WORKING PAPER SERIES

2016/01

# Modelling multiple measures of compositional effect: does factorisation simplify the picture in Belgium?

**JULIEN DANHIER**

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

Université libre de Bruxelles (ULB)

jdanhier@ulb.ac.be

**Abstract**

Both the Dutch- and French-speaking Belgian educational systems are characterized by a high level of segregation, placing a strong emphasis on academic, socioeconomic, and ethnic make-up of their student population. In this context, school composition is expected to have an effect, however the choice of measurement is not straightforward. Multilevel analysis was conducted on a Dutch-speaking subsample of 4,618 students in 156 schools using PISA 2012 data in order to explore different modelling options of multiple compositions. Factorisation is then explored, in combination with multilevel modelling, to see whether it provides an adequate strategy to deal with the limitations of multiple compositions variables. In addition, a French-speaking subsample of 2,759 students in 95 schools was used to test measurement invariance, check whether the construct has the same meaning in both groups, and assess whether it can be used in multilevel models.

**Keywords:** Education, Achievement gap, Compositional effect.

**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 Andrew Crosby, Perrine Devleeshouwer, Morgane Giladi, Marie Graux, Dirk Jacobs, Laurie Hanquinet, Robie Kaelen, Emilie Martin and Letitia Zwickert for helpful comments and assistance.

## 1. Introduction

In much of the literature, segregation is identified as a detrimental process that hinders optimal school achievement for all pupils. Its impact is usually considered to be linked to a compositional effect. Following the lead of Dumay and Dupriez (2008, p. 440), the compositional effect can be defined as the "impact of pupils' aggregated characteristics (socioeconomic status, sociocultural capital, prior achievement, etc.), once these variables have been taken into account at the individual level". In other words, the compositional effect is the measure of the way in which student grouping influences individual performance.

Researchers have repeatedly observed that an "unfavourable" composition (e.g. lower socioeconomic composition, lower averaged performances, or lower proportion of migrants) is significantly associated with lower academic performance (Condron, 2009; De Fraine, Van Damme, & Onghena, 2002; Dumay & Dupriez, 2008; Duru-Bellat, Le Bastard-Landrier, & Piquée, 2004; Opdenakker & Van Damme, 2001; Rumberger & Palardy, 2005; Sykes & Kuyper, 2013; Timmermans, Doolaard, & de Wolf, 2011), or lower aspirations (Dupont & Lafontaine, 2011; Van Houtte & Stevens, 2009), as well as other dependent variables such as, the quality of the friendship (Demanet, Agirdag, & Van Houtte, 2012). In other words, two similar students will have different results depending on the school they attend: the one in a less advantaged school will have lower results than the one in an advantaged school. Consequently, a strong composition effect means a large gap between students from schools with different compositions.

From a substantive point of view, the compositional effect remains vague. A school's as well as teacher's characteristics are linked to their school population (Thrupp, 1999). Pupils attending schools with high levels of working class and immigrant children collectively underachieve because they are confronted with a cocktail of potentially problematic influences within the classroom and school. In their literature review, van Ewijk and Sleegers (2010a) offer three categories of explanation: the compositional effect can result from direct peer interactions (discussions, motivation, disruptions, or, for ethnic composition, tensions between races or language difficulties), from teacher practices (adjustments in teaching style or expectations), and from the school quality (problems in human resources management or funding). In other words, school compositional effects actually refer to student characteristics and interactions and a range of factors associated with teachers and schools hosting a specific segment of the population.

Regarding its operationalisation, how composition should be measured and modelled is still discussed. Actually, Thrupp, Lauder, and Robinson (2002) have proposed a list of ten conditions that an ideal model should fulfil, among which we find a combination of different types of compositions (academic, socioeconomic, etc.) and different measures of these types. Although we subscribe to such a suggestion, its implementation is not straightforward. Strong inter-correlations between such measures force the researcher to test the effect of each

measure in separated models and conclusions are strongly dependent on the models he explored. In secondary education of the Netherlands, Sykes and Kuyper (2013) found, for example, a significant effect of socioeconomic composition on achievement while ethnic composition became nonsignificant when both types of composition were simultaneously modelled. However, in his inaugural lecture at Maastricht University, Dronkers (2010) suggested the exact opposite. Using a multilevel analysis of twelve countries of the PISA 2006 survey, he found ethnic diversity of secondary schools hampers the educational performances of both students with an immigrant background and native students.

Let us note that the latter results were widely reported by the Dutch and Flemish press, and had considerable impact on the public debate. Consequently, opposition party MPs in the Flemish Parliament extensively referred to this study to criticize any policy that attempted to regulate enrolment policies based on social or ethnic factors. Although this debate seems technical, identifying which composition matters is capital from a substantive point of view, especially as it can define public policy initiatives focusing on some specific desegregation. This paper seeks to investigate how composition should be operationalised and measured in the context of multilevel modelling. However, from a technical point of view, the exercise is far from easy. In this article, we make an attempt to combine different types of composition. We explore different methods of modelling and simultaneous modelling of multiple types of composition in the Dutch-speaking Community of Belgium with data from PISA 2012. This education system is particularly relevant as it is deeply marked by school segregation and a strong effect of composition have been observed (Danhier, Jacobs, Devleeshouwer, Martin, & Alarcon, 2014). Next, factorisation is explored, in combination with multilevel modelling, as a simple solution to provide an adequate strategy to deal with the limitations of multiple compositions variables. Measurement invariance is tested in a second Belgian linguistic community to check whether the construct has the same meaning in both groups and can be used in multilevel models. However, we will see that factorisation provides little improvement if one wishes to compare compositional effects between educational systems. Concerning which composition matters, our analysis requires us to consider the results with caution.

## 2. The effect of multiple compositions

Raudenbush and Willms (1995) have suggested that the measurement of composition requires the inclusion of a full set of student background variables, including prior achievement. On the basis of conceptual and methodological issues in measuring composition, Thrupp, Lauder, and Robinson (2002) have proposed a list of ten conditions that an ideal model should fulfil, among which we find: a full set of entry-level variables, a combination of different types (academic, socioeconomic, etc.), and measures of composition. Concerning the latter, Chan (1998) makes a distinction between several composition paradigms, including an additive paradigm (the summation of the lower-level units) and a dispersion paradigm (the dispersion among lower-level units). The distinction between different types and paradigms of composition

requires multiple variables to be entered at the aggregated level. The choice of including these variables is not straightforward and researchers have not been consistent with their selection.

When taking into account student characteristics and prior achievement, researchers have repeatedly observed significant negative effects of certain composition variables on students' results 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. In the French-speaking part of Belgium, at the end of primary education, Dumay and Dupriez (2008) observed effects of academic and sociocultural composition, even after controlling for individual abilities, socioeconomic background, and language characteristics. However, they acknowledged the difficulty of disentangling the composition with a relevant effect because of the high correlation between different types of composition. In the secondary education system 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 (HAVO). 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. On achievement gains in the United States, Condron (2009) observed an effect of ethnic composition, but not of socioeconomic composition, in primary education (ECLS-K).

