College Major Choice, Spatial Inequality and Elite Formation: Evidence from South Africa

Biniam E. Bedasso

ERSA working paper 547

September 2015

Economic Research Southern Africa (ERSA) is a research programme funded by the National Treasury of South Africa.

College Major Choice, Spatial Inequality and Elite Formation: Evidence from South Africa*

Biniam E. Bedasso†

September 10, 2015

Abstract

This paper explores the determinants of college major choice in the presence of significant inter-group and spatial inequalities. I combine four years of admissions application data at an elite university in South Africa with quarterly labor force data to trace the link between aptitude-weighted expected earnings, spatial inequality and the choice of college major. The results show that much of the effect of expected earnings on college major choice operates through the choice of high school curriculum. Black and white individuals respond to differentials in expected earnings differently. Spatial inequality influences major choice through high school curriculum, near-peer role models and relative achievement at high school level. Identification is achieved through the help of a rich set of academic and geographic information contained in the admissions database.

Keywords: College majors; Spatial inequality; Expected earnings; Non-market returns; South Africa

JEL classification: J24; R23

1 Introduction

In an era of increasing specializations and rising wage differentials, not all diplomas are created equal. There are a number of studies in the existing literature that show the link between the distribution of college majors and income inequality (Grogger and Eide (1995), Arcidiacono (2004)). In addition to vertical inequality, the heterogeneity of human capital is also used to explain gender and racial income gaps (Daymont and Andrisani (1984), Weinberger (1998)). From the point of view of aggregate efficiency, the allocation of talent across

*I would like to thank the Information and Communication Technology Services of the University of Cape Town for allowing me access to the admissions database. I am grateful to Kende Kefale for facilitating access to the database. I thank Chris Rooney and Callee Davis for excellent research assistance. Financial support for this research has been provided by Economic Research Southern Africa.

†Economic Research Southern Africa

various fields of specialization with differential impact on innovation and technological change will affect growth in the long run (Murphy et al., 1991). The effect of major choice on the allocation of talent could be far-reaching enough to influence institutional quality (Bedasso, 2015).

This paper examines the determinants of college major choice in a developing country that is characterized by significant inter-group and spatial inequalities.[1] Specifically, the paper attempts to answer two questions: 1). How do various groups respond to differentials in major-specific expected earnings against the backdrop of sizeable inequality in economic and socio-political endowments? 2) How does spatial inequality influence college major choice at the levels of specific neighborhoods and high schools? I exploit the extensive information contained in the admissions database of the University of Cape Town (UCT) between 2010 and 2013, jointly with the Quarterly Labor Force Survey conducted at a national level. The nobility of the dataset lies in the amount information it provides about the high school education and area of residence of college applicants. Moreover, the status of UCT as the best ranked institution of higher education on the African continent allows me to put the analysis in the context of elite formation in a society that has been undergoing social and political transformation.[2]

In standard theoretical frameworks, college major choice is analyzed in the context of a lifecycle model of stochastic career choice (Altonji et al., 2012). This approach establishes a link between educational choices at earlier stages in life and the choice of college major. To the extent that pre-college educational opportunities are determined by space, it is possible to draw a connection between spatial inequality and major choice. There are a number of channels through which spatial inequality may influence individual decisions in college. This includes the quality of schools in a given geographical area, the influence of near-peer role models and the effect of relative achievement in different schools. Essentially, individuals are constrained by all or some of these background factors as they optimize expected lifetime earnings from each major. The level of income expected from a certain major may, however, depend on non-market forces such as political connection which, in turn, is linked to social inequality. In this regard, the effect of inequality is not limited to the constraints of the optimization problem, but it also enters the objective function.

I estimate a random utility model of the determinants of choice between five faculties at UCT with nesting structure elicited from the data. I exploit the availability of two sets of national test scores with varying emphasis on measuring overall ability and college adaptability to identify the effect of aptitude-weighted expected earnings while controlling for academic ability directly. The

[1] The educational inequalities of the Apartheid era continue to persist in South Africa as manifested in disparities in schooling outcomes between historically black and historically white schools (Van der Berg, 2007). Regional inequality is rife in the schooling system in South Africa. As of 2013, less than 44 percent of public schools in one of the poorest provinces offer maths in Grades 10 to 12. The corresponding figure for the best served province is 91 percent.

[2] Two of the most widely cited university rankings – the Times Higher Education Ranking and the QS World University Ranking – consistently rank the University of Cape Town as the best university in Africa.

results show that much of the effect of expected earnings on major choice operates through the choice of high school curriculum. However, the choice of high school curriculum is often dictated by the place of residence of the student among the 242 municipalities in South Africa, indicating the relevance of regional inequality. As far as racial disparities are concerned, white applicants are shown to be, on average, 1.8 times more responsive to major-specific earnings differentials than black applicants. I also attempt to account for the effect of political connections by including an indicator of black applicants from middle-class households coming from municipalities that are electorally dominated by the ruling party. Accordingly, it is shown that those applicants who are likely to have come from a political elite family tend to choose majors in social sciences.

With regard to neighborhood effects, I exploit the variation in past admission trends to UCT across 3773 postcodes represented in the database to capture the influence of near-peer role models. Correcting for possible clustering of unobserved preferences along postcodes, a one standard deviation increase in the ratio of near-peers who were admitted to a certain faculty during the last three years is shown to increase the probability of choosing the same faculty by around 10 percent. I also endeavor to explore if there is a bright side to spatial inequality. Based on the assumption that applicants who belong to the high end of the grade distribution in a less competitive high school have better self-esteem than academically similar students who have gone to more competitive high schools, I test whether being a 'big frog in a small pond' induces applicants to choose more demanding majors. The results show that top ranking applicants from a bottom quartile high school are more likely to choose health sciences over humanities than applicants with higher average grades who nevertheless belong to a lower rung of the grade distribution in a top quartile high school.

