Ethnic Diversity and Labour Market Outcomes: Evidence from Post-Apartheid South Africa[*]

Sara Tonini[†]and Peng Zhang [‡]

Job Market Paper
Please click here for the latest version

February 1, 2018

Abstract. This paper investigates how ethnic diversity amongst black South Africans affects their labour market outcomes in the post-Apartheid era. We find that ethnic diversity has a positive impact on the employment rate of the black South Africans, and it only affects ethnic groups with relatively large population size. To address the endogeneity of ethnic composition, we explore the location of historical "black homelands" and argue that districts equally distant to multiple homelands are ethnically diverse. In our instrumental variable regressions, a one standard deviation increase in ethnic diversity index increases employment rate by 3 (5) percentage point in 1996 (2001), which is around 8% (13%) of the average employment rate. We also disentangle the two components in the ethnic diversity index and show that the variation in our diversity index comes from the dispersion of group size. We then propose a model of a coordination game to explain these findings. A more ethnically diverse place has less dispersion of group size, which implies a higher rate of inter-ethnic communication needed to maintain the overall level of social connection. As inter-ethnic communication requires more skills than intra-ethnic connection, people in ethnically diverse districts are motivated to invest more in social skills to be able to communicate with those outside their own group. The acquisition of these social skills makes them better equipped for the labour market. The key mechanism of the model is verified by both numerical simulation and empirical evidence.

Keywords: Ethnic diversity; Employment; Social skill; South Africa; Homeland.

JEL classification: O12, J40, Z13, N37.

[*] We are deeply indebted to Pramila Krishnan, Toke Aidt and Enrico Rettore for their continuous support and valuable suggestions throughout the project. We are also grateful to Kaivan Munshi, Hamish Low, Sriya Iyer, Elie Murard, Arnaud Chevalier, Ingo Isphording and Oskar Nordstrom Skans whose comments significantly improved the paper. We also thank participants in IZA Brown Bag seminar, IZA Labour and Development Workshop, AEHN 2017 Annual Meeting and the Empirical Microeconomics Seminar at the University of Cambridge for helpful suggestions and Jake Bradley, Yujiang Chen, Stephanie De Mel, Sihua Ding, Lu Han, Jiaqi Li, Simon Lloyd, John Spray, Alessandro Tondini, Chen Wang and Jasmine Xiao for insightful discussions. We thank researchers from the University of Cape Town for their support, especially Tim Brophy and Adrian Frith who kindly provide the map of Bantustans. Financial support from the PODER program (Policy Design and Evaluation Research in Developing Countries) as part of the Marie-Curie Fellowship by the European Commission is highly acknowledged.

[†] School of Economics, University of Cape Town, South Africa. Email: sara.tonini-1@unitn.it.

[‡] Faculty of Economics, University of Cambridge, UK. Email: pz247@cam.ac.uk.

1 Introduction

Many developing countries are characterised by a diverse composition of ethnic groups. A growing body of literature studies the link between ethnic diversity and economic performance, which is summarised in Alesina and La Ferrara (2005). The majority of the literature finds a negative association between ethnic diversity and many socio-economic indicators. More ethnically fractionalised communities can experience slower economic development as measured by GDP per capita (Easterly and Levine, 1997). They may also have higher social costs which are reflected in lower levels of trust and participation in social activities (Alesina and La Ferrara, 2000, 2002), inefficient public goods provision (Alesina et al., 1997; Gomes et al., 2016) and higher inequality (Alesina et al., 2016). The ethnic cleavage may also be detrimental to the establishment of a culture of inclusiveness and tolerance which is favorable to economic growth.

Much less is known on how ethnic diversity affects individual outcomes, especially labour market performance which is of great importance in driving economic development (Anand et al., 2016). This paper adds to this micro-level discussion by investigating how ethnic diversity amongst black South Africans (i.e. within-black ethnic diversity) affects their labour market outcomes in post-Apartheid era. We focus on the black people as the black population represents an overwhelming part of the whole South African labour force.[1] Moreover, the inter-ethnic relationship amongst the black can be more important than the black-white (or black-coloured) division in social interaction as the long-term Apartheid regime separated different racial groups and confined their choice of residence, which persists in the post-Apartheid years. It is therefore not common that the black and the white (or the black and the coloured) reside in the same community and have close interactions. Therefore, the coexistence of the black and white (or black and coloured) in the same district may not necessarily imply social interactions between them in reality.

We focus on how the employment rate of the black South Africans responds to the composition of black ethnic groups in the district of their residence.[2] Post-Apartheid South Africa provides a unique and interesting setting for the study of the diversity-labour market nexus. On the one hand, ethnic identity remains distinct even after generations of integration. This is because ethnicity became a salient concept during Apartheid (from 1948 to 1994) when the Apartheid government deteriorated inter-ethnic relationships by reinforcing the ethnic solidarity to prevent black ethnic groups from forming a coalition to fight against the white government (Gradin, 2014). On the other hand, the

[1] For example, according to the census data, the proportion of the white over the whole South African population decreases from 18.2% in 1980 to 10.9% in 1996 with the coloured population staying small and stable (13.7% in 1980 and 11.7% in 1996). Among the working age (15-64) population, the black South Africans make up 74.25% of the whole labour force participants. The majority of workers (i.e. those who have a job) are also black (58.81% black, 17.7% coloured and 22.7% white).

