacy. In addition, the information on cognitive achievements can be connected with features related to students' personal characteristics as well as family and school background. The Spanish data refers to 25 thousand students representing a population of more than 373 thousand youngsters. Given that schooling is mandatory in Spain for youngsters up to 16 years, there is no bias in the analysis related to school drop out.

PISA selects the participant students using a two-step sampling procedure. At the first step schools in each participating country are randomly chosen. At the second step fifteen-year-old students, attending grade 7 or higher are randomly selected from the schools. Sampling weights are provided by PISA to adjust the

[1]casilda.lassodelavega@ehu.es
[2]agurtzane.lekuona@ehu.eus
[3]susan.orbe@ehu.eus

results for the probabilities of being selected and the schools' and students' non-response. Accordingly, we consider the survey design of the data in the computations by using the sampling weights.

The educational achievements of the participant students are determined by using a complex technique based on the Item Response Theory (IRT).[4] Accordingly, instead of providing a single score point that represents academic performance, a set of five plausible values are randomly drawn for each student. In this study the educational achievements of students are measured using the first plausible value on mathematics (`PV1MATH`) to keep things simple.[5] Plausible values are standardized so that the average score of OECD countries is 500 and standard deviation equals to 100. The standardization process consists of both a translation by the difference between the new and the original means and a rescaling by the ratio between the new and the original standard deviations. Since the standardization is just a change in the metric, this process should not alter the inequality rankings one would obtain before and after the standardization. For instance, if the original achievements in country A are more unequally distributed than the standardized achievements in country B, then the standardized achievements in country A should also be more unequally distributed than the standardized achievements in country B. However, no relative inequality measures satisfy this requirement, meaning that if inequality is measured according to any relative inequality index, the ranking between two countries may reverse before and after standardization. Fortunately, the *variance* always preserves these rankings before and after the standardization process. More details on the reasons and the properties of the variance will be provided in the methodology section. Accordingly, the descriptive statistics and the inequality of opportunity measures presented through the paper are computed using the variance.

Another important issue related with the data is that there are missing values in the PISA 2012 dataset. Dropping students with missing values might lead to a sample selection bias if the values are not missing randomly. Indeed, the average achievement of students that have not answered the questions about their family background, such as their parents' education level or occupation, is lower than the overall average achievement. This fact indicates that missing values are more concentrated in students with lower average achievements, and so, if we drop these students the educational achievements might be upward biased. In order to deal with the missing values, we assume that they are not completely random and that the systematic differences between the missing and the observed values can be explained by differences in the observed data. Based on this assumption, we have carried out imputations of the missing values following the procedure introduced by Buuren et al. (1999), known as *Multiple Imputations Chained Equations* (MICE), implemented in Stata by Royston & White (2011).

Table 1 presents the sample information for educational achievements in the different regions of Spain.

[4] See Ferreira & Gignoux (2014)for detailed information on this method.

[5] Data Analysis Manual (OECD (2009)) suggests that the results should be estimated using each of the PVs separately and then averaged to obtain the final estimate. However, the same manual states that with a sample size larger than 6400 students, using one or five plausible values does not make any substantial difference in the mean estimates nor in the standard error estimates. Given the larger sample size of Spain, with 25313 students, we use a single plausible value.

It can be observed that the country average score is 484.62, which is below the OECD's average score (500). However, we find strong internal divergence across regions. In particular, the regions that are located in the North present mean achievements higher than 490 score points, (except Galicia with a slightly lower value) whereas the regions located in the South show mean values lower than 475. Accordingly, the regions have been sorted into two macro-regions, Northern regions with their overall mean (499.34) very close to the one of OECD countries, and the Southern regions whose overall mean (469.92) lies far below that average. Average scores also vary according to gender. In general girls obtain lower results than boys in mathematics with their average scores varying from 476.56 to 492.45, respectively.

Table 1: Sample statistics and descriptive statistics of educational achievements per region

	Math score		
Regions	**Mean**	**Variance**	**Sample**
North	**499.34**	**7474.02**	**186787**
Aragon	495.26	8626.32	9988
Asturias	500.04	8719.92	7125
Basque Country	505.02	7071.26	16143
Cantabria	491.46	8088.02	4334
Castile and Leon	509.65	7060.86	18422
Catalonia	493.69	7128.55	55833
Galicia	488.94	7366.18	18287
Madrid	503.40	7445.76	48845
Navarre	515.92	7199.21	5245
La Rioja	502.85	10197.53	2566
South	**469.92**	**7343.97**	**186904**
Andalusia	472.70	7204.33	75553
Balearic Islands	474.94	7583.10	8385
Extremadura	461.46	8624.18	10399
Murcia	461.89	8120.79	13115
Rest of Spain	469.18	7136.23	79452
Total	**484.63**	**7624.96**	**373691**

Rest of Spain comprises the regions Castile-La Mancha, Valencia, Canary Islands and Ceuta and Melilla.

We consider the four populations that are more homogeneous in term of educational achievements — girls in the North, girls in the South, boys in the North, boys in the South — and we analyse the inequalities for the entire population and then for the following sub-populations using the variance. Table 2 presents the descriptive statistics of these sub-populations. As can be observed in the table, only the sub-sample of boys in the Northern regions obtain higher average achievement than OECD's average. At the other end, the lowest performance is observed among the girls in the Southern regions. The variances indicate that achievements are more unequally distributed among boys in both macro-regions. This fact is reaffirmed in the absolute Lorenz curves in Figure 1.

Table 2: Sample statistics and descriptive statistics of educational achievements per region

	Girls	**Boys**	**Total**
North	491.43	507.01	499.34
	6745.13	8061.67	7474.02
	91917	94870	186787
South	461.71	477.90	469.92
	6429.02	8105.23	7343.97
	92073	94831	186904
Total	476.56	492.45	484.63
	6807.12	8294.34	7624.96
	183990	189701	373691

The *first row* presents the mean educational achievement; *second row* shows the variance and the *third row* provides the weighted number of observations.

Lorenz curves offer a visual representation of inequality and provide partial orderings of distributions. The horizontal axis indicates the cumulative percentage of students ranked from those with the lowest to those with the highest achievements. The vertical axis shows the average achievement that would be necessary in order to provide any student in that percentile with the population's mean achievement. The larger the distance from the horizontal line representing perfect equality, the larger the inequality. The curves always take negative values and they are decreasing when the cumulative outcomes of students are lower than the mean value of the distribution, and increasing subsequently up to when the value of the total population is again equal to zero.

It is observed in Figure 1 that the achievement distributions of girls both in the Northern and Southern regions are more equal than those of boys in these macro-regions. Given that the curves offer only a partial ranking, when Lorenz curves cross they offer no additional information to prefer one distribution over another. Accordingly, no clear conclusions can be drawn between the distributions of girls in the Northern and Southern regions nor between boys in these macro-regions. In particular, with regard to the achievement distribution of the girls, whereas the median suggests that the distribution of girls is more equal in the Southern regions than in the Northern ones, the upper quartile suggests that these distributions of girls are equal. When it comes to the distribution of boys, looking at the median suggests that again the boys in the Southern regions are more equal than boys in Northern regions. However, looking at the upper quartiles suggests the contrary.

This study is concerned with understanding the different sources of the inequalities observed in the given distributions. In particular, the focus is on quantifying how much of the overall inequality in the achievement distributions can be explained by differences in socially-inherited circumstances of students. In the following paragraphs we present the variables used for this purpose.

We select variables provided in the PISA 2012 database indicating the highest parental education, number

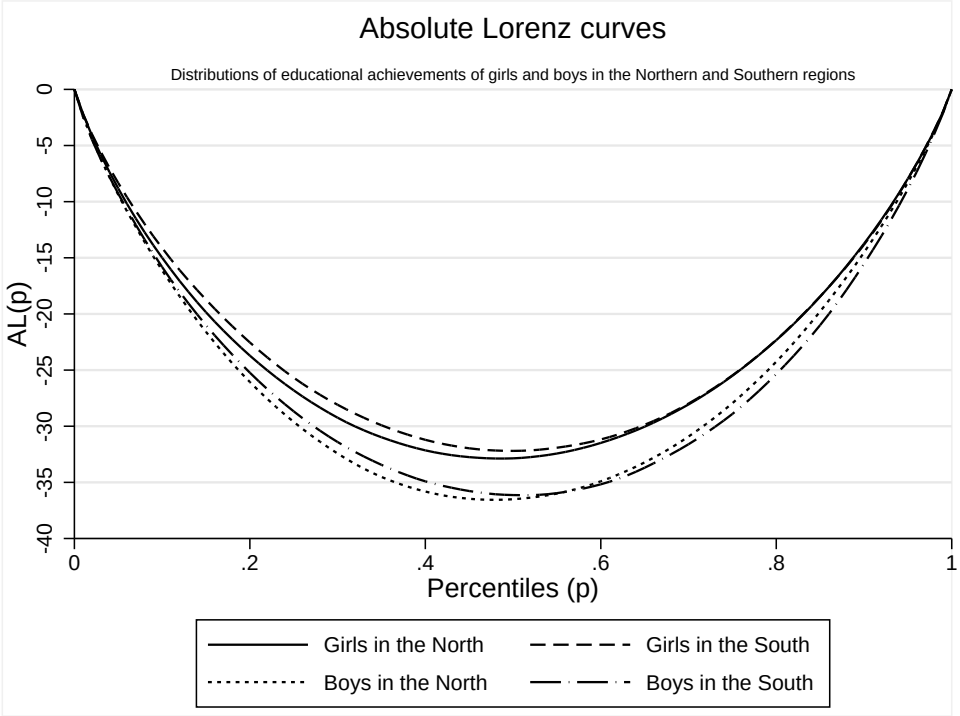

Figure 1: Absolute Lorenz curves of the distribution of educational achievements by gender and geographic location

of books at home, and the position on the parents' occupational status[6] and socio-economic and cultural status[7] indices' distribution within the sample. These variables are described in Table 3

Parental education is a standard measure of family background in studies on student background and educational achievement (see for instance,Ammermüller et al. (2005); Nonoyama-Tarumi (2008); Wößmann (2010)) as well as studies that measure the educational inequality of opportunity using either non-parametric approaches (Checchi & Peragine (2005) and Gamboa & Waltenberg (2012)) or parametric approaches (Ferreira & Gignoux (2014); Salehi-Isfahani et al. (2014); Waltenberg & Vandenberghe (2007)).

