British Journal of Educational Studies
Vol. 62, No. 2, June 2014, pp. 171–188

COMPARING COMPOSITIONAL EFFECTS IN TWO EDUCATION SYSTEMS: THE CASE OF THE BELGIAN COMMUNITIES

by JULIEN DANHIER and ÉMILIE MARTIN, *Université libre de Bruxelles*

ABSTRACT: The Belgian educational field includes separate educational systems reflecting the division of the country into linguistic communities. Even if the French-speaking and the Dutch-speaking communities keep sharing important similarities in terms of funding rules and structures, they present a huge gap between their respective pupils' achievement. The Belgian educational systems are then characterized by high levels of segregation; this paper aims to model the potential differential effect of school composition on pupils' achievement. Multilevel models are consequently tested on the Programme for International Student Assessment 2009 data (7184 respondents in 236 schools). Our results suggest that the position of pupils in the educational system still has a deep influence in a system that functions as a quasi-market and where grade repetition and tracking are widely used. Our analysis also confirms that academic and socio-economic compositions have an extra negative effect on pupils' achievement.

Keywords: education, achievement gap, compositional effect

1. INTRODUCTION

Since 1989 the Belgian educational field has included three separate educational systems reflecting the division of the country into three linguistic communities. The Dutch-speaking and the French-speaking communities provide schooling for the majority of the pupil population (respectively about 55% and 44% of the pupils in Belgian schools), while less than 1% of this population are in German-speaking schools. The two major educational systems are separately managed but continue to share important similarities in terms of funding rules and structures. Both systems share the same constitution, which defines education as free but also as a quasi-market where parents are free to choose their children's schools. The other side of the coin is that schools are in competition, implement strategies and specialize themselves to attract pupils (or specific pupils) on which public subsidization is mainly based (the education being free). In terms of structure, both systems were historically divided on grades and levels (nursery, primary and secondary education) and organized different tracks in the secondary level. An important difference lies, however, in the introduction of a more thoroughly integrated first grade in the French-speaking part of the country, where most of the pupils are supposed to follow the same curriculum up to age 14. Furthermore,

ISSN 0007-1005 (print)/ISSN 1467-8527 (online)

http://dx.doi.org/10.1080/00071005.2014.930091
http://www.tandfonline.com

different educational traditions and cultures have slowly developed. Consequently, this provides a unique opportunity to compare historically similar but increasingly different systems.

The existing comparative literature places particular emphasis on examining the attainment discrepancies between the two broadest communities. Indeed, it is widely known that there is a sizeable achievement gap between them. Various surveys, such as the Programme for International Student Assessment (PISA), have shown that the Dutch-speaking schools are at the top of the international rankings while the French-speaking schools do not perform better than the OECD mean (Hirtt, 2008; Hindriks and Verschelde, 2010; Jacobs and Rea, 2011; Vandenberghe, 2011). If various studies agree about the existence of a performance gap, they do not agree about the reason for its existence and propose different hypotheses. A first hypothesis concerns the differential population of pupils provided with schooling in the two communities, in terms of socio-economic or ethnic origins. It has been argued that the French-speaking average results are lower because of a wider proportion of migrant or disfavoured pupils who tend to have lower scores. However, the Dutch-speaking community performs better than the French-speaking one even when the socio-economic and ethnic origins are accounted for (Hirtt, 2008; Hindriks and Verschelde, 2010). Hirtt (2008), among others, defends a second hypothesis: the higher subsidization of Dutch-speaking schools could explain their better performance. Nevertheless, Vandenberghe (2011) shows that a gap exists prior to the 'communautarisation' and its consecutive distance in terms of subsidization. He points to the lack of coherence in school governance in the French-speaking community as an alternative explanation. Two additional reasons are worth noting: Hindriks and Verschelde (2010) put forward the role of school autonomy, and Hirtt (2008) proposes the accuracy and the consistency of pedagogical programmes.

Although no clear consensus has been found to explain this inter-community achievement gap, all researchers agree on one point: Belgium as a whole remains one of the most unequal educational systems among the OECD countries. Indeed, the Flemish community might do well in the ranking of educational attainment, but as far as equal opportunities are concerned both the French and the Flemish schools have poor results (Jacobs *et al.*, 2009). By making use of the PISA 2006 data, Hindriks and Verschelde (2010) demonstrate that, in the French-speaking community, the performance gap between pupils with the best and the lowest performance in scientific literacy is equal to 352 points, which is the widest gap among the OECD countries. In the Flemish community, this achievement gap is almost as important. More subtle analyses highlight that the communities' education suffers from a structural illness: segregation (Baye and Demeuse, 2008; Demeuse and Friant, 2010; Dupriez and Vandenberghe, 2004; Hindriks and Verschelde, 2010; Jacobs *et al.*, 2009). Therefore, in addition to being unequal, the Belgian educational system is deeply marked by academic and socio-economic segregation.

This article aims to model simultaneously both communities to test two different hypotheses. Firstly, the supposed differences of socio-economic background between the pupils from both communities cannot explain the achievement gap. Secondly, the school where pupils are registered has an impact in both communities but can play differently because of a different structure of segregation. Modelling simultaneously communities and not running separate analyses is an advantage because it allows us to observe the way in which the achievement gap is modified by the inclusion of new variables.

2. The Compositional Effect

Following Dumay and Dupriez, we will give a quite neutral but statistical definition of the compositional effect: the 'impact of pupils' aggregated characteristics (socioeconomic status, socio-cultural capital, prior achievement, etc.) when these variables have been taken into account at the individual level' (Dumay and Dupriez, 2008, p. 440).

