School choice determinants with preference data

Thomas Wouters
HIVA - KU Leuven, Belgium

April 12, 2015

Abstract

There is an ongoing debate in the economics of education literature on what school characteristics parents favour. The study of school choice determinants informs us about the dynamics of school segregation and the effectiveness of desegregation policies. Considering segregation, one of the key questions is to what extent preferences for school composition differ across social groups and induce a dynamic of self-segregation. This paper adds to the evidence that heterogeneous preferences for group composition strongly drive school segregation.

Using a unique dataset containing pure school preferences for a large Flemish city (Ghent), we study preferences for distance, socio-economic school composition and school quality, as well as the degree of heterogeneity. The determinants of school choice are usually estimated using conditional logit models on realised choice data. A first cause of bias are thus capacity problems. Preference data can remedy this. Although revealed preference models improve upon stated preferences models by eliminating socially desirable responding, simple discrete choice models are still likely to render biased results: constraints on choice sets, differences in information availability, unobserved heterogeneity and nonlinearities in demand all prevent us from observing true preferences. Our findings indicate that information availability (and group-dependent information), rather than differences in choice sets, drive apparent differences in preferences for school quality.

1 Introduction

How do parents make school choices? Which school characteristics do they take into account? These questions mater when we one wants to explain school segregation or in the evaluation of the role of free school choice in improving school quality (Hastings et al [7]). When school quality gets little attention in the school choice process (and when quality differences are sizable), free school choice would be inefficient from a production perspective.

School segregation is a second reason why we would like to infer as good estimates as possible on parental preferences for school characteristics. It is the perspective from which this papers considers school choice. School segregation and school choice are two sides of the same coin, with school choice being the micro-level process and segregation the macro-level outcome. In the economic literature, only a few papers combine both perspectives to explain (school) segregation. These papers either consider heterogeneous preferences or differences in willingness to pay for certain school characteristics (such as school quality) as the main reason why schools are socio-economically (and ethnically) segregated.

In this paper, we focus on three main determinants of school choice: distance to school, school composition and school quality. School composition can further be split into an ethnic part and a socio-economic part. We use discrete choice models to understand the trade-offs people face when making school choice, to see to what extent preferences are heterogeneous across groups, and to check whether these inferred preferences are robust for different specifications of the model. A main difficulty is the impossibility to directly observe an individual's choice set. We need to distinguish as good as possible between constraints and preferences.

We start by reviewing the literature on school segregation in section 2 and link this to the literature on school choice. We discuss the dataset in section 3. In section 4, we elaborate on the econometric model, for which we present the results in section 5. Section 6 concludes.

2 Literature

We shortly touch upon the most important theoretical models that help us understand the dynamics of neighbourhood or school segregation. We subsequently move on to the more empirical literature.

2.1 Social interaction models

We consider two classes of social interaction models: the Schelling model [12] and market models (in the tradition of Tiebout [14]). Schelling considers one neighbourhood and explains segregation through a preference for one's own group. The other models we consider are more in the Tiebout tradition: agents choose a location based on the bundle of public goods and taxes this choice implies. The Tiebout model stresses that one's residential location is not a given. Rational agents compare differences in neighbourhoods and pick the one that suits their preferences best. In such contexts, it makes little sense to look at school segregation as an isolated phenomenon. We thus explore the theoretical links between neighbourhood and school segregation. Furthermore, we will show that preference heterogeneity (i.e. preferences that differ across groups) is not a necessary condition to bring segregation about.

Schelling [12] starts from a number of weak assumptions about people's preferences for their immediate neighbours and shows how these preferences interact to create collective results as residential segregation. He models agents as belonging to an ethnic group and assumes a distribution about the maximum proportion of neighbours from the other ethnic group they can stand. This distribution is assumed to be uniform. The least tolerant individual only wants to live among people from her own group. The most tolerant individual does not mind being the only one of his type in a neighbourhood, as long as the number of people from the other type is not too high. The most striking result is that individual preferences with respect to segregation at the macro level seem to have little influence on the collective result.

We consider Schelling's bounded-neighbourhood model, in which a fixed geographic area is taken as a neighbourhood, and people only care about the relative presence of their own type within the neighbourhood. In contrast with the *market models* we consider next, Schelling uses distributions of tolerance levels towards people from another type instead of assuming that preferences across types are the same. The biggest difference, however, is that segregation dynamics are not modelled as the outcome of a market process, but as the interaction between groups' tolerance levels and distributions.

Concretely, it becomes possible to explain relatively extreme and short-run changes in segregation: in a neighbourhood that is originally dominated by type A people, a dynamic of tipping can be started where the composition of the neighbourhood may completely turn around as a result. This starts when the least tolerant individual from type A moves out of the neighbourhood because the amount of type B people has become too high. While the ratio of B to A people has increased, the ratio of A to B people has decreased, resulting in an even larger outflow of type W people. This process goes on until all individuals from type A have left. The resulting situation is one of complete segregation. The bounded-neighbourhood model also allows for an easy way to represent tipping behaviour (Card et al build on this and test the tipping hypothesis empirically for Chicago [4]).

The second class of social interaction models we consider are in the tradition of Tiebout's seminal work on the provision of local public goods [14]. We label them *market models*. These models typically require prices for an equilibrium to exist and neighbourhood and school segregation are considered together. The necessity of prices for the existence of an equilibrium reflects that preferences will be different from the Schelling model. Instead of preferences that differ across groups, all individuals now prefer neighbours (or peers) from one particular type (i.e. high-quality type).

Becker and Murphy [becker2000social] provide a simple market model for neighbourhood segregation. The model starts from a simple setup in which residents choose to live in one of two neighbourhoods that are identical in every dimension except for the presence of the preferred type

(type H). In such a situation, the equilibrium in a competitive housing market has a price differential, with houses in the neighbourhood with the highest ratio of type H people selling at the highest prices. This price differential reflects a difference in willingness to pay between the two neighbourhoods, which is equal for both types of people in equilibrium. Translating the insights from this simple (partial-equilibrium) of neighbourhood choice to school choice, we conclude that school segregation only results when at least one of the following three conditions are met:

- Differences in school quality and willingness to pay (private schools)
- Preference for own-group peers (Schelling model)
- Homogeneous preferences (across groups) for a certain type of peers, combined with cream skimming or tracking

In a general equilibrium setting, when neighbourhood and school choice are considered simultaneously, the picture becomes more complex. Nechyba's chapter in the Handbook of the Economics of Education provides an excellent overview to these models [11]. Although in general it would be nave to suppose that neighbourhood and school choice are independent, in Flanders this is more warranted. In principle, schools are open to everybody, independent of place of residence (although exceptions are beginning to emerge).

2.2 Revealed preferences

We now review the literature on school choice determinants. We neglect the research that only uses surveys (stated preference). Glazerman [5] was among the first to point at the risks of only relying on survey methodology. Especially with closed questions, wording is very important. In his study, educational quality comes out as a much less important determinant of school choice than can be expected on the basis of questionnaires. He concludes that variables as SES and ethnicity play a more important role than people dare to admit.