The number of books is also commonly used as proxy for educational, social and economic background of students' families (see for instance Wößmann (2004); Peterson & Wößmann (2007); Schütz et al. (2008)). Mullis et al. (2004) assert that a large number of books indicates a family background that appreciates education greatly and boosts children's academic effort. Peterson & Wößmann (2007) claim that it also captures social and economic aspects because it reflects the feasibility of a family and their decision on how much to invest in books. The same authors affirm that the number of books has repeatedly proved to be the single most important predictor of student performance on international achievement tests.

The *highest occupational level of parents* index, `HISEI`, is derived from the students' responses about their parents' occupation and working status and then codified according to the International Socio-Economic

[6]PISA 2012 index of highest occupational status (`HISEI`)
[7]PISA 2012 index of the family socio-economic and cultural status (`ESCS`)

Table 3: Description of independent variables

Variables used to divide the population into sub-samples	
Gender	Dummy variable equal to 1 if the student is a boy, 0 otherwise.
Geography	Dummy variable equal to 1 if the student lives in a region in the North, 2 if she lives in a region in the South. Northern regions include Aragon, Asturias, the Basque Country, Cantabria, Castile and Leon, Catalonia, Galicia, Madrid, Navarre and la Rioja. Southern regions comprise Andalusia, Balearic islands, Extremadura, Murcia and rest of Spain (Canary Islands, Castile-La Mancha, Ceuta and Melilla and Valencia).
Circumstances	
Pared	Categorical variable indicating parents' highest educational level coded according to ISCED codes into four categories: 1) no education (ISCED 0) or primary education (ISCED 1); 2) lower secondary (ISCED 2); 3)upper secondary(ISCED 3) or post-secondary non-tertiary education (ISCED4); 4) first stage of tertiary education (ISCED 5) or second stage of tertiary education (ISCED 6) .
Books	Categorical variable indicating number of books at home coded into four categories: 1) 0-25 books, 2) 26-100 books, 3) 101-200 books and 4) more than 200 books.
HISEI	Categorical variable indicating the quartile in the sample distribution of the index *Highest occupational status of parents*; 1 indicating the first quartile and 4 the fourth quartile. Higher values represent higher socio-economic status.
ESCS	Categorical variable indicating the quartile in the sample distribution of the index *socio-economic and cultural status.*

Index of Occupational Status (ISEI) codification (see Ganzeboom et al. (1992) for its construction). This index captures the attributes of occupations that convert education into income, that is, it maximizes the indirect effect of education on income through occupation and minimizes the direct effect of education on income. The index, HISEI, displays the higher of the two index values of either parent, or the index value itself in the single parent case. This index, hence, represents the socio-economic status of the family and it has been used in several studies such as Entorf & Minoiu (2005); Jenkins et al. (2008); Schnepf (2007).

The index of *Economic, Social, and Cultural Status*, `ESCS`, is a broader measure of students' socio-economic status that incorporates the previous variables in addition to the other variables that indicate the ownership of home durables related to family wealth, cultural possessions and educational resources. Positive values of this index indicate higher levels of socio-economic and cultural status. In the reports provided by PISA (see for instance OECD (2013)) this index is used to analyze how equitable a school system is, that is, the stronger the impact of the index of `ESCS` on the students' performance, the less equitable the school system. This index has been used in a similar way also in other studies such as Schütz et al. (2007); Causa & Chapuis (2011).

Descriptive statistics of the variables used to represent the circumstances are presented in Appendix A. As can be observed from Table 6 to Table 9, the means of achievements are increasing in the values of variables representing parental education, number of books and the position in the distribution of `HISEI` and `ESCS` indices. In this line, the values of these variables (1-4) are considered as being ordered from the least to the most favourable circumstantial environment. The mean scores are also higher for sub-samples of boys than for those of girls, and they are also higher in the North than in the South. Accordingly, the highest mean scores are found among boys in the Northern regions and in the fourth category (students whose parents have tertiary education or have more than 200 books or are in the fourth quartile of `HISEI` or `ESCS`). In contrast

the lowest average scores are obtained by girls in the Southern regions and in the first category (students whose parents have primary education or less, have 0-25 books, or are in the first quartile of the indices).

The sample shares of students that belong to each category of circumstance's variables are summarized in Table 4. As can be observed, girls and boys are equally distributed within the categories, that is, the shares of girls and boys are similar in all categories.

Table 4: Share of students that belong to each type in each sub-population

	North		South				
	Girls(%)	**Boys(%)**	**Girls(%)**	**Boys(%)**	**North(%)**	**South(%)**	**Total (%)**
Pared							
1	5.15	5.15	10.29	9.60	5.15	9.94	7.54
2	13.85	12.78	19.71	20.42	13.31	20.07	16.69
3	26.62	27.14	29.11	28.66	26.88	28.88	27.88
4	54.39	54.93	40.89	41.33	54.66	41.11	47.89
Books							
1	18.19	21.38	26.55	32.1	19.81	29.37	24.59
2	31.03	30.84	33.29	32.09	30.93	32.68	31.81
3	23.48	20.53	21.59	18.09	21.98	19.81	20.9
4	27.3	27.25	18.58	17.72	27.27	18.14	22.71
`HISEI`							
1	24.75	23.62	35.65	33.85	24.18	34.74	29.46
2	22.52	21.48	21.54	23.35	21.99	22.46	22.23
3	24.61	25.92	22.69	21.29	25.28	21.98	23.63
4	28.11	28.98	20.12	21.52	28.55	20.83	24.69
`ESCS`							
1	20.46	19.14	30.92	29.73	19.79	30.32	25.06
2	24.78	24.65	24.83	26.04	24.72	25.45	25.08
3	25.48	26.64	23.52	24.37	26.07	23.95	25.01
4	29.27	29.57	20.73	19.85	29.42	20.28	24.85

Girls(%) share of girls in Northern or Southern regions that belong to each type —; Boys (%) share of boys in Northern or Southern regions that belong to each type —; North (%) share of students in Northern regions that belong to each type —; South (%) share of students in Southern regions that belong to each type —; Total(%) share of whole population that belong to each type .

The category that includes the students whose parents have completed a tertiary education takes the largest share of the population in both macro-regions, with shares of 54.66% and 41.11% in the Northern and Southern regions, respectively. However, on average there are more highly-educated parents in the North.

The category in which students own 26-100 books presents the largest share of students in both macro-regions, with shares of 30.93% in the Northern sample and 32.68% in the Southern. However, in the Northern regions the second largest category is the one composed by students with more than 200 books (27.27% of students in the North) whereas in the Southern regions the second largest category is the one that gathers students that own less than 25 students (29.37% of students in the South). So, in general students in the Northern regions possess more books.

In the North there are more students that are located in the upper tail of the `HISEI` and `ESCS` indices' distribution. The opposite is true in the South, where the largest number of students is located at the bottom end of the distributions.

All in all, the students in the Northern regions present better a circumstantial environment than those in the Southern regions. However, there are no significant differences in the circumstances environment between boys and girls.

3. Methodology

3.1. Analytical framework

Consider a population of N students and we denote by $x_h \in R_+^N$ the educational achievement of student h. Suppose that for each student $h = 1, ..., N$, x_h is completely determined by two classes of characteristics. The first class includes factors beyond the individual's responsibility and is represented by a single *circumstance*, c_h, whose values belong to a finite set $\Omega = \{\omega_1, \cdots, \omega_i, \cdots, \omega_k\}$. The second class includes factors within an individual's responsibility and is represented by a scalar variable of *effort* $e_h \in \Theta$. We denote by $X = \{x_1, \cdots, x_N\}$, $C = \{c_1, \cdots, c_N\}$ and $E = \{e_1, \cdots, e_N\}$ the respective vector of achievements, circumstances and effort and we denote the population by $D = (X, C, E)$.

A measure of inequality of opportunities is a function $M : D \to R$. In order to ensure that the measure M reflects the inequality of opportunity, it needs to satisfy some principles that are basically classified as compensation or reward principles; In this paper we focus on *ex-post* formulation of the compensation principle.[8] This principle is a generalization to the measurement of inequality of opportunities of the Pigou-Dalton transfer principle, the key axiom in income inequality measurement. The intuition of the axiom is the following. We consider two populations with the same number of students. In the first population, students i and j, make the same effort, $e_i, e_j \in \Theta$, respectively, but the achievement of student i is higher than the achievement of student j, $e_i = e_j$ and $x_i^1 > x_j^1$. Since they exert the same effort, these differences are due to circumstances and are considered illegitimate. In the second population, students i' and j' have the same circumstances and efforts as i and j respectively. For some $\delta > 0$, the achievement of student i' is $x_i^2 = x_i^1 - \delta$ and the achievement of student j' is $x_j^2 = x_j^1 + \delta$. The rest of the students in the two populations are pairwise identical, which means that for each of the rest of students in the first population there is one student in the second population with exactly the same achievement, effort and circumstance variables. Then, according to the ex-post compensation principle the inequality of opportunity is lower in the second population than in the first, because the inequalities due to circumstances to be compensated are smaller in the second population. Formally this principle is stated as follows.