This paper belongs in a long vein of literature on career choice under uncertainty (see for example, Altonji (1993), Keane and Wolpin (1999), Arcidiacono (2004), Montmarquette et al. (2002), Wiswall and Zafar (2015)). For that matter, much of the theoretical framework I formulate in the next section is based on the above-cited papers. Most studies have, either implicitly or explicitly, dealt with the impact of major choice on income inequality. However, there has been little focus on the implications of overall inequality, let alone spatial inequality, for the pattern of major choice. The economics and sociology literatures are rather rich in empirical evidence on the impact of geographical segregation on educational outcomes (Braddock II (1980), Card and Rothstein (2007), Goldsmith (2009)). But there is still little evidence on the effect of neighborhood characteristics on educational outcomes in the context of heterogeneous human capital. This paper contributes to the literature in four ways. First, it takes advantage of extensive geographic variation in the data to analyze the determinants of college choice in perspective with spatial inequality. Second, it refines the measurement of expected earnings by explicitly incorporating individual-specific proxy for the ex-ante probability of success. Third, it takes account of the effect of potential non-market returns on the choice of college major. Fourth, it exploits big administrative data to estimate the determinants of major choice in a developing country, which would otherwise have been impossible due to the absence relevant survey data.

The remainder of the paper is organized as follows. Section two outlines the theoretical framework and the empirical strategy I apply to answer the questions raised in this paper. Section three provides an overview of the data and summary statistics. Section four is devoted to presentation of results. Section five concludes.

2 Theoretical framework and empirical strategy

The choice of college major is part of a dynamic lifecycle decision making process. In most cases, the sequence of choices starts as early as high school where students and their parents may have to choose the type of curriculum to be pursued. This will then be followed by a series of choices on whether to go to college or join the labor market, on which college to attend, and what major to choose. The last stage of decision making often consists of the choice of occupation conditional on education. Since I am dealing with the choice of college major at the point of admission, I will start with specifying the value function of individual i who has already chosen to go to college and picked out major j from a set of available alternatives $j = 1, 2, ..., J$,

$$v_{1ij} = E_1(u_{1ij}|a, \theta) + \beta \int E_1(v_{2ij}|a)da \tag{1}$$

where u_{1ij} is the flow utility in period 1 (i.e. college) conditional on ability, a, and preference, θ, both of which might not be directly observable. The second term in equation (1) comprises the terminal value of lifetime earnings expected in period 2 (i.e. work) conditional on ability. β is the discount rate. The utility received by the student while attending college in major j is determined by the psychic and pecuniary costs of major j as well as individual preference for that particular major. This can be written as:

$$u_{1ij} = \alpha_1 C_{ij} + \alpha_2 X_i + \varepsilon_{1ij} \tag{2}$$

where X_i is a vector of individual, school and neighborhood characteristics that may influence the inclination of student i towards major j. ε_{1ij} is the unobserved preference of the individual. C_{ij} is a combination of psychic and pecuniary costs specific to the individual. The individualized cost function is given by,

$$C_{ij} = \gamma_1 A_i(h) + \gamma_2 Z_{ij} \tag{3}$$

where $A_i(h)$ is the observed academic ability of the individual which is assumed to affect the psychic cost of major j through the intensity of effort required to complete college. At the point of college admission, observed ability is a function of high school preparation, h, and - therefore - conditional on choices made earlier in life. Z_{ij} is the pecuniary cost of major j, which is assumed to be specific to individual i because the relative cost of college varies across socioeconomic background. This is because credit markets are imperfect. Substituting

(3) in (2), the flow utility of attending college in major j can be parameterized as follows,

$$u_{1ij} = \alpha_1\gamma_1 A_i(h) + \alpha_1\gamma_2 Z_{ij} + \alpha_2 X_i + \varepsilon_{1ij}$$

The second component of the value function in equation (1) can be simplified into a discrete form based on the following assumption: unobserved ability affects expected earnings through only the probability of successfully completing college in major j. This assumption is innocuous, as far as the present paper is concerned, as the measurement of expected earnings I use in the empirical section is averaged over individuals. Accordingly, the expected earnings of individual i in major j can be written as,

$$E_1(v_{2ij}) = p_{ij}(a)e_j^g + (1 - p_{ij}(a))e_j^d \tag{4}$$

where $p_{ij}(a)$ is the probability of success which is assumed to be a function of ability, a. e_j^g and e_j^d represent expected income after graduating in major j and expected income after dropping out of college, respectively.

I can now combine equations (4) and (5) to arrive at a reduced form equation that can be estimated empirically. This means the indirect utility for individual from majoring in can be written as a linear function of observed academic ability, individual and neighborhood characteristics, and expected earnings:

$$v_{1ij} = \alpha_1\gamma_1 A_i(h) + \alpha_1\gamma_2 Z_{ij} + \alpha_2 X_i + \mu_\beta[p_{ij}(a)(e_{ij}^g - e_{ij}^d) + p_{ij}e_{ij}^d] + \varepsilon_{1ij} \tag{5}$$

Before moving on to the issue of estimation, it remains to be shown, in the context of the present paper, how this model will be employed to explain the relationship between college major choice and socioeconomic and political attributes that usually shape individual behavior in a place like South Africa. First, the expected return to a given major does not come solely from labor market earnings. In many cases, there is social and political benefit to be reaped from majoring in certain fields. It is difficult to directly measure the non-market returns that will accrue to individuals due to major j. However, it can be assumed that the non-market returns to a college major depend on the social and political capital of the individual. Therefore, using individual and group characteristics in vector X_i that can be taken as indicators of social and political capital, it is possible to indirectly measure the effect of non-market returns on the choice of college major. Second, the choices made in high school, which are expected to influence major choice through A_i, are constrained by geographical space. In countries such as South Africa that are characterized by significant spatial inequality, neighborhood characteristics affect major choice both directly through shaping preferences and indirectly through . Hence, provided that there are appropriate constructs to measure socio-political capital and neighborhood characteristics, the model specified in equation (6) can be adapted to explain the choice of college major in South Africa.