Other researchers made different attempts to model multiple composition without being able to control prior achievement. In Norway, Fekjær and Birkelund (2007) were surprised to find that, when controlling for academic composition, the effect of ethnic composition on students' achievement becomes significantly positive. Szulkin and Jonsson (2007) have shown, in a Swedish context, that ethnic density in schools had a negative effect on educational achievement, but also socioeconomic composition when it is measured by the proportion of students with at least one parent with a university degree. In a comparison study of twelve OECD countries (including Belgium), Dronkers (2010) identified significant effects of additive socioeconomic and ethnic composition, but also different negative effects of diversity. Agirdag, Van Houtte and Avermaet (2011) have found that the ethnic compositional effect vanished in the Flemish Community when prior academic achievement and socioeconomic background were taken into account. In Spain and Italy, Azzonlini *et al.* (2012) have shown significant effects of socioeconomic composition, but not of ethnic composition. With data from PISA 2009, Danhier and Martin (2014) have shown that academic and socioeconomic compositions have

an effect in Belgium, but have observed an interaction effect between composition and community membership signifying the effect is different depending on the Belgian linguistic communities.

This overview of the different attempts to model multiple compositions highlights the lack of a common conclusion about which composition matters, especially when different national contexts are considered. In the following chapter we will explore how to deal with multiple compositions in the Belgian Dutch-speaking Community and test factor models in order to use a statistical construct in an international comparative multilevel framework. As the Belgian French-speaking Community system shares similarities with the Dutch-speaking one, it will serve as a test for the invariance of the construct.

## 3. Multiple compositions in the Dutch-speaking Community

### THE BELGIAN CASE

Since 1989, the Belgian educational field includes three separate educational systems, reflecting the division of the country into three linguistic communities. The two major communities (Dutch and French-speaking) provide schooling for the majority of the student population (respectively about 55 and 44 percent of the pupils in Belgian schools). The existing national literature places particular emphasis on examining the attainment discrepancies between these two communities. Various surveys, such as 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. Although no clear consensus has been found to explain this intercommunity achievement gap (Hindriks & Verschelde, 2010; Hirtt, 2008; Vandenberghe, 2011), some researchers agree on one point: the Dutch-speaking Community might do well in the educational attainment rankings, but as far as equal opportunities are concerned, both the French and the Dutch-speaking schools have poor results.

Even though they are managed separately, analyses highlight that both educational systems are characterized by an important level of segregation (Baye & Demeuse, 2008; Danhier et al., 2014; Verschelde, Hindriks, Rayp, & Schoors, 2010) originating in the structure of the systems (Demeuse & Baye, 2008). The Belgian constitution, which applies to both systems, guarantees freedom of education. In practice, this leads to the existence of a quasi-market where parents are free to choose their children's schools and where schools are in competition with each other, implementing strategies to attract pupils, whose the number determines the amount of public subsidization. In addition, both systems organize different tracks in secondary education and widely use grade repetition to manage pupils' academic heterogeneity. This particular combination probably explains the high level of segregation in Belgian communities (Lafontaine & Monseur, 2011).

## SAMPLE AND MULTILEVEL MODELLING OF COMPOSITION

PISA (standing for “Programme for International Student Assessment”) is a research project led by the OECD which aims to assess students’ ability “to use their knowledge and skills to meet real-life challenges” (OECD, 2014). For our purpose, the Belgian Dutch-speaking subsample was selected. In line with the two-stage stratified sampling design used by OCDE, schools were sampled according to their size (after being separated into explicit strata, namely form of education, tracks, funding and ISCED levels, and ordered by implicit strata, namely grade repetition and percentage of girls) and students (15-year-old students from grade 7 or higher) were randomly sampled in selected schools to obtain 35 respondents per school (or less if there were not enough valid students) (OECD, 2014). In our sample, regular full-time education was selected. Consequently, part-time vocational education (2 schools) and education for students with special needs (6 schools) were excluded. The rate of item nonresponse was 2.3% at the student level. Due to the limited non-response rate, listwise deletion was used in order to make the analysis simpler without introducing too much bias (Graham, 2009). It is important to note that, in order to assure a stable basis to compute compositional effects, schools with fewer than 10 respondents were excluded (namely, 10 schools, mainly screened as outliers).The final sample covered 4,618 respondents in 156 schools.

Multilevel modelling is a technique used to analyse hierarchical data. PISA data is hierarchical, not only because educational data is typically hierarchical (students are clustered in schools), but also because of its two-stage sampling design. Consequently, students in the same school are likely to be more similar to each other than to students from other schools. In the absence of the assumed independence of observations, standard statistical tests lead to a strong underestimation of the parameters’ standard errors and consequently to spurious significant effects (Hox, 2010). Multilevel techniques are not the only way to deal with such a structure, but they allow for the modelling of the effects 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[1]) has been included, the extra effect of composition is simply measured by the effect of the aggregation of individual variables. Technically, MLwiN (Rasbash, Steel, Brown, & Goldstein, 2012) was used to perform the multilevel analyses in R (inspired from Zhang, Charlton, Parker, Leckie, & Brown, 2012). Students were modelled as the first level and

[1] With grand mean centring, the level-one coefficient is a blend of intra- and inter-school relations that cannot be disentangled. This feature is an advantage for testing whether the compositional effect is significant, that is, whether composition has an additional effect. Due to the correlation between 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).

schools as the second one.

The plausible values (PV) for mathematical achievement were used as dependent variables. In order to reconcile the limited time that is available to test each student and the need to cover a wide range of domain knowledge, five plausible values need to be used. Actually, PISA provides a battery of items that are characterized by a binary result (success/failure). Each student is tested on one of the subsamples of the whole battery. Based on the students' results, an items' difficulty and students' ability can be computed, using a method called Rasch model. Because only an incomplete item subsample is administered, scores are computed with a relative uncertainty. After the analyses have been separately conducted on each PV, the results must be properly combined so as to obtain unbiased estimates (Rubin, 1987; Schafer, 1999).

Following the previously presented literature, the composition modelling requires a full set of student background variables. However, PISA does not provide any measure of prior achievement. Nevertheless, in systems where grade repetition and orientation largely define the pupils' position in the hierarchy of the educational system, these variables are expected to hold information about prior achievement. In our sample, almost 27% of the students have at least repeated a grade and more than 53% are enrolled in vocational education. In fact, this position is a direct result of decisions based on previous achievement. This is not perfect, however, because the decision also depends on class structure and teacher subjectivity. In the Dutch-speaking Community, Agirdag, Van Avermaet and Van Houtte (2013) made such an assumption. A dummy variable for vocational education and a delay variable (the distance in years between a 15-year-old's theoretical grade and his actual one) were entered at the student level in the analyses. Other background variables are available in PISA in order to construct the student-level model. Students' first or second generation migrant status is controlled, as well as their origin. In the operationalization, each student with at least one foreign parent was defined as having an immigration background. Nine different categories of migrant origins are defined in the database: France, Germany, Netherlands, Turkey, Sub-Saharan country, North Africa country, Other western European countries, and European countries. Some origins are at the country level while others are pre-grouped or put in the "other" category. Thus, such a categorization lacks coherence. A socioeconomic status variable (student index of economic, social and cultural status - ESCS) is available in the database. This index summarizes the information from three sources: the highest level of parental occupation, the highest level of parental education, and the number and kinds of home possessions. A reverse coded version of the index was used to assess socioeconomic origin (disadvantaged students having a high

***Table 1: Bivariate correlations between compositional variables (Dutch-speaking Community)***

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1. Academic density** | 1 | | | | | |
| **2. Academic diversity** | **0.639** | 1 | | | | |
| **3. Socioeconomic density** | *0.700* | 0.541 | 1 | | | |
| **4. Socioeconomic diversity** | 0.291 | 0.257 | **0.341** | 1 | | |
| **5. Ethnic density** | *0.567* | 0.575 | *0.438* | 0.392 | 1 | |
| **6. Ethnic diversity** | 0.560 | 0.593 | 0.404 | 0.360 | **0.872** | 1 |

ESCS). The language spoken at home[2], and the gender are also included.