Another possibility is to consider search behaviour. Schneider and Buckley [13] analyse data from a website where parents can compare schools from Washington DC. They find that school composition is the school attribute most searched for. Hastings et al [7] use a mixed logit model to derive school determinants and find slight heterogeneity in preferences for distance (with whites being more sensitive to distance than non-whites). They also find evidence for a preference for one's own race at school, with 70% of students belonging to the same group as the preferred school composition. Preferences for quality (higher test scores) are shown to be stronger for high SES than for low SES students.

Bayer et al [1] show that socio-demographic characteristics (income, language, immigrant status, etc.) have the potential to largely explain racial segregation between some groups, but much less so between others (especially blacks and whites). In other words, when we control for a host of other variables, ethnicity will not matter much for school choice, except for segregation between blacks and whites. This constitutes a warning that one must be careful to extrapolate findings from one region to another. Racial tensions between blacks and whites in the US have a long history, which cannot be compared with the arrival of other ethnic groups in the US or in Europe.

Jacob and Lefgren [9] provide evidence that goes against some of the conclusions reached above (of important differences between parental preferences). Looking at preferences for types of teachers, they find that, parental preferences for student satisfaction relative to academic achievement are similar across socio-economic backgrounds within schools. Between poor and rich schools, however, important differences exist. And in general, low-income parents are less likely to actively request teachers; they are less assertive. The authors point out that their overall findings are consistent with a model in which people have similar preferences for school quality but, depending on other circumstances such as school composition, prefer other types of teachers to reach similar objectives. This makes it less likely that differences in parental preferences for school quality cause segregation. Of course, different preferences for group composition might still play a role in the sorting of students between schools, as well as differences in sensitivity to distance. This conforms with the findings by Hastings et al [6] that minority parents (when the minority is poorer on average, as is often the case) face a trade off between school quality and the presence of peers from their own (minority) group.

For the UK, Burgess et al [3] find considerable heterogeneity across social groups, especially for school composition. The heterogeneity in preferences for school quality is mainly ascribed to differences in choice sets rather than differences in preferences per se.

2.3 Information

Lastly, we mention some papers that challenge the often implicit assumption of complete information about school characteristics. We already mentioned the work by Burgess et al ([3]) who showed that choice may be more restricted for low SES groups than for high SES groups. We also discussed a study by Jacob and Lefgren [9], indicating that low-income parents are less likely to actively request teachers; they are less assertive. Although preferences between groups are similar, various groups may obtain different outcomes depending on the size of their network or whether they understand how to play the system. Other evidence for the idea that the choice process, rather than preferences themselves, might differ across social groups is provided by Kristen [10] in a study of primary school choice in Germany. She asserts that Turkish parents have a different perception of school alternatives than native German parents. In short, they are more likely to consider only one school since they are more unfamiliar with the school system. This is reminiscent of the study by

Hastings and Weinstein [8], which found that low SES parents do not necessarily place less weight on academic achievement, but that apparent differences in sensitivity to quality between high and low SES parents can (partly) be remedied by providing them with easily accessible information on test scores. This suggests that part of the preference differential is in fact due to differences in access to information. Certainly, schools may also erect barriers to entry, preventing the execution of free choice, but it is hard to move beyond anecdotal evidence on this issue.

3 Data

We use two datasets on Flemish primary schools and students. Our dataset contains administrative data from the Flemish Department of Education, covering all students in the compulsory education system over a period of 12 years (from academic year 2001-2002 up to 2012-2013). Information on students' socio-economic background (mother's education, financial situation and home language) and place of residence are available too. The level of detail on the students' socio-economic background varies. For mother's education, we can distinguish between five levels of education: none, primary, lower-secondary, upper-secondary and tertiary. We translate this variable into years of study [1]. For financial background and home language, information is limited to a dummy variable. Household income is either low enough to receive a study grant, or not. Home language is either Dutch or not. We use home language as a proxy for ethnicity. In this dataset, we observe realised choices. Since we are interested in school choice, we only retain those students at the start of their school careers, choosing a school for the first time. We thus focus on initial school choice.

We limit ourselves to a subset of Dutch speaking schools, namely those located in the province of Flemish Brabant [2]. This subset of the dataset contains 62445 observations for 8259 students.

A second administrative dataset is available in the context of centrally-administered registration platforms (CAR, Dutch: "Centrale aanmeldingsregisters"). This type of data is unique in Belgium and made available for the first time. We use data for students that have to be allocated to a primary school in Ghent in the school year 2014-2015. The data contain information on parental socio-economic background, home and work address and school capacity. Most importantly, parents also indicate their order of preference for schools in Ghent. As such, pure preferences can be observed, instead of only the outcomes of the school choice process (as is the case with the first dataset), where school capacity constraints and possibly discriminatory practices are mixed up with preferences.

[1] The translation is done as follows: 0 years (no primary education), 6 years (primary education), 9 years (lower secondary education), 12 years (upper secondary education) and 15 years (tertiary education)

[2] Given that our second dataset contains data for Ghent, it would be better to compare with a dataset with realised choice data that also considers Ghent. Unfortunately, the geolocation information for Ghent is, at this point in time, not precise enough to obtain precise distance measurements

[merged with school information from other dataset] Our second dataset contains observations for 2493 students and 92 (primary) schools.

3.1 True preferences?

The student allocation mechanism in Ghent is strategy-proof: ranking schools in a way different from one's real preference ordering can never lead to a better allocation. Parents submit their preferences for schools. Subsequently, the mechanism allocates each student to her most preferred place from the set of schools where she is eligible for a place. All other places then become availabe for other students. The algorithm terminates when no one is changing schools anymore. The place where a school is ranked in an individual's preference ordering has no effect on the probability the student obtains a place in that particular school. This is why the mechanism is strategy proof. Nevertheless, it cannot be excluded that some parents do not understand the allocation mechanism and assume they can outsmart the system by submitting *strategic* preferences, even though the website[3] clearly mentions that parents should submit their true preferences and have nothing to gain by doing otherwise.

This does not imply that parents will submit a complete ranking of schools, not even of those schools they did consider. If a certain school is considered no better than any other school that was not considered, parents might as well not include it in their submitted preference ranking. However, there is also no clear reason why not to rank a school one is not interested in, since the allocation mechanism may guarantee a place at a school, but will never force parents to redeem this *ticket*.

4 A model for school choice

In this section, we discuss the model used to estimate the coefficients for the different determinants of school choice. Since we only observe realised school choices, the model needs to handle a binary response variable. For each combination of a school and a student, we either have a match or not. For the second dataset, we need a model incorporating an ordered response variable. This will be an rank-ordered logit model.