The parameters in equation (6) can be estimated using standard techniques as long as certain assumptions are made regarding the distribution of unobservables. First, I assume that ε_{1ij} is independently distributed across individuals once relevant neighborhood characteristics are controlled for. Second, I assume ε_{1ij} is correlated across majors sharing certain characteristics. For example, the unobserved preference of a student for a physics degree is likely to be correlated with the unobserved preference for a chemistry degree. The same might not hold between physics and history. Formally, suppose there are N groups of academic streams across which the J majors are distributed. This means a student will have to choose an academic stream $n = 1, ..., N$ before deciding on which specific major to pursue. This implies that the final choice set can be written as follows:

$$j\epsilon\{(j_{11}, ..., j_{J_1 1})(j_{12}, ..., j_{J_2 2}), ...(j_{1N}, ..., j_{J_N N})\}$$

The above type of structure leads to nested logit probabilities with J alternatives and N nests. Suppose the full set of explanatory variables can be split into explanatory variables of nest choice, W_n, and explanatory variables of the choice of specific alternatives, S_{nj}. Note that I have dropped the individual index i and the period index 1 for simplicity. This means equation (6) can be re-written as

$$v_j = W_n\kappa_n + S_{nj}\lambda_j + \varepsilon_{nj} \tag{6}$$

Provided that ε_{nj} is drawn from a generalized extreme value distribution and following Amemiya (1985), the probability of choosing the j^{th} alternative belonging to the n^{th} nest is given by,

$$\begin{aligned} \Pr(j) &= \Pr(j|n)\Pr(n) \\ &= \frac{\exp(S_{nj}\lambda_j/\tau_n)}{\sum_{j\epsilon J_n}\exp(S_{nj}\lambda_j/\tau_n)}\left(\frac{\exp[W_n\kappa_n' + \tau_n IV_n)]}{\sum_{n\epsilon N}\exp[W_n\kappa_n' + \tau_n IV_n)]}\right) \end{aligned} \tag{7}$$

where $IV_n = \ln\left(\sum_{j=1}^{J_n}\exp(S_{nj}\lambda_j)\right)$ and τ_n measures the correlation of ε_{nj} within a given nest n.

The coefficients in (8), including the dissimilarity coefficient , can be estimated using full information maximum likelihood method. I will return to the issue of identification once I have presented the data in the next section.

3 Data

I use the admissions application database of the University of Cape Town (UCT) to estimate the model of major choice specified in the previous section. The data is available for the entire population of applicants, which ranges between 13897 and 16077 applicants per year, for the four years from 2010 until 2013. In addition to basic demographic information and academic record, UCT has started collecting data on family background and high school characteristics from all applicants as of 2013. Therefore, I use the population of applicants in 2013 to estimate the model. Moreover, information from previous years is employed to identify neighborhood trends based on patterns of admission of applicants from a given neighborhood across the four years. The application database is rich in geographic information as applicants could be traced to the level of residential postcode. Moreover, the names of the high schools of the applicants are given in the dataset. This means I can exploit spatial variations across 3773 postcodes and 1577 high schools. Much of the analysis is based on the entire population of applicants, regardless of admission status. This is done in the interest of minimizing the selection bias that may arise due to the university admission process as well as the decision of applicants on whether or not to accept offers. The first choices of students at the point of application are supposed to reflect the revealed preferences of individuals before the imposition of supply constraints. Results based on the admitted sub-population will be presented to check robustness.

Estimation of the model specified in the previous section requires data on expected earnings as well as socio-political capital. The combination of earnings and college major data is available for the 3rd and 4th quarters of the Quarterly Labor Force Survey (QLFS) in 2012. Ideally, the expected lifetime earnings of applicants in each field of study are estimated using a sample of current workers of various age groups who share similar characteristics as the applicant. However, this approach can be problematic in the case of South Africa because the earnings trajectories of mature workers in 2012 are likely to have been shaped by a distorted system during apartheid. This dynamic is unlikely to hold in the post-apartheid period. Therefore, I have restricted the age limit of the sample of workers that are used to calculate expected earning to 30 years.[3] Another challenge, in this case, is that there are not enough observations of college educated under 30 years-olds in the QLFS data to estimate earnings on a full range of personal and group characteristics. Therefore, I have chosen to use the median earnings of the under 30 years old sample for each major to approximate expected earnings of college applicants.

One of the most potent non-market forces that are believed to shape earnings trajectories in today's South Africa is political connection (Southall, 2004). Unfortunately, there is no direct measure of political connection in the admissions data. Hence, I try to construct an indicator of the potential political capital of individuals using triangulation of three established facts. First, the current political elites in South Africa belong to the officialdom in administrative regions governed by the African National Congress (ANC). Second, the ANC employs cadre deployment to fill in senior public service positions in local administration (Cameron, 2010). Third, civil service jobs are one of the surest ways to join the middleclass in South Africa (Herbst and Mills, 2015). Therefore, a black middleclass college applicant from ANC-dominated municipality can be judged to have a potentially high political capital. Accordingly, I use data on the percentage of votes received by the ANC in the 2009 general elections across the 242 local and metropolitan municipalities. Political capital is measured by the interaction of ANC votes in the municipality where the applicant resides in and dummy variables for private school attendance and black.

[3] This choice makes more practical sense particularly if one assumes that college students are able to observe the incomes of those who are the in same generation as they are more closely than that of older people.

