How big is the handicap for disadvantaged pupils in segregated schooling systems?

JULIEN DANHIER

Affiliations

Group for research on Ethnic Relation, Migration and Equality (GERME)

Université libre de Bruxelles (ULB)

Email addresses

jdanhier@ulb.ac.be

Abstract

The effect of composition may be strong in systems where students are systematically sorted based on their socioeconomic background. This paper aims to model the differential effect of class composition on pupils' achievement in Belgium (French-speaking Community), France, Spain, and Portugal. Multilevel models are consequently tested on the PIRLS 2011 data (20830 pupils in 1139 classes). Our results suggest that socioeconomic composition does not have an equivalent effect on pupil achievement in the four countries included in our analysis: its effect is strong in the French-speaking Community of Belgium and France but smaller in Spain and Portugal.

Keywords: Compositional effect, Cross-country comparison, PIRLS, Multilevel modelling

Introduction

The compositional effect defined as the additional effect of aggregated characteristics modelled at once at the pupil level (Dumay & Dupriez, 2008), has been increasingly included in studies on student achievement, particularly in the wake of the development of multilevel techniques. This has proven to be essential to grasp more accurately how schools are entangled in processes of inequality reproduction: socioeconomically disadvantaged students may indeed face a double handicap because they are in classes with other students from similar socioeconomic backgrounds. This has an impact on their achievement, already lower than that of socioeconomically advantaged students. Three categories of explanations have been offered (Harker and Tymms 2004, Van Ewijk and Sleegers 2010a). The compositional effect can result from direct peer interactions (discussions, lack of motivation, disruptions or, for ethnic composition, tensions between students from different sociocultural backgrounds, or language difficulties), teacher practices (adjustments in teaching style or expectations) and school quality (problems regarding human resources management or funding). In other words, composition measurement remains of prime importance because it allows us to extend the equity discussion from the individual effect of socioeconomic background to also taking into account that the way in which students are grouped can hamper the progress of disadvantaged students.

At the same time, criticisms about required essential variables to be included in the model and statistical methods to be used to avoid biased measures have cast doubt on the existence of the concept, suggesting its significance could merely be a statistical artefact. Actually, the strong requirement that a full set of individual variables (including a measure of prior achievement) is needed to correctly model composition exclude most of the available databases. Notably, as cross-country large-scale surveys as those from the OECD (Organization for Economic Co-operation and Development) or the IEA (the International Association for the Evaluation of Educational Achievement) do not include any measure of prior achievement, their use is inappropriate to model compositional effect. As these databases offer a unique opportunity to compare countries not only in terms of performance but also in terms of equity and as compositional effect is an important component of the process of inequality reproduction, this unsuitability is unsatisfactory. The paper attempts to mobilize these types of cross-country surveys in studying the influence of composition by strongly assuming that grade repetition can be used as a proxy for prior achievement. As this practice of grade repetition is common only in some countries, this requires us to cautiously select countries and, consequently, limit the possible comparisons.

The strategy followed in this paper consists of testing the naïve expectation that the effect of composition is similar and significant in countries that are highly and similarly segregated. If composition is the way segregation has a detrimental effect on achievement, especially for the most disadvantaged students, compositional effects may be able to express themselves strongly in systems where students are systematically sorted based on their socioeconomic background. The aim of this article is to test this hypothesis exactly. After selecting a subsample of segregated countries in Western Europe (French-speaking Community of Belgium, France, Spain, and Portugal), multilevel modelling is applied in order to compare the strength of the compositional effect in these countries. This article does hence not address the way composition influences achievement but addresses the potential differential effect that composition may have in segregated countries.

The Compositional Effect

Measuring and modelling composition is not straightforward. An argument, largely developed in the literature (Gorard, 2006; Harker & Tymms, 2004; van Ewijk & Sleegers, 2010; Willms & Raudenbush, 1989), states that measuring composition requires a rich set of individual-level data, including a measure of prior achievement and other variables known to be linked to achievement, e.g., socioeconomic background, language spoken at home or ethnic origin but also some measures of non-cognitive characteristics as, for example, student motivation. This requirement refers to a phenomenon called omitted-variable bias in regressions-like statistical models. The absence of variables highly determinant for achievement and correlated with explanatory compositional variables results in the attribution of the effect of the omitted variables to the currently included compositional variables. In other words, without such a full set of individual variables, the compositional effect would be largely overestimated as van Ewijk and Sleegers (2010a, 2010b) have shown with meta-analyses. This argument is the principal methodological criticism that has emerged from scholars, arguing that the compositional effect might be a statistical artefact. Harker and Tymms (2004) showed that the effect did indeed disappear when certain important student characteristics like prior achievement were entered into the model. In other words, the compositional effect may capture what is not adequately captured by the level-one model.

Because of the conceptual and methodological issues in measuring composition, Thrupp, Lauder, and Robinson (2002) proposed a list of ten features that an ideal model should fulfil, including a full set of student variables (including prior achievement from longitudinal data and a robust measure of socioeconomic background) but also different measures of different composition types (for example, socioeconomic, academic or ethnic composition). In other words, a full set of student variables is required not only to correctly specify the individual level but also to be able to measure the effect of different compositions.

When taking these methodological requests into account and using a multilevel approach, researchers have nevertheless repeatedly observed significant negative effects of certain composition variables on students’ achievement and achievement gains in different countries with different datasets. Negative effect in this context means that a more “favourable” composition (higher socioeconomic composition, higher averaged performances, and lower proportion of migrants) is associated with higher students’ performances. Opdenakker and Van

Damme (2001) mentioned that academic and socioeconomic composition has an effect on mathematical achievement in secondary schools of the Belgian Dutch-speaking community (LOSO data), but that only the effect of academic composition was significant when both variables were entered together. On the same data, De Fraine et al. (2002) have found a significant negative effect of class composition (measured by average prior cognitive ability and socioeconomic background) on language and mathematical achievement. In the French-speaking part of Belgium, at the end of primary education, Dumay and Dupriez (2008) observed effects of academic, language and sociocultural composition. In France, Duru-Bellat et al. (2004) found a significant negative effect of socioeconomic class composition in CE1 (second grade of elementary education) while this effect is not significant in CM1 (fourth grade of elementary education). In secondary education of the Netherlands (VOCL data), Timmermans et al. (2011) have measured the effect of additive and dispersion paradigms of academic and socioeconomic composition. When all the composition measures are simultaneously modelled, only socioeconomic density has a significant negative effect on overall achievement for the students in prevocational track while no variable remains significant for the ones in general education. Using the same data, Sykes and Kuyper (2013) found a significant effect of socioeconomic composition on achievement while ethnic composition became nonsignificant when both types of composition were simultaneously modelled. In the United States, Rumberger and Palardy (2005) have found a significant effect of socioeconomic composition in high schools (NELS data), while Condron (2009) observed an effect of ethnic composition – but not of socioeconomic composition – in primary education (ECLS-K). In secondary schools in Australia, Darmawan and Keeves (2006) found a significant effect of academic composition on science achievement at the class level.