In order to model different types (academic, socioeconomic, and ethnic) and paradigms of composition (additive and dispersion), six variables have been considered. These variables are classically used in the literature presented previously. First, to measure socioeconomic composition, we computed the school's mean and variance on the ESCS, provided in the database. Second, two variable to measure academic density and academic diversity. Delay average and variance at the school level were used to measure both additive (density) and dispersion (diversity) paradigms for academic composition. Although the delay variable has only few categories, its aggregation, used to measure composition, is expected to have a significant effect in Belgian systems (Danhier & Martin, 2014). For the measure of academic diversity, the limited number of categories could be, however, problematic. Finally, ethnic compositions were also explored. The proportion of immigrants with non-European origins were used to measure the ethnic density. Because the proportion has a skewness of 1.41 and a kurtosis of .66 with scores collapsed around .1 (but with several scores ranging from .3 to .8), arcsin squared transformations were advised. Following Dronkers and van der Velden (2013), the Herfindal index for ethnic diversity was used as a measure of ethnic dispersion. This index varies between zero and one, with zero indicating perfect homogeneity and one, perfect heterogeneity.

Let us note that the PISA database is provided with a set of sampling weights in order to deal with the over- 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 nonresponse (OECD, 2014). The literature emphasizes that a proper use of weight needs some scaling of the conditional level-one weights (Asparouhov, 2006; Pfeffermann, Skinner, Holmes,

[2] The Flemish dialects were recoded as "the same language", since they are considered to be closely related to the school language. It should also be noted that the variable has a substantial proportion of missing values (over 11%). The main reason is that multiple languages are coded as invalid. According to preliminary analyses, the influence of the "invalid" category is quite similar to the "different language" category. Consequently, two dummies were included in the model: one for students who speak another language at home and a second one for the invalid answers.

Goldstein, & Rasbash, 1998). Method 2 consists in making the sum of weights equal to the number of students in each school. Actually, this method is the most suitable when the analyst is interested in point estimates, when the cluster size is larger than 20 (Carle, 2009), and is used by default by MLwiN. For the level one, the latter conditional weights were used and for the level two, standardized weights were included. MLwiN provides Sandwich estimators and performs weighted multilevel analysis using the IGLS algorithm (Centre for Multilevel Modelling, 2011).

Some limitation regarding the PISA database and its use to measure composition with multilevel techniques are worth noting. Firstly, misspecifications of the models are possible. The measurement of composition requires a full set of individual variables, including prior achievement, which implies longitudinal data most of the time. The underlying idea is to measure the effect of the student grouping, one has to control for all the potential individual characteristics that are associated with achievement. The absence of variables highly determinant for achievement and correlated with included explanatory variables results in the attribution of the effect of the omitted variables to the currently included variables. In other words, without such a full set of individual variables, the compositional effect might be a spurious effect due to the omission of individual characteristic. In a meta-analysis, van Ewijk and Sleegers (2010a, 2010b) have found that the effects of socioeconomic and ethnic composition are greatly overestimated when prior achievement is not included, although the effect is large but not significant for ethnic composition. As stated before, only delay and tracking are available to control for prior achievement. Additionally, other variables known to be linked to achievement have to be considered. ESCS, language spoken at home, ethnic origin, and sex have been included in all models. In order to discuss results concerning composition, we have to explicitly assume that student background variables account for prior achievement and at least limit the omission bias in the measure of composition.

Secondly, the measurement of the compositional effect requires the identification of all relevant levels. Students are clustered in classes, which in turn are clustered in schools. Ignoring this hierarchical structure could have important consequences for the analysis. When intermediate levels are ignored, the variance is complexly distributed at the other levels which can result in false positives when trying to identify significant variables at these levels (Meyers & Beretvas, 2006; Opdenakker & Van Damme, 2000; Van den Noortgate, Opdenakker, & Onghena, 2005). As a consequence, working with analyses from a hierarchical modelling framework is particularly recommended. The lack of a class identification does not allow to model this level. The potential effect of class composition will be consequently caught by variables at the student and the school levels.

Finally, let us note that the compositional effect is essentially a statistical concept and its measure remains in the centre of an open technical debate. Researchers have recently claimed that both sources of error have to be taken into account by applying what they call “doubly latent” models (Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009, 2012). Such a modelling uses multilevel structural equation modelling combining two statistical frameworks,

namely, multilevel modelling and structural equation modelling. Pokropek (2015) found that the use of multilevel modelling in large scale assessments without taking measurement errors into account leads to an overestimation of the compositional effect. Televantou et al. (2015) estimated the advised “doubly latent” models on mathematical achievement in fourth grade in Cyprus but did not find any significant compositional effect, in contrast with the significant effect found when they followed the classical approach. These recent developments uncover promising methods that need to be explored on different data and grades in different countries. On the other hand, they contribute to making the statistical approach of composition more complex. In this article, we limit ourselves to the classical approach in the multilevel framework and incorporate well-known factor analyses, but recognize the interest of other approaches.

## CORRELATIONS BETWEEN THE MULTIPLE MEASURES OF COMPOSITION

Table 1 presents correlations between the six composition variables. Some compositions are highly correlated as eight of the fifteen correlations are about .5 or higher. Some high correlations are observed between paradigms of the same type. There is a high correlation between ethnic density and diversity (.85). The crude categorization of origins may explain the lack of difference between both paradigms, but the reason may also be more fundamental: it may be that both variables measure the same concept. This is what Schaeffer (2013) has observed based on German data: it may be that the Herfindal index simply reflects minority concentration, in other words, the density paradigm. For academic composition, the correlation between density and diversity reaches .64. Again, one might argue whether these two variables actually measure two separate paradigms. Delay variance is clearly a weak measure of academic diversity, given that this variable only has two major occurrences (0 and 1 year of delay). However, although in the majority of schools, the higher the mean of delay is, the higher is its variance, there are some schools with a high average but a low variance and vice-versa. By contrast, for socioeconomic composition, the correlation between density and diversity is low, although schools with a more disadvantaged population tend to be a bit more diverse.

Another facet of the problem concerns correlations between types of composition. Although socioeconomic and ethnic densities are only weakly correlated, they are both highly correlated with academic composition. In other words, we observe higher means of delay in schools with a more disadvantaged population or with a higher proportion of non-European migrants. The low correlation between socioeconomic and ethnic densities is due to disadvantaged schools with low proportions of migrants, probably in places other than big cities. A consequence of this problem is that the choice of measures for composition included in the model can dramatically change the coefficients at the second level.