This effect has been studied for decades in the sociology of education with equivocal results. Quite recently, some methodological criticisms have arisen arguing that compositional effect could mostly be a statistical artefact. Harker and Tymms (2004) shows that the observation of this effect was linked with modelling problems (namely, measurement errors and model misspecifications). Actually, the effect disappears when some important pupil characteristics, such as prior achievement, were entered in the model and when more reliable and valid variables were used to measure these characteristics. In other words, the compositional effect would capture what is not so well captured by the level-one model. However, the methodology used by Harker and Tymms has been criticized in its turn, among others, for sampling and reliability issues (Lauder *et al.*, 2010).

A second axe of criticism relates to the concept of composition itself and its validity (Gorard, 2006; Harker and Tymms, 2004). In short, the question is what is really measured by composition. There is some evidence that we might be (at least partially) confronted with a spurious effect between the pupil composition and their educational performance. Composition could then function as a proxy for other school characteristics not included in the model.

Thrupp (1999) suggests that there might be an accumulation of small effects (categorized as reference group, instructional and organizational processes) that were linked with the pupil composition. Moreover, Opdenakker and Van Damme (2001) find an important joint effect between school composition and school processes as a high level of cooperation between teachers, which are correlated with composition and had a positive impact on pupils' achievement. Rumberger and Palardy (2005) observe that the estimated effect of socio-economic composition was explained by four school process variables, namely teachers' expectation, hours of homework, number of advanced courses taken and feeling unsafe at school. Research in Flanders shows that migrant pupils (but also pupils with a working-class background) could develop a futility culture (referring to the belief that they have no control over their success), which may explain the compositional effect for these subpopulations (Agirdag *et al.*, 2011, 2013). In their literature review, Van Ewijk and Sleegers (2010a) sum up three categories of explanations. Next to statistical misspecifications, compositional effect can result from direct peer interactions (discussions, motivation, disruptions or, for ethnic composition, tensions between races or language difficulties), teachers' practices (adjustments in teaching style or expectations) and school quality (problems in human resources management or funding). We see that school compositional effects actually refer to a black box including, on the one hand, pupil body characteristics and peer influences, and on the other a range of features associated with teachers and with schools hosting specific groups of the population. As a consequence, when focusing on the effects of school composition, the danger exists that one might forget or remove the effects of other (correlated) school characteristics. Inversely, one might overestimate the impact of school processes when not including school composition into the analysis.

Raudenbush and Willms (1995) make a strong contribution to the literature on composition modelling. The first effect (type A) concerns the school influence, but the second (type B) allows isolating the influence of school practices (net of the school context). To be able to properly measure the type A effect, the authors provide the foundation for a statistical model. According to them, this measurement requires the inclusion of a full set of pupil background variables, including prior achievement. On the basis of conceptual and methodological issues in measuring composition, Thrupp *et al.* (2002) propose a list of 10 features that an ideal compositional model should fulfil, among which we find a sample with both ends of the socio-economic spectrum, a full set of entry-level variables (including prior achievement), measures covering the three dimensions of composition (peer group effect, instructional processes, school organization), a combination of different types (academic, socio-economic, etc.) and measures of composition.

In addition, deciding which composition should be taken into account is not straightforward. Opdenakker and Van Damme (2001) show that, in the Flemish educational system, academic and socio-economic compositions had an effect on pupils' achievement but only the academic one was significant when both variables were entered together. In the French-speaking part of Belgium, Dumay and Dupriez (2008) observe effects of academic and socio-cultural compositions even after controlling for individual abilities, socio-economic background and language characteristics. They acknowledge, however, the difficulty to disentangle the composition with a relevant effect because of the high correlation between different types of composition. In a comparative study of 12 OECD countries (including Belgium), Dronkers (2010) identifies significant effects of additive socio-economic and ethnic composition but also different negative effects of diversity. Agirdag *et al.* (2011) find that an ethnic compositional effect vanished in the Flemish community where prior academic achievement and socio-economic background were taken into account. Danhier (2013) shows that in the Flemish community a combination of problems (collinearity, a small sample size at the school level, a small part of variance remaining to explain and some outliers) turns modelling of composition into a complex puzzle.

3. Data and measures

Database

PISA is a research project led by the OECD that aims to assess pupils' ability 'to use their knowledge and skills to meet real-life challenges' (OECD, 2012, p. 22). The most recent survey for which data were available was conducted in 2009. This survey focused specifically on reading skills over 65 participating countries. The survey relied on a two-stage stratified sampling procedure: schools were sampled according to their size after having been separated between explicit strata (form of education, public/private) and ordered by implicit strata; pupils were randomly sampled in selected schools to obtain 35 respondents by schools (or less if there are not enough valid 15-year-old pupils).

As a main advantage, PISA allowed comparing the Belgian Dutch-speaking and French-speaking communities. No common database was actually available in Belgium to compare pupils from different communities but PISA is the only survey studying educational achievement in which both participate. For our purpose, we extracted the Belgian Dutch-speaking and French-speaking subsamples. The initial sample covered 7705 pupils grouped in 268 schools.

From this initial sample, we firstly chose to exclude some schools. We only selected regular full-time education and consequently excluded part-time vocational education (eight schools) and education for pupils with special needs (15 schools), which are organized in different structures. Secondly, we chose to exclude some pupils. Variables have missing values and their weighted rates of missing values are presented in Table 1. The overall rate of missing values at the pupil level (namely, the weighted proportion of pupils with at least one missing value) reached 1.7%. Due to the limited rate of missing values, we used the listwise deletion method to make our analysis simpler without introducing too much bias (Graham, 2009). Let us thirdly note that in order to assure a stable basis to compute compositional effects, we chose to exclude schools with less than 10 respondents (nine schools). Our final sample covered 7184 respondents in 236 schools (respectively 141 and 95 schools in the Dutch-speaking and the French-speaking communities).