As we stated before, the requirement of a measure of prior achievement cannot be fulfilled in most of the databases including cross-country large-scale surveys. As a result, several researchers have modelled composition without including prior achievement. Doing this implies making the strong assumption that the correlation pattern among student-level variables and the omitted variable of prior achievement is sufficient to catch the information of the latter. However, using this type of approach can lead us to severely overestimate composition effects. Acknowledging this limitation, some researchers explicitly include variables highly correlated with prior achievement. Dumay and Dupriez (2007) have used background variables (socioeconomic background, language and educational expectations) to compensate for the lack of prior achievement measure. Using 2003 TIMSS eighth-grade data, they observed important net effects of composition (whose inclusion increases between-class variance explained from 19 % to 28 %) in the Dutch-speaking Community of Belgium, in the Netherlands, in England, and in the United States. Following the same logic, grade repetition and orientation are assumed to capture information about prior achievement in educational systems in which those variables largely define the pupils’ position in the system. In a study on primary schools of the Dutch-speaking Community (SIPEF project), this assumption was followed by Agirdag et al. (2013; 2011) who observed a significant effect of socioeconomic composition on mathematical achievement. With data from PISA 2009, Danhier and Martin (2014) have shown that academic and socioeconomic compositions have a negative effect on students' performances in secondary schools of the largest Belgian communities, but that the effect is different depending on the community.

Acknowledging that measuring compositional effect remains an open debate, we will use PIRLS to measure the socioeconomic compositional effect in primary education. Without any measure of prior achievement, the use of such a database to assess compositional effect remains problematic. Actually, we had to assume that delay, in combination with other background and attitudinal variables, can absorb at least a major part of the bias due to the omission of prior achievement. Such an assumption can be indirectly evaluated as the literature provides some indications regarding the amount of variance that a full set of individual characteristics is supposed to explain, although with a very large range depending on the data. Indeed, researchers have observed a reduction of variance at the student level reaching 20.6% in secondary schools (8th grade) of the Belgian Dutch-speaking community (Opdenakker & Van Damme, 2001), 23.8% in the sixth grade of Dutch education (Van der Slik, Driessen, & De Bot, 2006), 36.3% in the secondary education (9th grade) of the Netherlands (Sykes & Kuyper, 2013), 44.8% at the end of primary education (6th grade) in the French-speaking part of Belgium (Dumay & Dupriez, 2008), and 65.2% in the 4th grade of French education (Duru-Bellat, Le Bastard-Landrier, & Piquée, 2004). As our best model we will be presenting explains 24.0 % of the variance at the student level, we can assume that our student-level model is probably imperfect but acceptable.

PIRLS Data

PIRLS (standing for "Progress in International Reading Literacy Study") is a research project led by the IEA that aims to assess students' reading literacy, i.e. their "ability to understand and use those written language forms required by society and/or valued by the individual" (I. V. Mullis, Martin, Kennedy, Trong, & Sainsbury, 2009, p. 11). This large-scale survey has been conducted every five years since 2001. This article focusses on PIRLS 2011, which includes 48 countries. Following a two-stage stratified sampling design used by IEA, schools were sampled according to their size (after being separated into explicit strata and ordered by implicit strata) and (usually one or two) entire classes were randomly sampled in each selected school (Joncas & Foy, 2012). The population covered by the survey is students in the fourth grade of formal education. As educational systems differ between countries, this grade is defined as the fourth year after the beginning of level 1 of primary education schooling as defined by UNESCO's International Standard Classification of Education (ISCED) (or the next grade if the average age in the latter grade does not reach 9.5).

In this paper, we focus on a subset of countries - not only in order to limit the relevant interactions to be tested in the multilevel modelling - but also because of our modelling strategy. This selection goes through three steps. Firstly, we chose the countries in a specific geographical area, namely, the Western European countries as they share a common history but present a real diversity in terms of educational systems. Among them, England was excluded because there the home questionnaire was not administered (Mullis, Martin, Foy, & Drucker, 2012). The remaining countries or educational systems are the following: Wallonia-Brussels Federation (FWB - the French-speaking Community of Belgium as the other communities are not available in the database), Denmark, Finland, France, Germany, Ireland, Italy, Netherlands, Northern Ireland, Norway, Portugal, Spain, and Sweden. Secondly, with our naïve hypothesis stating that the effect of composition is similar and significant in highly segregated countries, we selected only countries where pupils accumulate such a school delay. Thirdly, assuming that delay, as well as other individual variables, can compensate for the lack of prior achievement, only countries where grade repetition is largely used are selected. For the last two steps, a measure of segregation and delay is necessary and will be presented in the following sections.

In the PIRLS database, both class and school identifications are available. Unfortunately, however, it is impossible to model both levels at the same time because of sampling choices. Indeed, in 39.2 % (in Norway) to 98.0 % (in Germany) of schools, only one class was sampled. When only one class per school is selected, it is impossible to correctly partition the variance in multilevel modelling between the class and the school level. The class was chosen as the cluster in the following lines.

The five plausible values (PV) for overall reading were used as dependent variables in multilevel modelling. The complete PIRLS achievement test consists of ten reading passages accompanying questions covering two reading purposes (reading for literary experience and reading to acquire and use information) and lasts for more than 6 hours (necessary to obtain a reliable measure). Because of time constraints, the passages are distributed in booklets of 40 minutes and each student fills in two. Multiple imputation[1] is then used to provide five plausible values for each student. In order to transform the battery of items into one continuous score, IRT models have been mobilized. Three- or two-parameter models were used for items with

[1] Multiple imputation is a method for handling missing data by imputing them on the basis of a model using available data and taking into account the uncertainty of the imputation. Practically, it consists in generating multiple datasets, conducting analyses on each set, and finally combining the estimators following Rubin's rule (Rubin, 1987). While the combined estimator is simply the average estimator across imputed datasets, its combined variance is the average variance plus a measure of the dispersion of the estimators from the imputed datasets. In other words, when the model produces datasets with computed estimators that are very different, the combined estimator will be less accurate and stands the risk of being non-significant.

only two response options (right/false). This type of modelling allows us to simultaneously estimate the discriminating power and the difficulty of the parameter but also to control for guessing in the case of multiple-choice items. For items with more than two options, a partial credit model was used (Martin & Mullis, 2012).

The PIRLS dataset also contains variables to measure the socioeconomic background in both the student and the parent questionnaires. A socioeconomic index (SES), namely, the Home Resources for Learning Scale, is based on the number of books and home study support (internet connection and own room) at home (retrieved from the students' questionnaire), but also the number of children's books at home, the highest level of parental education, and the highest type of parental occupation (retrieved from the parents' questionnaire). The age variable allows to measure delay provided that we can identify the age that the pupil is supposed to have in the fourth grade. The dataset also contains extra sociodemographic variables as the gender or the language spoken at home. Following Dumay and Dupriez (2007), we consider attitudinal variables that are supposed to be correlated with prior achievement to compensate for the omission of this variable. Finally, compositional variables can be computed at the class level. Although Thrupp, Lauder, and Robinson (2002) advocate using multiple measures of composition, collinearity can limit the feasibility of this advice (Danhier, 2016). We chose to only use the average SES as a measure of socioeconomic composition since this variable has a coherent effect in the literature and can be constructed on a reliable measure in this data.