First, we specify the utility function underlying the model. Individual i chooses school s, based on three main determinants: (straight-line) distance between the school and her place of residence (D_{is}), school socio-economic composition ($\overline{SES_s}$) and school quality ($SchoolQual_s$). In the absence of scores on standardized tests, our school quality measure is based on the percentage of students progressing to the academic track in secondary school. This does not make for an objective

[3] https://meldjeaan.gent.be/faq (only in Dutch)

assessment of school quality. Nevertheless, it is probably one of the best proxies for *perceived* school quality we can construct from our data[4]

These three variables are also interacted with her own socio-economic status (SES_i), which is a binary variable. We consider an individual as disadvantaged when she meets at least one of the following three conditions: (1) mother's highest qualification is secondary education or lower, (2) language spoken at home is not Dutch or (3) she receives a study grant (because family income is below a certain threshold). The term Z_s includes other school-level controls, possibly interacted with individual characteristics. Examples are school size, a dummy indicating whether the school is Catholic, etc. We also include squared terms to capture nonlinear relationships. We discuss the variables in more detail in section 5. All observable factors make up the part V_{is} of utility. The error terms are ϵ_{is}, which are iid extreme value distributed. The extent to which the coefficients α, β and γ vary over individuals and schools will depend on the specification of the model, whether it is the conditional, mixed or exploded logit.

$$\begin{aligned} U_{is} = V_{is} + \epsilon_{is} = \alpha_i + \beta_1 * D_{is} + \beta_2 * D_{is} * SES_i + \gamma_1 * \overline{SES_s} + \gamma_2 * \overline{SES_s} * SES_i \\ + \delta_1 * SchoolQual_s + \delta_2 * SchoolQual_s * SES_i + \zeta * Z_s + \epsilon_{is} \end{aligned} \tag{1}$$

4.1 A conditional (fixed effects) logit model

In the conditional logit model, each individual has a different reference value (fixed effect) for utility: α_i. The choice probability π_{is} (the probability that individual i chooses school s), which is independent of α_i, can be obtained as follows:

$$\pi_{is} = \frac{e^{V_{is}}}{\sum_{j \in S} e^{V_{ij}}} \tag{2}$$

As Train [15] argues, one of the key restrictions of the (conditional) logit model is the assumption that the error terms are iid. When this is not the case, for instance when tastes vary with other, unobserved, parameters, the model is misspecified. Another restriction, following from the absence of correlation over alternatives, is the *Independence from Irrelevant Alternatives* (or IIA) property. This implies that the elasticities of substitution are constant and thus independent of the other alternatives.

[4] We constructed two other measures, both based on the prevalence of grade retention: the percentage of students that did not experience grade retention at the primary school under consideration, and the percentage of students from that school that have not yet been retained in secondary school. Both measures, however, perform less well than the one based on secondary school tracks.

4.2 A rank-ordered logit model

When the same individual makes multiple choices, we are dealing with panel instead of cross-sectional data. The probability π_{is} then becomes the product of logit formulas, integrated over the distribution of θ (the coefficients). Ranked data can also be represented as a series of choices by the same decision maker. First, the most preferred option is chosen from the whole choice set. The second option is the one that would be chosen from a set containing all but the most preferred one, etc. This idea, originally developed by Beggs et al [2], is called rank-ordered (or exploded) logit. Take student i, who has schools A, B and C in her choice set, with $C \succeq B \succeq A$. Then, again following Train [15], the probability becomes:

$$Prob(ranking\ C, B, A) = \int \left(\frac{e^{V_{iC}}}{\sum_{j=A,B,C} e^{V_{ij}}} \frac{e^{V_{iB}}}{\sum_{j=A,B} e^{V_{ij}}} \right) f(\theta) d\theta \tag{3}$$

4.3 The choice set

After the model is specified, we still need to define the choice set (or consideration set), i.e. the set of schools parents actually take into consideration and compare to each other. While in principle any Belgian public school can be chosen, parents are unlikely to pick one located further than a couple of kilometers away from their home or work place.

With realised choice data, we only observe the school the student attends. We do not know which schools her parents initially *considered*.

To limit the size of the choice sets, we use a distance criterion. Most students choose schools fairly close to their home address (see table 1). We specify that choice sets only contain those schools within a 5 km range of the individual's home address. The further away parents choose schools, the more likely this behaviour will be explained by factors unknown to the researcher (work address, other places of residence, family member address, etc).

Distance	Percent	Cum. Percent
$0 <= D < 1$	46.8	46.8
$1 <= D < 2$	22.8	69.6
$2 <= D < 3$	11.4	80.9
$3 <= D < 4$	6.4	87.4
$4 <= D < 5$	3.6	91.0
$5 <= D < 6$	2.4	93.4
$6 <= D < 7$	1.6	95.0
$7 <= D < 8$	1.2	96.2
$8 <= D < 9$	0.8	97.0
$9 <= D < 10$	0.6	97.6
$D >= 10$	2.4	100.0

Table 1: Distribution of home-school distance (Flemish Brabant)

When we have ranked data, the easiest option is to include only the schools appearing in the individual's ranking in the choice set, and to limit ourselves to the variance between these schools. However, the observed rankings will be incomplete. Their ranking will only contain those schools parents consider, not those they do not want to attend. What if some nearby school (e.g. one more nearby than the most preferred school) is not ranked by this individual? It is unlikely that parents were not aware of this school's existence. One could argue that this school should be ranked last, as the individual apparently (weakly) prefers all other schools in the area above it.

5 Estimating school choice determinants

We first consider results from the realised choice data. As mentioned in section 4.3, we do not know whether a student attends its favourite school or had to settle with his second choice because of capacity constraints. Nevertheless, the region we consider here (Flemish Brabant) does not include large cities where capacity problems are likely to be of greater significance. We subsequently move towards rank-ordered data for Ghent, representing parental preferences instead of realised choices.

5.1 Realised choice data, Flemish Brabant

The results for the realised choice data in Flemish Brabant are in table 2 below. We only observe one school per individual, which is assumed to be preferred over all others in her choice set. The choice set is defined as the set of schools within 5 km of the student's home address. All tables report coefficients. These are not directly interpretable, but can be compared to each other (for instance to derive rates of substitution for school characteristics). Further on in this paper, we will present these estimates graphically, as probabilities that a certain school will be chosen over another school, to make preferences and the degree of heterogeneity more tractable.

We start from the most simple model and subsequently add new terms. In column 1, only distance (and distance squared) is included. Judging from the values for the pseudo-R^2, this is by far the most important determinant we have information on.

The second column represents the basic model, including the 3 main determinants. Distance is expressed in kilometers, school average years of study ($School_YoS$) in years, school average low income ($School_LowInc$) in percentages (the percentage of students that receive study grants), school average foreign language ($School_ForLang$) in percentages too (the percentage of students that do not speak Dutch at home). School quality ($SchoolQual$) is the percentage of students from that school continuing into the academic track in secondary education (determined in the third year of secondary education, since tracking de iure does not exist during the first two years). To correct somewhat for differences in school composition, we take this percentage both for disadvantaged and for advantaged students and apply an equal weight. We get: $SchoolQual = 0.5 * SchoolQual_{Adv} + 0.5 * SchoolQual_{Disadv}$, where $SchoolQual_{Adv}$ is the percentage of advantaged students continuing into the academic track and and $SchoolQual_{Disadv}$ the percentage of disadvantaged students. We categorize a student as disadvantaged when she does not speak Dutch at home, receives a study grant, *or* has a mother who did not complete (upper) secondary education. The category of advantaged students is made up of all other students (who speak Dutch, do not receive a study grant, *and* have a mother who completed (upper) secondary education.