Reading Proficiency

To conciliate the limited time that is available to test each pupil and the need to cover a wide range of domain knowledge, a specific methodology was used to

TABLE 1: Descriptive statistics

Variable	Missing rate	Mean	Minimum	Maximum	Standard deviation	Skewness	Kurtosis	Mean, Dutch/French
Dependent								
Reading (first plausible value)	0	515.7	154.4	779.1	95.54	–0.340	–0.277	527.3/501.2
Reading (second plausible value)	0	516.1	166.7	781.0	95.01	–0.360	–0.227	528.0/501.2
Reading (third plausible value)	0	514.9	108.2	818.5	96.56	–0.370	–0.141	526.9/499.9
Reading (fourth plausible value)	0	515.9	93.0	776.8	95.13	–0.350	–0.172	527.9/501.0
Reading (fifth plausible value)	0	515.4	105.9	791.5	95.96	–0.360	–0.173	527.1/500.7
Student level								
Gender (ref. female)	0	0.500	0	1				0.503/0.496
Vocational (ref. general)	0	0.455	0	1				0.550/0.336
Delay	0	0.390	–2	3	0.610	1.030	1.100	0.269/0.541
Economic, social and cultural status (–)	1.25	–0.234	–2.7	5.7	0.920	0.230	0.039	–0.242/–0.224
Language at home (other, ref. same)	0.66	0.100	0	1				0.055/0.156
Language at home (invalid, ref. same)	0.66	0.052	0	1				0.088/0.007
Non-European origin (ref. European origin)	0.75	0.136	0	1				0.086/0.200
School level								
Community (ref. French-speaking)	0	0.609	0	1				
Academic composition (additive)	0	0.454	–0.120	2.100	0.360	1.420	3.652	0.355/0.638
Socio-economic composition (additive)	0	–0.122	–1.590	1.600	0.520	–0.150	–0.167	–0.139/–0.096

assess pupils' ability using 'plausible values'. Such a statistical construction requires a brief explanation. PISA provided a battery of items that are characterized by a binary modality of result (right/false). Each pupil was tested on one of the subsamples of the whole battery. On the basis of the pupils' results, items' difficulty and pupils' ability were simultaneously computed using a method called the Rasch model. Because only an incomplete item subsample is administered, scores were computed with a relative uncertainty. This is why PISA did not provide a single pupil achievement estimate but five plausible values, which 'represent the range of abilities that a student might reasonably have' (Wu, 2005, p. 115). The reading plausible values were used as dependent variables. After the analyses had been separately conducted on each plausible value, the results were properly combined in order to obtain unbiased estimates (Rubin, 1987).

Pupils' Characteristics

Grade repetition and tracking are the key strategies that allow the Belgian educational systems to manage pupils' heterogeneity. First, it has been shown that being in vocational education could have a very large negative impact on pupil achievement (Jacobs and Rea, 2011). The logic behind this is well known. Unsuccessful pupils are redirected from general to vocational education following a sort of waterfall system. A dummy variable for vocational education was then created. Secondly, since failing pupils have to repeat the same grade, the 15-year-old pupils are scattered over several grades. A delay variable (the distance in years between the theoretical grade for a 15-year-old pupil and his actual one) was entered at the pupil level in the analyses. All of these variables enable us to assess the position of the pupils in the hierarchy of the educational system. Let us note that these two tools are not similarly used in both communities (Hindriks *et al.*, 2009). A quick look at the pupils' distribution regarding grade repetition and tracking (last column in Table 1) revealed that the mean delay of the French-speaking pupils reached 0.54 (against 0.27 for the Dutch-speaking one) while 55% of the Dutch-speaking pupils were in a vocational track (against 34% for the French-speaking one).

The influence of the socio-economic background on pupils' achievement has been proven many times. PISA provided a statistical construct called the index of economic, social and cultural status. This index synthesized the information from three sources: the highest level of parental occupation, the highest level of parental education, and the number and kind of home possessions. We used the reverse-coded version of this index to assess disfavoured socio-economic origin.

The use of a different language at home and at school has a negative impact on pupil achievement in Belgium (Jacobs *et al.*, 2009). PISA provided the variable 'language spoken at home'. We recoded the Flemish and Walloon dialects as 'same language', because dialects should be considered closely related to the school language. One should also note that the variable had a substantial proportion of missing values. The main reason was that multiple languages in the Flemish subsample were coded as invalid. As the French-speaking community insisted that only one response was to be obtained in 2009, the proportion of missing values decreased sharply to 1% while the aforesaid proportion reaches about 9% in the Dutch-speaking community (Table 1). According to preliminary analysis, the influence of the invalid category was quite similar to the 'different language' category. Two dummies were included in the model: one for pupils who speak another language at home and a second for the invalid ones.

It has been frequently shown that migration status can have an effect on educational outcomes. Typically, two variables are considered to account for the migrant situation: ethnic group and generation. Portes and MacLeod (1996) found for the United States that different migrant groups can present different advantages or disadvantages and that these effects remain after controlling for the socio-economic background. Levels and Dronkers (2008) showed that the immigrants' origins can have divergent impacts in different host countries. In contrast, Kristen and Granato (2007) observed in Germany that the socio-economic background removes the negative effect of origin. Next to ethnic group, one can also distinguish migration generation. Equally, the effect of generation is not clear cut. Controlling for socio-economic and socio-cultural factors, both between-school and within-school differences in performance using PISA 2000, Marks (2005) showed that only the second generation of migrant pupils who speak another language at home present significantly lower scores in Belgium. On PISA 2009, Jacobs and Rea (2011) found a significant effect only for the first generation on reading achievement in the Dutch-speaking community, while the first and second generations are significant in the French-speaking one. In our operationalization we define each pupil with at least one foreign parent as having an immigration background. Preliminary analyses have showed that crossing generation with origin fails to provide sufficient power to test each group effect even when origins are regrouped in only five groups (Western European countries, Eastern European countries, sub-Saharan African countries, non-European Mediterranean countries and others). As non-European migrants displayed comparable effects, we decided to enter one dummy to account for non-European origin and two extra dummies to model the migrant generation. However, we did not show generation in our results because it remains not significant when origin is modelled.