Ex-post Compensation Principle A measure of inequality of opportunity M satisfies ex-post compensation if, for all $d^1 = (X^1, C, E)$, $d^2 = (X^2, C, E) \in D$, such that there are $\delta \in R_{++}$ and $i, j \in \{1, \ldots, N\}$ with $e_i = e_j$ and $x_i^2 = x_i^1 - \delta \geq x_j^2 = x_j^1 + \delta$, and for all $h \notin \{i, j\} : x_h^2 = x_h^1$, then $M(d^2) < M(d^1)$.

Then, we define the function by which the educational achievements are generated.

$$X = g(C, E), X \in R_+^N \tag{1}$$

From this model, it is possible to build two kinds of counterfactual distributions. The first counterfactual reflects only differences due to circumstances while all the differences due to efforts are removed. Thus, given

[8] For a detailed description of these principles see, for instance, Fleurbaey & Peragine (2013); Van de gaer & Ramos (2015).

a population $d = (X, C, E)$, we denote by $X^C(X, C, E)$ the counterfactual distribution of this type. Then, a *direct* measure of inequality of opportunity, denoted by M^D, evaluates inequality of opportunity as follows:

$$M^D(X, C, E) \quad = \quad I(X^C(X, C, E)) \tag{2}$$

where I is an inequality measure.

The second kind of counterfactual reflects only fair inequalities whereas all the unfair inequality is removed, and it is denoted by $X^E(X, C, E)$. An *Indirect* measure, M^I, evaluates inequality of opportunities as the difference between the inequality in the actual distribution X and the inequality in the counterfactual distribution, X^E as follows

$$M^I(X, C, E) \quad = \quad I(X) - I(X^E(X, C, E)) \tag{3}$$

The selection of the inequality measure, I, depends on the nature of the outcome variable. As it was mentioned in the dataset section, PISA's original achievement data is transformed through a standardization procedure. As a consequence of this metric change, only the variance is suitable because it is the only measure that guarantee that the rankings before and after the standardization are preserved. Another reason for choosing the variance is that, I needs to be necessarily additively decomposable, such that the total inequality of a population can be divided into illegitimate and legitimate inequalities. Correspondingly, the variance is the only absolute decomposable measure that preserves the rankings before and after the standardization of the original data. The reasons for the uniqueness of the variance are that, firstly, it is invariant under translation; secondly, it is unit-consistent ; and finally, it is additively decomposable. This fact allows us to divide the total inequality of a population into a weighted average of the inequality existing within its subgroups and the inequality existing between them (see Zheng (2007) for further details). Therefore, the equations (2) and (3) are reformulated as follows,

$$M^D(X, C, E) \quad = \quad var(X^C(X, C, E)) \tag{4}$$

and

$$M^I(X, C, E) \quad = \quad var(X) - var(X^E(X, C, E)) \tag{5}$$

Another advantage of using the variance is that the inequalities applied to counterfactual distributions in equations (4) and (5) coincide.

In this study we follow the non-parametric procedure introduced by Checchi & Peragine (2010) in order to construct counterfactual distributions and to measure the inequality of opportunity. In particular, we follow the *ex-post* approach that satisfies the ex-post compensation.

In the ex-post approach, the individuals that share the same efforts belong to the same group called

tranches. There is equality of opportunities if individuals that make the same effort have equal opportunities to reach the same outcomes. Therefore, the inequality of opportunity should reflect the outcome inequalities between students that exert the same effort, i.e., within tranches. Ex-post compensation suggests that within-tranches inequalities should be compensated.

Therefore, we start dividing the population in groups, then we construct counterfactual distributions and finally we apply the inequality measure to the counterfactual distributions directly or indirectly. In the following section this procedure is explained in detail.

3.2. Ex-post approach

Since this approach is focused on the outcome differences found among the individuals who exert the same effort, the population should be partitioned according to their efforts. At this point, there are two key issues to be considered. On the one hand, effort is not directly observable and generally no appropriate proxies are provided in the datasets. On the other hand, effort is not uniquely determined by individuals' autonomous choices but also by their circumstances. According to Roemer (1998), this effort should be cleaned from any influence of circumstances. Taking these two facts into account, we follow the alternative proposed by this author that consists of two steps. In the first step, the population of students is divided according to their circumstances into circumstances-groups, so-called *types*. In the second step, the effort is proxied by the rank of each individual in the outcome distribution of their own type. The intuition behind this proposal is that, under the assumption that all circumstances are considered and achievements are monotonically increasing in effort, individuals are located in the same quantile in both, effort and achievement distributions of their types. Thus, individuals in the same position of their own type distribution have exerted the same *degree* of effort. This is known as *Roemer's Identification Assumption* (Van de gaer & Ramos (2015)). Under this assumption, it is possible to divide the population based on autonomously undertaken effort, whether it is observed or not.

Since the effort is considered as a characteristic of a type, first of all, the population is partitioned into $K = \{1, ..., i, ..., k\}$ types. Then, each type is divided in $M = \{1, ..., j, ...m\}$ tranches, where each tranche j is composed of individuals who are placed in the j^{th} quantile in the outcome or effort distribution of their type $i \in K$. The overall outcome distribution can be written as,

$$\chi = \{\chi_1, ..., \chi_j, ..., \chi_m\} \in R_+^N \tag{6}$$

where $\chi_j = \{X_{1j}, ..., X_{ij}, ..., X_{kj}\} \in R^{N^j}$ is the outcome distribution of the individuals located in the j^{th} effort-group across types, N^j is the number of individuals in effort-group j. $X_{ij} = x_{ij}^1, ..., x_{ij}^h, ..., x_{ij}^{N_i^j} \in R^{N_i^j}$ is the distribution of achievements of those who have exerted effort-degree j and are in type i, and it is denoted as *cell* ij. N_i^j is the number of individuals in cell ij.

Within the distribution X_{ij} the individual outcomes, from x_{ij}^1 to $x_{ij}^{N_i^j}$, are likely to present some inequalities. However, in Roemer's framework, the achievements are determined only by circumstances and

autonomous efforts. Therefore, in order to decompose the total inequality into inequalities by sources, first of all the residual inequality within X_{ij} should be eliminated. A way of doing this is to apply the *smoothing* process to the distribution of interest; X_{ij}, in this case. The process consists of constructing a hypothetical distribution where individuals in cell ij obtain the average achievement of their own cell, denoted by $\bar{X}_{ij}$. That is, the outcome of individual $h \in N_i^j$, $x_{ij}^h \in X_{ij}$, $\forall i \in K$ and $\forall j \in M$ is substituted by the arithmetic mean of that distribution, $\bar{X}_{ij}$. Hence, we obtain the smoothed distribution $X_{ij}^S = \{\bar{X}_{ij} 1_{N_i^j}\}$ where $1_{N_i^j}$ is the unit vector of length N_i^j. The corresponding distribution of effort group j is now denoted as $\chi_j^S = \{X_{1j}^S, ..., X_{ij}^S, ..., X_{kj}^S\} \in R_+^{N^j}$, and the initial ex-post achievement distribution (6) is rewritten as,

$$\chi^S = \{\chi_1^S, ..., \chi_j^S, ..., \chi_m^S\} \in R_+^N \tag{7}$$

We now proceed to decompose the overall inequality by sources.

Counterfactual reflecting inequalities only due to efforts

In this distribution, the inequalities due to circumstances are eliminated by substituting each outcome of individuals in tranche j across types, X_{ij}^S, by the mean achievement of that tranche, $\bar{X}_j$.

$$\chi_B^S = \{\bar{\chi}_1 1_{N^1}, ..., \bar{\chi}_j 1_{N^j}, ..., \bar{\chi}_m 1_{N^m}\} \in R_+^N \tag{8}$$

where $\bar{\chi}_j$ is the mean achievement of students in tranche j and 1_{N^j} is the unit vector of length N^j. In this distribution, every individual that exert the same degree of effort obtains the same achievement, so there is no inequality of opportunity.

Counterfactual reflecting inequalities only due to circumstances

In this distribution inequalities between tranches are eliminated by means of a rescaling process applied to the smoothed distribution χ^S.

$$\chi_W^S = \{\widetilde{\chi}_1, ..., \widetilde{\chi}_j, ..., \widetilde{\chi}_m\} \in R_+^N \tag{9}$$

where $\widetilde{\chi}_j$ is obtained by rescaling the mean achievement of students in cell ij, $\bar{X}_{ij}$, $\forall i \in K$ and $\forall j \in M$ in the following way,

$$\bar{X}_{ij} \to \bar{X}_{ij} + \bar{X} - \bar{\chi}_j \tag{10}$$

where $\bar{X}$ is the overall average achievement and $\bar{\chi}_j$ is the average achievement of tranche j. By means of this transformation, the new means of all tranches are set to be the population mean. So, when an absolute inequality index is used, the influence of belonging to a particular effort-group is eliminated while the inequalities within these groups are preserved. This way, χ_W^S reflects inequalities emerging only from circumstances differentials.