***Table 2: Multilevel models***

| *Parameters* | *Model 1* | *Model 2* | *Model 3* | *Model 4* | *Model 5* | *Model 6* | *Model 7* | *Model 8* | *Model 9* |
|---|---|---|---|---|---|---|---|---|---|
| **Fixed part** | | | | | | | | | |
| Intercept | 533 (3.34) *** | 538.7 (2.59) *** | 535.8 (2.88) *** | 538.8 (2.62) *** | 533.4 (3.29) *** | 535.7 (3.26) *** | 535.4 (3.13) *** | 539.8 (2.38) *** | 539. (2.36) *** |
| Female | -21.9 (2.36) *** | -22.23 (2.40) *** | -22.09 (2.35) *** | -22.48 (2.33) *** | -21.94 (2.35) *** | -21.73 (2.35) *** | -21.7 (2.35) *** | -22.49 (2.36) *** | -22.49 (2.35) *** |
| ESCS (-) | -10.7 (1.59) *** | -10.32 (1.60) *** | -10.59 (1.59) *** | -9.4 (1.65) *** | -10.64 (1.60) *** | -10.73 (1.60) *** | -10.79 (1.59) *** | -9.59 (1.66) *** | -9.64 (1.66) *** |
| First generation | -13.7 (6.65) * | -13.19 (6.59) * | -12.97 (6.55) * | -13.44 (6.61) * | -13.55 (6.64) * | -13.36 (6.66) * | -13.44 (6.67) * | -13.15 (6.58) * | -12.88 (6.55) * |
| Origin: France | -24.15 (13.62) | -22.42 (13.23) | -23.34 (13.68) | -24.97 (13.2) | -23.95 (13.51) | -23 (13.37) | -22.67 (13.49) | -23.47 (13.15) | -23.26 (13.23) |
| Origin: Germany | -16.98 (11.19) | -15.55 (11.24) | -16.06 (11.21) | -15.64 (11.22) | -16.5 (11.16) | -16.16 (11.19) | -15.28 (11.27) | -15.15 (11.24) | -14.98 (11.24) |
| Origin: Netherlands | 4.82 (5.20) | 5.57 (5.37) | 5.47 (5.25) | 5.19 (5.16) | 4.9 (5.21) | 5.06 (5.18) | 5.94 (5.22) | 5.57 (5.31) | 5.76 (5.32) |
| Origin: Turkey | -28.41 (9.28) ** | -25.86 (9.23) ** | -25.74 (9.20) ** | -25.46 (9.19) ** | -28.07 (9.25) ** | -25.04 (9.31) ** | -25.95 (9.19) ** | -24.72 (9.2) ** | -23.95 (9.21) ** |
| Origin: Sub-Saharan | -27.35 (8.51) ** | -25.79 (8.46) ** | -26.48 (8.45) ** | -27.45 (8.33) *** | -27.26 (8.49) ** | -26.41 (8.41) ** | -25.85 (8.41) ** | -26.35 (8.38) ** | -26.13 (8.37) ** |
| Origin: North African | -11.06 (7.30) | -7.19 (6.94) | -9.27 (7.22) | -9.16 (7.2) | -10.07 (7.3) | -7.27 (7.76) | -7.49 (7.69) | -7.18 (7.03) | -6.97 (7.01) |
| Origin: Other W-European | -13.96 (8.43) | -11.14 (8.20) | -13.18 (8.47) | -12.59 (8.23) | -13.58 (8.38) | -12.01 (8.35) | -11.76 (8.43) | -11.13 (8.19) | -11.2 (8.21) |
| Origin: Eastern European | -8.09 (8.51) | -6.41 (8.50) | -6.79 (8.50) | -6.56 (8.4) | -7.94 (8.50) | -6.73 (8.42) | -6.3 (8.36) | -5.96 (8.45) | -5.66 (8.47) |
| Origin: Other | -8.32 (8.75) | -7.11 (8.37) | -6.84 (8.35) | -8.2 (8.62) | -8.18 (8.72) | -6.57 (8.65) | -6.22 (8.63) | -7.42 (8.42) | -6.92 (8.3) |
| Other Language | -22.87 (4.44) *** | -22.82 (4.38) *** | -22.59 (4.35) *** | -22.84 (4.35) *** | -22.71 (4.45) *** | -21.79 (4.44) *** | -21.95 (4.41) *** | -22.81 (4.34) *** | -22.69 (4.31) *** |
| Invalid Language | -18.44 (6.23) ** | -18.89 (6.10) ** | -18.21 (6.14) ** | -18.68 (6.35) ** | -18.42 (6.28) ** | -17.08 (6.48) ** | -17.31 (6.41) ** | -18.91 (6.23) ** | -18.74 (6.19) ** |
| Delay | -54.18 (2.93) *** | -51.31 (2.98) *** | -53.42 (2.93) *** | -52.85 (2.89) *** | -54.13 (2.93) *** | -54.09 (2.92) *** | -54.02 (2.96) *** | -51.36 (2.95) *** | -51.45 (2.95) *** |
| Vocational | -75.72 (4.74) *** | -71.95 (4.58) *** | -73.45 (4.56) *** | -68.9 (4.67) *** | -75.58 (4.73) *** | -76.71 (4.63) *** | -75.86 (4.64) *** | -68.87 (4.72) *** | -68.68 (4.67) *** |
| Academic density | | -81.6 (12.0)*** | | | | | | -51.5 (15.8) | -40.4 (14.9) |
| Academic diversity | | | -140 (23.3)*** | | | | | | -49.1 (19.8) |
| Socioecon. density | | | | -49.8 (7.0)*** | | | | -28.7 (9.6) | -26.2 (9.8) |
| Socioecon. diversity | | | | | -36.0 (17.0)* | | | | - |
| Ethnic density | | | | | | -45.2 (12.9)*** | | - | - |
| Ethnic diversity | | | | | | | -62.3 (14.9)*** | | - |
| **Random part** | | | | | | | | | |
| Student variance | 3706.9 (194) | 3700.6 (193) | 3702.8 (193) | 3698.7 (193) | 3706.1 (193) | 3707.3 (193) | 3706.3 (193) | 3698.5 (193) | 3698.2 (192) |
| School variance | 987.4 (164.9) | 570.4 (107.7) | 699.0 (132.8) | 599.3 (115.34) | 945.2 (170.6) | 830.8 (171.5) | 807.8 (162.2) | 501.4 (93.26) | 478.8 (94.64) |
| **Goodness of fit** | | | | | | | | | |
| Deviance | 52007.6 | 51926.5 | 51955.7 | 51930.5 | 52000.5 | 51984.3 | 51979.3 | 51907.5 | 51901.3 |
| AIC | 52053.6 | 51974.5 | 52003.7 | 51978.5 | 52048.5 | 52032.3 | 52027.3 | 51957.5 | 51953.3 |
| Level-two $R^2$ | 80.53 | 88.75 | 86.21 | 88.18 | 81.36 | 83.62 | 84.07 | 90.11 | 90.55 |

Significance: *** = .001, ** = .01, * = .05, "-"= nonsignificant and removed from de model.
Student outliers (35541, 33946) and school outliers (3145, 3080) on multiple plausible values were modelled.