Academic, Socio-economic and Ethnic Compositions

In segregated educational systems, such as the Belgian system, we can expect some effects of school composition on the pupil's achievement. Following Thrupp *et al.* (2002), different types of composition were considered and measured but, at least in the case of Belgium, we could not model all of the possibilities (Dumay and Dupriez 2008). As stated earlier, compositional effects were measured by the aggregation of individual characteristics when the latter are previously entered in the model. We decided to account for the academic and socio-economic densities. The socio-economic composition variable was measured by the average index of economic, social and cultural status of pupils for each school. In order to create the academic composition variable, we compute the school average of the pupil delay variable. The proportion of pupils in vocational track was not used here as a measure of academic composition because of its low power in predicting achievement and its bimodal distribution.[1]

Let us note two weaknesses of this type of modelling. Firstly, if a full set of entry-level variables were not included in the model, PISA did not provide any measure of prior achievement. This is a clear drawback in the use of such databases to study composition. In a meta-analysis, Van Ewijk and Sleegers (2010a, 2010b), found that the compositional effects are greatly overestimated when prior achievement is not included, although the effect is large but not significant for ethnic composition. Nevertheless, in systems where repetition and orientation are largely used to deal with heterogeneity of pupils, we expected that the place in the educational system captures information about prior achievement. The place in the educational system results from decisions based on previous achievement, although not perfectly; the decision depends also on the class structure and the teacher's subjectivity. In Flanders, such an assumption was equally followed by Agirdag *et al.* (2013). Secondly, researchers have argued that compositional effect needs to be measured against the progress from an initial level and not against the final level itself. In line with Duru-Bellat *et al.* (2004), we supposed that a pupil evolves mainly in the same school context, although the argument is relevant for the assessment of teachers' effect.

4. Description of the Technique

Educational data are typically hierarchical: pupils are clustered in schools. Moreover, the PISA sampling design includes two stages: firstly the schools are drawn, and pupils are then sampled in each of the selected schools. Consequently, pupils in the same school are likely to be more similar to each other than to those from other schools. In fact, they tend to share some features due to the school selection processes (e.g. they more probably belong to the same social class) and to a common school environment (e.g. they share the same teachers, the same pedagogic resources and methodologies).

In such a situation, the assumed independence of observations cannot hold. The violation of independence in standard statistical tests leads to a clear underestimation of the parameters' standard errors and consequently to the discovery of spurious significant effects (Hox, 2010, pp. 4–5). In order to provide correct standard errors, we use multilevel modelling, a statistical technique that allows us to deal with the clustering in our data. Such a method has also some extra interesting features: it permits us to include variables at both pupils' and schools' levels and to model intra-school and inter-school relations. We run our analyses on MLwiN (Rasbash *et al.*, 2012) in the R environment (Zhang *et al.*, 2012). MLwiN provides Sandwich estimators and performs weighted[2] multilevel analysis using the IGLS algorithm. We model pupils as the first level and schools as the second level. All variables are grand mean centred.

With only two communities, it is clearly impossible to consider them as a third level. Another approach is preferred here: communities are entered as a school-level dummy and interactions allow us to test whether a variable has a different effect in the two communities. Such an approach can take place in a larger debate about how to compare regions and countries although it is beyond the scope of our article. Actually, if the level-one sample size is generally large enough at the lowest level, this is rarely the case for the highest-level sample size. In the case of comparison between countries, modelling the latter as the highest level is clearly problematic. With only 30 groups, Maas and Hox (2005) did not find non-negligible bias for fixed and random estimates but found that the confidence intervals (CIs) were clearly too small for the random slope and the variance at the group level. In other words, a significant group effect could be found mistakenly.

The intercept-only model allows us to compute the intraclass correlation and observe the way the variance is distributed at both levels. Since the variance reaches 4869.5 (standard error 236.5) for the first level and 4132.0 (standard error 361.7) for the second, the intraclass correlation is 0.46, which means that 46% of the variance occurred at the school level. Such a value justifies the use of multilevel modelling.

Let us note that such a value is quite high. Actually, this is due to specificities of the Belgian system, which we will discuss later, but is also partly connected with the sampling structure. In the database, pupils are gathered in schools but actually a pupil belongs to a specific class and is taught by a specific teacher. Consequently, a pupil–school structure is clearly too simple to adequately represent the clustered reality. Such a simplification could be trivial but simulation studies showed that ignoring intermediate levels could have important consequences for the analysis (Opdenakker and Van Damme, 2000; Van den Noortgate *et al.*, 2005). When intermediate levels are ignored, the variance is distributed at the other levels. The variance repartition in our sample is near to the repartition obtained in the simulations ignoring intermediate levels. For a model with predictors, the bias due to forgotten levels is more complex but significant overestimations and underestimations of the model parameters can occur. Such technical consideration has implications, to which we will return. Some effects could come into play at the class level and, consequently, some pupils' or schools' parameters could be different. For example, it is likely that compositional effects are present at the class level. Concretely, this signifies that compositional effects at the school levels will be minimal because a part of the class effect is distributed between schools and pupils.