However, the data contains a lot of missing values. There is indeed a lot of incomplete data linked to the parents' questionnaire in some countries, like the Netherlands and Northern Ireland. Using only the socioeconomic index, 5.0 % (in Finland) to up to 43.9 % (in the Netherlands) of students would be deleted if listwise deletion was chosen. With a proportion of missing values of over 5 %, listwise deletion would introduce considerable biases (Graham, 2009). Moreover, the consecutive reduction of the sample size would decrease the power of the analyses. Consequently, the use of multiple imputation appears to be a valid choice. The MICE R package has been used to generate datasets (van Buuren & Groothuis-Oudshoorn, 2011). All the variables used in the multilevel modelling were included in the imputation, namely, reading score, age, gender, language spoken at home and home resources for learning. Let us note that achievement variable is complete and is only used to help impute the other variables. To better impute the SES variables, composite variables that are supposed to be correlated with SES were entered into the model (as to like reading, engagement in reading lessons, motivation to read, confidence in the reading, bullying, and some variables measuring parents' investment). Finally, at the class level, we entered class size, average SES and average age. Predictive mean matching, logistic regression and a two-level linear model were used as imputation techniques depending on the scale, the level and the intraclass correlation of the variable. The highest proportion of incomplete cases (47.9 % in the Netherlands) was used to select the number of imputations (m=50) (White, Royston, & Wood, 2011). With 20 iterations, the solutions seem to be proper (plots show a healthy convergence of the Gibbs sampler and imputed data are comparable to non-imputed data). In summary, we imputed 50 datasets for each of the five plausible values for reading achievement in each country (the relation among variables not necessarily being the same), namely 3250 datasets. All the cases being included in our analyses sample sizes range from 3190 pupils in Norway to 8580 in Spain (see Mullis, Martin, Foy, & Drucker, 2012: 270).

Socioeconomically Segregated Countries

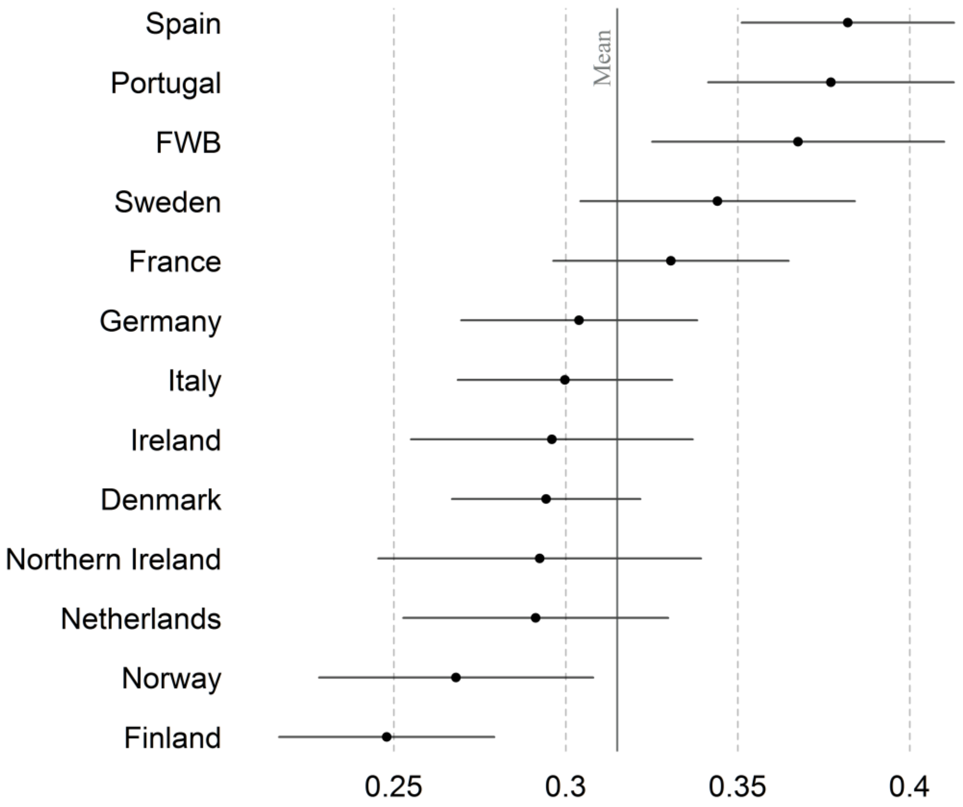
The aim of this section is to rationally select a limited number of segregated western European countries. Several numerical indexes have been developed to measure segregation. Selecting an index requires defining what segregation is (Massey & Denton, 1988) and assumes a measurement theory (see Hutchens, 2004; James & Taeuber, 1985). Consequently, in some cases, different theoretical bases will produce different rankings, in others, some indexes with different theoretical backgrounds will produce very similar results and rankings (Massey & Denton, 1988; M. J. White, 1986). In other words, some choices of indices will lead to different conclusions in terms of segregation evolution, while others will not. For this study, we define segregation as the spatial separation of students endowed with characteristics differently valued by the society (Delvaux, 2005). Here, the characteristic of interest is the socioeconomic background and the spatial separation defined by students repartition in classes. We will limit ourselves to measure the evenness dimension of segregation (Massey & Denton, 1988), for which the dissimilarity index[2] (D) has been largely used. The latter measures how far the composition of a class differs from the overall average composition. It can be interpreted as the proportion of disadvantaged students who should change classes to reach an even distribution of these students among classes (Duncan & Duncan, 1955). Strictly speaking, it is the proportion of students that have to be moved without replacement to reach an even distribution (Cortese, Falk, & Cohen, 1976). A closer look at the formula tells us that the index ranges between 0 for minimum segregation and 1 for maximum segregation (the weighted sum of the class deviations from the overall composition being divided by its maximum).

From a technical point of view, such an index requires a dichotomous variable. We arbitrarily define disadvantaged students as the 20 % of students with the lowest socioeconomic index in each country. The final student weights (computed by multiplying the base weights and the weight adjustment at the pupil, class and school level) were used in computation[3]. The

[2] $D = \sum_{j}^{m} t_j \left| p_j - P \right| / (2TP(1-P))$, where p_j and t_j are respectively the proportion of disadvantaged students in the j^{th} class and the total enrolment in this class. *P* and T are the overall aforesaid proportion and the total number of students.

[3] In order to include the weights (w_i), the elements of the dissimilarity formula become $t_j = \sum_i w_i$, $p_j = \sum_i w_{i|Disfavoured} / t_j$, $T = \sum_j t_j$ and $P = \sum_i w_{i|Disfavoured} / T$.

Figure 1: Dissimilarity indexes



standard errors (and the confidence intervals) were obtained by applying the jackknife repeated replication technique chosen by the IEA. For each pair of schools defined in the 75 sampling zones, one school has been removed while the weights of the others have been doubled. The indexes computed on each replicated sample are then combined to obtain standard errors. The index computed on multiple imputed datasets are presented in Figure 1. The analyses were replicated with listwise deletion. The results are only slightly different and reveal that multiple imputation penalizes particularly the precision of the estimation in Northern Ireland where a lot of missing values were observed.

While indices range from .25 (95% CI .22-.28) for Finland to .38 (CI .35-.41) for Spain, most confidence intervals overlap. Two scenarios can be used to select European segregated countries. The first one consists of comparing systems' segregation to the European mean and excluding systems whose segregation is significantly lower than the mean. Among our selection, three countries present a significantly higher segregation than the mean (Spain, Portugal and Wallonia-Brussels Federation) while two present a lower segregation (Norway and Finland).

The second scenario consists in comparing[4] the systems' segregation to the segregation of the system with the highest level of segregation. When we compare the indexes with Spain, the system with the highest segregation, Wallonia-Brussels Federation, Portugal, and Sweden are not significantly different at a .05 alpha level, while France is not significantly different at a .01 level. The more conservative .01 alpha level has been preferred, meaning that Wallonia-Brussels Federation, France, Portugal, Spain, and Sweden were considered as the most segregated countries among the western European countries and are candidates for the multilevel modelling.