When considering at the size of the coefficients, it is important to keep in mind that distance and years of schooling are not expressed in percentages. The coefficients for distance and quality have the expected sign: the longer one has to travel or the worse students performed at the school before, the less likely the school will be chosen. The coefficients for school composition (and for its interaction terms) are a first indication for self-segregation. The higher a student's mother is educated, the more positively she reacts to the average years of schooling (by the mothers) at the school. However, even at maximum schooling, the coefficients indicate that a higher school average years of schooling i not valued positively. Still, at this stage we only allow for linear preferences. For income and home language, parents also seem interested in a school composition that resembles their own background. From this model, the school composition effect seems sizable. Advantaged parents are willing to choose a school located 1 km further (-1.242 in utility terms) for a school with a more favourable student composition (an increase of about 21 percentage points in advantaged students or $0.21 * (-2.434 - 3.528)$ in utility terms).

We add quadratic terms in the third column to allow for nonlinear preferences. We also allow for differences across groups in preferences for distance and school quality. The impact of distance is negative but diminishing (as the quadratic term has a positive sign). We also add an interaction term with socio-economic status (SES, which is a dummy taking the value 1 for a disadvantaged student) is not really significant but goes in the expected direction (lower SES implies a higher sensitivity for distance).

The school socio-economic composition, expressed as the proportion of students from disadvantaged backgrounds, plays an important role too. A tendency for self segregation along socio-economic lines appears again. Advantaged students are more likely to pick schools with more advantaged students, and vice versa for disadvantaged students. The more disadvantaged a school gets, the stronger the tendency for advantaged students to avoid the school, and the weaker the tendency for disadvantaged students to prefer it over schools with a less disadvantaged mix. School average years of schooling now has the expected term. Together with the quadratic term, and for realistic values of average years of schooling (somewhere between 9 and 13 years), it implies a preference for more educated peers, shared across groups with different levels of schooling. For language and income, self-segregation is strong.

Somewhat surprisingly, the coefficient for school quality has turned negative, even when we take into account that the base category is a parent without schooling. Especially lowly educated parents seem to prefer schools where fewer students continue into the academic track. This result does not necessarily amount to anything more than a correlation. It does not in itself imply a preference for bad quality. Parents from different socio-economic backgrounds may be looking for different school characteristics, in the first place because their children may have different needs (cfr Hastings et al [6]).

Of course, these results are not insensitive to the choice set specification. When we take choice sets with a 10 km radius (instead of 5 km), for instance, the distance coefficient becomes less negative and the pseudo-R^2 increases. It is, however, not logical to assume everyone considers all schools within 10 km. The higher pseudo-R^2 (up to 0.5) is then artificially high.

	Cond. logit 1 coeff. (std err)	Cond. logit 2 coeff. (std err)	Cond. logit 3 coeff. (std err)
$Distance$	-2.033***	-1.242***	-1.983***
	(0.04)	(0.01)	(0.05)
$Distance^2$	0.180***		0.167***
	(0.01)		(0.01)
$SES * Distance$			-0.165
			(0.09)
$SES * Distance^2$			0.046*
			(0.02)
$\overline{YoS_s}$		-0.762***	2.178***
		(0.06)	(0.66)
$\overline{YoS_s}^2$			-0.121***
			(0.03)
$YoS * \overline{YoS_s}$		0.045***	-0.123*
		(0.00)	(0.05)
$YoS * \overline{YoS_s}^2$			0.007***
			(0.00)
$\overline{ForLang_s}$		-2.434***	0.456
		(0.17)	(0.42)
$\overline{ForLang_s}^2$			-3.473***
			(0.63)
$ForLang * \overline{ForLang_s}$		3.972***	4.494***
		(0.23)	(0.71)
$ForLang * \overline{ForLang_s}^2$			0.535
			(0.89)
$\overline{LowInc_s}$		-3.528***	-3.198***
		(0.29)	(0.62)
$\overline{LowInc_s}^2$			1.653
			(1.22)
$LowInc * \overline{LowInc_s}$		3.286***	6.186***
		(0.37)	(1.18)
$LowInc * \overline{LowInc_s}^2$			-4.210*
			(1.87)
$SchoolSize$			0.004***
			(0.00)
$Catholic$			-0.103***
			(0.03)
$SchoolQual$		0.608***	-3.355***
		(0.12)	(0.66)
$SchoolQual^2$			0.781
			(0.45)
$YoS * SchoolQual$			0.182***
			(0.04)
$Forlang * SchoolQual$			0.870**

			(0.29)
pr2	0.3680	0.3867	0.4147
N	153064	123337	123337

Table 2: Estimation results for conditional logit models on realised choice data, choice sets with 5 km radius

5.2 Preference data, Ghent

We now consider the second dataset and switch from realised choices to preference data. The rank-ordered logit model now becomes more appropriate. For the sake of comparison, we add tables for the conditional logit models in the appendix (section 7). From a comparison with the conditional logit models, our first striking result is that the pseudo-R^2 have tumbled. Ghent is much more densely populated, with more schools within a given distance. Our model will explain less of the observed variation with the same variables. At first, we limited the choice set to those schools that were submitted by parents in the student allocation mechanism. This renders counterintuitive results, which are at odds with the results obtained for Flemish Brabant. We include the these results in the appendix (table 7). These results follow from the way our choice sets are defined, i.e. only containing schools that were ranked positively by parents (see section 3.1). We are left with less variation to derive preferences from.

In table 3, we switch to a new choice set definition, where we try to take into account the schools that parents explicitly do not mention in the allocation procedure. We do not want to impose identical choice sets, as we did with the realised choice data. Instead, we look at the most distant school on each student's preference list and use it as a boundary on the choice set [5]. We exclude schools further than 5 km from the home address.

In the first column of table 3, only distance (and distance squared) are included in the model. As mentioned above, these coefficients are smaller because more schools are available for a given distance.

In column 2, we again add the variables representing the three main school characteristics (distance, composition and quality), including interaction effects for quality. The degree of heterogeneity is strong for language, but less so for income. Heterogeneity in preferences for average school income was much stronger in Flemish Brabant, most likely because differences in average income between schools are more outspoken.

[5] When work address is available, it is also used to calculate distances. The shortest distance (from home, first or second work address) is then used to determine which schools should be included in the one's choice set.

In column 3, we add quadratic effects. We do not find any effect for average school income. For all other determinants, preferences are clearly nonlinear. This is something we do not observe for the conditional logit on the same data (cfr table 9) (we did observe nonlinear relationship to some extent in the data for Flemish Brabant).