Inequality of opportunity measurement

The indirect measure of inequality of opportunity is computed as the difference between the inequality in the

actual distribution and the inequality in the distribution reflecting only fair inequalities,

$$M^I(X) = var(\chi^S) - var(\chi_B^S) \tag{11}$$

The counterfactual χ_B^S in the indirect approach satisfies ex-post compensation. The direct measure of inequality of opportunity is given by the following expression,

$$M^D(X, C, E) = var(\chi_W^S) \tag{12}$$

Given the decomposability property of the variance, the direct and indirect approaches provide the same results, $M^D(X) = M^I(X)$.

3.3. Empirical application

In order to divide the students according to their circumstances in types, we need to take into account at least two key factors: Firstly, Gamboa & Waltenberg (2012) state that "any choice of circumstances will be questionable" and highlight the importance of working with different alternative definitions of types. Secondly, Ferreira & Gignoux (2011) point out the inconveniences of the excessive division of the population, given that a small number of students within groups leads to larger sampling variances, and this, in turn, leads to biased estimations of inequality of opportunity. Therefore, in this study we consider both the need for alternative definitions of types as well as the sample size of the groups. Accordingly, we choose four different circumstance variables, each with four categorical values, and then, the population is divided into four types according to each circumstance variable.

In order to divide the students into tranches, first of all we define degrees of efforts under the assumption that students in the same percentile of the achievement distribution, conditional on each type definition, have exerted the same degree of effort. Thus, we have partitioned the conditional achievement distributions into 10 deciles, representing 10 degrees of effort, in each type and in the four sub-samples. All in all, according to each type definition there are 40 groups of students that share the same circumstances and degrees of effort, or *cells* (10 degrees of effort x 4 types of family backgrounds), in the North and in the South and for each gender.

Tables 10- 13 in Appendix B show the average educational achievements and the weighted number of observations in these cells. Each table presents the types constructed upon the four variables of circumstances presented in the dataset section.

4. Results

The results presented in this section must be regarded as very preliminary and require further analysis.

Figures 2, 3, 4 and 5 in Appendix C present the graphs of distributions χ^S, χ_W^S and χ_B^S defined in equations (7), (9) and (8) according to parental education, number of books, parents' occupational status and

socio-economic and cultural status, respectively. The first graphs in the figures display the mean educational achievements of students in the four types and in each degree of effort. It can be observed that mean achievements are increasing in circumstances and degrees of effort. This means that the lowest mean performance is observed among students with the lowest circumstantial environment and that exert the lowest degree of effort, whereas the best mean achievements are observed among students that enjoy the best circumstances and that exert largest degree of effort. It is also observed that the four lines never cross meaning that among the students that exert the same degree of effort, and hence, belong to the same tranche, the students with more favourable circumstances are likely to obtain better average achievements.

The second graphs in the figures show the counterfactual distribution χ_W^S where the effect of the effort-group is eliminated. This fact is perceived in the four flat lines that are not increasing in the degrees of effort. The only remaining differences are within each decile due to belonging to a particular circumstance type.

The third graphs present the counterfactual distribution χ_B^S where the differences emerged from belonging to different circumstance-groups have been removed. This fact is observed in the four overlapping lines in the graph. The only inequalities that remain are found between deciles, due to exerting different degrees of effort.

Table 5: Inequality of Opportunity according to each definition of types, by gender and geographical location—Ex-post approach

	Pared		Books		HISEI		ESCS		Total
	IOp	Smooth	IOp	Smooth	IOp	Smooth	IOp	Smooth	inequality
Entire population	550.35	7358.34	1530.43	7375.90	884.78	7355.76	1100.44	7360.97	7624.97
Geography									
North	485.88	7226.99	1452.12	7236.97	866.24	7220.085	1072.57	7227.06	7474.02
South	579.34	7061.32	1540.51	7085.62	867.81	7064.044	1077.02	7065.86	7343.97
Gender									
Girls	541.01	6568.59	1475.56	6590.59	899.27	6580.93	1130.60	6575.34	6807.12
Boys	558.77	8003.85	1576.34	8020.43	867.57	7984.944	1067.42	8003.18	8294.34
Geog+Gender									
North-Girls	477.12	6530.36	1401.04	6554.21	889.31	6537.96	1102.47	6538.54	6745.13
North-Boys	493.49	7786.69	1495.95	7784.01	841.56	7765.11	1043.26	7779.81	8061.68
South-Girls	572.59	6170.62	1486.96	6190.88	876.24	6187.78	1107.60	6176.14	6429.02
South-Boys	582.28	7805.56	1574.85	7841.65	854.07	7790.688	1034.00	7809.55	8105.23

Pared,Books, HISEI,ESCS=types defined according to parental education, number of books, highest parental occupation and the index of Socio Economic and Cultural Status—; $Smooth = var(\chi^S)$—; Total inequality= $var(X)$.

The indices of inequality of opportunity are presented in Table 5. The results are provided for the entire country and for each population sub-group. The column blocks are associated to each variable used to define types and they are composed by two indices, the first indicates the inequality of opportunity and the second the inequality that remains after the residual inequality within cells is eliminated, i.e.,inequality in the smoothed distribution ($var(\chi^S)$). The last column shows the overall achievement inequality.

The value of inequality of opportunity varies according to the variables used to construct the types. The smallest value is found between types defined according to parental education and it is followed by types built upon HISEI, ESCS and the number of books at home, subsequently. This means that the distributions of scores for types defined according to parental education are closer to each other and that the distributions that are more distant from each other are those defined upon the number of books at home. Accordingly the

magnitudes of unfair inequalities for the entire country vary from 550.35 to 1530.43 variance points. This last definition provides values about 1000 variance points higher than the first definition. In fact, the share of unfair inequalities explained by differentials in parental education takes 7.48% from the overall inequalities (after eliminating the residual inequalities within cells) whereas the share explained by differentials in the number of books takes 20.75%.

According to the geographic location, in the same way as in the overall inequalities, the Southern regions exhibit slightly larger values of inequality of opportunity than the Northern ones; however the difference is not very substantial. Indeed, the indices of overall inequalities are more discrepant between the macro-regions than the indices of inequalities of opportunities are.

When it comes to gender, the conclusions are more contrasting across definitions of types. When the types are defined according to the PISA indices `HISEI` and `ESCS`, girls suffer larger inequalities between types than boys do. Hence, despite the fact that boys exhibit larger inequalities in achievements than girls, these girls suffer larger inequalities of opportunities to obtain the same educational achievements. Nevertheless, when the types are built upon parental education and number of books, boys show larger inequalities of opportunities than girls do, just as in the case of overall inequality.

When it comes to the four sub-populations, the rankings of the unfair inequalities vary according to the variables used to define types. According to parental education, the largest inequality of opportunity is suffered by students in the Southern regions and the lowest by those in the Northern regions. Then, within both macro-regions, boys show larger inequalities between types than girls do.

With regard to the types defined using the number of books at home, the boys in both macro-regions suffer larger inequality of opportunities than girls do. Similarly, boys and girls in Southern regions show larger illegitimate inequalities than those in the Northern regions. Correspondingly, the highest unfair inequalities are found among boys in the South and North, and then among girls in the South and North, respectively.

Regarding the types built upon the highest occupational status and the socio-economic and cultural status, both sub-samples of girls present larger inequality of opportunities than those of boys.

References

Ammermüller, A., Heijke, H., & Wößmann, L. (2005). Schooling quality in Eastern Europe: Educational production during transition. *Economics of Education Review*, *24*, 579–599.

Buuren, S. V., Boshuizen, H. C., & Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, *18*, 681–694.

Causa, O., & Chapuis, C. (2011). Equity in student achievement across OECD countries. *OECD Journal: Economic Studies*, *1*, 1–50.

Checchi, D., & Peragine, V. (2005). Regional disparities and inequality of opportunity: The case of Italy. *IZA Discusion Paper No.1874*, (pp. 1–32).

Checchi, D., & Peragine, V. (2010). Inequality of opportunity in Italy. *The Journal of Economic Inequality*, *8*, 429–450.

Entorf, H., & Minoiu, N. (2005). What a difference immigration policy makes: A comparison of PISA scores in Europe and traditional countries of immigration. *German Economic Review*, *6*, 355–376.

Ferreira, F. H. G., & Gignoux, J. (2011). The measurement of inequality of opportunity: Theory and an application to Latin America. *Review of Income and Wealth*, *57*, 622–657.

Ferreira, F. H. G., & Gignoux, J. (2014). The measurement of educational inequality: Achievement and opportunity. *The World Bank Economic Review*, *28*, 210–246.

Fleurbaey, M., & Peragine, V. (2013). Ex-ante versus Ex-post equality of opportunity. *Economica*, *80*, 118–130.

Gamboa, L. F., & Waltenberg, F. D. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006–2009. *Economics of Education Review*, *31*, 694–708.

Ganzeboom, H., Graaf, P. D., & Treiman, D. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*, 1–56.

Jenkins, S. P., Micklewright, J., & Schnepf, S. V. (2008). Social segregation in secondary schools: how does England compare with other countries? *Oxford Review of Education*, *34*, 21–37.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Boston: TIMSS & PIRLS International Study Center.

Nonoyama-Tarumi, Y. (2008). Cross-national estimates of the effects of family background on student achievement: A sensitivity analysis. *International Review of Education*, *54*, 57–82.

OECD (2009). *Data Analysis Manual*. Paris: OECD Publishing.

OECD (2013). *Results: Excellence Through Equity: Giving Every Student the Chance to Succeed (volume II)*. Paris: OECD Publishing.