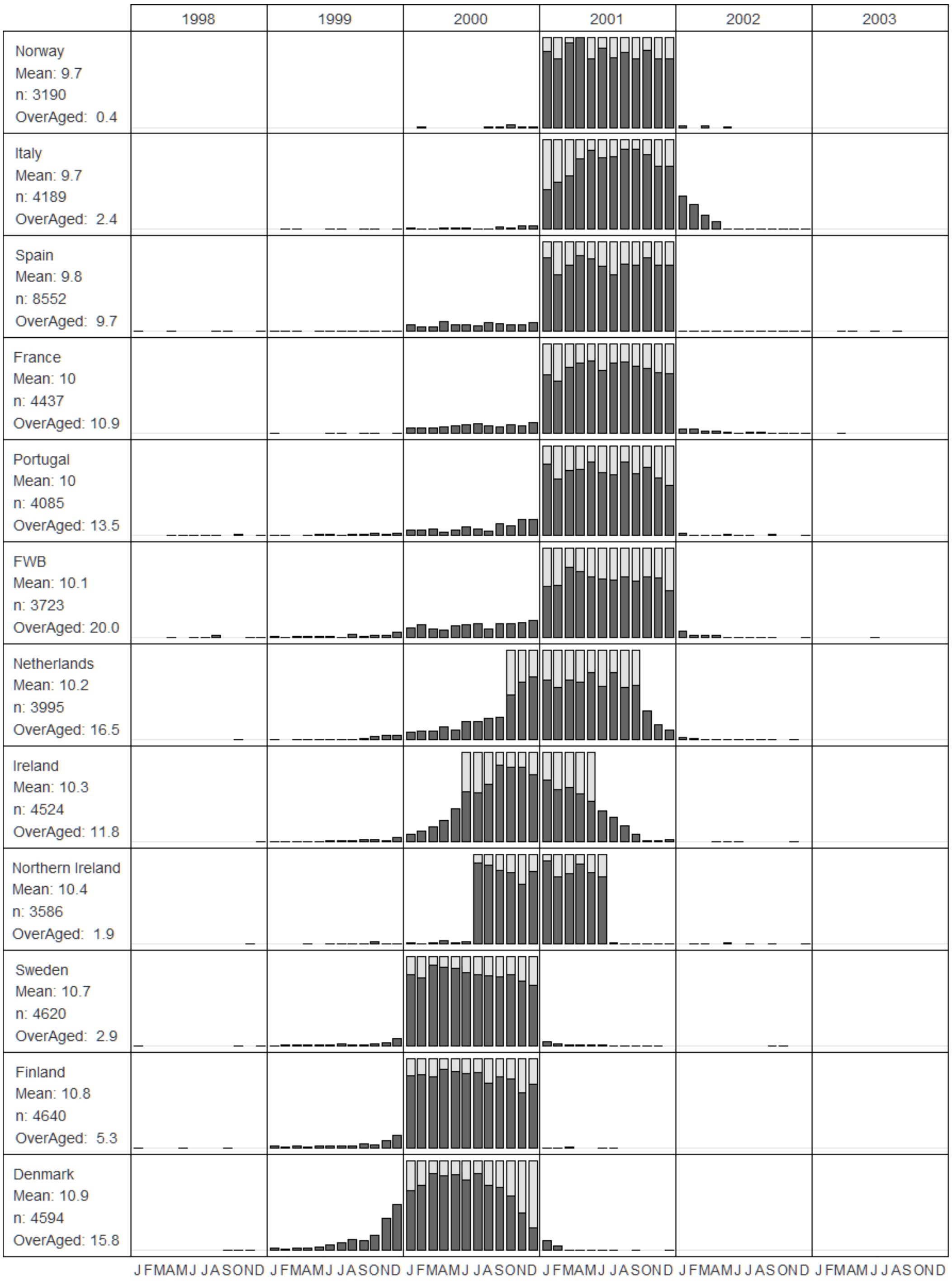
Delay at School

Modelling composition requires a measure of prior achievement. As we do not have such a measure in PIRLS, we decided to use the fact that some pupils have fallen behind in their career as a proxy for prior achievement. Strictly speaking, delay is not recorded in PIRLS. However, the age distribution of pupils can provide an interesting piece of information and is presented in Figure 2 (updated Martin, Mullis, & Foy, 2011). Because the age distribution is generally spread over more than 12 months, its comparison with the "predominant cohort" of fourth-grade pupils provides a measure of the delay that a pupil has accumulated during primary education and before. The definition of such a cohort depends on the age of admission while delay depends on rules defining student progression and their particular application (Eurydice, 2011, 2015; Martin et al., 2011; Mullis, Martin, Minnich, Drucker, & Ragan, 2012). We can distinguish a modal cohort (cohort with the most pupils) and a theoretical cohort (the position where the student should be according to the rules). Sometimes, modal and theoretical cohorts do not coincide and determining the predominant cohort becomes difficult.

The grade surveyed by IEA being the fourth one after the beginning of primary education (or the following grade if the average age in it does not reach 9.5), we need to define the entry age at that level in each educational system. In the first group of countries (Wallonia-Brussels Federation, France, Italy, Norway, Portugal, and Spain), pupils begin primary education around September of the calendar year when they reach the age of six. Some specificities are worth noting. In France, pupils can begin primary school early if they are ready and if their parents or teachers request it. In Italy, pupils can go to primary school from the age of five and half years, that is, they go to primary school if they reach the age of five before the 30th of April. Actually, in Portugal, pupils begin primary education in September if they are six years old or if they reach this age before the end of December when places are sufficient in the school. In practice, the

[4] The dissimilarity index is distributed normally for values moving away from the 0 and 1 limits (Ransom, 2000). To compare two indexes, we chose to use two-tailed student's t-tests for independent samples assuming an unequal variance.

Figure 2: Percentage of students by the month of birth (with predominant cohorts highlighted) ordered by mean age in the fourth grade



modal age of pupils in the fourth grade is similar to that of other countries of this group. In the second group of countries (Denmark, Finland, and Sweden), the pupils begin primary education around September of the calendar year when they reach the age of seven. In Denmark, pupils go to the pre-primary grade of Folkeskole when they are 6 years of age, and the first primary grade one year later. The three remaining countries present different entry rules. In the Netherlands, the pupils can begin kindergarten from the age of four but must begin it the first school day after they turn five. They generally enter primary education when they are six with the main cohort being born between October 2000 and September 2001. In Northern Ireland, the pupils enter primary education in the month of September following their fourth birthday if their birthday is before the 1st of July. Finally, in the Republic of Ireland, pupils can join the pre-primary classes (junior infants', then senior infants' classes) of primary schools from the first month of September following their fourth birthday. The age distribution is particularly wide but a predominant cohort can be identified on the basis of the modal age, namely, the 12 months with the highest number of pupils (beginning in June 2000). Let us note that, since in Germany the limit defining the entry year is a competence of the Lander, and since the database does not provide information to identify the latter, we exclude this country from our analyses.

Once the predominant cohort has been defined, pupils that are behind at school can be qualified. Delay can come both from retention in pre-primary education and grade repetition in primary education. It consists in holding the student back at the end of the school year, and making him (her) repeat the year, most of the time because he (she) does not fulfil certain academic criteria. Although in most countries, age is the only criterion for admission, in Wallonia-Brussels Federation, and especially in Denmark, maturity can justify a later entry in primary school for a significant part of pupils. In Finland and in Sweden, pupils can enter primary education later on parents' request but it concerns less than 2 % of pupils. In primary education, the profiles are varied, ranging from automatic promotion to a large use of grade repetition. When the pupils are retained, this is mainly based on their academic progression. In practice, the use of grade retention is limited in Denmark, Finland, Ireland, Italy, Northern Ireland, Norway, and Sweden. In Norway, pupils automatically progress from one grade to the next. The situation is similar in Sweden though pupils can be retained in some cases. It is worth noting the difference between law and practice regarding grade retention in Denmark, Finland, Ireland, Italy, and Northern Ireland, where grade repetition exists but is rarely used. Conversely, grade retention is largely used in Wallonia-Brussels Federation, Spain, France, the Netherlands, and Portugal. In most of these countries, the number of repetitions is limited. Pupils can be retained once by stage in Spain and France (primary education consisting in one stage), in Wallonia-Brussels Federation (primary education being separated into two stages). In Portugal, progression depends on academic progress except in the first year where it is automatic. In the Netherlands, the retention decision lies at the school level and depends on pupils' attainment.