	RO logit 1	RO logit 2	RO logit 3
	coeff. (std err)	coeff. (std err)	coeff. (std err)
$Distance$	-1.143***	-0.296***	-1.366***
	(0.05)	(0.02)	(0.07)
$Distance^2$	0.235***		0.282***
	(0.01)		(0.02)
$SES * Distance$			-0.018
			(0.11)
$SES * Distance^2$			-0.013
			(0.03)
$\overline{YoS_s}$		-0.186***	0.533***
		(0.03)	(0.13)
$\overline{YoS_s}^2$			-0.035***
			(0.01)
$YoS * \overline{YoS_s}$		0.017***	-0.016
		(0.00)	(0.01)
$YoS * \overline{YoS_s}^2$			0.002**
			(0.00)
$\overline{ForLang_s}$		-2.229***	-2.300***
		(0.18)	(0.45)
$\overline{ForLang_s}^2$			1.255**
			(0.45)
$ForLang * \overline{ForLang_s}$		1.637***	2.644***
		(0.14)	(0.56)
$ForLang * \overline{ForLang_s}^2$			-1.023
			(0.56)
$\overline{LowInc_s}$		-0.095	-0.060
		(0.15)	(0.48)
$\overline{LowInc_s}^2$			-0.714
			(0.51)
$LowInc * \overline{LowInc_s}$		0.179	0.786
		(0.24)	(1.12)
$LowInc * \overline{LowInc_s}^2$			0.294
			(1.22)
$SchoolSize$			0.001***
			(0.00)
$Catholic$			-0.565***
			(0.03)
$SchoolQual$		-0.386**	4.318***

		(0.15)	(1.04)
$SchoolQual^2$			-6.298***
			(1.30)
$YoS * SchoolQual$			-0.095
			(0.08)
$YoS * SchoolQual^2$			0.210*
			(0.10)
$LowInc * SchoolQual$			2.278*
			(1.06)
$LowInc * SchoolQual^2$			-0.672
			(1.12)
pr2			
N	37933	33865	33865

Table 3: Estimation results for rank-ordered logit models on realised choice data

5.3 Preference data, Ghent, controlling for other choice set constraints

In their 2014 paper, Burgess et al [3] argue that differences in preferences for school characteristics between socio-economic groups are often due to differences in constraints. The choice sets of advantaged an disadvantaged students may well look very different. Although individual-level fixed effects are already included in the model, this may not prevent our estimates from being biased if relations are non-linear. For instance, if sensitivity for quality depends on the average school quality in one's choice set and choice sets for advantaged and disadvantaged students differ in that respect, our linear model will indicate heterogeneity in preferences for school quality. In order to moderate this, we add interaction terms with average school quality in the choice set ($SchoolQual * AvgSchoolQual$) (and its quadratic term) to the model.

We refine our measure for heterogeneity in preferences for distance, by interacting with an income and language effect instead of the general dummy for disadvantage. We now find that low-income parents are less sensitive to distance. It could be that for these parents, the opportunity costs of bringing their children to a more distant school is lower. We also notice that those speaking a foreign language are more likely to choose a nearby school (in line with the work by Kristen for Germany [10]).

We also add an another interaction term, for school composition and neighbourhood composition ($SchoolComp * NeighComp$). Neighbourhood composition is derived from one's closest (30) neighbours of school-going age. We compute these interaction terms based on the percentage of one's neighbourhood low-income share ($\overline{LowInc_n} * \overline{LowInc_s}$), foreign language share ($\overline{ForLang_n} * \overline{ForLang_s}$), and education level ($\overline{YoS_n} * \overline{YoS_s}$). The subscript n indicates that averages are taken at neighbourhood level. A positive interaction term $\overline{LowInc_n} * \overline{LowInc_s}$, for instance, shows that students are on average more likely to attend a poor school (i.e. a school with a relatively poor student composition) when they live in a poor neighbourhood themselves. Including these interaction terms diminishes the importance of the composition terms, suggesting that part of the preferences for school composition are determined (or at least correlated) with the place where people live.

Including the interaction term with average school quality in the model does not make the heterogeneity for school quality go away. It seems that differences in choice sets, at least to the extent that we capture them, does not explain why various social groups react differently to school quality. Nevertheless, the coefficient for this interaction term is significantly negative, implying that parents pay more attention to school quality when average school quality is low.

	Cond. logit coeff. (std err)	RO logit coeff. (std err)
$Distance$	-2.043***	-1.373***
	(0.11)	(0.06)
$Distance^2$	0.377***	0.282***
	(0.03)	(0.01)
$ForLang * Distance$	-0.021	-0.229***
	(0.08)	(0.05)
$LowInc * Distance$	0.691***	0.428***
	(0.12)	(0.07)
$\overline{YoS_s}$	-0.297	0.075
	(0.24)	(0.14)
$\overline{YoS_s}^2$	-0.030*	-0.036***
	(0.01)	(0.01)
$YoS * \overline{YoS_s}$	-0.019	-0.016
	(0.02)	(0.01)
$YoS * \overline{YoS_s}^2$	0.002*	0.002**
	(0.00)	(0.00)
$\overline{ForLang_s}$	0.973	-1.337**
	(0.93)	(0.48)
$\overline{ForLang_s}^2$	-1.338	0.882
	(0.95)	(0.47)
$ForLang * \overline{ForLang_s}$	2.876*	1.490*
	(1.16)	(0.62)
$ForLang * \overline{ForLang_s}^2$	-1.265	-0.385
	(1.16)	(0.62)
$\overline{LowInc_s}$	-0.776	-0.233

	(0.94)	(0.49)
$\overline{LowInc_s}^2$	0.188	-0.940
	(0.98)	(0.52)
$LowInc * \overline{LowInc_s}$	1.425	0.676
	(1.75)	(1.08)
$LowInc * \overline{LowInc_s}^2$	-0.549	-0.063
	(1.94)	(1.18)
$\overline{YoS_n} * \overline{YoS_s}$	0.062***	0.043***
	(0.01)	(0.01)
$\overline{ForLang_n} * \overline{ForLang_s}$	-2.515*	-1.301*
	(1.19)	(0.59)
$\overline{LowInc_n} * \overline{LowInc_s}$	0.350	1.979*
	(1.54)	(0.81)
$SchoolSize$	0.003***	0.001***
	(0.00)	(0.00)
$Catholic$	0.301	-0.035
	(0.21)	(0.12)
$YoS * Catholic$	-0.070***	-0.046***
	(0.02)	(0.01)
$ForLang * Catholic$	0.232	0.221***
	(0.12)	(0.07)
$LowInc * Catholic$	-0.275	-0.220*
	(0.15)	(0.09)
$SchoolQual$	9.159**	8.547***
	(3.14)	(1.76)
$SchoolQual^2$	-8.342*	-8.342***
	(3.38)	(1.85)
$YoS * SchoolQual$	-0.110	-0.128
	(0.17)	(0.09)
$YoS * SchoolQual^2$	0.120	0.214*
	(0.19)	(0.11)
$ForLang * SchoolQual$	-2.270	-0.292
	(1.69)	(0.91)
$ForLang * SchoolQual^2$	1.100	-0.767
	(1.62)	(0.87)
$LowInc * SchoolQual$	1.901**	1.304**
	(0.71)	(0.41)
$\overline{SchoolQual_{CS}} * SchoolQual$	-12.084*	-9.053**
	(6.11)	(3.32)
$\overline{SchoolQual_{CS}} * SchoolQual^2$	8.204	5.512
	(5.55)	(3.07)
pr2	0.1227	
N	29474.000	33574.000

Table 4: Estimation results for conditional and rank-ordered logit models on realised choice data, controlling for choice set constraints

These estimations can be represented graphically. We graph the tradeoff faced between two schools by advantaged and disadvantaged parents. In each chart, we let one determinant vary. In our stylized example, advantaged and disadvantaged parents differ on a couple of characteristics, as shown in table 5. On all other dimensions, the environment and background of advantaged and disadvantaged parents are assumed to be the same.