Peterson, P., & Wößmann, L. (2007). *Schools and the Equal Opportunity Problem*. Cambridge, MA: MIT Press.

Roemer, J. E. (1998). *Equality of Opportunity*. Cambridge, MA: Harvard University Press.

Royston, P., & White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *Journal of Statistical Software*, *45*, 1–20.

Salehi-Isfahani, D., Hassine, N. B., & Assaad, R. (2014). Equality of opportunity in educational achievement in the Middle East and North Africa. *The Journal of Economic Inequality*, *12*, 489–515.

Schnepf, S. V. (2007). Immigrants' educational disadvantage: An examination across ten countries and three surveys. *Journal of Population Economics*, *20*, 527–545.

Schütz, G., Ursprung, H. W., & Wößmann, L. (2008). Education policy and equality of opportunity. *Kyklos*, *61*, 279–308.

Schütz, G., West, M. R., & Wößmann, L. (2007). *School Accountability, Autonomy, Choice, and the Equity of Student Achievement*. OECD Publishing.

Van de gaer, D., & Ramos, X. (2015). Approaches to inequality of opportunity: Principles, measures and evidence. *Journal of Economic Surveys*, *0*, 1–29.

Waltenberg, F. D., & Vandenberghe, V. (2007). What does it take to achieve equality of opportunity in education?: An empirical investigation based on Brazilian data. *Economics of Education Review*, *26*, 709–723.

Wößmann, L. (2004). How equal are educational opportunities? Family background and student achievement in Europe and the US. *CESifo Working Paper Series 1162*, (pp. 1–42).

Wößmann, L. (2010). Families, schools and primary-school learning: Evidence for Argentina and Colombia in an international perspective. *Applied Economics*, *42*, 2645–2665.

Zheng, B. (2007). Unit-consistent decomposable inequality measures. *Economica*, *74*, 97–111.

5. Appendix A

Table 6: Descriptive statistics of the educational achievements for each type defined according to parental education

Highest educational	**North**		**South**		**Total**
attainment among parents	**Girls**	**Boys**	**Girls**	**Boys**	
Primary	431.35	446.82	421.16	446.69	435.56
	7138.74	8136.57	5639.85	6225.22	6633.83
	4731	4886	9476	9101	28 194
Lower Secondary	463.87	473.84	445.76	451.01	456.54
	5887.74	7857.30	5422.05	6498.92	6431.62
	12 728	12 126	18 148	19 360	62 362
Upper Secondary	484.22	498.05	461.58	469.99	478.10
	6445.28	7570.86	5797.72	8204.38	7206.06
	24 469	25 743	26 802	27 176	104 190
Tertiary	507.67	524.79	479.68	503.91	505.94
	6252.92	7398.22	6714.98	7977.96	7309.35
	49 990	52 114	37 646	39 194	178 944
Total	491.43	507.01	461.71	477.90	484.63
	6745.13	8061.68	6429.02	8105.23	7624.97
	91 917	94 870	92 073	94 831	373 691

The *first row* presents the mean educational achievement; the *second row* presents the variance of educational achievements and the *third row* presents the number of observations.

Table 7: Descriptive statistics of educational achievements for each type defined according to the number of books at home

Number of books	**North**		**South**		**Total**
at home	**Girls**	**Boys**	**Girls**	**Boys**	
<25	425.20	440.91	410.52	427.12	425.40
	5659.14	6975.63	5214.25	6255.42	6138.01
	16 722	20 286	24 443	30 443	91 895
25-100	481.42	501.65	461.34	479.10	480.63
	5346.93	6419.78	4526.99	6268.36	5838.09
	28 519	29 256	30 651	30 427	118 853
100-200	510.11	528.30	480.56	509.82	507.06
	5215.21	6367.35	5869.12	5911.62	6112.94
	21 582	19 479	19 875	17 155	78 091
>200	530.88	548.89	513.60	535.10	533.72
	5485.91	6498.48	5396.22	8065.22	6437.63
	25 093	25 849	17 104	16 806	84 852
Total	491.43	507.01	461.71	477.90	484.63
	6745.13	8061.68	6429.02	8105.23	7624.97
	91 917	94 870	92 073	94 831	373 691

The *first row* presents the mean educational achievement; the *second row* presents the variance of educational achievements and the *third row* presents the number of observations.

Table 8: Descriptive statistics of the educational achievements for each type defined according to highest parental occupation (quartiles of HISEI)

Highest occupational	**North**		**South**		c **Total**
status among parents	**Girls**	**Boys**	**Girls**	**Boys**	
1^{st} **quartile**	455.62	468.19	434.94	446.20	449.26
	6981.14	7666.09	5893.39	6631.04	6835.81
	22 752	22 406	32 825	32 100	110 083
2^{nd} **quartile**	472.94	492.14	449.07	461.96	469.02
	5813.81	8204.52	5820.09	7223.82	7016.15
	20 704	20 377	19 835	22 140	83 056
3^{rd} **quartile**	506.02	514.26	482.23	505.26	502.51
	5832.30	6985.79	5235.03	7427.03	6513.38
	22 624	24 594	20 892	20 186	88 296
4^{th} **quartile**	525.03	543.17	499.52	517.97	523.75
	5370.96	6501.89	6055.18	7877.83	6633.46
	25 836	27 492	18 521	20 406	92 255
Total	491.43	507.01	461.71	477.90	484.63
	6745.13	8061.68	6429.02	8105.23	7624.97
	91 917	94 870	92 073	94 831	373 691

The *first row* presents the mean educational achievement; the *second row* presents the variance of educational achievements and the *third row* presents the number of observations.

Table 9: Descriptive statistics of the educational achievements for each type constructed according to Socio-Economic and Cultural background of the family (quartiles of ESCS)

Socio-Economic and	**North**		**South**		**Total**
Cultural Status	**Girls**	**Boys**	**Girls**	**Boys**	
$1^{st}quartile$	447.27	459.49	429.34	441.28	442.38
	6560.69	8079.18	5507.56	6529.04	6635.01
	18 811	18 156	28 468	28 195	93 630
$2^{nd}quartile$	471.95	488.45	454.69	463.69	469.68
	6013.46	7213.40	5713.94	7234.64	6710.92
	22 776	23 389	22 861	24 699	93 724
$3^{rd}quartile$	500.23	514.15	472.20	488.08	494.50
	5497.20	7049.01	5579.72	7391.98	6702.71
	23 424	25 273	21 656	23 112	93 465
$4^{th}quartile$	531.15	546.80	506.48	538.86	532.37
	5256.62	6299.62	5857.71	6395.62	6074.87
	26 906	28 052	19 087	18 825	92 871
Total	491.43	507.01	461.71	477.90	484.63
	6745.13	8061.68	6429.02	8105.23	373 690.70
	91 917	94 870	92 073	94 831	7625

The *first row* presents the mean educational achievement; the *second row* presents the variance of educational achievements and the *third row* presents the number of observations.

6. Appendix B

Table 10: Mean of educational achievements and number of observations by circumstances and effort groups in the population sub-groups generated according to geographic location and gender

Region→	**North**								**South**							
Gender→	**Girls**				**Boys**				**Girls**				**Boys**			
Circumstances→ Effort↓	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	298.09 603	322.78 1093	330.09 2056	351.96 3923	281.22 514	323.50 1328	334.62 1933	357.78 3880	302.32 1140	315.76 1733	329.09 2733	347.30 5756	276.42 570	315.57 2137	323.18 3702	346.60 4337
2	344.60 339	373.80 1090	386.64 1787	415.01 4870	354.45 372	371.16 808	386.87 2200	415.23 4631	355.16 1284	371.72 2160	388.68 3855	417.67 4734	355.59 871	371.52 2146	389.07 2656	414.56 3799
3	382.42 595	400.42 964	417.92 2383	447.41 4563	384.13 325	399.30 987	418.84 1854	448.43 4389	382.12 1267	399.75 2032	419.04 2794	448.63 4441	385.39 608	398.97 2234	416.90 3317	448.38 4393
4	405.70 543	425.74 1438	444.99 2476	474.53 5600	406.11 602	425.42 1013	445.53 2344	474.58 3643	406.48 927	424.36 2453	444.20 3428	473.13 4750	408.57 724	427.29 1379	443.86 2334	473.53 3946
5	424.97 447	448.43 1167	468.47 2482	497.52 5329	424.72 410	447.18 1129	468.57 2117	498.66 4645	425.62 982	448.11 1782	470.12 3208	497.67 4111	424.55 1093	446.39 2112	468.65 2489	497.33 3788
6	442.60 546	467.23 1395	491.55 2853	519.66 5705	445.10 366	466.73 648	491.65 2446	520.38 4933	442.27 780	465.51 1907	491.12 2692	520.22 3363	443.49 1026	467.37 2293	491.38 2396	519.64 3999
7	465.38 269	486.24 1313	514.09 2714	542.19 5595	462.88 378	487.01 1372	513.14 2941	542.08 5699	459.91 865	486.26 1595	513.64 2320	542.72 3604	464.35 1322	485.39 1953	515.74 2477	543.41 2969
8	489.65 489	510.13 1703	536.88 2579	567.07 5512	487.72 602	512.03 1270	538.50 2596	568.48 6079	488.32 778	510.02 1715	538.50 2757	566.31 2600	488.08 953	512.78 1553	536.78 2450	566.20 3628
9	522.72 444	541.86 1273	567.51 2909	596.55 4889	519.10 612	543.99 1407	570.87 3065	599.56 6420	519.38 831	539.31 1738	567.11 1824	597.64 2829	519.52 967	540.23 1824	566.74 2648	597.02 3822
10	587.58 457	598.82 1291	621.13 2230	645.80 4004	587.55 704	600.67 2165	624.22 4247	649.54 7795	573.68 623	592.68 1034	620.67 1191	636.30 1458	586.70 968	598.62 1729	623.05 2707	647.53 4514

Four sub-populations basing on the geographic location and gender (girls in the North, boys in the North, girls in the South and boys in the South); circumstances= circumstances-groups (types) basing on parental education, $C_i, \forall i = \{1, ..., 4\}$, where 1 corresponds to "Primary or lower education", 2 to "lower secondary", 3 to "upper secondary" and 4 to "tertiary"; Efforts= the degree of effort exerted by students, $E_j, \forall j = \{1, ..., 10\}$, where j corresponds to the decile of outcome distribution of the circumstances-groups.