As we can see in Figure 2, seven countries present a high rate of overaged pupils, namely pupils lagging behind the predominant cohort: Wallonia-Brussels Federation, France, Spain, the Netherlands, Portugal, Ireland, and Denmark. Let us note that in Denmark, although grade retention is rarely used in primary education, pre-primary retention is largely used, especially for younger pupils. As a consequence, the modal cohort does not match the theoretical one. However, the latter is used to compute the high rate of overaged students. Finally, the approach seems to be less relevant for Ireland where both modal and theoretical cohorts greatly exceed one year and provide little information to identify the predominant cohort.

Multilevel Modelling of the Compositional Effect

Once we have identified four segregated countries where the delay is "sufficiently" widespread to model composition (Wallonia-Brussels Federation, France, Spain, and Portugal), we can apply multilevel modelling. The sample size reaches 20830 pupils in 1139 classes (3727 pupils in 218 classes in Wallonia-Brussels Federation, 4438 pupils in 276 classes in France, 8580 pupils in 403 classes in Spain and 4085 pupils in 242 classes in Portugal).

Multilevel modelling is a technique used to examine hierarchical data. The PIRLS data are hierarchical, not only because educational data are typically hierarchical (students are clustered in classes), but also because of the two-stage sampling design. Consequently, students in the same class are likely to be more similar to each other than to students from other classes. Since the assumed independence of observations does not hold, standard statistical tests lead to a strong underestimation of the parameters' standard errors and consequently to discover spurious significant effects (Hox, 2010). Multilevel techniques are not the only way to deal with these kinds of datasets but allow modelling the effect of variables at different levels. Such a feature is useful to test the compositional effect. Once a full set of student variables (centred around the grand mean[5]) are included in the model, the additional effect of composition is simply measured by the effect of the aggregation of individual variables. At the technical level, MLwiN (Rasbash, Steel, Brown, & Goldstein, 2012) was used to perform the multilevel analyses in R (inspired from Zhang, Charlton, Parker, Leckie, & Brown, 2012). MLwiN provides sandwich estimators and performs weighted multilevel analyses using the IGLS algorithm. We chose to model students as the first level and classes as the second one.

The strategy followed in this article aims to test whether the effect of socioeconomic composition is different between a limited number of segregated educational systems. In this case, it is clearly problematic to consider the country as a third level. With only 30 groups, Maas and Hox (2005) did find that the confidence intervals were clearly too small for the random slope and the variance at the group level. In other words, significant group effects could be found mistakenly. Another approach is preferred here: countries are entered as a class-level dummy, and interactions enable us to test if a variable has a different effect in the four education

[5] With grand mean centring, the level-one coefficient is a blend of intra- and interschool relations that cannot be disentangled. At first sight this may seem problematic, but this feature is an advantage if one wished to test whether the compositional effect is significant, i.e. whether the composition has an additional effect. Due to the correlation between the level-one variables and their compositional effect, the coefficients of the latter can be viewed as partial regression coefficients. That is to say, it measures the effect of a composition variable when the level-one variable and its (unequal) repartition are taken into account. In other words, the coefficient is equal to 0 if composition doesn't explain any extra variance (Enders & Tofighi, 2007).

Table 1: Descriptive statistics

Variables	*Mean*	*Min.*	*Max.*	*S.D.*	*Skew.*	*Kurt.*
Dependent variables						
Reading (1st plausible value)	519.71	278.8	750.28	67.80	-0.23	-0.308
Reading (2d plausible value)	518.46	263.38	766.57	68.38	-0.25	-0.256
Reading (3rd plausible value)	518.98	247.5	744.34	68.21	-0.26	-0.253
Reading (4th plausible value)	518.53	234.36	740.28	68.11	-0.24	-0.283
Reading (5th plausible value)	518.67	283.1	734.36	67.90	-0.23	-0.300
Student level						
Delay (ref. on time)	0.12	0	1			
Home Resources for Learning	0.00	-8.05	8.26	1.79	0.15	0.150
Language at home (ref. same)	0.25	0	1			
Gender (ref. female)	0.51	0	1			
Student like reading	0.00	-8.43	8.21	2.00	0.32	0.506
Students motivated to read	0.00	-7.61	2.31	1.94	-0.16	-1.092
Students confident in their reading	0.00	-7.84	4.55	1.82	0.69	0.455
School level						
Socioeconomic composition	0.00	-3.76	3.45	1.13	-0.11	-0.084

systems. However, this type of strategy can become costly when the number of composition variables or the number of educational systems increases. Five models will be successively presented. In order to assess the difference in terms of reading performance between the four countries, three dummy variables are entered into the analysis (Model 1). Pupil characteristics are added in Model 2 and socioeconomic composition in Model 3. Finally, in order to investigate whether the socioeconomic composition has a different effect in each country, we added three interaction terms to our analysis. Descriptive statistics are available in Table 1 and the multilevel models can be found in Table 2.

Let us note that the PIRLS database is provided with a set of sampling weights and adjustments in order to deal with the over- and undersampling of some strata of the population and to adjust for different patterns of nonresponse (Martin & Mullis, 2012). Rescaled conditional level-one weights and rescaled level-two weights were used in multilevel modelling. The literature emphasizes that a proper use of weight needs some scaling of the conditional level-one weights (Asparouhov, 2006; Pfeffermann, Skinner, Holmes, Goldstein, & Rasbash, 1998). The second method proposed by the authors has to be preferred when the analyst is interested in point estimates (Carle, 2009). In this case, it amounted to using a constant level-one weight (due to PPS sampling) since the non-response adjustment has been dropped off by the rescaling. Most of the variation lies at the class level (class weights being computed by multiplying the base weights and the weight adjustment of the classes and the schools) that we rescaled in order for their sum to reach 250 in each country, which balances the influences of classes from each country.

Table 2: Results of multilevel modelling on reading performance (Standard errors given in brackets)

Parameters	***Model 1***	***Model 2^a^***	***Model 3^a^***	***Model 4^a^***
		Fixed part		
Intercept	505.57 (1.62) ***	508.17 (1.35) ***	505.99 (1.28) ***	505.04 (1.33) ***
France	16.30 (1.90) ***	13.36 (1.62) ***	13.59 (1.52) ***	13.07 (1.57) ***
Portugal	34.91 (1.88) ***	34.55 (1.58) ***	40.34 (1.56) ***	38.64 (1.55) ***
Spain	7.36 (1.90) ***	7.16 (1.66) ***	10.73 (1.61) ***	11.55 (1.62) ***
Home Resources for Learning		8.48 (0.58) ***	7.40 (0.59) ***	7.41 (0.59) ***
Language at home (ref. same)		-9.76 (1.06) ***	-9.80 (1.06) ***	-9.66 (1.05) ***
Delay (ref. on time)		-31.21 (1.39) ***	-30.62 (1.39) ***	-30.49 (1.39) ***
…				
Socioeconomic composition			9.3 (0.96) ***	12.65 (1.29) ***
Composition*France				2.95 (1.53)
Composition*Portugal				-9.09 (1.49) ***
Composition*Spain				-4.53 (1.54) **
		Random part		
Level-one variance (σ^2)	3527.6 (6.47) ***	2680.6 (5.93) ***	2683.6 (5.96) ***	2683.4 (5.96) ***
Level-two variance (τ_{00})	730.0 (6.93) ***	332.1 (5.28) ***	254.3 (4.91) ***	238.1 (4.71) ***
		Goodness of fit		
Deviance	231274	225125	224944	224894
AIC	231286	225151	224972	224928
BIC	231317	225216	225042	225014
Level-one R^2	0.0	24.0	23.9	23.9
Level-two R^2	14.8	61.3	70.3	72.2

Significance for Wald test: *** = .001, ** = .01, * = .05, non-significant = ‘-‘

[a] Gender, Student like reading scale, students motivated to read scale and students confident in their reading scale are modelled as control variables.