## RESULTS OF MULTILEVEL ANALYSES

Table 2 presents the results from the multilevel analyses. The null model is not shown, but let us note that 52 % of the variance lays at the school level as the student-level and the school-level variances reach 5072.2 (SD 281.6) and 5550.6 (SD 704.2), respectively. Since the article focuses on compositional effects, only the coefficients of these measures are discussed. Outliers have been modelled in all the models. The model 1 is the baseline model to which the fit of the successive models are compared. Due to an important differential recruitment between schools, the school-level variance decrease drastically once student-level variables are entered in the model. Now, 21 % of the variance lays at the school-level. Let us note that after adding individual variables, the level one pseudo-$R^2$ reaches 26.4. Compared with what is often observed when prior achievement is used in the literature, this value is acceptable but not very high. For example, researchers have observed a reduction of variance at the student level reaching 20.6% in secondary schools of the Belgian Dutch-speaking community (Opdenakker & Van Damme, 2001), 36.3% in the secondary education of the Netherlands (Sykes & Kuyper, 2013), and 44.8% at the end of primary education in the French-speaking part of Belgium (Dumay & Dupriez, 2008). In other words, although the level-one pseudo-$R^2$ is not disproportionally, some biases due to the misspecification of the student-level model cannot be excluded and an overestimation of compositional effect is possible.

When each composition is assessed separately, they all have significant negative effects. In other words, being in a school with a lower socioeconomic composition, a lower academic composition or a higher proportion of immigrant students is associated with lower results. Moreover, being in a school that is more heterogeneous on one of these dimensions is also associated with lower scores. Let us note that socioeconomic diversity is only significant at the .05 alpha level when outliers are modelled.

When the three densities are simultaneously considered, the picture is different. In the Model 8, only the additive paradigm is explored. When academic composition and socioeconomic composition are added, the effect of ethnic composition becomes non-significant. When all the compositional effects (from both paradigms – model 9) are included, ethnic compositions but also socioeconomic diversities are not significant anymore. In both models, composition variables together explain 10% of the variance at the school level. The models fit the data well, but the latter presents a better statistical fit with one extra parameter. Nevertheless, the academic diversity renders these results difficult to interpret and takes up an important part of the effect of academic density. As stated before, the high correlation between both paradigms of academic composition challenges the ability of the variable to correctly measure academic diversity.

Then, we would conclude that academic and socioeconomic densities have an extra negative effect on achievement, controlling for individual characteristics. That is, being in a school with a more disadvantaged population and with a higher delay mean is associated with lower performance. Conclusions about the other compositions are more sensitive. According to the fifth model, ethnic density has a positive effect but fails to be significant when it is modelled with other densities. Concerning the heterogeneity of the population provided for schooling, caution is required. The more faithful one, namely the socioeconomic diversity, is barely significant when modelled alone and became nonsignificant when modelled with other compositions. Both other diversities are maybe too crude to really measure diversities.

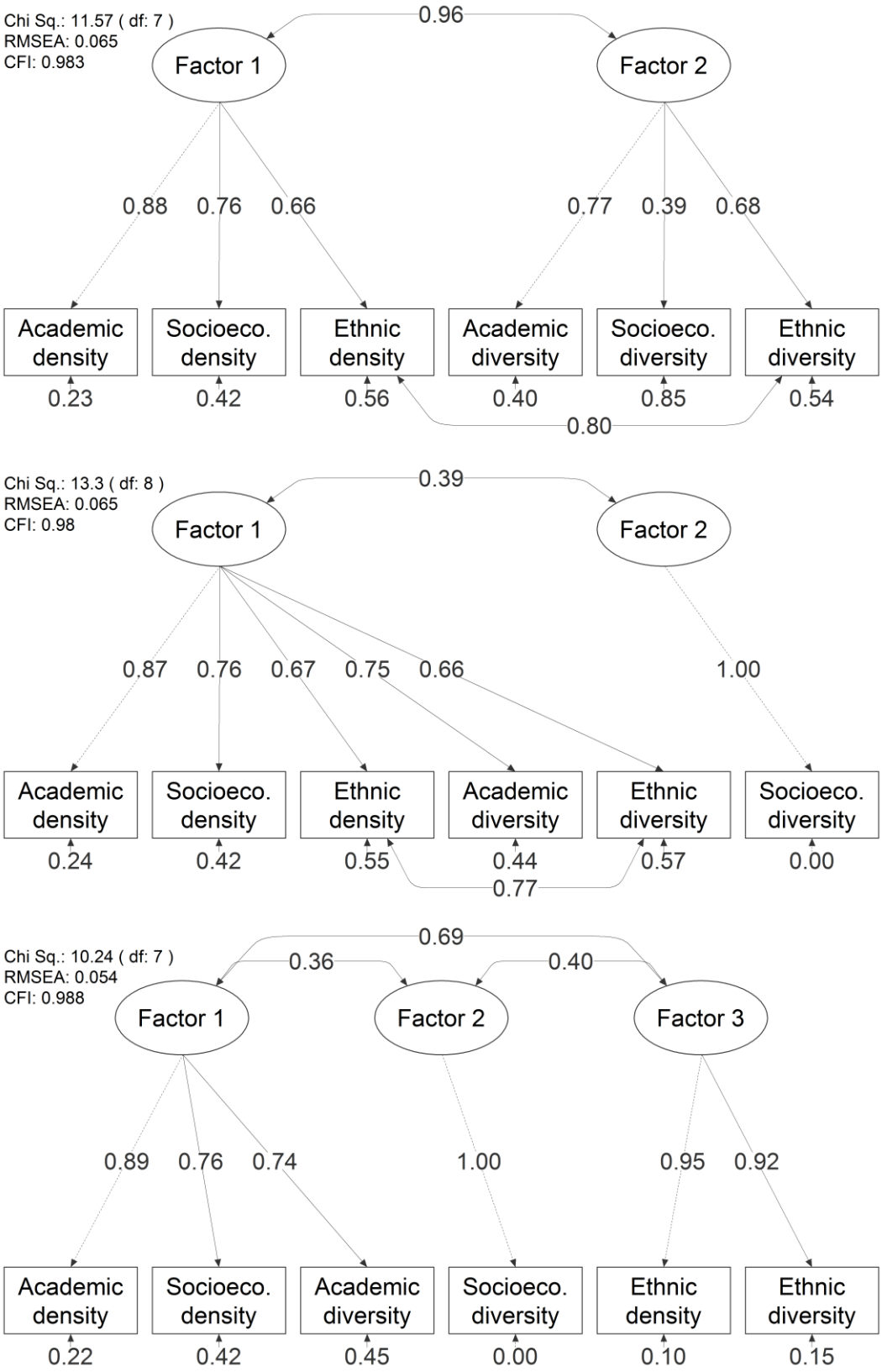
A cocktail of reasons make composition difficult to approach in the PISA context. One of the potential reasons for this is collinearity: because of the strong links between types of composition in the Dutch-speaking Community, they are particularly difficult to disentangle. Secondly, the remaining variance to be explained is limited because of the important number of student characteristics, which means the remaining effect size is not expected to be high. Student background and school segregation, together, may explain what renders modelling of composition so difficult. The variance to be explained is reduced because background has a deep effect on achievement, especially at the second level because of the high level of segregation. Thirdly, the sample size is small at the second level. Finally, the presence of some highly influential values can switch the significance of some coefficient when they are modelled. In conclusion, as we said before, although we have subscribed to the proposition of Thrupp, Lauder, and Robinson (2002) and have tried to model multiple compositions, the implementation remains problematic. However, some of the encountered difficulties could be resolved by resorting to statistical techniques such as factor analysis.