	Advantaged	Disadvantaged
$ForLang$	No (Dutch)	Yes (non Dutch)
YoS	15 (Tertiary)	9 (Lower-secondary)
$LowInc$	No (no school grant)	Yes (school grant)
$\overline{ForLang_n}$	10%	20%
$\overline{YoS_n}$	13	10
$\overline{LowInc_n}$	10%	30%

Table 5: Characteristics of advantaged and disadvantaged parents (example)

The first figure (figure 1) represents preferences for distance, which we let vary from 0 to 5 kilometers. The outside option (school 2) is located at 1 km. Parents prefer closer school, and advantaged even more so than disadvantaged parents. For greater distances (¿ 3 km) the figure insinuates the relationship changes. However, this is most likely the result of the quadratic function we imposed, in combination with the majority of observations situated closer than 3 km from people's place of residence. For school quality, we encounter a similar phenomenon. The results are also less probable for high values of school quality. But we should keep in mind that less than 10% of schools obtain a "'quality score"' higher than 70%. It may also be that our quality index captures a degree of social elitarism as well, which could deter some parents from listing the school on top of their ranking. For the other school characteristics, heterogeneity is much more significant.

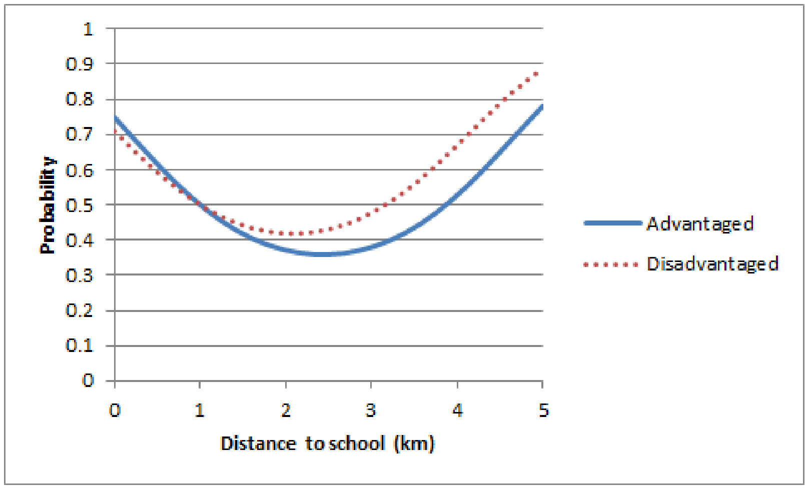

Figure 1: Probability of choosing a school by distance, advantaged versus disadvantaged students

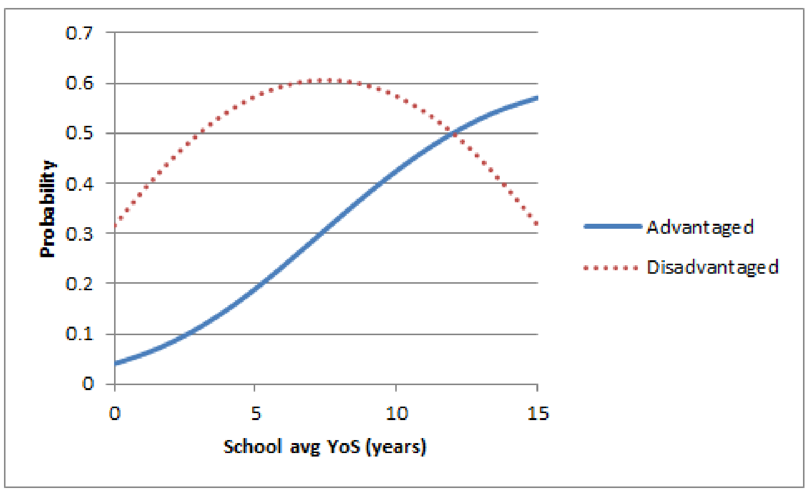

Figure 2: Probability of choosing a school by its average mother's YoS, advantaged versus disadvantaged students

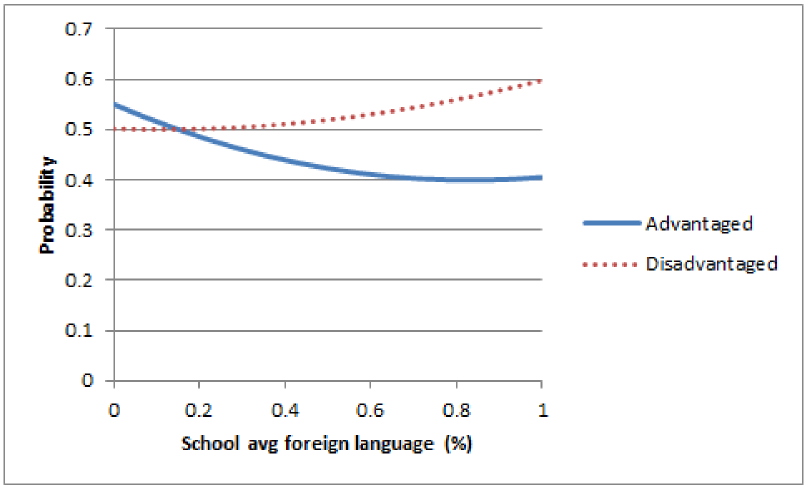

Figure 3: Probability of choosing a school by its percentage foreign language speakers, advantaged versus disadvantaged students

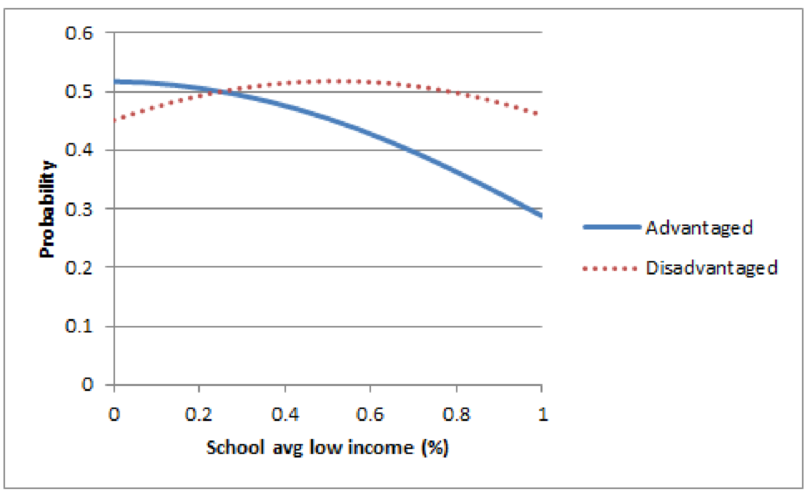

Figure 4: Probability of choosing a school by its percentage low income students, advantaged versus disadvantaged students