5. Results

With grand mean centring, the intercept in a specific school can be viewed as the 'adjusted mean' for this school, namely the mean when the effects of all the explanatory variables have been removed and the overall intercept is the averaged 'adjusted mean'. In other words, it is the expected score in reading for an 'averaged' pupil (a pupil with a mean score on all independent variables). Clearly it is difficult to grasp what an average score is for a dummy like gender, but if we apprehend the intercept as an adjusted mean, we can consider the intercept for a specific school as the mean of this school if the proportion of boys and girls has been equal across schools (Enders and Tofighi, 2007). The regression coefficients represent the increase in reading proficiency that is associated with a one-unit increase in the given predictor, controlling for the other variables included in the model. According to the OECD, a 39-point difference in reading is more or less the results of one year of schooling (OECD, 2010, p. 8).

To assess the difference in terms of reading performance between the two communities, a dummy variable is entered into the analysis (Model 1 in Table 2). The significant coefficient indicates that the two communities present a reading achievement gap equal to 25.8 points (CI 6.1; 45.2).[3] Unsurprisingly, this gross gap is in favour of the Dutch-speaking schools. Attending a school in the French-speaking part of the country has a negative influence on reading achievement. By modelling both communities simultaneously, we can observe the proportion of the variance that is explained by the community membership and put its extent in perspective. The 3.7 level-two pseudo-R^2 value indicates that the model significantly reduces the variance at the school level. Such a reduction is at once big and small: big according to the similarity of both systems, but small since the major part of the variation remains to be explained.

Pupils' characteristics are entered in two steps. Firstly, the non-academic ones – namely gender, socio-economic background, language spoken at home and ethnic origin – are modelled. These characteristics have the expected effect on achievement. Not surprisingly, being male, being from a disadvantaged socio-economic background and not speaking the school language at home are associated with weaker performance in reading. Moreover, the non-European origin has a significant impact. Indeed, controlling for the other pupils' non-academic characteristics, being a pupil with a non-European background keeps an extra explanatory power to predict reading performance: there remains a difference of 15.6 points (CI 8.6; 22.7).

Model 2 allows testing whether the achievement gap is due to differential recruitment processes in both communities. The model provides an achievement gap net of non-academic characteristics. The French-speaking community could provide more disfavoured pupils with schooling and consequently obtain lower results. Descriptive statistics showed that the French-speaking subsample includes more pupils with a migration background, fewer pupils speaking the language of the test at home but, nevertheless, slightly more favoured pupils.

TABLE 2: Results of multilevel modelling on reading performance

Parameter	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Fixed part						
Intercept	**503.0 (4.90)**	**506.4 (4.07)**	**513.6 (2.22)**	**518.1 (1.70)**	**517.8 (1.71)**	**515.0 (2.29)**
Flemish community	**25.8 (10.1)**	*23.9 (8.4)*	**25.0 (4.55)**	*15.6 (5.46)*	*15.3 (5.07)*	**18.1 (4.68)**
Male		**–16.9 (2.31)**	**–12.1 (1.94)**	**–12.6 (1.88)**	**–12.4 (1.86)**	**–12.2 (1.84)**
Socio-economic status		**–14.6 (1.4)**	**–8.72 (1.09)**	**–7.09 (1.11)**	**–6.79 (1.10)**	**–6.8 (1.10)**
Other language		**–31.8 (3.82)**	**–22.6 (3.62)**	**–21.6 (3.63)**	**–21.7 (3.66)**	**–21.4 (3.62)**
Invalid language		**–32.2 (5.97)**	**–23.3 (4.59)**	**–21.9 (4.46)**	**–21.7 (4.44)**	**–21.6 (4.45)**
Non-European origin		**–15.7 (3.58)**	**–15.4 (3.06)**	**–12.9 (2.96)**	**–12.4 (2.93)**	**–12.4 (2.93)**
Delay			**–52.2 (1.92)**	**–49.9 (1.90)**	**–48.4 (1.97)**	**–48.4 (1.97)**
Vocational			**–68.4 (3.79)**	**–61.6 (3.86)**	**–61.4 (3.69)**	**–62.0 (3.71)**
Academic composition				*–32.5 (12.2)*	*–32.6 (10.8)*	*–31.3 (10.5)*
Socio-economic composition				**–27.5 (6.23)**	**–29.3 (5.87)**	**–30.5 (5.34)**
Composition × delay					–	
Composition × vocational					–	
Composition × academic composition						–48.0 (19.7)
Composition × socio-economic composition						**28.3 (9.9)**
Random part						
Level-one variance (σ^2)	**4870 (236.5)**	**4528 (211.6)**	**3599 (162.4)**	**3589 (161.6)**	**3477 (158.9)**	**3475 (158.6)**
Level-two variance (τ_{00})	**3978 (342.5)**	**2743 (219.1)**	**658.1 (90.6)**	**329.8 (59.9)**	**254.7 (70.0)**	**234.6 (63.5)**
Vocational slope					**597.3 (186.1)**	**607.8 (182.2)**
Delay slope					*161.1 (63.9)*	*160.5 (62.8)*
Goodness of fit						
Deviance	83,032	82,442	80,536	80,386	80,308	80,296
AIC	83,054	82,474	80,572	80,426	80,358	80,350
BIC	83,092	82,529	80,634	80,495	80,444	80,443
Level-one R^2	0.0	7.0	26.1	26.3	28.6	28.6
Level-two R^2	3.7	33.6	84.1	92.0	93.8	94.3

Standard errors given in parentheses. Significance for Wald test: bold = 0.001, italic = 0.01, normal = 0.05, – = non-significant.

Actually, the drop in achievement gap is very limited. The hypothesis of different populations is not really supported here.