[2] There is literature about ethnic diversity at the workplace level, which shows the complementarities between workers from different cultural backgrounds as a rationale for the existence of a global firm (Lazear, 1999b). We argue that it is better to focus on the ethnic diversity in places of residence than places of work in our setting. Firstly, the model we propose in the paper links ethnic diversity to interactions an individual has ever had in his daily life, which is better captured at places of residence. Secondly, we later on show that in the data around 60% of the whole black population do not have a job, which means the information on their place of work is not available. Thirdly the overwhelming majority of black South Africans live and work in the same magisterial district (i.e. the geographical unit in our analysis). For example, in 1996 census data, the correlation between district of work and district of residence among the whole black population who are employed is 0.98. Therefore it does not make too much sense to distinguish the two concepts.

Apartheid regime has largely destroyed both the regional path dependence in demand of black labour and the intergenerational occupational persistence in labour market outcomes by compressing the educational and job opportunities of the black South Africans universally. The Apartheid government imposes strict labour regulations to prevent the black South Africans from performing semi-skilled and skilled jobs or running their own business in "white" areas. Therefore the post-Apartheid era is the first time since the early 20^{th} century when the majority of the black South Africans could freely make decisions on occupations or set up their own business. Thus contemporaneous labour market outcomes of the black might convey less information on the persistence in regional labour demand and inherited abilities which are the confounders in our analysis.

Baseline results, based on 1996 and 2001 census data, show that black individuals are more likely to be employed in a more ethnically diverse district. Especially they are more likely to work as an employee, as opposed to setting up their own business.

One challenge in interpreting this as a causal relationship is that the formation of ethnic diversity in a district may not be random. For example, if a district has more job opportunities or higher levels of development, it will attract people from diverse ethnic backgrounds. These people will be more likely to be employed simply due to the higher labour demand in those districts. Or if people with some specific characteristics (i.e. higher ability) are attracted by more ethnically diverse districts, they might also perform better than their counterparts wherever they go. A simple OLS regression will generate biased estimates of the effect of ethnic diversity on employment.

We therefore turn to an instrumental variable strategy, which relies on the location of historical black settlements (known as "homelands"). Following the standard assumption in the literature about migration (Alesina et al., 2015), we assume that the magnitude of migration decreases with the distance between the original homelands and the destination districts. In particular, our instrument exploits the fact that a district tends to host a more diverse population if it is equally distant to multiple homelands. On the contrary, a district becomes more homogeneous if it is relatively close to one homeland but far away from the rest. Importantly, the equidistance to multiple homelands remains a strong predictor of ethnic diversity even after controlling for the proximity of the district to the closest homeland. This further confirms that what can be captured by this instrument is not purely the absolute distance to these homelands but the equidistance to multiple homelands.

In our main IV regressions, a one standard deviation increase in ethnic diversity index in 1996 (2001) increases employment rate by 2.98 (4.56) percentage points, which is 8.12% (13.04%) of the average employment rate in 1996 (2001). This positive effect only holds for the black ethnic groups with relatively large population size.

We further decompose ethnic diversity index into the inverse of the number of different ethnic groups and the dispersion of group size. A clear investigation of the mechanism through which ethnic diversity works on labour market outcomes requires disentangling these two components. This can also be solved with our instrumental variable approach. By construction, the number of ethnic groups is fixed in our instrumental variable, which is exactly the number of historical homelands. Therefore the only variation in the instrumental variable comes from the difference in the distance between the destination district and different homelands, which captures the difference in the population size among different ethnic groups in the destination. Both OLS and IV regressions based on this decomposition shows that when the number of ethnic groups is fixed, a more even distribution of group size (which leads to a higher degree of ethnic diversity) in the district of our interest can increase the employment rate of the black South Africans.

We propose a model of a coordination game in the spirit of the literature on social interaction to explain these findings. Utility comes from both intra- and inter-group communication. We assume inter-ethnic communication is more costly than intra-ethnic connection (because one needs to overcome barriers such as language). Given the number of ethnic groups, a more ethnically diverse place has less dispersion of group size, which implies a higher rate of inter-ethnic communication needed to maintain the overall level of social connection. Therefore it is more necessary for people in ethnically diverse districts to invest in social skills to be able to communicate with those outside their own groups. Their labour market outcomes will improve accordingly as these additional social skills can help them in finding jobs, either by reducing search cost or by improving their productivity.

Our key mechanism can also explain why only groups with large population size respond to ethnic diversity. Starting from the situation where everyone in the district invests in social skills in order to participate in inter-ethnic communication, groups with larger size are more likely to deviate from this coordination because they can get enough social connection by intra-ethnic communications. This is especially the case in an ethnically homogeneous place where these are the dominant groups, but is less likely to be the case if the district is more diverse as their population share becomes smaller. For groups with smaller size who heavily rely on inter-ethnic connection, they do not have the incentive to deviate and will always participate in inter-ethnic interaction and invest in social skills regardless of the diversity level.