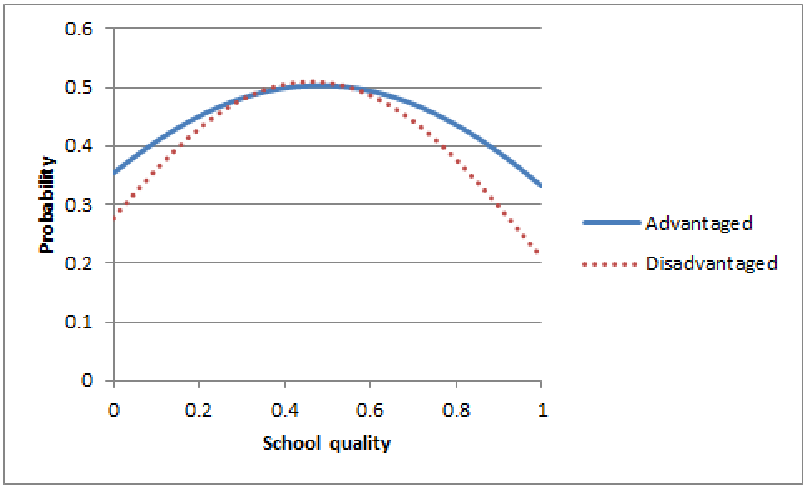

Figure 5: Probability of choosing a school by school quality, advantaged versus disadvantaged students

5.4 Preference data, Ghent, allowing for group-specific information

Lastly, we consider the role of information. More specifically, we want to know whether information is group-specific. Instead of a uniform school quality measure, we adopt a group-specific one. Low-educated parents only consider the performance of children from other low-educated parents that attended the school, while high-educated parents only consider children from other high-educated parents. A school where all children from low-educated parents did well in secondary school, while those from high-educated parents performed less well later, will be considered a high-quality school by low-educated parents and a low-quality school by high-educated parents.

The interactions terms between school quality and education group are no longer significant. Also, the interaction term with average school quality in the choice set has become insignificant. This is a first indication that apparent heterogeneity in preferences for school quality may be the result of group-specific information, rather than differences in preferences per se.

	Cond. logit coeff. (std err)	RO logit coeff. (std err)
main		
$Distance$	-2.010***	-1.372***
	(0.11)	(0.06)
$Distance^2$	0.373***	0.283***
	(0.03)	(0.01)
$LowInc * Distance$	0.746***	0.449***
	(0.12)	(0.07)
$ForLang * Distance$	-0.062	-0.252***
	(0.08)	(0.05)
$\overline{YoS_s}$	-0.056	0.270
	(0.25)	(0.15)
$\overline{YoS_s}^2$	-0.043***	-0.048***
	(0.01)	(0.01)
$YoS * \overline{YoS_s}$	-0.040*	-0.030**
	(0.02)	(0.01)
$YoS * \overline{YoS_s}^2$	0.004***	0.003***
	(0.00)	(0.00)
$\overline{ForLang_s}$	1.548	-1.144*
	(0.91)	(0.47)
$\overline{ForLang_s}^2$	-2.059*	0.482
	(0.90)	(0.45)
$ForLang * \overline{ForLang_s}$	3.352**	2.281***
	(1.06)	(0.56)
$ForLang * \overline{ForLang_s}^2$	-1.036	-0.687
	(1.08)	(0.57)
$\overline{LowInc_s}$	-0.876	-0.148
	(0.93)	(0.49)
$\overline{LowInc_s}^2$	0.324	-1.029*
	(0.99)	(0.52)
$LowInc * \overline{LowInc_s}$	0.682	0.327
	(1.74)	(1.07)
$LowInc * \overline{LowInc_s}^2$	-0.935	-0.540
	(1.95)	(1.19)
$\overline{YoS_n} * \overline{YoS_s}$	0.058***	0.040***
	(0.01)	(0.01)
$\overline{ForLang_n} * \overline{ForLang_s}$	-2.584*	-1.405*
	(1.20)	(0.59)
$\overline{LowInc_n} * \overline{LowInc_s}$	0.413	2.050*
	(1.53)	(0.81)
$SchoolSize$	0.003***	0.001***
	(0.00)	(0.00)
$Catholic$	0.254	-0.069
	(0.21)	(0.11)

$YoS * Catholic$	-0.069***	-0.045***
	(0.01)	(0.01)
$ForLang * Catholic$	0.218	0.197**
	(0.12)	(0.07)
$LowInc * Catholic$	-0.248	-0.218*
	(0.15)	(0.09)
$SchoolQual_{educ}$	10.746**	6.047**
	(3.67)	(1.85)
$SchoolQual^2_{educ}$	-10.134**	-4.468**
	(3.34)	(1.69)
$LowEduc * SchoolQual_{educ}$	-6.099*	0.087
	(2.81)	(1.45)
$LowEduc * SchoolQual^2_{educ}$	5.614	-3.351
	(3.40)	(1.85)
$\overline{SchoolQual_{educ,CS}} * SchoolQual_{educ}$	-10.273	-5.404
	(7.10)	(3.58)
$\overline{SchoolQual_{educ,CS}} * SchoolQual^2_{educ}$	8.530	2.622
	(5.93)	(3.07)
pr2	0.1231	
N	29407	33535

Table 6: Estimation results for conditional and rank-ordered logit models on realised choice data, controlling for other choice set constraints and allowing for group-specific information

6 Conclusion

We set out to capture parental preferences for different school characteristics. We used a unique dataset, containing school preferences parents submitted to a student allocation mechanism in Ghent. We find considerable heterogeneity in preferences for school composition on all dimensions we consider (language, income and mother's education level). These preferences give rise to self segregation: students speaking a foreign language at home are more likely to choose a school where a large part of students also do not speak Dutch as a native language, for instance. Preferences for school quality also vary across social groups, but they go in the same direction. On average, schools with higher-educated parents and higher-quality schools are preferred over others. Linking this to theoretical models on neighbourhood and school segregation, we tentatively conclude that heterogeneous preferences are a main driving force for school segregation.

We also explored where the apparent heterogeneity in preferences for school quality may stem from. We find little evidence that it is caused by differences in choice sets. Instead, it seems that the information parents have at their disposal is group-dependent.

We are aware that an endogeneity problem remains. Especially school composition may be correlated with other unobserved school characteristics. As a result of this omitted variables bias, our coefficients for school composition are likely to be an overestimation. Our future research will give an idea about the size of this bias.

References

[1] P. Bayer, R. McMillan, and K. S. Rueben. What drives racial segregation? new evidence using census microdata. *Journal of Urban Economics*, 56(3):514–535, 2004.

[2] S. Beggs, S. Cardell, and J. Hausman. Assessing the potential demand for electric cars. *Journal of econometrics*, 17(1):1–19, 1981.

[3] S. Burgess, E. Greaves, A. Vignoles, and D. Wilson. What parents want: school preferences and school choice. *The Economic Journal*, 2014.

[4] D. Card, A. Mas, and J. Rothstein. Tipping and the dynamics of segregation. *The Quarterly Journal of Economics*, 123(1):177–218, 2008.

[5] S. M. Glazerman. School quality and social stratification: the determinants and consequences of parental school choice. In *Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998)*, 1998.

[6] J. Hastings, T. Kane, and D. Staiger. Heterogeneous preferences and the efficacy of public school choice. Technical report, National Bureau of Economic Research, 2008.