## FACTOR ANALYSIS

Factorisation is classically used to deal with collinearity, either to produce components aggregating correlated variables, or to measure factors, that is, the underlying processes that cause the observed variables (Tabachnick & Fidell, 2007, p. 609- 610). The important correlations between composition measures suggest that these dimensions overlap. Theoretically, we can argue that composition is multidimensional (Thrupp et al., 2002). In our case, we expect a two-factor structure to be extracted, one for the density and one for diversity. As our modelling strategy is essentially exploratory, a simpler one-factor structure and alternative solutions are also estimated.

With only six manifest variables and 156 schools, the conditions are far from perfect. Indeed, inadmissible solutions and non-convergence issues may occur more frequently. Consequently, models were kept as simple as possible and were cautiously checked (Kline, 2010). RMSEA, SRMR and CFI were systematically reported. Hu and Bentler (1999) have advised combining fit indexes as SRMR and CFI to evaluate models fit (with SRMR lower than

***Figure 1: confirmatory factor analysis (top: model 1, middle: model 2, bottom: model 3)***



.09 and with CFI higher than .96). Moreover, CFI is known to perform well with small sample sizes (Bentler, 1990). Hu and Bentler (1999) have also suggested a combination with the RMSEA index that should be lower than .06. Nevertheless, the latter is not advised in small samples or in non-robust conditions. Satorra-Bentler corrections were then used to correct RMSEA and CFI. A confidence interval was computed for the RMSEA index in order to test a poor fit hypothesis (when the upper limit exceeds .08) (Hooper, Coughlan, & Mullen, 2008) but the small sample size limits the power of such a test (MacCallum, Browne, & Sugawara, 1996). In addition, correlation residuals exceeding .1 are signs of poor fit (Kline, 2010). The analyses were run in R with the lavaan.survey package (Oberski, 2014), which allows school weights and stratification to be taken into account (Rosseel, 2012). Robust ML (PLM) with a Satorra-Bentler correction was specified for the estimation.

As a baseline model, we used a model with only one factor and correlated errors between ethnic density and ethnic diversity, two variables with a .87 correlation (see Table 1). Correlating errors, we observe a significant improvement of the model fit (29.8 Chi-square difference of 1 degree of freedom; p-value < 0.000). With a SRMR of .050 and a CFI of .965, the model has a good fit. Nevertheless, the RMSEA of 0.065 (CI .007; .110) means that the poor fit hypothesis cannot be rejected. Moreover, three correlation residuals exceed .1. The loading inspection reveals that VSES is barely explained by the factor (15 % of the variance) and both ethnic density and ethnic diversity are poorly explained (respectively 45 % and 43 % of the variance). Thus, some improvements are required. The following models are presented in Figure 1.

Model 1 allows testing a two-factor structure: one factor for the density paradigm and the other for the diversity paradigm. Correlated errors between ethnic density and diversity have been allowed to solve an improper solution. With a SRMR of.049 and a CFI of .968, the model has a good fit. Nevertheless, the RMSEA reaching 0.065 (CI .000; .110), the poor fit hypothesis cannot be rejected. Only one correlation residual exceeds .1. As Model 1 and the Baseline model are nested (correlation between both factors is set to 0 in the baseline model), we used the Chi-square difference to compare them. With a Chi-square of 1.73 with one extra degree of freedom, the difference is not significant. Consequently, we decided to keep the most parsimonious baseline model.

Model 2 is equivalent to the baseline model. The socioeconomic diversity is, however, considered as a proxy for a latent factor since this variable is poorly explained by the unique factor of the baseline model. This alternative presentation gives the possibility to test a three-factor model (Model 3). The latter model has a good fit, as shown by the SRMR of .038 and the CFI of .976. Moreover, the RMSEA goes down to 0.054 (CI .000; .103), although the poor fit hypothesis cannot be rejected. Only one correlation residual exceeds .1. With one degree of freedom missing, the Chi-square reduction (3.06) is not significant (a .080 p-value). All the variances explained are higher than 50 %. In short, although the Chi-square reduction is just below significance, the model is acceptable.

***Table 3: Bivariate correlations between composition variables (French-speaking Community)***

| | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **1. Academic density** | 1 | | | | | |
| **2. Academic diversity** | **0.254** | 1 | | | | |
| **3. Socioeconomic density** | *0.698* | 0.366 | 1 | | | |
| **4. Socioeconomic diversity** | 0.184 | 0.139 | **0.251** | 1 | | |
| **5. Ethnic density** | *0.261* | 0.424 | *0.215* | 0.175 | 1 | |
| **6. Ethnic diversity** | 0.325 | 0.361 | 0.173 | 0.277 | **0.735** | 1 |

From a substantive point of view, we expected to choose the Model 1, but it actually does not seem to be an improvement over the simplest one. Based on the Chi-squared differences, Model 2 (or the equivalent baseline model) is preferred. The high correlated errors between the two paradigms of ethnic composition remain problematic. Moreover, both factors are very highly correlated. This confirms what the pattern of correlations in Table 1 seemed to indicate, that is that the different, relatively crude, indicators do not measure different paradigms of composition. However, socioeconomic diversity (a more reliable indicator of diversity) appears to measure a separate dimension. Model 3, which proposes to model ethnic compositions as a different factor, is mainly guided by statistical motives in order to obtain acceptable loadings. The solution is indeed acceptable but the Chi-square difference fails to be significant. Unfortunately, the interpretation of the model is not straightforward.

After computing factor scores, we can enter them in a multilevel model. Factor scores from Model 2 were computed and used in multilevel models. The first factor has a highly significant negative effect (-109.2, SD 12.5), while the second one is not significant. In other words, being in a school with a disadvantaged composition (a school with a high proportion of students having repeated a grade, with students from lower socioeconomic backgrounds, and with a high proportion of non-European migrant students, but also with a higher dispersion in terms grade repetition and migration origin) is associated with lower scores in math. The model explains 92.1 % of the variance at the school level. Compared with the multilevel Model 9, it provides a small improvement in fit based on the BIC criterion, but the conclusion of a better fit goes in the opposite direction when based on the AIC criterion. The advantage of the model with the factors is then balanced.

## 4. Multiple compositions in the French-speaking Community

Although the preceding section does not provide evidence of the superiority of a model with composition factors, the approach, however, an important advantage for international comparison. It allows testing easily whether the effect of composition is different between educational systems. It can be done with a limited number of interactions to test, namely the interaction between the factor(s) and the dummy for the considered educational system.

The Belgian context enables us to compare results obtained from the Dutch-speaking subsample to those obtained in the French-speaking one. Both systems have a similar structure (but dissimilar performance) and are deeply segregated. In both systems, composition has an effect on students' results, but the effect can be different depending on the composition considered. Including interaction between the dummy specifying Community membership, Danhier and Martin (2014) observed significant interactions for academic and socioeconomic densities in PISA 2009, which means that these compositions influence the students' results differently in both communities. However, the strategy used by the authors can become costly when the number of composition variables or the number of educational systems increases.