Table 1 presents a summary of variables that are used to estimate the model of major choice. The categories of choice are organized in terms of official faculties at UCT. I have combined the faculties of humanities and law to simplify the choice set into to five faculties: three in science and technology (health sciences, engineering, science) and two in social sciences (commerce, humanities and law). The admission rates show that health sciences and engineering are by far the most selective faculties. Academic preparation is measured by six variables displayed in the second panel of Table 1. Firstly, scores in the three compulsory subjects of the National Senior Certificate (NSC) examination - English, mathematics and life orientation - are applied to measure basic academic ability. Secondly, scores in the National Benchmark Test (NBT), which is being used as an additional requirement for admission into major universities, is employed to measure the probability of success in college. The NBT is specifically designed to gauge the adaptability of students to college curriculum. The test is given in three modules: academic literacy, quantitative literacy and mathematics. Thirdly, the number of science courses a student has taken in high school is used as a proxy for the level of preparation in science and technology. The high standard deviation, relative to mean, of the number of science courses indicates that high school graduates vary significantly in terms of their preparation in science.

The third panel in table 1 presents variables that are used to measure socioeconomic background. A majority of the applicants attended private high schools, signifying the middleclass status of most applicants. In terms of the educational background of families, over 47 percent of applicants would be the first ones in three generations to have earned a college degree. The fourth panel in Table 1 contains demographic characteristics of applicants. Women constitute the majority of applicants. Black applicants form 51 percent of the applicant pool. Whites constitute 24 percent while Colords and Indians make up 25 percent.

The bottom panel in Table 1 presents summary of expected earnings for each faculty. The first row contains the median monthly earnings of workers under 30 years of age with 3- or 4-year college degree in each field. The second row presents the median income of workers in the same age group who have completed only a technical school diploma in one of the five fields. The incomes of technical school graduates are utilized to approximate the expected earnings of applicants in case they drop out of college. I apply the formula in equation (5) to calculate the average expected income for each applicant weighted by the probability of success. I use the NBT scores described above to predict the probability of success. It is to be expected that each faculty has a specific requirement of skill sets determining success in that particular field. In order to determine which NBT module to use as a weight to calculate expected earnings in a given field, I run a probit regression of admission probability on the three types of NBT scores for the years prior to 2013. This is done under the assumption that university officials know the skill sets that are required to succeed in a given field and have already incorporated that information in the admission process. Based on the results of the probit regression (not reported here), math scores are used to weight expected earnings in engineering, health sciences and science, whereas academic literacy scores are used to weight expected earnings in commerce and humanities and law. The last row in Table 1 presents the weighted-average expected earnings of applicants should they choose a given faculty.

The next step concerns identification of the nesting structure that is supposed to govern the choice process of applicants. Intuitively, one can assume that applicants differentiate between the broad categories of science and technology on the one hand and social sciences on the other hand before picking specific faculties and departments within those categories. This assumption is corroborated by the clustering of skill sets required for admission along the science and technology vs. social sciences divide. On top of all this, I am able to explicitly show the correlation of choices across faculties because there is data on the first and second choices of most applicants. Table 2 shows that an overwhelming number of applicants choose two majors in the same faculty as first and second choices. When applicants decide to choose a second major in another faculty, they mostly choose within the nests of social sciences and science and technology. These results justify the assumption that unobserved preference for specific majors is correlated within nests. Fig 1 displays the nesting structure that will be used to implement the nested logit estimation in the next section.

At this point, I can turn to the model specified in equation (7) to discuss the sources of empirical identification. Note that the main coefficients of interest in the vectors of parameters, and , relate to the effects of expected earnings, political capital and neighborhood characteristics. The fact that expected earnings are measured by the median income of contemporary workers can be used as an exclusion restriction to identify the effect of expected earnings on the choices of college applicants. However, the effect of median income is conditional on the probability of success which is measured by NBT results. The test scores of applicants in relevant modules are likely to influence major choice through channels other than the probability of success. Therefore, in equation (6) may not be identified independently of unless there is exogenous variation in test results. The strategy I follow to identify involves using the NSC scores in math, English and life orientation to control for unobserved preference that would otherwise be attributed to NBT results. For instance, a high score in NSC math is sufficient to predict whether an applicant would enjoy attending college in engineering. Once that effect is controlled for, the residual effect of NBT, which is designed to measure adaptability of prior knowledge to college curriculum, can be assumed to be affecting major choice through only the probability of success.

The effect of political capital is identified because individual variation is pivoted on the variation in ANC votes across municipalities. Likewise, the coefficients of neighborhood effects are identified because they are estimated based on geographic variations. There is a valid concern that unobserved preference for educational outcomes could be correlated with the choice of residence. In other words, parents choose decent neighborhoods and good schools in the interest of better educational outcomes for their children. However, it is unlikely that parents would choose one neighborhood over the other because they want their children to study, say engineering, in college. The exclusion restriction becomes stronger particularly once I control directly for high school preparation.

4 Results

4.1 Determinants of high school curriculum choice

Both theory and existing empirical evidence suggest that the choice of college major is conditional on high school preparation. More accurately, the choice of high school curriculum is often made in an anticipation of a certain college major and career path. This means a major part of the decision is already made in high school. However, individuals make two more nontrivial decisions at the point of college application. First, they can choose whether or not to switch the broad field they pursued in high school based on the information they have discovered about their own aptitude during high school. Second, whether or not they are switching fields, college applicants need to choose majors that are more distinctive than high school curriculums. I begin the presentation of results with the determinants of the choice of high school curriculum.

Table 3 presents the determinants of the number of high school science courses students take. All explanatory variables are basic demographic and socioeconomic variables that are also included in the college major choice estimations to be presented later. Both black and white students are found to take less science courses than Colored and Indians. The dummies for female and first-generation college are also negative while public high schools are associated with more science courses. The positive coefficient of public high school is probably because those students who apply to an elite university such as UCT from public high schools self-select themselves on their inclination for technical subjects. At the level of high school, there is no correlation between potentially high political capital and the choice of curriculum. More importantly, there are 240 municipality fixed-effects that are estimated as part of the parameters in Table 3. These fixed-effects are used to capture the impact of one's location in the country on the number of science courses taken by the student. Given the significant level of regional inequality in South Africa, the variation is likely to originate from the differential availability of science education across the country. I employ the municipality fixed-effects alternately with the number of high school science courses to control for secondary school preparation in estimating the determinants of major choice in the following sections.