Data is provided for each cell ij (which belongs to students with circumstances i and effort-degree j). The values in the first and the second rows of the cells indicate the mean educational achievement and the number of observations in each cell, respectively.

Table 11: Mean of educational achievements and number of observations by circumstances and effort groups in the population sub-groups generated according to geographic location and gender

Region→	**North**								**South**							
Gender→	**Girls**				**Boys**				**Girls**				**Boys**			
Circumstances→ Effort↓	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	298.69	347.02	371.53	386.98	290.77	351.88	372.74	387.72	295.61	352.43	362.39	386.03	283.23	338.92	361.85	375.54
	1866	2597	1786	2043	1657	2327	1384	2295	3034	3779	3183	1912	2732	3265	1480	2295
2	345.04	401.58	424.98	448.90	345.24	399.70	425.92	450.03	341.87	399.09	425.50	448.06	344.19	403.17	425.01	451.36
	1561	2629	1945	2672	1643	2088	1261	2004	2794	4065	2801	2458	3138	3046	1799	1305
3	371.75	427.90	454.54	482.61	372.62	428.35	453.10	482.07	370.64	427.78	454.90	478.31	372.47	428.68	456.17	482.34
	1518	2948	1948	3118	1651	2296	1869	1821	3235	3828	2283	2385	2759	2814	1690	1175
4	392.14	449.95	477.47	506.71	393.70	449.26	478.67	507.04	394.32	450.09	475.89	504.96	394.66	450.17	479.20	507.28
	1548	2746	2165	2844	2121	2312	1649	2130	2574	3762	2512	2022	2937	3096	1649	1543
5	414.29	470.42	499.63	527.64	412.80	471.81	499.07	526.88	415.38	471.05	497.59	526.95	413.54	469.63	497.14	525.52
	1560	3054	2377	2462	1779	2187	1732	2323	2686	3188	1860	1597	3184	3500	1734	2123
6	434.43	491.04	519.09	547.29	434.25	491.51	518.73	546.94	434.22	491.55	517.52	548.16	435.52	491.95	519.02	547.60
	1890	3143	2725	2579	1931	2568	1702	2426	2193	3082	1882	2135	3201	3024	1504	1258
7	455.02	511.83	540.99	568.16	454.65	511.91	540.70	569.24	454.46	511.65	540.57	570.05	454.58	512.62	540.57	567.95
	1579	2618	2363	2797	2028	3530	2034	2845	2394	2837	1640	1445	3155	2912	1708	1542
8	477.87	531.61	562.18	591.75	477.91	532.63	563.93	592.78	476.46	534.03	562.49	594.18	476.06	532.16	561.93	594.13
	1888	3314	2328	2603	1672	3051	2064	2982	1932	2835	1407	1224	3735	2645	2024	1529
9	506.41	561.58	589.63	618.78	506.77	560.67	591.00	618.45	507.13	558.30	590.15	619.55	508.99	559.91	592.23	618.94
	1837	2980	2011	2139	2495	3851	2278	3414	2191	2240	1387	1345	2629	2813	2156	1609
10	558.91	610.69	633.37	665.58	570.07	617.33	640.16	662.16	553.60	601.89	625.86	666.35	572.19	614.78	640.99	662.39
	1475	2489	1934	1836	3310	5048	3506	3607	1410	1035	920	580	2972	3312	1412	2428

Four sub-populations basing on the geographic location and gender (girls in the North, boys in the North, girls in the South and boys in the South); circumstances= circumstances-groups (types) basing on number of books at home, $C_i, \forall i = \{1, ..., 4\}$, where 1 corresponds to "less than 25 books", 2 to "25-100 books", 3 to "101-200 books" and 4 to "more than 200 books"; Efforts= the degree of effort exerted by students, $E_j, \forall j = \{1, ..., 10\}$, where j corresponds to the decile of outcome distribution of the circumstances-groups.

Data is provided for each cell ij (which belongs to students with circumstances i and effort-degree j). The values in the first and the second rows of the cells indicate the mean educational achievement and the number of observations in each cell, respectively.

Source: Authors' analysis based on data from PISA 2012.

Table 12: Mean of educational achievements and number of observations by circumstances and effort groups in the population sub-groups generated according to geographic location and gender

Region→	**North**								**South**							
Gender→	**Girls**				**Boys**				**Girls**				**Boys**			
Circumstances→ Effort↓	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	310.73 2261	324.58 1587	361.21 2004	379.02 1804	310.46 1751	322.45 1549	359.71 2338	380.62 2250	308.54 3784	324.36 2475	356.44 2396	368.70 2569	305.05 3214	319.42 2711	354.73 2133	363.45 2667
2	362.07 1841	380.06 1731	417.84 2168	437.62 2692	365.53 1908	381.96 1927	420.57 2162	440.83 1994	365.17 4155	381.43 2586	420.08 2465	437.31 2572	362.61 3281	381.59 2055	418.44 2009	435.66 1953
3	391.18 2070	411.18 1703	448.15 1737	471.52 2688	393.64 1686	410.97 1420	448.33 2126	470.39 1853	392.10 3930	410.62 2566	447.26 2680	471.39 2759	390.92 3306	412.10 2649	447.60 2315	468.84 1904
4	415.15 2334	435.85 2370	471.30 2084	495.16 3071	414.49 1988	437.25 1685	471.57 1544	496.93 2371	415.00 3451	436.18 2178	470.43 3236	494.79 1869	414.08 3110	438.12 2049	468.55 1973	494.96 1944
5	436.86 2081	459.92 2528	493.07 2505	518.51 2971	437.19 2076	457.41 1615	493.16 2393	517.98 2494	437.04 3708	457.07 1938	493.19 2488	517.92 1518	437.87 3210	460.86 2223	494.18 1458	517.31 2219
6	458.15 2208	480.49 2170	514.01 2476	538.47 2753	457.69 2034	480.67 1425	514.56 2637	538.87 2749	456.79 3124	481.74 1950	513.71 1948	538.90 1687	457.64 3551	482.12 2747	516.32 1711	536.87 2034
7	480.98 2302	500.49 2613	535.66 2694	560.27 2731	480.96 2354	501.36 1922	536.56 2570	559.46 2782	479.05 2809	502.55 2108	536.40 2055	557.52 1882	479.32 3578	500.46 1708	535.90 1541	559.78 1819
8	506.89 2437	525.64 2153	558.35 2783	582.53 2794	506.85 2524	525.09 2108	560.15 2435	582.35 3148	505.61 2878	525.08 1788	558.33 1268	581.21 1658	506.45 3158	525.03 2296	558.28 2371	584.01 1619
9	538.31 2750	556.01 1977	588.59 2174	609.08 2313	537.23 2418	555.16 2925	589.49 2998	610.10 3581	537.75 3244	551.06 1452	588.94 1509	609.73 1446	536.81 2626	558.67 1871	590.04 2086	607.15 1908
10	597.59 2468	609.54 1874	636.99 1999	656.34 2019	603.37 3667	619.44 3801	639.57 3391	655.71 4271	588.96 1742	618.12 795	634.80 846	649.34 563	593.04 3065	619.37 1830	640.09 2588	655.22 2339

Four sub-populations basing on the geographic location and gender (girls in the North, boys in the North, girls in the South and boys in the South); circumstances= circumstances-groups (types) basing on HISEI, $C_i, \forall i = \{1, \ldots, 4\}$, where 1 indicates that student belong to the "1^{st} quantile", 2 to the "2^{nd} quantile", 3 to the "3^{rd} quantile" and 4 to the "4^{th} quantile" of the overall distribution of HISEI; Efforts= the degree of effort exerted by students, $E_j, \forall j = \{1, \ldots, 10\}$, where j corresponds to the decile of outcome distribution of the circumstances-groups.
Data is provided for each cell ij (which belongs to students with circumstances i and effort-degree j). The values in the first and the second rows of the cells indicate the mean educational achievement and the number of observations in each cell, respectively.