Secondly, variables defining the position of the pupil in the educational system are added in the third model. The predicted score for an average pupil reaches 513.6 (CI 509.3; 518.0). Among the predictors modelled at the pupil level, the coefficients associated with the position in the educational system are significant. Being a year behind is associated with a drop of 52.2 points in reading performance (CI 48.4; 55.9). Likewise, to be in vocational education is associated with a fall of 68.4 points (CI 61.0; 75.8). It is worth noting that the effects of grade repetition and the track can be cumulated for a specific pupil. Such a result confirms previous findings and can be explained, on the one hand, by reasons that cause a lower acquisition in this specific schooling (lower exposition to schooling, stigma, discouragement, etc.) and, on the other, by the lower prior achievement of these pupils that conducted them to this position. Let us note that according to the decline in the coefficients of the non-academic characteristics, the position in the system plays as an intermediary between the latter ones and reading achievement, except for ethnic origin.

In comparison with the first model, both the second and third models represent a large improvement.[4] The 26.1 level-one pseudo-R^2 value of the model indicates that the model significantly reduces the variance at the pupil level. Pupil characteristics play equally at the school level: with a pseudo-R^2 value equal to 84.1, they sharply reduce the variance of this level. Further examinations of the R^2 variations confirm previous analyses (Monseur and Lafontaine, 2009) and show that 22.4% of the variance at the school level is explained by the joint effect of the academic and non-academic characteristics, while 50.4% and 7.5% are explained by their respective unique effects. In other words, an important part of the variation of pupils' achievement between schools can be explained by differential recruitments, especially on academic bases. A major part of the school variance is then due to the structure of the educational system. Such a feature explains also why we can observe a high intraclass correlation in regions where tracking is used.

With grand mean centring, the level-one coefficient is a blend of intra-school and inter-school relations that cannot be disentangled. At first sight this may seem problematic, but this feature is an advantage for testing whether the compositional effect is significant. 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, the effect of a specific composition variable when the level-one variable and its (unequal) repartition is taken into account. In other words, the coefficient is equal to zero if composition does not explain any extra variance (Enders and Tofighi, 2007).

In the fourth model, the academic and socio-economic compositions appear to have significant extra effects on reading performance, when controlled for individual academic and socio-economic characteristics. This means that being in a school with a population from a low socio-economic background or with a low academic attainment is respectively associated with a 27.5 (CI 15.3; 39.8) or a 32.5 (CI –56.3; –8.7) decrease in reading performance. This is a quite strong effect if we interpret a 39-point difference as the result of one year of schooling. Let us note that academic and socio-economic compositions are highly correlated. With a 0.66 correlation, instability due to collinearity may be a problem. The model has been re-run with each of the composition to check whether it invalidates our results. Both compositions present a higher effect (because of the joint effect where they are modelled together) but the direction of the effect remains the same. Instability is therefore not a problem.

In Model 5 we decide to treat the delay and the vocational variables as a random effect. This specification allows us to test whether the different use of grade repetition and tracking in the two communities explains the achievement gap. For that purpose, we allow the slope of both delay and vocational variables to vary randomly across schools and we specify cross-level interactions between the community dummy and these variables. According to AIC and BIC, the random specification holds. Nevertheless, the cross-level interactions are not significant. Consequently, if the effects of those two predictors are significantly different across schools, this difference is not explained by the membership of one community rather than the other.

Finally, we investigate whether the compositional academic and socio-economic effects are different in the French-speaking and the Dutch-speaking parts of the country. In order to do so, we add two interaction terms to our analysis: an interaction effect between community and the academic composition, and an interaction effect between community and socio-economic composition. Both effects present a statistically significant coefficient. The negative interaction coefficient for academic composition indicates that the effect of the academic composition is significantly larger in schools from the Dutch-speaking community while the positive coefficient of the socio-economic composition points out that it has a larger effect in the French part of the country. Because, the interactions are only significant when they are modelled together, the results of this latter model have to be considered with caution. Nevertheless, the model has been run separately on each community and gives similar results.

The centring of variables makes the interactions difficult to interpret. We then provide the coefficients of the compositional effects for each community. Actually, only the socio-economic composition has an effect in the French-speaking community (–47.5, CI –32.4; –62.7), while, in the Dutch-speaking community, academic composition (–50.4, CI –21.1; –79.7) has a higher effect than socio-economic composition (–19.2, CI –5.8; –32.7). Such results are read as follows: a one-point change on the socio-economic composition scale (which goes from –1.6 to 1.6) is associated with a 47.5-point drop in the French-speaking community and a 19.2-point drop in the Dutch-speaking one.

Different justifications can be mobilized here. Firstly, the high correlation between the two types of composition in the French-speaking community (0.76) but also the very strong association between socio-economic composition and the achievement at the school level can explain why the effect of school compositions is entirely caught by the socio-economic one. In the Dutch-speaking community, such correlation is a little bit lower (0.67). Secondly, the use of mean delay as a proxy to measure academic composition provides another explanation. Mean delay could be more discriminating in the Dutch-speaking community where grade repetition is used differently and to a more limited extent, although we do not measure a different effect at the pupil level. Thirdly, from a more substantive point of view, the magnitudes of academic and socio-economic segregations are not the same in both communities. A look at the dissimilarity index shows that while socio-economic segregation is more salient in the French-speaking community, academic segregation is more striking in the Dutch-speaking one. The different effects of composition types across communities can then be due, at least partially, to the different segregations pupils experience in both communities.

Moreover, we observe a quite important decrease of the coefficient associated with the community predictor between the third model and the following models. This means that a part of the performance gap between the two communities could be due to the differential effect of academic and socio-economic schools' compositions (or by characteristics associated with this composition if the variable is viewed as a proxy). However, the fact that a very large gap still remains (especially in the last model), indicates that we need to consider other reasons to explain it.