4.2 The effects of market and non-market returns

This section focuses on the role of expected labor market earnings and potential non-market returns, as measured by an indicator of political capital, on the choice of college major. The choice between science and technology on the one hand and social sciences on the other hand, which represents the first-level decision according to the nesting structure in Fig 1, is specified to be a function of high school curriculum and political capital. The working hypothesis regarding the effect of political capital is that the children of the political elites tend to choose majors in social sciences because they expect more returns in those fields from non-market sources than in science and technology. The choice of specific faculties, which is modeled as the second-level decision, is a function of major-specific expected earnings, family background and demographic controls

Table 4 presents the coefficient estimates of the random utility model of major choice. Since the magnitude of the coefficients in Table 4 is not directly interpretable, average marginal effects are presented in Table 5 to provide intuitive interpretation of the estimates. A comparison of columns (1) and (2) in Table 4 shows that expected earnings have a positive overall effect on major choice until high school curriculum is directly controlled for. There are two potential explanations for the vanishing of the statistical significance of the expected earnings coefficient in column (2). First, expected income influences the choice of college major through an earlier choice in the form of high school curriculum. Second, the coefficient in column (1) is picking up the correlation between expected earnings and unobserved preferences, which is removed once high school curriculum is controlled for. However, the second explanation can be ruled out on the basis that the endogenous component of expected earnings is already removed by controlling for the same effect in the NSC scores. So long as the two sets of national tests, NSC and NBT, are interchangeable in terms of capturing preferences, there would be no significant correlation left between expected income and unobserved preferences when NSC scores are accounted for.[4]

Apart from income, many of the conventional controls such as math score and gender appear with counterintuitive signs when high school curriculum is directly included for the full population. Therefore, it is fair to judge that a major part of the decision about college major takes place in high school. That is why it is important to understand how regional inequality impacts the type of human capital students receive in college through the persistent effect of disparities in high school curriculum. This effect is captured in Table 4 using the municipality fixed-effects, kept from estimations in Table 3. The results show that individuals who come from municipalities with positive impact on the number of science courses students take in high school tend to choose science and technology in college.

The first two specifications in Table 4 are repeated for the black and white sub-populations in columns (3) to (6). The most notable difference from the full population results is that the coefficient of expected earnings is marginally significant even after controlling for high school curriculum for the white sub-population. The differential response to expected earnings between black and white applicants is illustrated further in the average marginal effects presented in Table 5. In the case of black applicants, a one standard deviation increase in

[4] Estimating specification (1) without NSC maths scores increases the coefficient of expected earnings by more than a third.

expected earnings in engineering and science raises the probability of choosing those fields by 1.4 percent and 0.7 percent, respectively. The corresponding marginal effects are 2.1 percent and 1.2 percent for white applicants. In commerce, the marginal effect for white applicants is double that of the marginal effect for blacks. The marginal effect of the municipality fixed-effects on the choice of science and technology for both population groups is significantly higher than the effect of expected earnings on specific majors. A one standard deviation increase in municipality fixed-effects raises the probability of choosing science and technology by 23 percent for blacks and 20 percent for whites.

The results regarding the effect of non-market returns suggest that potentially high political capital draws college applicants away from science and technology. Columns (7) and (8) show that black applicants who have gone to private schools choose social sciences over science and technology as the percentage of ANC votes increases in their municipality. This result holds strongly with or without controlling for high school curriculum. Note that the same indicator has no statistically significant effect on the number of science courses students take in high school as shown in Table 3. This indicates that the rise in the importance of political capital at the point of college application is likely to be driven by a strategic choice in favor of certain career paths. The elite status of the applicants makes a critical difference here. As a comparison between columns (7) and (9) or columns (8) and (10) shows, black applicants who hail from ANC-dominated municipalities, taken as a whole, tend to choose science and technology over social sciences. The effect becomes negative when the domain is restricted to private school graduates, increasing the likelihood that those applicants are the children of the local political elites. The marginal effects in Table 5 demonstrate that a one standard deviation increase in the percentage of votes won by ANC in a given municipality reduces by 10 percent the probability that a black middleclass applicant chooses science and technology.

4.3 Spatial inequality and neighborhood effects

I have begun to account for the effect of spatial inequality on major choice through the use of municipality fixed-effects in the previous section. Nevertheless, spatial inequality in South Africa is much finer than differences at the municipality level. This section focuses on neighborhood effects down to the postcode and specific high school levels. First, I construct a measure of neighborhood-level admission trends for each of the 3773 postcodes. This variable measures the ratio of students admitted to UCT to a given program from a single postcode out of the total number of students admitted from the same geographical area in the same year. This is calculated using admissions data between 2010 and 2012. I expect this variable to capture the influence of near-peer role models in a given neighborhood on the career decisions of college applicants. The concept of nearness in this context has both generational and spatial dimensions. The neighborhood-level admission variable might be picking up a correlation between the unobserved preferences of applicants in the same neighborhood. Hence, in order to isolate the temporal near-peer effect, I correct for clustered errors at the postcode level.

In addition to influencing current decisions through past role models, the neighborhood effect might impact major choice by shaping the beliefs of individuals about their own ability. This hypothesis draws on established arguments about the 'frog-pond' effect in the sociology of education literature (Davis (1966), Espenshade et al. (2005)). The hypothesis predicts that a high-achieving student from a relatively less competitive school, i.e. a Big Frog in a Small Pond (BFSP), tends to choose a more demanding major in college than a relatively low-achieving student from a more competitive school, i.e. a Small Frog in a Big Pond (SFBP). This effect, if it holds empirically, might contribute to counterbalance the impact of spatial inequality on major choice. In order to assign frog-pond identification to the population of applicants in the data, I create a pool of 1557 high schools with at least five applicants to UCT between 2010 and 2013. Then I calculate a three-subject average NSC score for each school based on average scores in math, English and life orientation. This makes it possible to allocate schools across a distribution of average grades divided into four quartiles. The next step is creating school-level quartiles of students using the same measurement as above. Table 6 displays the average grades for the 16 student-school quartile combinations. The tightness of the distribution indicates that applicants to UCT already self-select themselves on their high school grades. I have identified one pair of BFSP and SFBP based on the criteria that the 'big frogs' belong to the top quartile in their school even if their grades on average are lower than the 'small frogs' in a more competitive school. The average score in the SFBP cell in Table 6 is statistically greater than the average grade in the BFSP cell. I compare the effect of belonging to the BFSP cell as opposed to belonging to the SFBP cell on major choice, controlling for all other cell categories.