Table 13: Mean of educational achievements and number of observations by circumstances and effort groups in the population sub-groups generated according to geographic location and gender

Region→	**North**								**South**							
Gender→	**Girls**				**Boys**				**Girls**				**Boys**			
Circumstances→ Effort↓	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	309.33	327.85	347.95	392.86	299.73	335.09	350.20	388.86	306.57	325.08	343.55	390.27	296.57	323.74	349.85	376.47
	1972	2004	1549	2370	1686	1449	1954	2376	3225	2574	2933	2988	2607	3390	2999	1577
2	358.84	382.69	408.53	450.42	356.00	384.40	409.42	450.95	360.11	383.47	406.77	451.15	357.69	384.61	409.13	452.12
	1561	1852	2172	2573	1563	2803	1897	2554	3234	2880	2454	2620	2881	1799	2736	1618
3	386.84	414.68	439.61	481.89	386.80	413.45	438.80	482.46	386.26	411.52	439.62	479.60	387.71	412.28	439.51	481.63
	1801	2702	2132	3128	1120	1743	2042	1673	3459	2492	2854	2746	3006	2469	2317	1690
4	409.40	437.46	464.88	505.85	408.45	437.54	461.94	505.35	409.36	436.93	465.26	504.72	409.87	437.23	463.81	504.64
	1571	2136	2340	2655	1710	2074	1890	2524	3274	2427	2819	2113	2795	2757	2355	1988
5	431.84	459.02	485.40	523.74	431.88	457.37	485.09	525.21	430.10	457.76	485.30	525.40	431.47	458.56	484.07	523.48
	1888	2341	2459	2827	1530	1669	2173	2557	3397	2612	2571	1564	2672	2806	2097	2343
6	450.88	481.83	506.71	544.60	451.16	480.43	507.63	545.64	451.99	479.97	507.74	546.51	450.45	481.76	507.15	547.53
	1739	2478	2815	2973	1623	2166	2652	2716	3030	2412	1939	1962	2897	2250	1958	1594
7	472.20	502.89	528.46	566.01	472.34	504.40	529.01	566.84	472.46	503.76	528.25	568.19	471.92	501.17	528.45	564.62
	2130	2499	2966	2985	1533	2576	2640	2802	2006	2216	1981	1722	3667	2071	1796	1798
8	497.32	527.01	551.38	588.46	496.11	527.94	551.26	588.52	496.69	526.94	548.91	590.19	495.45	525.75	551.31	590.23
	2201	2134	2852	2936	2226	2201	2782	3294	2713	2169	1621	1262	2194	2844	2050	1777
9	527.52	556.72	580.91	614.44	528.54	556.52	581.54	615.39	528.80	550.65	580.50	612.80	527.42	556.76	582.78	611.59
	2068	2504	2192	2449	2187	2760	3051	3589	2487	2085	1533	1494	2746	2031	2610	1754
10	591.93	604.15	633.19	661.14	594.90	615.98	634.41	660.42	579.36	610.34	626.38	653.69	587.30	615.91	634.76	657.33
	1881	2126	1947	2010	2978	3948	4192	3968	1643	994	950	616	2732	2281	2196	2686

Four sub-populations basing on the geographic location and gender (girls in the North, boys in the North, girls in the South and boys in the South); circumstances= circumstances-groups (types) basing on HISEI, $C_i, \forall i = \{1, ..., 4\}$, where 1 indicates that student belong to the "1^{st} quantile", 2 to the "2^{nd} quantile", 3 to the "3^{rd} quantile" and 4 to the "4^{th} quantile" of the overall distribution of HISEI; Efforts= the degree of effort exerted by students, $e_j, \forall j = \{1, ..., 10\}$, where j corresponds to the decile of outcome distribution of the circumstances-groups.
Data is provided for each cell ij (which belongs to students with circumstances i and effort-degree j). The values in the first and the second rows of the cells indicate the mean educational achievement and the number of observations in each cell, respectively.

7. Appendix C

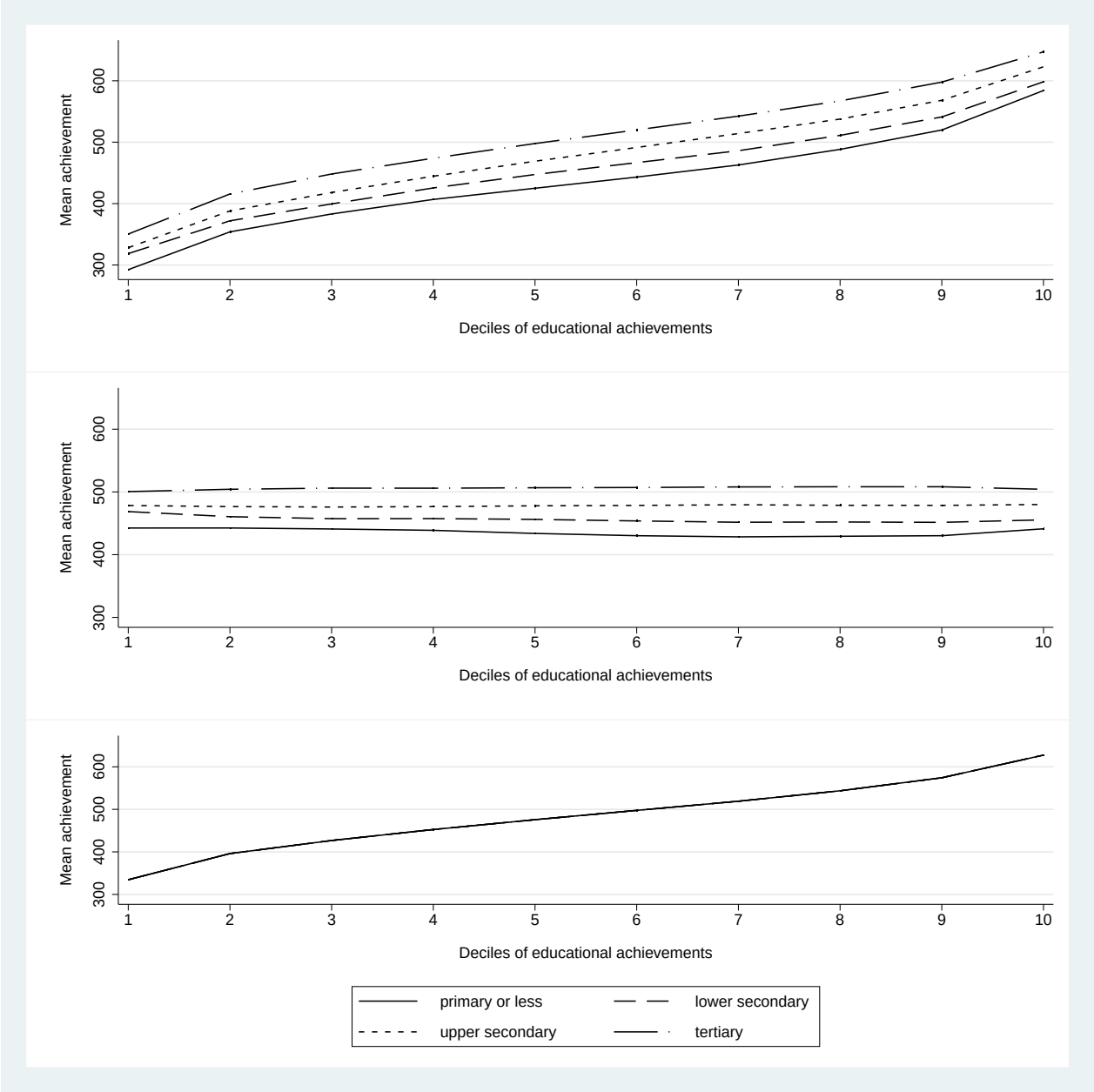

Figure 2: Distributions χ^S, χ^S_W and χ^S_B when circumstances are defined according to parental education