Results

The intercept-only model is used to compute the intraclass correlation (ICC) and observe the way the variance is distributed at both levels. Since the variance reaches 3527.3 (SE 6.5) for the first level and 730.0 (SE 6.9) for the second, the ICC is .172, which means that about 17 % of the variance occurred at the class level. Such value justifies the use of multilevel modelling.

In order to assess the difference in terms of reading performance between the four countries, three dummy variables are entered into the analysis (Model 1). While the intercept gives the average achievement score for a Belgian pupil (505.6, CI 502.4;508.8), the regression coefficients represent the increase in reading proficiency that is associated with a one-unit increase in the given predictor, controlling for other variables included in the model. Here, the significant coefficient indicates that, compared with a Belgian pupil, an “average” pupil studying in France, Spain or Portugal, presents higher achievement scores. A gross estimation of this gap is 7.4 points (CI 3.6; 11.1) in Spain, 16.3 points (CI 12.6; 20.0) in France, and 34.6 points

(CI 30.9; 38.3) in Portugal, what is equivalent to the classical ranking of countries by reading achievement (Mullis et al., 2012). By modelling the four countries simultaneously, we can observe the proportion of the variance that is explained by the country membership and put its extent in perspective. The 14.8 level-two pseudo-R^2 indicates that the model significantly reduces the variance at the class level. This is equivalent to a 2.9% reduction of the total variance. In other words, a major part of the variation remains for us to explain.

Pupils' characteristics are added in Model 2. In comparison to the first model, this model represents an improvement[6]. All variables have been grand mean centred. Using such a centring method changes the meaning of the 508.2 intercept (CI 505.5; 510.8) which has become an averaged "adjusted mean". The intercept in a specific class can be viewed as the 'adjusted mean' for this class, namely the mean when the effects of all the explanatory variables have been removed. In other words, it is the expected score in reading for an "average" pupil (a pupil with a mean score on all the independent variables). In the case of dummies like gender, the intercept for a specific class can be considered as the mean of this class if the proportion of boys and girls were equal across classes (Enders & Tofighi, 2007).

Let us turn to the regression coefficients. Only the coefficients for home resources for learning, language spoken at home, and delay are displayed in Table 2 even though gender, students like reading scale, students motivated to read scale, and students confident in their reading scale were modelled as control variables. Being from a disadvantaged socioeconomic background, not speaking the school language at home, and having repeated a grade are associated with a weaker performance in reading. The coefficients of the country dummies have moved a bit because the population differs slightly between countries. However, all differences remain significant.

The evolution of the random part is interesting. The 24.0 level-one pseudo-R^2 of the model indicates that the model significantly reduces the variance at the pupil level. In preliminary analyses, only SES, language spoken at home, and delay were considered and produced a 11.6 level-one pseudo-R^2. Introducing attitudinal scales as control variables improved the pupil level model. Although this value is in the range of values observed in the literature for models with prior achievement (from 20 to 60%), we expected a higher value. With this relatively low value, we cannot exclude the presence of biases in the coefficient of the compositional effect. Pupil characteristics also play a role at the class level: with a pseudo-R^2 equal to 61.3, they

[6] Changes in deviance, AIC and BIC have been systematically observed. The averaged deviance is used as a combination of the analyses conducted on the different plausible values. Deviance difference has to exceed a Chi-square distribution with the number of extra parameters as degree of freedom. AIC = Deviance + 2p and BIC = Deviance + ln(n)*p. In multilevel modelling, it is not clear which population size should be used. We chose to use the smallest population, namely the level-two population, in order to limit differences between the indices. Differences in AIC and BIC are significant if they exceed at least 2 units by extra parameter. The residuals were also screened.

sharply reduce the variance of this level. In other words, an important part of the variation of pupils' achievement between classes can be explained by differential recruitment of pupils.

In the third model, socioeconomic composition appears to have a significant effect on reading performance, when controlling for individual characteristics. This means that being in a class with a population from a 1-point-scale lower socioeconomic background is associated with a 9.3 (CI 7.4; 11.2) decrease in reading performance. Composition alone explains 9 % of the variance at the class level. This effect seems not large but consists, however, in about a 67-point difference between students in the most favoured and the most disadvantaged classes.

In the last model, we added three interaction terms to investigate whether socioeconomic composition has a different effect in each country. The 12.7 coefficient (CI 10.1; 15.2) for composition is the effect of composition on the achievement of Belgian pupils. In France, the effect of composition is slightly higher, but is not significantly different from the one in Wallonia-Brussels Federation. However, the compositional effect is lower in Spain, and even more in Portugal, but remains significantly different from 0. Let us note that the effects of being in classes in Portugal or in Spain have significantly increased between the first and the last model, meaning that for two students in classes with an equivalent composition, the advantage for the one in Portugal or Spain is even larger as classes tend to be slightly more disadvantaged and composition has a lower effect in these countries. The difference in achievement is not significantly different between France and Spain anymore. According to the goodness of fit indexes, this model is the best among the five and 72.2 % of the class variance is explained.

As socioeconomic origin could have a differential effect in classes, we put home resources for learning at random. The specification does not hold, indicating that the effect of socioeconomic origin does not have a different effect in the four countries, but also that its effect does not change in function of the class's composition.

Conclusion

In this article we have tried to check if socioeconomic composition influences pupil achievement equally in segregated Western European countries. As modelling of composition requires a rich set of variables to be entered at the pupil level, and especially a measure of prior achievement, international survey data like those from PIRLS could be inappropriate if one wishes to model composition without heavy biases. However, these types of datasets provide a large number of variables and make it possible to compare the compositional effect between countries. Like other researches before (Agirdag et al., 2013, 2011; Danhier & Martin, 2014; Dumay & Dupriez, 2007), to use such possibilities, we explicitly assume that delay and other background variables can account for prior achievement and at least limit the omission bias in the measure of composition. Such an assumption excludes de facto the countries where pupils do not accumulate delay. After comparing segregation and delay accumulation, we decided to keep the Wallonia-Brussels Federation, France, Spain, and Portugal for our analysis.

Before summarizing the findings, some limits have to be discussed. Firstly, the use of delay as a proxy for prior achievement is obviously not perfect because they do not cover exactly the same piece of information. Grade retention in primary education or before does not depend only on an objective decision based on performance, but is often the result of a negotiation marked by parent and teacher subjectivity (including the reference to a specific average class level). Furthermore, it also depends on country practices and the structure of its educational system. Secondly, grade retention remains limited (in terms of the number of pupils concerned and the number of grades that are repeated) at the fourth grade of primary education. With around 10 % or more of pupils with delay and the use of a dummy to represent this information, there is considerable doubt as to whether prior achievement is entirely represented. Model 2 confirms this doubt. Only 11.6 % of the pupil variance is explained when home resources for learning, language spoken at home, and delay are modelled. After adding gender and three attitudinal scales available in PIRLS (students like reading, students motivated to read and students confident in their reading) in order to improve the level-one models and decrease the potential omission bias at the second level, the level-one pseudo-R^2 reaches 24.0. Compared with what is often observed when prior achievement is used in the literature, this value is acceptable but not so high. In other words, in order to minimize biases, we recommend considering socioeconomic, delay and attitudinal variables when prior achievement is not available, although we can doubt that these strategies could entirely compensate for the lack of prior achievement.