Table 7 presents the estimated coefficients of the random utility model with neighborhood and school effects described above. The results show that the impact of neighborhood-level admission trend on the choice for the respective major is highly significant under most specifications. The higher the number of students that were admitted to a certain field in the previous three years, the more likely it is for current applicants to choose the same field. The marginal effects obtained in Table 8 show that the highest impact of near-peer role models is manifested in the choice of commerce. A one standard deviation increase in the ratio of near-peers who were admitted to commerce faculty during the last three years increases the probability of choosing commerce by 18 percent. On the contrary, the same increase in the case of science results in only 5 percent rise in the probability of choosing science.

Column (3) in Table 7 includes the frog-pond indicators generated based on Table 6. The frog-pond effect is mainly about the self-evaluation of students in high school in comparison with their schoolmates. Therefore, the effect is more accurately measured after controlling for choices in high school such as high school curriculum. The results show that there is an indication of a frog-pond effect in the choice of commerce, health sciences and science. The effect is statistically significant at 4 percent level with regard to the choice of health sciences whereas the significance is limited to 6 percent and 9 percent in the case of science and commerce, respectively.

These results are unlikely to be spurious because, out of 12 student-school quartile combinations that are characterized by average NSC score lower than the SFBP cell, the only category that has a positive and significant effect on any choice is the BFSP cell. It might be interesting to put these results in perspective with the effect of near-peer role models. The path-dependent effect of neighborhood-level admission trends may suggest that applicants from less competitive schools might end up choosing low return majors. The frog-pond effect adds a twist to that story. After all, the career destiny of some students from less competitive and, presumably, under-resourced schools might be improved by the fact that they have maintained relatively high self-esteem coming out of high school.

5 Conclusion

This paper has set out to examine the determinants of college major choice against the backdrop of inter-group and spatial inequalities. I have used the extensive set of academic, geographic and socioeconomic information contained in the admissions application database of the University of Cape Town to estimate random utility models of major choice. In line with the predictions of the lifecycle model of career choice, the choice of high school curriculum is shown to be crucial for the choice of college major. Accordingly, the choice of high school curriculum is one of the channels through which spatial inequality shapes college level decisions. At a neighborhood level, the influence of near-peer role models on the choices of college applicants is found to be large and significant. Applicants who are likely to have come from political elite families tend to choose social sciences over science and technology, possibly due to higher returns in career in government.

The dynamics of major choice at a selective institution such as UCT is likely to have significant implications for economic and political transformation in the long run through the composition of elites who will be spearheading the process. Primarily, potential inefficiencies in the allocation of talent that might be caused by persistent inequalities will hamper innovation at the top of the socioeconomic pyramid. Moreover, the gravitation of the children of the political elites towards less technical majors may deprive the political class of sufficient interest in productive activities. This, in turn, is likely to leave the elites with little incentives to respect property rights in the future. Hence, policy measures that will improve the availability of science education at high school level or account for the effect of near-peer role models in college admissions may go a long way in terms of affecting economic development.

References

[1] Altonji, Joseph G. 1993. "The demand for and return to education when education outcomes are uncertain." *Journal of Labor Economics* 11: 48–83.

[2] Altonji, Joseph G., Erica Blom, and Costas Meghir. 2012. "Heterogeneity in Human Capital Investments: High School Curriculum, College Major, and Careers." *Annual Review of Economics*, Annual Reviews, 4(1): 185-223.

[3] Amemiya, Takeshi. 1985. Advanced Econometrics. Cambridge, MA: Harvard University Press.

[4] Arcidiacono, Peter. 2004 "Ability Sorting and the Returns to College Major." *Journal of Econometrics* 121: 343–375.

[5] Bedasso, Biniam E. 2015. "Educated Bandits: Endogenous Property Rights and Intra-Elite Distribution of Human Capital." *Economics & Politics*, forthcoming, DOI: 10.1111/ecpo.12063.

[6] Braddock II, Jomills Henry. 1980. "The Perpetuation of Segregation across Levels of Education." *Sociology of Education* 53(3): 178-86.

[7] Cameron, Robert. 2010. "Redefining political–administrative relationships in South Africa." *International Review of Administrative Sciences* 76(4): 676–701.

[8] Card, David and Jesse Rothstein. 2005. "Racial Segregation and the Black-White Test Score Gap." *Journal of Public Economics* 91(11-1 2): 1 258-84

[9] Davis, James A. 1966. "The Campus as a Frog Pond: An Application of the Theory of Relative Deprivation to Career Decisions of College Men." *American journal of Sociology* 72:17-31.

[10] Daymont, Thomas N. and Paul J. Andrisani. 1984. "Job Preferences, College Major, and the Gender Gap in Earnings." *Journal of Human Resources* 19(3): 408-428

[11] Espenshade, Thomas J., Lauren E. Hale, and Chang Y. Chung. 2005. "The Frog Pond Revisited: High School Academic Context, Class Rank, and Elite College Admission." Sociology of Education 78(4): 269-293.

[12] Goldsmith, Pat Rubio. 2009. "Schools or Neighborhoods or Both? Race and Ethnic Segregation and Educational Attainment." *Social Forces* 87(4): 1913-1941.