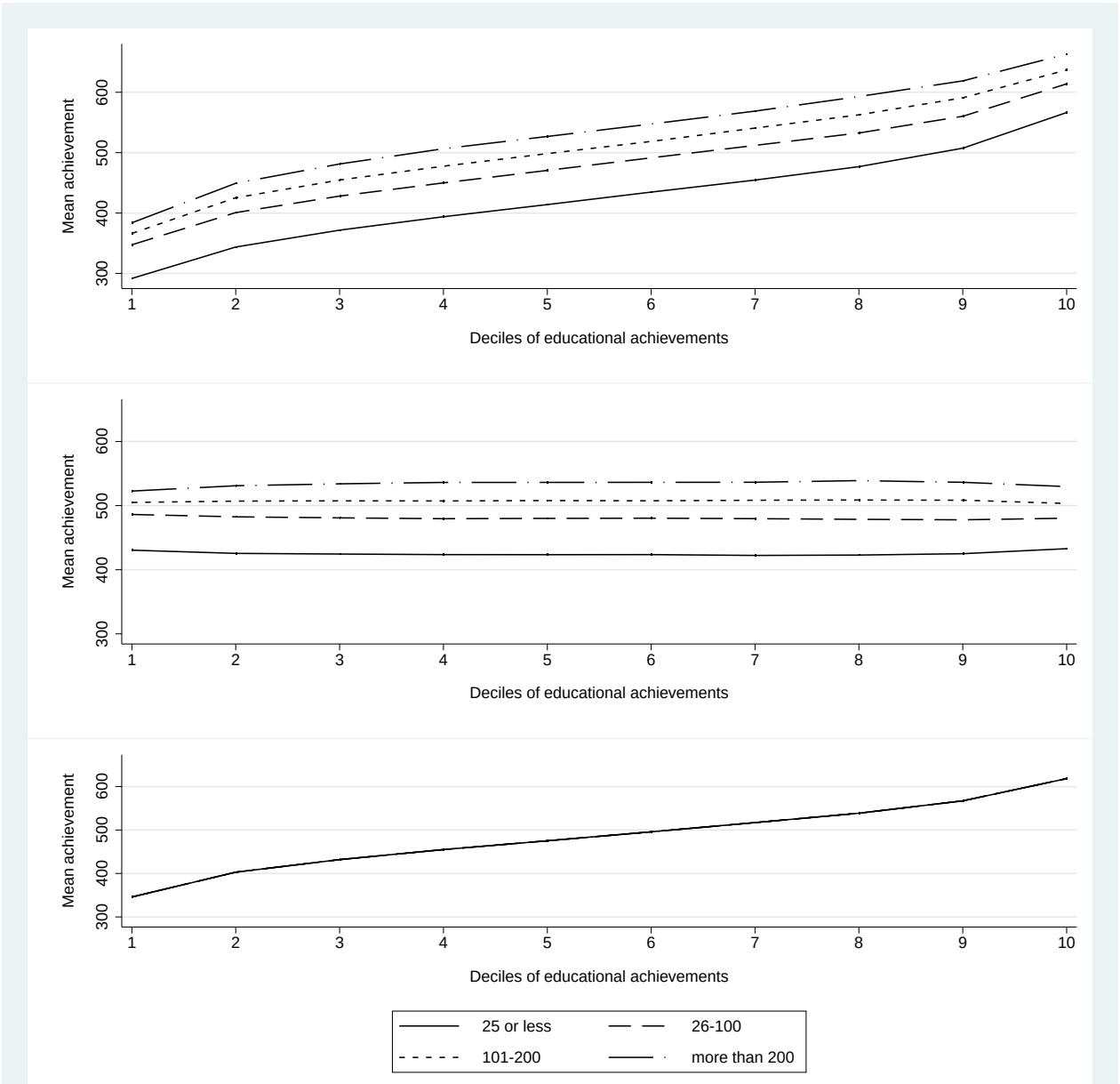

Figure 3: Distributions χ^S, χ^S_W and χ^S_B when circumstances are defined according to number of books

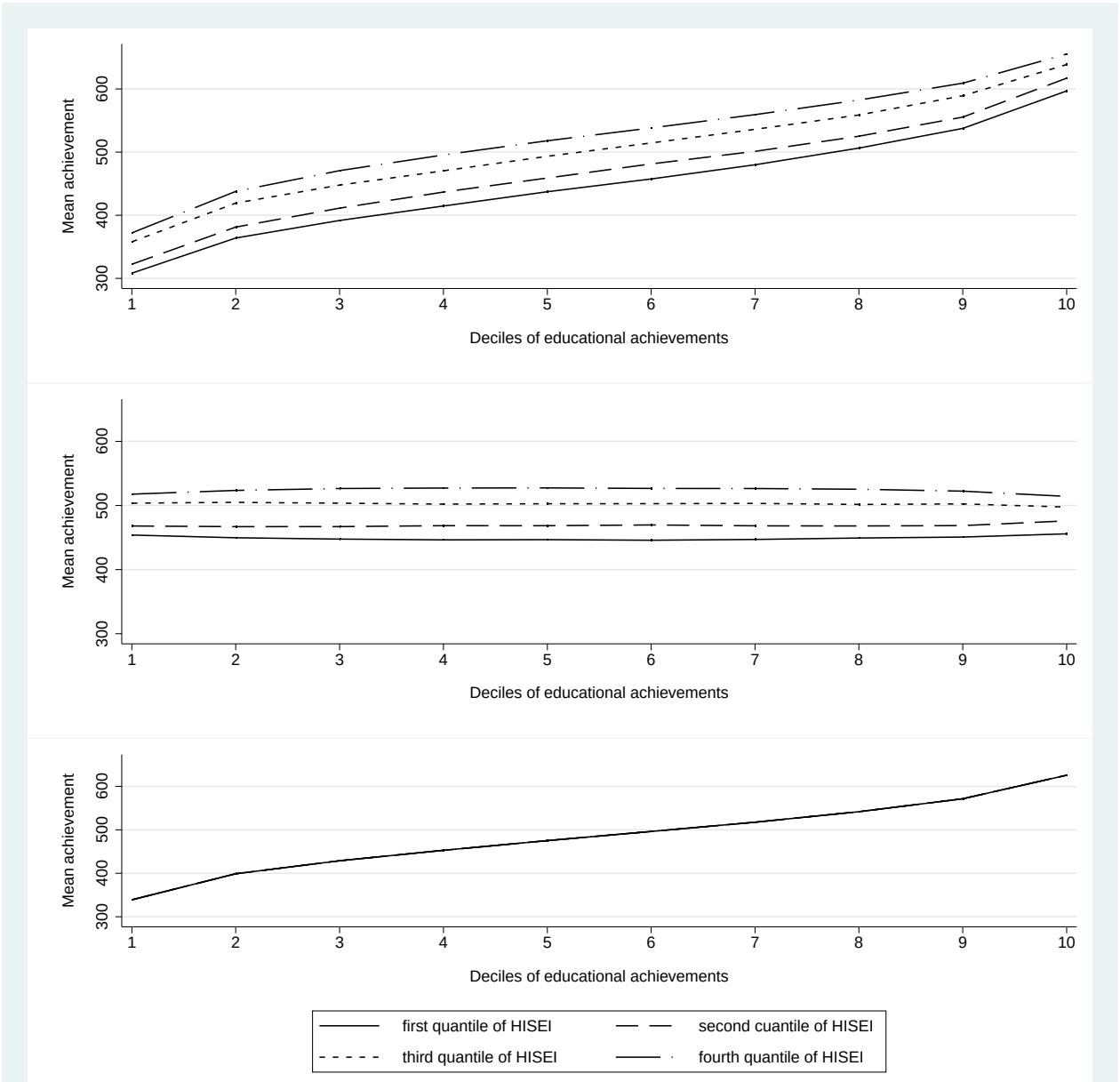

Figure 4: Distributions χ^S, χ^S_W and χ^S_B when circumstances are defined according to highest parental occupation

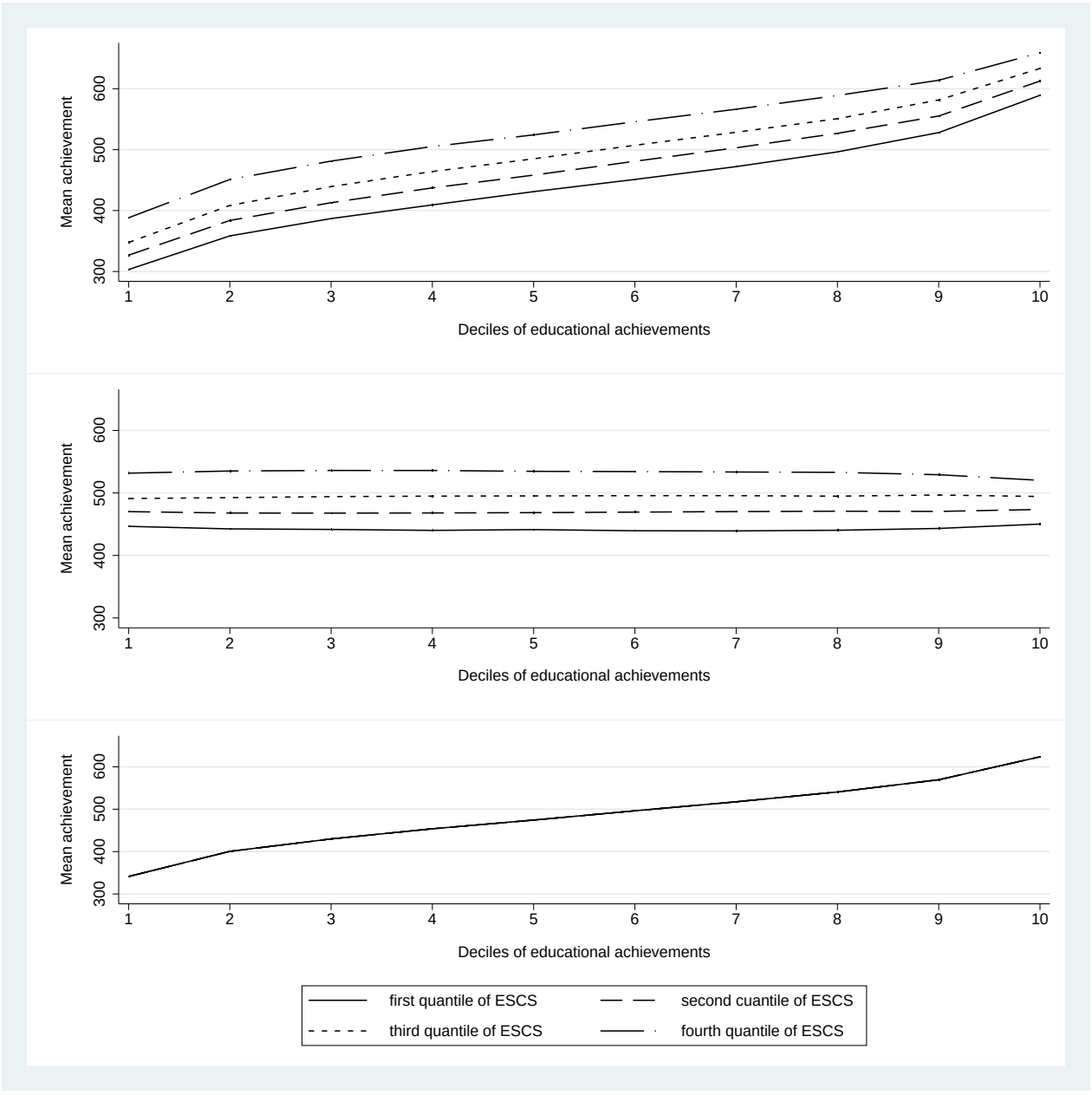

Figure 5: Distributions χ^S, χ^S_W and χ^S_B when circumstances are defined according to socio-economic and cultural status