We therefore added the French-speaking subsample to our sample. In line with the two-stage stratified sampling design used by OCDE, schools were sampled according to their size (after being separated into explicit strata, namely form of education, ISCED levels and tracks, and ordered by implicit strata, namely grade repetition, percentage of girls and school types) and students were randomly sampled in selected schools. As for the Dutch-speaking subsample, only regular full-time education was selected. Part-time vocational education (3 schools) and education for students with special needs (5 schools) were excluded. The rate of item missingness was 2.6% at the student level. Due to the limited missingness, listwise deletion was used. The final French-speaking subsample covered 2,759 respondents in 95 schools. Let us, however, note that the definition of school differs between both communities as the whole schools are sampled in the French-speaking Community and the implantations (tracks taught on a single address) are sampled in the Dutch-speaking Community (OECD, 2014).

Before using factors measuring compositional effects to compare communities, we have to test the measurement invariance of the factor model across the two communities. Invariance refers to the idea that the construct measures something similar across sub-populations (Kline, 2010). Different types of invariance can be progressively assessed, from configural invariance (same factors and pattern of loadings across sub-populations) to metric invariance (equal loadings), and to more restricted model (with equal means of factors or with equal residual variance and covariance). These forms of invariance were tested with multiple-group confirmatory analyses with the lavaan R package (Hirschfeld & von Brachel, 2014).

The results were opposite to what we expected. Preliminary observation of correlations (Table 2) reveals quite a different pattern, compared to the Dutch-speaking situation. All the correlations are lower and only two are around .5 or higher. However, as it was the case in the Dutch-speaking Community, there is a high correlation between ethnic density and ethnic diversity (.74), and also between academic and socioeconomic density (.69). The multilevel modelling of French-speaking data reveals a different pattern as well. Only the academic and socioeconomic densities have a significant negative effect when entered separately, and only the socioeconomic one remains significant when entered together.

Regarding factor analyses, no convincing factor model emerges. The same models as those tested for the Dutch-speaking Community were explored. The model 2 seems to be the

best one but fails to be a good fit, with a SRMR of .088, a CFI of .90 and a RMSEA of 0.12 (CI .064; .177). Moreover, some loadings are around or below .4 (academic diversity, ethnic density and diversity) while two correlation residuals exceed .2. Models 1 and 3 do not provide better solutions. With fewer than 100 schools and a limited number of variables, more limited models (with, for example, socioeconomic diversity excluded) do not converge or present improper solutions. Multiple-group confirmatory analyses were run on each of the model presented in figure 1, but none converge to create a proper solution.

In other words, factorisation provides little improvement if one wishes to compare compositional effects between similar educational systems, at least for the two main Belgian systems. However, even with this kind of approach, we cannot assure that the composition variables measure the same thing in both systems, especially when we know that the available tools to gather similar students are used in different ways in both communities (Delvaux, 1998).

## 5. Conclusion

Recently, compositional effect has been more and more frequently included when studying student achievement. This inclusion is essential as it allows assessing the effect of student grouping on their achievement. The use of multilevel techniques to measure such an effect is relevant as it allows to distinguish school effects from the individual ones. With regards to scholarly achievement, students from disadvantaged background would suffer from both their own background and their gathering in the same schools. This modelling is of prime importance as it participates to a major debate exploring whether some students must be separated or not from others. Based on Belgian data, two components of this debate have been challenged: the literature invites us to consider different types and paradigms of composition and a comparative framework offers a way to explore the consequences of different choices regarding student groupings. In this article, we illustrated that, regarding those components, modelling composition is not easy, somewhat tricky and interpretation of concurrent models is not straightforward.

A preliminary issue concerns the requirements regarding the data. The literature states that a full set of student characteristics and the identification of all relevant levels are required. In this article, we used PISA, a rich database that permits us to compare educational systems. However, PISA does not meet the necessary requirements. First, it does not provide any measure of prior achievement. As Belgium is a system where grade repetition and orientation define the pupils' position in the hierarchy of the educational system, we assumed that these variables hold information about prior achievement. Nevertheless, such an assumption still needs to be validated and some overestimation of the compositional effect is expected. Moreover, this assumption is limited to systems that separate students according to their achievement and, consequently, limits possible comparisons. Second, only the student and the school levels are available. The class level, where a lot of processes explaining the compositional effect can have a strong influence is not available. As a consequence, a part of the class compositional effect is measured as a student effect and the compositional effect can be underestimated. In other words, PISA may not be the best database to measure multiple compositional effects. However, there is no alternative data for international comparison that meets all the requirements.

The main issue concerns the multiple measures of composition and their simultaneous inclusion in a model. Valid measurements of types and paradigms of composition are required. In the context of PISA, the available measurements are not convincing. The high correlation between ethnic density and diversity suggests that both variables measure the same concept. With regards to academic composition, the correlation between density and diversity reaches .64 in the Dutch-speaking Community. Again, it is questionable whether these two variables actually measure two separate paradigms. Delay variance is clearly a weak measure of academic diversity, given that delay has only two major occurrences (0 and 1 year of delay). However, although the majority of schools tend to have simultaneously high averages and high

variances concerning delay, there are some schools with a high average but a low variance and vice-versa. By contrast, for socioeconomic composition, the correlation between density and diversity is low, although schools with a more disadvantaged population tend to be a bit more diverse. In conclusion, the availability of refined measures of student characteristic is capital to measure the effect of composition. However, even with more valid measures, collinearity could remain a problem, however, at least we will be able to distinguish it from measurement errors. Let us note also that even with refined measures, measurement errors, but also sampling errors, will cause overestimation of the compositional effect (Lüdtke et al., 2011).

Besides collinearity, other reasons make the compositional effect difficult to assess. The remaining variance left to explain is limited because of the huge number of student characteristics taken into account in the analyses. Consequently, we do not expect the remaining effect size to be high. The variance left to explain is reduced because background has a deep effect on achievement, especially at the second level, due to the important segregation. Next, the sample size is generally small at the second level. This made the model more sensitive to the presence of some highly influential values that can switch the significance of some coefficient when they are modelled. This also limits the power to identify significance of some measures.

We expected factorisation to simplify the picture. This was not necessarily the case. On the theoretical basis, we expected to choose the two-factor model, one for the density paradigm, the other for the diversity paradigm. Such a model does not seem to be an improvement compared to the simpler one-factor model. Moreover, both factors are very highly correlated. This confirms what we have observed in the pattern of correlations, that is, the different crude indicators of composition do not measure different paradigms of composition. However, the one-factor model is usable in multilevel modelling. Scores from factorisation were used in the previous multilevel modelling and provide coherent results: being in a school with a disadvantaged composition is associated with lower scores in mathematics.

A final issue concerns the use of composition in a comparative framework. Multiple-group confirmatory analyses were run in order to use factor scores to compare the compositional effect across the two main Belgian communities. Unfortunately, we did not find any equivalent model allowing a comparison. In this sense, the factorisation failed to simplify the picture. It provides little assistance to compare compositional effects between similar educational systems, at least in the Belgian case. We have to admit that the complexity of the two Belgian education systems are resistant to the combination of multilevel and factor analyses that we proposed in this article.

So, what could we advise from this work? The major claim of this article is that caution is required when multiple variables measuring composition are used, especially, when they are used in a comparative framework and with large scale assessment as PISA. The statement that one type of composition can dominate over others thus depends on theoretical and methodological choices. At least, some reserves regarding political implications are required

and limitations have to be explicitly presented.