[13] Grogger, Jeff and Eric Eide. 1995. "Changes in College Skills and the Rise in the College Wage Premium." *Journal of Human Resources* 30(2): 280-310.

[14] Herbst, Jeffrey and Greg Mills. 2015. How South Africa Works. Johannesburg: Pan Macmillan.

[15] Keane, Michael P. and Kenneth I. Wolpin. 1999. "The Career Decisions of Young Men." *Journal of Political Economy* 105(3): 473-522.

[16] Montmarquette, Claude, Kathy Cannings, and Sophie Mahseredjian. 2002. "How do young people choose college majors?" *Economics of Education Review* 21: 543–556.

[17] Murphy, Kevin. M, Andrei Shleifer, and Robert W Vishny. 1991. "The Allocation Talent: Implications for Growth." *Quarterly Journal of Economics* 106 (2): 503-530.

[18] Southall, Roger. 2004. "The ANC & black capitalism in South Africa." *Review of African Political Economy* 31: 313-328

[19] Van der Berg, Servaas. 2007. "Apartheid's Enduring Legacy: Inequalities in Education." *Journal of African Economies* 16(5): 849-880.

[20] Weinberger, Catherine. 1998. "Race and Gender Wage Gaps in the Market for Recent College Graduates." *Industrial Relations* 37(1): 67–84.

[21] Wiswall, Matthew and Basit Zafar. 2015. "Determinants of College Major Choice: Identi?cation using an Information Experiment." *Review of Economic Studies* 82: 791–824.

Table 1: Summary statistics of college applications and labor market data

	Commerce	Engineering	Health sciences	Science	Humanities and law	Total
Number of applicants	3359	2600	3738	1263	3257	14217
Ratio of admitted	0.22	0.12	0.08	0.21	0.20	0.15
Academic record						
NSC Math	67.8 (17.2)	69.0 (16.4)	65.9 (17.5)	64.1 (17.1)	59.5 (18.1)	65.4 (17.7)
NSC English	70.5 (9.9)	68.5 (9.8)	71.2 (10.2)	68.8 (9.8)	66.6 (10.5)	69.3 (10.3)
NSC Life Orientation	79.6 (9.3)	78.3 (9.3)	80.8 (8.9)	77.9 (9.2)	75.1 (10.0)	78.4 (9.6)
NBT Math	46.9 (17.3)	48.2 (17.2)	43.6 (16.8)	43.5 (16.4)	36.0 (12.7)	44.7 (17.0)
NBT Academic literacy	63.5 (12.7)	60.2 (14.3)	60.2 (13.9)	61.1 (14.1)	60.9 (13.7)	61.2 (13.8)
Number of high school science courses	1.5 (1.1)	2.4 (0.81)	2.3 (0.71)	2.3 (0.78)	1.1 (0.90)	1.8 (1.0)
Socioeconomic background						
Ratio of public school	0.27	0.33	0.31	0.31	0.27	0.29
Ratio of private school	0.73	0.67	0.69	0.69	0.73	0.71
Ratio of first generation college	0.42	0.47	0.48	0.51	0.47	0.47
Demographic variables						
Male	0.48	0.70	0.31	0.47	0.34	0.44
Female	0.52	0.30	0.69	0.53	0.66	0.56
Black	0.50	0.55	0.53	0.51	0.48	0.51
White	0.24	0.22	0.17	0.26	0.25	0.24
Colored and Asian	0.25	0.21	0.28	0.22	0.25	0.25
Labor market variables						
Undergraduate median earnings: under 30 years old (in Rand)	16000	20000	18000	16000	15000	15000
Technical school median earnings: under 30 years old (in Rand)	10200	11000	10000	10000	9000	10500
Aptitude-weighted expected earnings	11939 (738)	13195 (1978)	12582 (1364)	10545 (1375)	9449 (555)	11487 (1879)

Note: Standard deviation is given in parentheses in case of mean

Table 2: Correlation between first and second choice faculties

	Second choice				
First choice	Commerce	Engineering	Health sciences	Science	Humanities and law
Commerce	0.22	-0.09	-0.12	-0.08	0.02
Engineering	-0.04	0.32	-0.06	0.10	-0.22
Health sciences	-0.11	-0.02	0.27	0.21	-0.22
Science	-0.03	0.08	0.03	-0.11	-0.05
Humanities and law	-0.04	-0.23	-0.12	0.44	-0.15

Table 3: Determinants of science curriculum in high school

Dependent variable: Number of science electives in high school	(1)	(2)
Female dummy	-0.215 (0.02)	-0.216 (0.02)
Public school dummy	0.070 (0.02)	0.107 (0.03)
First generation college dummy	-0.064 (0.02)	-0.061 (0.02)
Black dummy	-0.359 (0.02)	-0.402 (0.03)
White dummy	-0.086 (0.02)	-0.084 (0.02)
Percentage of ANC votes in a municipality		2.66 (0.31)
ANC votes × private school × black		0.098 (0.06)
Municipality fixed effect (240 municipality dummies)	Yes	Yes
Number observations	11644	11644

Note: Standard errors are given in parentheses.