Secondly, compositional effect can act on student achievement at both the class and the school levels. Simulation studies have shown that ignoring levels can have important consequences for the analysis (Opdenakker & Van Damme, 2000; Van den Noortgate, Opdenakker, & Onghena, 2005). When intermediate levels are ignored, the variance is distributed at the adjacent levels. For a model with predictors, the bias due to forgotten levels is complex, but significant over- and underestimations of the model parameters can occur. Concretely, it means that if we model compositional effect only at the school level (as it would be the case if we used PISA data), this effect will be downplayed because a part of the class effect is measured at pupil level. On the opposite, by ignoring school level, as we did in our analyses, we obtained an unbiased composition measure, but we cannot disentangle the school-level and class-level effects of composition.

Finally, the measure of composition suffers from two types of errors that make it unreliable: measurement and sampling errors. Although sampling error is not problematic since PIRLS samples complete classes, measurement error could be. Harker and Tymms (2004) showed that the effect did indeed disappear when more reliable variables were used to measure individual characteristics as socioeconomic background. Let us note that the methodology used by Harker and Tymms has in turn also been criticized due to sampling and reliability issues (Lauder, Kounali, Robinson, & Goldstein, 2010). Recently, Marks (2015) found a stronger effect of school socioeconomic composition when adding measurement errors in the socioeconomic measure. Televantou et al. (2015) have shown that increasing measurement errors causes a positive bias in estimating compositional effect. Following the claim made by Marsh et al. (2009) stating that models of school contextual effect required taking both sampling and measurement errors into account, Televantou et al. (2015) estimated the advised “doubly latent” models on mathematical achievement in fourth grade of Cyprus education within the multilevel structural equation modelling framework. They have found a non-significant compositional effect, in contrast with the significant effect found while they followed the classical approach. Comparing methods to correct the overestimated compositional effect when measurement error is ignored, Pokropek (2015) showed that the use of plausible values can provide accurate results when the reliability of level-one variable is high. New methodological approaches will be explored but at present, we can highlight that the reliability of the SES indicator is high (the Cronbach’s alpha ranging between .71 in Wallonia-Brussels Federation to .77 in Portugal).

Having stated all these caveats, let us assume our modelling strategy is acceptable. When we focus on our main results, we can observe that socioeconomic composition does not have an equivalent effect on pupil achievement in the four segregated countries included in our analysis. Indeed, its effect is important in Wallonia-Brussels Federation and France but lower in Spain and Portugal. While the compositional effect is a useful concept to describe the detrimental effect of segregation on disfavoured pupils, it seems that socioeconomic segregation is not as such as sufficient condition to observe a large socioeconomic compositional effect. These results are striking but not totally unexpected as they are in line with recent works of Le Donné (2014) who showed (using multilevel analyses of PISA 2009 data) that, in Spain and Portugal, individual and school characteristics explain a smaller part of variance compared to Wallonia-Brussels Federation or France. In other words, one should not simply assume that socioeconomic segregation leads to a high compositional effect. This is the main point we wish to make in this contribution, as it shows that the link between segregation and compositional effect still needs to be fully understood. We therefore advocate that similar analyses are also to be extended to countries with a low level of segregation.

On a more substantive level, let us note that the reason why the compositional effect is larger in France and Wallonia-Brussels Federation than elsewhere for the time being remains obscure. A first avenue to explore concerns academic segregation that could affect student performance in addition to socioeconomic segregation. However, the four countries also appear as highly segregated on the basis of academic results. Then, the way the entire school career is structured could already influence the way composition plays in primary education. Based on a factor analysis of PISA 2000 data, Mons (2007) proposed a typology to synthesize the way the heterogeneity of a school’s student body is managed. However, the four countries belong to the same type defined by Mons, the “uniform integration model”, which is defined as maintaining a common core until a certain age, while repetition works mainly as a mechanism to differentiate pupils. Dubet, Duru-Bellat, and Vérétout (2010) used the concept of cohesion to characterize educational systems. They defined cohesion as a set of attitudes or values conducive to co-operation, confidence, and tolerance. This axis allows distinguishing France and Wallonia-Brussels Federation – where school cohesion is weaker – from Portugal and Spain – where it is higher. Two educational styles were also identified by the authors. The “benevolent community” (Portugal and Spain) is marked by a confidence in the school, an attachment to the family links, and a relaxed atmosphere in classes, while the “school of knowledge” (France and Wallonia-Brussels Federation) is marked by a priority given to the transmission of a corpus of academic knowledge (clearly set apart from a familiar knowledge). Although the latter proposition is promising, it has been confirmed by few studies. A more qualitative approach is required to identify why we observed such a difference and how composition affects student outcomes in these countries.

In other words, there may be variables linked to the school organization or principals that probably mediate the effect of composition. In their literature review, van Ewijk and Sleegers (2010a) offer three categories of explanations. The compositional effect can result from direct peer interactions (discussions, motivation, or disruptions; or, for ethnic composition, tensions between races or language difficulties), teacher practices (adjustments in teaching style or expectations), and school quality (problems in human resources management or funding). In other words, school compositional effects actually refer to a black box including student body characteristics and peer influences, and a range of factors associated with schools hosting a specific public and the teachers working at these schools. Actually, each of the preceding categories of explanations needs to be compared between the segregated countries in order to assess whether the compositional effect expresses itself similarly. The difference in compositional effect could be a sign that some of the aforementioned potential influences are smaller in Spain and Portugal.

Acknowledgement

The research leading to these results has received funding from the European Research Council, under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement 28360, for the EQUOP-project “Equal opportunities for migrant youth in educational systems with high levels of social and ethnic segregation - assessing the impact of school team resources”.

The author wishes to express his gratitude to Laurie Hanquinet, Dirk Jacobs and Rob Kaelen for helpful comments and assistance.

References

Agirdag, O., Van Avermaet, P., & Van Houtte, M. (2013). School Segregation and Math Achievement: A Mixed-Method Study on the Role of Self-Fulfilling Prophecies. *Teachers College Record*, *115*(3), 1–50.

Agirdag, O., Van Houtte, M., & Van Avermaet, P. (2011). Why does the ethnic and socio-economic composition of schools influence math achievement? The role of sense of futility and futility culture. *European Sociological Review*, *28*(3), 366–378.

Asparouhov, T. (2006). General multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, *35*(3), 439–460.

Carle, A. C. (2009). Fitting multilevel models in complex survey data with design weights: recommendations. *BMC Medical Research Methodology*, *9*(1), 9–49.

Condron, D. J. (2009). Social class, school and non-school environments, and black/white inequalities in children’s learning. *American Sociological Review*, *74*(5), 685–708.

Cortese, C. F., Falk, R. F., & Cohen, J. K. (1976). Further Considerations on the Methodological Analysis of Segregation Indices. *American Sociological Review*, *41*(4), 630–637.

Danhier, J., & Martin, É. (2014). Comparing Compositional Effects in Two Education Systems: The Case of the Belgian Communities. *British Journal of Educational Studies*, *62*(2), 171–189.