[7] J. S. Hastings, T. J. Kane, and D. O. Staiger. Parental preferences and school competition: Evidence from a public school choice program. Technical report, National Bureau of Economic Research, 2005.

[8] J. S. Hastings and J. M. Weinstein. Information, school choice, and academic achievement: Evidence from two experiments. *The Quarterly Journal of Economics*, 123(4):1373–1414, 2008.

[9] B. A. Jacob and L. Lefgren. What do parents value in education? an empirical investigation of parents' revealed preferences for teachers. *The Quarterly Journal of Economics*, 122(4):1603–1637, 2007.

[10] C. Kristen. Primary school choice and ethnic school segregation in german elementary schools. *European Sociological Review*, 24(4):495–510, 2008.

[11] T. J. Nechyba. Income and peer quality sorting in public and private schools. *Handbook of the Economics of Education*, 2:1327–1368, 2006.

[12] T. Schelling. Dynamic models of segregation. *Journal of mathematical sociology*, 1(2):143–186, 1971.

[13] M. Schneider and J. Buckley. What do parents want from schools? evidence from the internet. *Educational evaluation and policy analysis*, 24(2):133–144, 2002.

[14] C. M. Tiebout. A pure theory of local expenditures. *The Journal of Political Economy*, 64(5):416–424, 1956.

[15] K. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

7 Appendix

	RO logit 1	RO logit 2	RO logit 3
	coeff. (std err)	coeff. (std err)	coeff. (std err)
$Distance$	-0.490***	-0.460***	-0.492***
	(0.03)	(0.03)	(0.04)
$Distance^2$	0.008***		0.008**
	(0.00)		(0.00)
$SES * Distance$			-0.087
			(0.06)
$SES * Distance^2$			0.001
			(0.00)
$\overline{YoS_s}$		-0.023	0.095
		(0.04)	(0.19)
$\overline{YoS_s}^2$			-0.009
			(0.01)
$YoS * \overline{YoS_s}$		0.008**	-0.010
		(0.00)	(0.02)
$YoS * \overline{YoS_s}^2$			0.001
			(0.00)
$\overline{ForLang_s}$		-0.387	1.415*
		(0.26)	(0.64)
$\overline{ForLang_s}^2$		0.354	0.084
		(0.21)	(0.73)
$ForLang * \overline{ForLang_s}$			-1.719**
			(0.64)
$ForLang * \overline{ForLang_s}^2$			0.345
			(0.76)
$\overline{LowInc_s}$		0.455*	-0.488
		(0.22)	(0.72)
$\overline{LowInc_s}^2$		-0.421	-4.288*
		(0.40)	(1.73)
$LowInc * \overline{LowInc_s}$			0.897
			(0.73)
$LowInc * \overline{LowInc_s}^2$			4.426*
			(1.99)
$SchoolQual$		0.212	1.139
		(0.23)	(0.84)
$YoS * SchoolQual$			-0.087
			(0.06)
$SchoolSize$			0.001***
			(0.00)
$Catholic$			-0.114**
			(0.04)

pr2			
N	7248.000	6724.000	6724.000

Table 7: Estimation results for rank-ordered logit models on realised choice data, choice sets including only ranked schools

	Cond. logit 1 coeff. (std err)	Cond. logit 2 coeff. (std err)	Cond. logit 3 coeff. (std err)
$Distance$	-0.787***	-0.695***	-0.830***
	(0.05)	(0.05)	(0.07)
$Distance^2$	0.015***		0.024***
	(0.00)		(0.01)
$SES * Distance$			-0.020
			(0.12)
$SES * Distance^2$			-0.011
			(0.01)
$\overline{YoS_s}$		-0.062	-0.146
		(0.08)	(0.33)
$\overline{YoS_s}^2$			0.005
			(0.02)
$YoS * \overline{YoS_s}$		0.012*	-0.020
		(0.01)	(0.03)
$YoS * \overline{YoS_s}^2$			0.002
			(0.00)
$\overline{ForLang_s}$		-0.707	2.772*
		(0.45)	(1.13)
$\overline{ForLang_s}^2$			-3.487**
			(1.15)
$ForLang * \overline{ForLang_s}$		0.749	0.725
		(0.39)	(1.28)
$ForLang * \overline{ForLang_s}^2$			0.223
			(1.35)
$\overline{LowInc_s}$		0.354	-0.862
		(0.39)	(1.25)
$\overline{LowInc_s}^2$			1.504
			(1.27)
$LowInc * \overline{LowInc_s}$		0.004	-2.488
		(0.65)	(2.64)
$LowInc * \overline{LowInc_s}^2$			2.902
			(3.09)
$SchoolQual$		0.090	0.435
		(0.40)	(1.58)

$YoS * SchoolQual$			-0.072
			(0.12)
$SchoolSize$			0.002***
			(0.00)
$Catholic$			-0.109
			(0.07)
pr2	0.0747	0.0779	0.0938
N	6583.000	5771.000	5771.000

Table 8: Estimation results for conditional logit models on realised choice data, choice sets including only ranked schools

	Cond. logit 1 coeff. (std err)	Cond. logit 2 coeff. (std err)	Cond. logit 3 coeff. (std err)
$Distance$	-1.709***	-0.616***	-1.969***
	(0.09)	(0.04)	(0.13)
$Distance^2$	0.321***		0.343***
	(0.02)		(0.03)
$SES * Distance$			0.056
			(0.20)
$SES * Distance^2$			0.072
			(0.05)
$\overline{YoS_s}$		-0.223***	0.284
		(0.05)	(0.21)
$\overline{YoS_s}^2$			-0.023
			(0.01)
$YoS * \overline{YoS_s}$		0.022***	-0.018
		(0.00)	(0.02)
$YoS * \overline{YoS_s}^2$			0.002*
			(0.00)
$\overline{ForLang_s}$		-2.554***	-0.050
		(0.34)	(0.87)
$\overline{ForLang_s}^2$			-1.393
			(0.90)
$ForLang * \overline{ForLang_s}$		1.948***	3.593***
		(0.27)	(1.04)
$ForLang * \overline{ForLang_s}^2$			-1.396
			(1.07)
$\overline{LowInc_s}$		0.204	-0.768
		(0.28)	(0.92)
$\overline{LowInc_s}^2$			0.347
			(0.95)
$LowInc * \overline{LowInc_s}$		-0.106	1.639
		(0.41)	(1.82)
$LowInc * \overline{LowInc_s}^2$			-0.606
			(2.03)
$SchoolQual$		-0.134	2.121
		(0.27)	(1.82)
$SchoolQual^2$			-4.440
			(2.41)
$YoS * SchoolQual$			-0.008
			(0.15)
$YoS * SchoolQual^2$			0.108
			(0.18)
$LowInc * SchoolQual$			2.491
			(1.89)
$LowInc * SchoolQual^2$			-0.210

			(1.92)
$SchoolSize$			0.003***
			(0.00)
$Catholic$			-0.522***
			(0.05)
pr2	0.0412	0.0678	0.1136
N	35366.000	29610.000	29610.000

Table 9: Estimation results for conditional logit models on realised choice data, choice sets adjusted for non-ranked schools