6. Conclusion

What did we observe in our analysis and what can we conclude? Firstly, we observed that the position in the hierarchy of the educational system in Belgium has to be modelled. The tracks and grades can be viewed as tools to manage pupil heterogeneity and indicate a different exposure to school subjects. Both variables appear to have a significant effect. Furthermore, it is likely that this position in the school system plays as a mediator between pupil background characteristics and achievement. Secondly, traditional pupils' characteristics as socio-economic origin, ethnic origin and language spoken at home still remain relevant. All of these variables need to be included in research on educational outcomes. Moreover, the continuously observed ethnic gap should be tackled in order to assure equity. Nevertheless, the achievement gap cannot be explained by differences of population provided for schooling in both communities. Indeed, the gap barely goes down when pupils' characteristics are included in the model. Moreover, we found that the effect of grade repetition and tracking was significantly different across schools but this difference is not explained by school membership. The hypothesis that the different use of grade repetition and tracking in the two communities may explain the achievement gap has not been verified.

With regard to the school composition, we can say that both the academic composition and the socio-economic composition have additional effects on pupils' achievement. Our model shows that pupils tend to achieve better in schools attended by a pupil population with higher socio-economic background and higher academic performance. It is worth noting that these compositional effects are not trivial (according to the 39-point rule, we are speaking in years of schooling). In terms of community comparison, it seems that a part of the achievement gap between the French-speaking and the Dutch-speaking schools could be explained by the academic and socio-economic compositions of schools. Finally, by adding an interaction term to our analysis, we observed that the academic and socio-economic compositions have a different effect in both parts of the country. In fact, while both academic and socio-economic compositions mostly affect the French-speaking schools, the academic one has a more important effect in the Dutch-speaking schools.

Caution is required, however, in interpreting our results because some limitations are worth noting. Although we use the position in the educational system as a proxy to measure prior achievement, this proxy is far from a perfect tool. Moreover, the simplistic structure of the data (pupils in schools) can make composition effects appear at the school when they can play at least partially at the class level. Finally, the high correlations between types of compositions make it difficult to disentangle their respective effects. In spite of these limitations, PISA remains the only database that enables us to compare effect between Belgian communities. However, further research is required to strengthen our results and to explore whether the analogy between segregation and composition is useful elsewhere. PISA might not be appropriate for systems where the absence of educational stratification removes such an imperfect proxy for pupil prior achievement.

We can conclude that, in both Belgian communities, stimulating academic and social mix seems to remain a legitimate policy objective although the type of school mix to stimulate can be different. Firstly, segregation hampers the situation of pupils from disfavoured classes by adding a negative compositional effect. Secondly, the ensuing achievement gap reinforces and legitimizes the stigmatization of low achievers, which includes migrant pupils. At the same time, we can stress the necessity to break the cascade logic in school choice and to limit the reproduction of social stratification caused by early tracking and the systematic use of grade repetition. The waterfall structure of both educational systems produces a pupil population with extreme differences in skills and reproduces social inequality. The Belgian educational system as a whole does not perform well when promoting equal opportunities.

7. Acknowledgements

The authors wish to express their gratitude to Andrew Crosby, Morgane Giladi, Laurie Hanquinet, Mathieu Ichou, Dirk Jacobs, Noémie Le Donnée and Meenakshi Parameshwaran for helpful comments.

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'.

8. Notes

1 As said before, the trajectories of Belgian students at school are characterized by grade repetition and tracking. Academic composition should consequently cover both dimensions. However, the distribution of the variable and the organization of the school institution make difficult to include it. Actually, students can follow general or vocational streams and these streams are organized in the same schools or in separated schools. In the database we found a majority of schools organizing only one stream, and the vocational density measured as a proportion of students in this stream presents consequently two modes (0% or 100%). Nevertheless, even though grade repetition and tracking are differently used, vocational and delay densities are moderately to highly correlated in both communities.

2 The PISA database is provided with a set of sampling weights in order to deal with the over-sampling and under-sampling of some strata of the population, to take the potential lack of accuracy in sampling frame into account and to adjust for school and student non-response (OECD, 2012). The literature emphasizes that a proper use of weight needs some scaling of the conditional level-one weights (Pfeffermann *et al.*, 1998). When the analyst is interested in point estimates and when the cluster size is larger than 20, 'method 2' seems to be the more suitable (Carle, 2009). In this method, the weight of the student i in the school j is rescaled with a factor of $n_j / \sum w_{i|j}$, where n_j is the number of students in school j and $w_{i|j}$ their weights. Such a method is used as default by MLwiN. For the level-two, standardized weights will be included.

3 The 95% confidence intervals are given in parentheses.

4 Change 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. The deviance difference has to exceed a chi-square distribution with the number of extra parameters as degrees of freedom. AIC = deviance + 2p and BIC = deviance + ln(n) *p. Which population size has to be used is not clear in multilevel modelling. 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 two by extra parameters. The residuals have also been screened.

9. References

Agirdag, O., Van Houtte, M. and 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.

Agirdag, O., Van Avermaet, P. and 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.

Baye, A. and Demeuse, M. (2008) Indicateurs d'équité éducative. Une analyse de la ségrégation académique et sociale dans les pays européens [Indicators of educational equity. An analysis of the academic and social segregation in European countries], *Revue française de pédagogie*, 165 (4), 91–103.

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

Danhier, J. (2013) Simultaneous effects of different school compositions on student achievement: the case of Flanders (Brussels, Université libre de Bruxelles). Unpublished work.

Demeuse, M. and Friant, N. (2010) School segregation in the French Community of Belgium. In *International Perspectives on Countering School Segregation* (Antwerpen/Apeldoors, Garant), 169–187.