Danhier, J. (2016). Modelling multiple measures of compositional effect: does factorisation simplify the picture in Belgium? *GERME Working Paper Series*, (1).

Darmawan, I. G. N., & Keeves, J. P. (2006). Accountability of teachers and schools : a value-added approach. *International Education Journal*, *7*, 174–188.

De Fraine, B., Van Damme, J., & Onghena, P. (2002). Accountability of Schools and Teachers: What Should Be Taken into Account? *European Educational Research Journal*, *1*(3), 403–428.

Delvaux, B. (2005). Ségrégation scolaire dans un contexte de libre choix et de ségrégation résidentielle. In M. Demeuse, A. Baye, M.-H. Straeten, J. Nicaise, & A. Matoul (Eds.), *Vers une école juste et efficace* (pp. 275–295). Bruxelles: De Boeck.

Dubet, F., Duru-Bellat, M., & Vérétout, A. (2010). *Les sociétés et leur école: emprise du diplôme et cohésion sociale*. Seuil.

Dumay, X., & Dupriez, V. (2007). Accounting for class effect using the TIMSS 2003 eighth-grade database: Net effect of group composition, net effect of class process, and joint effect. *School Effectiveness and School Improvement*, *18*(4), 383–408.

Dumay, X., & Dupriez, V. (2008). Does the school composition effect matter? Evidence from Belgian data. *British Journal of Educational Studies*, *56*(4), 440–477.

Duncan, O. D., & Duncan, B. (1955). A Methodological Analysis of Segregation Indexes. *American Sociological Review*, *20*(2), 210–217.

Duru-Bellat, M., Le Bastard-Landrier, S., & Piquée, C. (2004). Tonalité sociale du contexte et expérience scolaire des élèves au lycée et à l'école primaire. *Revue française de sociologie*, *45*(3), 441–468.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*(2), 121–138.

Eurydice. (2011). *Grade retention during compulsory education in Europe: Regulations and statistics*. Brussels: Publications Office of the European Union.

Eurydice. (2015). Eurypedia. Retrieved 2 August 2015, from https://webgate.ec.europa.eu/fpfis/mwikis/eurydice/index.php/Countries.

Gorard, S. (2006). Is there a school mix effect? *Educational Review*, *58*(1), 87–94.

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, *60*(1), 549–576.

Harker, R., & Tymms, P. (2004). The effect of student composition on school outcomes. *School Effectiveness and School Improvement*, *15*, 177–199.

Hox, J. (2010). *Multilevel analysis. Techniques and applications* (2nd ed.). New York:

Routledge.

Hutchens, R. (2004). One Measure of Segregation. *International Economic Review*, *45*(2), 555–578.

James, D. R., & Taeuber, K. E. (1985). Measures of Segregation. *Sociological Methodology*, *15*, 1–32.

Joncas, M., & Foy, P. (2012). Methods and procedures in TIMSS and PIRLS 2011 - Sample design in TIMSS and PIRLS. Retrieved from http://timssandpirls.bc.edu/methods/

Lauder, H., Kounali, D., Robinson, T., & Goldstein, H. (2010). Pupil composition and accountability: An analysis in English primary schools. *International Journal of Educational Research*, *49*(2–3), 49–68.

Le Donné, N. (2014). European Variations in Socioeconomic Inequalities in Students' Cognitive Achievement: The Role of Educational Policies. *European Sociological Review*, jcu040.

Maas, C. J. M., & Hox, J. (2005). Sufficient Sample Sizes for Multilevel Modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *1*(3), 86–92.

Marks, G. N. (2015). Are school-SES effects statistical artefacts? Evidence from longitudinal population data. *Oxford Review of Education*, *41*(1), 122–144.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-Latent Models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error. *Multivariate Behavioral Research*, *44*(6), 764–802.

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M. O., Mullis, I. V. S., & Foy, P. (2011). Age distribution and reading achievement configurations among fourth-grade students in PIRLS 2006. *IERI Monographic Series: Issues and Methodologies in Large-Scale Assessments*, *4*, 9–33.

Massey, D. S., & Denton, N. A. (1988). The Dimensions of Residential Segregation. *Social Forces*, *67*(2), 281–315.

Mons, N. (2007). *Les nouvelles politiques éducatives : La France fait-elle les bons choix ?* PUF.

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 International Results in Reading.* ERIC.

Mullis, I. V., Martin, M. O., Kennedy, A. M., Trong, K. L., & Sainsbury, M. (2009). *PIRLS 2011 Assessment framework.* ERIC.

Mullis, I. V. S., Martin, M. O., Minnich, C. A., Drucker, K. T., & Ragan, M. A. (Eds.). (2012). *PIRLs 2011 Encyclopedia: Education Policy and Curriculum in Reading*. Chesnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

Opdenakker, M.-C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: an illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, *11*, 103–130.

Opdenakker, M.-C., & Van Damme, J. (2001). Relationship between school composition and characteristics of school process and their effect on mathematics achievement. *British Educational Research Journal*, *27*(4), 406–428.

Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *60*(1), 23–40.

Pokropek, A. (2015). Phantom Effects in Multilevel Compositional Analysis Problems and Solutions. *Sociological Methods & Research*, 44(4), 677-705.

Ransom, M. R. (2000). Sampling Distributions of Segregation Indexes. *Sociological Methods & Research*, *28*(4), 454–475.

Rasbash, J., Steel, F., Brown, W. J., & Goldstein, H. (2012). *A user's guide to MLwiN, v2.26*. University of Bristol: Centre for Multilevel Modelling.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rumberger, R. W., & Palardy, G. J. (2005). Does the segregation still matter ? The impact of student composition on academic achievement in high school. *Teachers College Record*, *107*(9), 1999–2045.

Sykes, B., & Kuyper, H. (2013). School Segregation and the Secondary-School Achievements of Youth in the Netherlands. *Journal of Ethnic and Migration Studies*, *39*(10), 1699–1716.

Televantou, I., Marsh, H. W., Kyriakides, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement*, *26*(1), 75–101.

Thrupp, M., Lauder, H., & Robinson, T. (2002). School composition and peer effects. *International Journal of Educational Research*, *37*(5), 483–504.

Timmermans, A. C., Doolaard, S., & de Wolf, I. (2011). Conceptual and empirical differences among various value-added models for accountability. *School Effectiveness and School Improvement*, *22*(4), 393–413.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3).

Van den Noortgate, W., Opdenakker, M.-C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*(3), 281–303.

Van der Slik, F. W. P., Driessen, G. W. J. M., & De Bot, K. L. J. (2006). Ethnic and Socioeconomic Class Composition and Language Proficiency: a Longitudinal Multilevel Examination in Dutch Elementary Schools. *European Sociological Review*, *22*(3), 293–308.

van Ewijk, R., & Sleegers, P. (2010a). Peer ethnicity and achievement: a meta-analysis into the compositional effect. *School Effectiveness and School Improvement*, *21*(3), 237–265.

van Ewijk, R., & Sleegers, P. (2010b). The effect of peer socioeconomic status on student achievement: A meta-analysis. *Educational Research Review*, *5*(2), 134–150.

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399.

White, M. J. (1986). Segregation and Diversity Measures in Population Distribution. *Population Index*, *52*(2), 198–221.

Willms, J. D., & Raudenbush, S. W. (1989). A Longitudinal Hierarchical Linear Model for Estimating School Effects and Their Stability. *Journal of Educational Measurement*, 26(3), 209–232.

Zhang, Z., Charlton, C. M. J., Parker, R. M. A., Leckie, G. B., & Brown, W. J. (2012). R2MLwiN (Version v0.1). University of Bristol: Centre for Multilevel Modelling.