With multiple compositions, some precautions can be followed. Firstly, the models with each composition variable have to be separately displayed before displaying a model combining some or all of those variables in order to let the reader able to detect some strange behaviour. Secondly, a sustained attention should be given to the construction of the composition variables, regarding both theoretical and methodological considerations. When a certain amount of doubt hangs on the construction of some variables, the number of composition types and paradigms should be reduced. Regarding the Belgian context, academic and socioeconomic densities should be considered for inclusion as a proxy in the model when a limited number of variables have to be selected. Indeed, they seem to present a coherent effect: they have an extra negative effect on achievement, controlling for individual academic and socioeconomic characteristics. Conclusions about the other compositions are more sensitive. Thirdly, the differential effect of composition types and paradigms should be systematically explored. The previous analyses showed that even in two similar systems, the effects of multiple compositions are unexpectedly complex.

Finally, we conclude by highlighting the need for further development in the area of compositional effect. More research is needed to explore in detail how multiple compositions can be modelled in a comparative framework. Cross-country validation of the models is also essential. Suitable datasets are also required. Such datasets should include precise student background variables, including prior achievement. Nevertheless, let us note that these changes in and of themselves cannot address the benefits of the inclusion a more qualitative approach to explaining the processes that lay behind these statistical measures.

## 6. Reference

Agirdag, O. (2011). *De zwarte doos van schoolsegregatie geopend* (Proefschrift voorgelegd tot het behalen van de graad van doctor in de sociologie). Univeriteit Gent, Gent.

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.

Azzolini, D., Schnell, P., & Palmer, J. R. B. (2012). Educational Achievement Gaps between Immigrant and Native Students in Two « New » Immigration Countries: Italy and Spain in Comparison. *The ANNALS of the American Academy of Political and Social Science*, *643*(1), 46‑77.

Baye, A., & Demeuse, M. (2008). Indicateurs d'équité éducative. Une analyse de la ségrégation académique et sociale dans les pays européens. *Revue française de pédagogie*, *165*(4), 91‑103.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological bulletin*, *107*(2), 238.

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

Centre for Multilevel Modelling. (2011). *Weighting in MLwiN*. Consulté à l'adresse http://www.bristol.ac.uk/cmm/software/support/support-faqs/weighting.pdf

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: a typology of composition models. *Journal of Applied Psychology*,

*83*(2), 234‑246.

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.

Danhier, J., Jacobs, D., Devleeshouwer, P., Martin, E., & Alarcon, A. (2014). *Vers des écoles de qualité pour tous ? Analyse des résultats à l'enquête PISA 2012 en Flandre et en Fédération Wallonie-Bruxelles*. Bruxelles: Fondation Roi Baudouin.

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.

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. (1998). L'échec scolaire en Belgique. *European Journal of Teacher Education*, *21*(2-3), 161‑198.

Demanet, J., Agirdag, O., & Van Houtte, M. (2012). Constrict in the School Context. *The Sociological Quarterly*, *53*(4), 654–675.

Demeuse, M., & Baye, A. (2008). Mesurer et comparer l'équité des systèmes éducatifs en Europe. *Éducation et formations*, *78*, 137–149.

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

Dronkers, J., & van der Velden, R. (2013). Positive but also negative effects of ethnic diversity in schools on educational performance? An empirical test using PISA data. In *Integration and Inequality in Educational Institutions* (p. 71‑98). Springer. Consulté à l'adresse http://www.springer.com/education+%26+language/book/978-94-007-6118-6

Dumay, X., & Dupriez, V. (2008). Does the school composition effect matter? Evidence from

Belgian data. *British Journal of Educational Studies*, *56*(4), 440‑477.

Dupont, V., & Lafontaine, D. (2011). Les choix d'études supérieures sont-ils liés à l'établissement secondaire fréquenté? *Schweizerische Zeitschrift für Bildungswissenschaften*, *3*(33), 461–478.

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.

Fekjær, S. N., & Birkelund, G. E. (2007). Does the Ethnic Composition of Upper Secondary Schools Influence Educational Achievement and Attainment? A Multilevel Analysis of the Norwegian Case. *European Sociological Review*, *23*(3), 309‑323.

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

Hindriks, J., & Verschelde, M. (2010). L'école de la chance. *Regards économiques*, *77*.

Hirschfeld, G., & von Brachel, R. (2014). Multiple-Group confirmatory factor analysis in R–A tutorial in measurement invariance with continuous and ordinal indicators. *Practical Assessment, Research & Evaluation*, *19*(7), 2.

Hirtt, N. (2008). *Pourquoi les performances PISA des élèves francophones et flamands sont-elles si différentes?* Bruxelles: Aped. Consulté à l'adresse http://www.skolo.org/spip.php?article452&lang=fr

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*(1), 53‑60.

Hox, J. (2010). *Multilevel analysis. Techniques and applications* (2e éd.). New York: Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A*

*Multidisciplinary Journal*, *6*(1), 1‑55.

Kline, R. B. (2010). *Principles and Practice of Structural Equation Modeling* (3rd edition.). New York: The Guilford Press.

Lafontaine, D., & Monseur, C. (2011). Quasi-marché, mécanismes de ségrégation sociale et académique. Une approche comparative. *Education comparée / Nouvelle série*, *6*, 69–90.

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2x2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological methods*, *16*(4), 444.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*(2), 130‑149.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124.

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.

Meyers, J. L., & Beretvas, S. N. (2006). The Impact of Inappropriate Modeling of Cross-Classified Data Structures. *Multivariate Behavioral Research*, *41*(4), 473‑497.

Oberski, D. (2014). lavaan. survey: An R Package for Complex Survey Analysis of Structural Equation Models. *Journal of Statistical Software*, *57*(1), 1–27.

OECD. (2014). *PISA 2012 technical report*. Paris: OECD Publishing.

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.

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.

Raudenbush, S. W., & Willms, J. D. (1995). The Estimation of School Effects. *Journal of Educational and Behavioral Statistics*, *20*(4), 307‑335.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36.

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.

Schaeffer, M. (2013). Can competing diversity indices inform us about why ethnic diversity erodes social cohesion? A test of five diversity indices in Germany. *Social Science Research*, *42*(3), 755‑774.

Schafer, J. L. (1999). Multiple Imputation: A Primer. *Statistical Methods in Medical Research*, *8*(1), 3‑15.

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.

Szulkin, R., & Jonsson, J. O. (2007). *Ethnic segregation and educational outcomes in Swedish comprehensive schools* (SULCIS Working Paper No. 2007:2). Stockholm University Linnaeus Center for Integration Studies - SULCIS.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn and Bacon.

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. (1999). *Schools Making A Difference*. Buckingham: Open University Press.

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.

Vandenberghe, V. (2011). Inter-regional educational discrepancies in Belgium. How combat them? In *Educational Divergence - Why do pupils do better in Flanders than in the French community?* (Re-Bel Initiative., p. 5‑25). Brussels.

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 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.

Van Houtte, M., & Stevens, P. A. J. (2009). School ethnic composition and aspirations of immigrant students in Belgium. *British Educational Research Journal*, *36*(2), 209‑237.

Verschelde, M., Hindriks, J., Rayp, G., & Schoors, K. (2010). Explaining social segregation in Belgium: an index decomposition approach. Consulté à l'adresse http://feb.kuleuven.be/eng/ew/papers_edupol/verschelde-hindriks-rayp-schoors.pdf

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.