Table 4: Estimates of random utility model of major choice with market and non-market returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Population	Full	Full	Black	Black	White	White	Full	Full	Full	Full	Admitted	Admitted
Expected earnings (‘000)	0.127 (0.024)	-0.010 (0.023)	0.131 (0.040)	0.020 (0.060)	0.245 (0.045)	0.100 (0.054)	0.124 (0.025)	-0.015 (0.023)	0.129 (0.024	-0.013 (0.023)	0.191 (0.054)	0.021 (0.067)
Level 1 choice: Science and technology (Reference: Social sciences)												
Municipality fixed-effects	1.032 (0.087)		1.120 (0.126)		0.901 (0.172)		0.967 (0.132)		0.992 (0.132)		1.052 (0.390)	
Number of high school science courses		1.257 (0.032)		1.427 (0.051)		1.047 (0.059)		1.246 (0.032)		1.243 (0.032)		1.072 (0.076)
Percentage of ANC votes in a municipality							2.749 (0.247)	0.862 (0.156)	2.302 (0.267)	0.379 (0.199)	2.321 (0.653)	0.543 (0.379)
ANC votes × black									0.931 (0.263)	0.846 (0.295)		
ANC votes × private school × black							-0.344 (0.137)	-0.564 (0.153)			-0.685 (0.346)	-1.081 (0.371)
Level 2 choice: Specific faculties (Reference: Humanities and law)												
NSC math												
Commerce	0.064 (0.014)	0.000 (0.000)	0.089 (0.033)	0.123 (0.034)	0.107 (0.035)	0.101 (0.041)	0.066 (0.013)	0.001 (0.000)	0.068 (0.011)	0.000 (0.000)	0.141 (0.049)	0.083 (0.049)
Engineering	0.069 (0.013)	-0.020 (0.004)	0.091 (0.032)	0.089 (0.025)	0.117 (0.033)	0.084 (0.036)	0.061 (0.011)	-0.021 (0.004)	0.062 (0.011)	-0.019 (0.004)	0.156 (0.050)	0.073 (0.042)
Health sciences	0.042 (0.008)	0.001 (0.002)	0.064 (0.023)	0.088 (0.023)	0.054 (0.018)	0.044 (0.014)	0.049 (0.008)	0.001 (0.003)	0.049 (0.008)	0.000 (0.000)	0.117 (0.034)	0.069 (0.035)
Science	0.041 (0.009	0.002 (0.004)	0.061 (0.022)	0.088 (0.023)	0.062 (0.022)	0.051 (0.023)	0.049 (0.008)	0.003 (0.003)	0.049 (0.008)	0.002 (0.004)	0.137 (0.041)	0.071 (0.038)
Other case-specific controls: NSC English, NSC life orientation, Public school dummy, First-generation college dummy, Black dummy, Female dummy												
τ_n science and technology	1.56	-1.17	1.66	0.069	2.43	1.38	0.737	-1.26	0.823	-1.14	0.664	0.067
τ_n social sciences	1.20	0.006	1.89	2.63	1.73	1.61	1.28	0.009	1.30	0.008	1.29	0.755
Log-likelihood	-13144	-12214	-6027	-5532	-3534	-3372	-13148	-12162	-13145	-12164	-2084	-1965
Number of cases	9239	9267	4295	4312	2459	2467	9239	9239	9239	9239	1567	1567

Note: Standard errors are given in parentheses. Coefficients for seven additional controls listed in the table are not reported here.

Table 5: Average marginal effects[1]

Choice	Expected earnings (own effect)		Municipality fixed-effects		Percentage of ANC votes[4]
	Black[2]	White[3]	Black[2]	White[3]	
Science and technology			0.229 (0.040)	0.207 (0.019)	-0.104 (0.032)
Commerce	0.022 (0.004)	0.044 (0.008)			
Engineering	0.014 (0.005)	0.021 (0.012)			
Health sciences	0.020 (0.004)	0.026 (0.011)			
Science	0.007 (0.002)	0.012 (0.002)			
Humanities and law	0.005 (0.004)	0.021 (0.012)			

Note: 1). Standard errors are given in parentheses. 2). Calculation is based on col (3) in Table 4. 3). Calculation is based on col (5) in Table 4. 4). Calculation is based on col (8) in Table 4.

Table 6: Average NSC scores in student-school quartiles

Quartiles of students within a school	Quartiles of schools			
	1st	2nd	3rd	4th
1st	54.5	59.5	62.2	65.2
2nd	63.6	68.4	70.9	73.8
3rd	71.1	75.3	77.8	80.2[1]
4th	78.3[2]	82.9	85.5	87.1

Notes: 1). Small Frog in a Big Pond (SFBP) 2). Big Frog in a Small Pond (BFSP)

Table 7: Estimates of random utility model of major choice with neighborhood and school effects

	(1)	(2)	(3)	(4)
Population	Full	Full	Full	Admitted
Expected earnings ('000)	0.124 (0.027)	0.013 (0.027)	0.065 (0.029)	0.129 (0.060)
Ratio of near-peers admitted to the same faculty	0.884 (0.139)	0.678 (0.194)	0.487 (0.170)	1.135 (0.508)
Level 1 choice: Science and technology (Reference: Social sciences)				
Municipality fixed-effects	0.998 (0.099)			
Number of high school science courses		1.246 (0.041)	1.242 (0.042)	1.071 (0.078)
Level 2 choice: Specific faculties (Reference: Humanities and law)				
Dummy of top-quartile NSC score in a bottom-quartile school (Reference: Dummy of 3rd quartile NSC score in a top-quartile school)				
Commerce			0.717 (0.419)	0.987 (0.757)
Engineering			0.502 (0.422)	1.184 (0.851)
Health sciences			0.926 (0.452)	1.665 (0.860)
Science			0.860 (0.459)	1.284 (0.775)
Other case-specific controls: NSC math, NSC English, NSC life orientation, Public school dummy, First-generation college dummy, Black dummy, Female dummy, 14 school-student quartile combinations				
τ_n science and technology	1.59	0.865	0.687	0.809
τ_n social sciences	0.953	1.49	0.603	0.680
Log-likelihood	-13108	-12244	-10887	-1830
Number of cases	9239	9267	8334	1518

Note: Standard errors are given in parentheses. Coefficients for 21 additional controls listed in the table are not reported here.

Table 8: Average marginal effects[1]

Choice	Ratio of near-peers admitted to the same faculty[2]
Commerce	0.180 (0.028)
Engineering	0.096 (0.041)
Health sciences	0.137 (0.031)
Science	0.053 (0.009)
Humanities and law	0.081 (0.051)

Note: 1).Standard errors are given in parentheses. 2). Calculation is based on col (1) in Table 4.

Figure 1: Nesting structure