Dronkers, J. (2010) Positive but also negative effects of ethnic diversity in schools on educational performance? An empirical test using cross-national PISA data. Paper presented at the Integration and Inequality in Educational Institutions Conference, University of Bremen, 24 September 2010.

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

Dupriez, V. and Vandenberghe, V. (2004) L'école en Communauté française de Belgique: de quelle inégalité parlons-nous ? [Schooling in the French Community of Belgium: what kind of inequality are we talking about ?], *Les cahiers de recherche du Girsef*, 27, 1–28.

Duru-Bellat, M., Le Bastard-Landrier, S. and Piquée, C. (2004) Tonalité sociale du contexte et expérience scolaire des élèves au lycée et à l'école primaire [Social 'school mix' and the experience of high school and primary school pupils], *Revue française de sociologie*, 45 (3), 441–468.

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

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

Hindriks, J. and Verschelde, M. (2010) L'école de la chance [The school of the chance], *Regards économiques*, 77.

Hindriks, J., Verschelde, M., Rayp, G. and Schoors, K. (2009) Analyse des disparités régionales d'éducation en Belgique [Analysis of regional educational disparities in Belgium]. Paper presented at the 18e Congrès des Economistes belges de Langue française: Quel Etat pour quelles performances économiques? Conference, Université libre de Bruxelles, 26 November 2009.

Hirtt, N. (2008) *Pourquoi les performances PISA des élèves francophones et flamands sont-elles si différentes?* [Why are the PISA performance of French-speaking and Dutch-speaking students so different?] (Bruxelles, Aped). Available at: http://www.skolo.org/spip.php?article452&lang=fr (accessed 23 July 2013).

Hox, J. (2010) *Multilevel Analysis. Techniques and Applications* (2nd edn) (New York, Routledge).

Jacobs, D. and Rea, A. (2011) *Gaspillage de talents: Les écarts de performances dans l'enseignement secondaire entre élèves issus de l'immigration et les autres d'après l'étude PISA 2009* [Waste of talent: The performance gap in secondary education among migrant students and others according to PISA 2009] (Bruxelles, Fondation Roi Baudouin).

Jacobs, D., Rea, A., Teney, C., Callier, L. and Lothaire, S. (2009) *L'ascenseur social reste en panne* [The social elevator is still broken] (Bruxelles, Fondation Roi Baudouin).

Kristen, C. and Granato, N. (2007) The educational attainment of the second generation in Germany Social origins and ethnic inequality, *Ethnicities*, 7 (3), 343–366.

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

Levels, M. and Dronkers, J. (2008) Educational performance of native and immigrant children from various countries of origin, *Ethnic and Racial Studies*, 31, 1404–1425.

Maas, C. J. M. and 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. (2005) Accounting for immigrant non-immigrant differences in reading and mathematics in twenty countries, *Ethnic and Racial Studies*, 28 (5), 925–946.
Monseur, C. and Lafontaine, D. (2009) L'organisation des systèmes éducatifs: quel impact sur l'efficacité et l'équité? [The organization of education systems: what impact on efficiency and equity?]. In Dumay, X. and Dupriez, V. (Eds) *L'efficacité dans l'enseignement* (Bruxelles, De Boeck), 141–163.
OECD (2010) *PISA 2009 at a glance* (Paris, OECD Publishing).
OECD (2012) *PISA 2009 technical report* (Paris, OECD Publishing).
Opdenakker, M.-C. and Van Damme, J. (2000) The importance of identifying levels in multi-level 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. and 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. and 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.
Portes, A. and MacLeod, D. (1996) Educational progress of children of immigrants: the roles of class, ethnicity, and school context, *Sociology of Education*, 69 (4), 255–275.
Rasbash, J., Steel, F., Brown, W. J. and Goldstein, H. (2012) *A user's guide to MLwiN, v2.26* (University of Bristol, Centre for Multilevel Modelling).
Raudenbush, S. W. and Willms, J. D. (1995) The Estimation of School Effects, *Journal of Educational and Behavioral Statistics*, 20 (4), 307–335.
Rubin, D. B. (1987) *Multiple Imputation for Nonresponse in Surveys* (New York, Wiley).
Rumberger, R. W. and 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.
Thrupp, M. (1999) *Schools Making a Difference* (Buckingham, Open University Press).
Thrupp, M., Lauder, H. and Robinson, T. (2002) School composition and peer effects, *International Journal of Educational Research*, 37 (5), 483–504.
Van den Noortgate, W., Opdenakker, M.-C. and Onghena, P. (2005) The effects of ignoring a level in multilevel analysis, *School Effectiveness and School Improvement*, 16 (3), 281–303.
Van Ewijk, R. and 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. and Sleegers, P. (2010b) The effect of peer socioeconomic status on student achievement: a meta-analysis, *Educational Research Review*, 5 (2), 134–150.
Vandenberghe, V. (2011) Inter-regional educational discrepancies in Belgium. How combat them?, *Educational Divergence – Why Do Pupils Do Better in Flanders than in the French Community?* (e-book) (Brussels, Re-Bel Initiative), 5–25.
Wu, M. (2005) The role of plausible values in large-scale surveys, *Studies In Educational Evaluation*, 31 (2–3), 114–128.
Zhang, Z., Charlton, C. M. J., Parker, R. M. A., Leckie, G. B. and Brown, W. J. (2012) *R2MLwiN* (University of Bristol, Centre for Multilevel Modelling).

Correspondence
Julien Danhier
ULB - GERME (CP.124)
44 avenue Jeanne
B-1050 Brussels
Belgium
Email: jdanhier@ulb.ac.be