We conduct both numerical simulation and analysis based on real data to verify the key mechanism of the model. We fix the number of different ethnic groups and explore the dispersion of group size. The results consistently show that holding other parameters constant, less dispersion of group size (i.e. larger diversity) incentivises people to invest in social skills. Our numerical simulation also shows that our results can be reconciled with papers finding the negative correlation between ethnic diversity and economic development, as the level of investment in social skills can potentially decrease with diversity when per unit cost of investment is too high. IV regressions similar to our main analysis based on 1996 census data also find that our proxy of social skills increases with ethnic diversity and this effect only exists among groups with large size.

Contributions This paper contributes to the literature in four ways. Firstly, we find an innovative way to capture the exogenous variation in ethnic diversity. Our instrumental variable has advantages over instruments exploring simple geographical features. For example, distance to certain places is commonly used as an instrument for migration but whether this is orthogonal to economic conditions has been challenged.[3] By construction we control for the distance to the closest homeland and explore the remaining variation in equidistance to multiple homelands, which could be less problematic than

[3] For example, a place close to an economic centre might get the positive spillover from the centre, or a place close to the road might perform better than others simply because the demand for road is higher in a better place.

the simple distance measures. Alternatively, one can use the historical ethnic diversity directly as an instrument for contemporary diversity, as explored in Miguel and Gugerty (2005) who use the historical distribution of ethnic residence in two districts in Kenya as an instrumental variable to study ethnic diversity and public goods provision. Such a historical distribution of ethnic settlements might also be correlated with other factors. For example, they find that places where several settlements intersect are in lack of sufficient public goods provision. This might however just be because public policies are less effective at the border between different districts in general, whether or not these districts represent a diverse composition of ethnic groups. Our instrument mitigates this violation of exclusion restriction by focusing on districts outside these settlements instead of the settlements themselves. More importantly, by construction we can have places relatively far from all homelands but still with reasonably high ethnic diversity level as long as they are equidistant to all homelands. These places are less likely to be affected by the initial conditions of original homelands.

Secondly, our instrumental variable also manages to disentangle the two components in diversity index: number of groups and dispersion of group size. In our instrument the number of different groups is fixed (i.e. the number of homelands is fixed), so the variation only comes from the dispersion of group size. Therefore ethnic diversity has a clear interpretation in our story: a more diverse place means the distribution of group size is more even. Accordingly the employment opportunity is driven by the degree of dispersion of group size, which is directly related to our theoretical model.

Thirdly, we contribute theoretically to the mechanism through which ethnic diversity affects economic performance. Traditional network literature emphasises the importance of group size. In particular, network literature shows that social connection increases with the size of own group, which means the network effect decreases with the degree of ethnic diversity. This indicates a negative association between ethnic diversity and socio-economic outcomes and contradicts our empirical findings. We propose that what drives our whole story is not the absolute amount of social connection but the composition of social interaction. A more ethnically diverse place does not necessarily have more total amount of social interaction, but it has more skill investment because a larger proportion of communication takes place across ethnic lines, which is more challenging than intra-ethnic communication and therefore motivates people to invest more in skills. Furthermore, traditional explanations on why diversity improves labour market performance, such as knowledge spillover, skill complementarity and discrimination, are not completely compatible with our empirical evidence.[4] Our model of coordination game provides a new perspective on how ethnic diversity positively affects labour market outcomes.

Moreover, our mechanism expands the literature on the importance of skill composition in labour market by linking skill mix to ethnic relations. Labour economists have highlighted the importance of skill mix in the labour force (Acemoglu and Autor, 2011). In particular, higher social skills in the workplace can facilitate people's trading of tasks based on each other's comparative advantage, therefore increasing overall productivity (Deming, 2017). Taking a step back, we provide some insight on how to motivate the acquisition of these social skills in preparation for the labour market. Our mechanism shows that this could potentially be achieved by encouraging ethnic diversity of their communities

[4] Detailed discussion is in the theoretical section of the paper.

and stimulating inter-ethnic communication.

Fourthly, we contribute to the literature on South African labour market by emphasising another dimension of inter-group relations in addition to black-white divisions, and showing this also has important implications on labour market outcomes of the black. Studies on South Africa have been focusing on the segregation between black and white while each group within the black population is implicitly seen as being homogeneous. What we show in this paper is that each black ethnic group has distinct features and the inter-ethnic relationship amongst the black population is important in their economic opportunities.

Focusing on the within-black ethnic diversity can also deal with the major obstacles to contemporary unemployment amongst the black South Africans. Banerjee et al. (2008) propose that the stagnancy of the high unemployment rate among the black in post-Apartheid South Africa might be mainly due to high search cost in job hunting and little growth in the informal sectors. On the one hand, social skill acquisition in an ethnically diverse district can reduce this high search cost. On the other hand, as the informal sector is not powerful enough to provide more employment opportunities, black South Africans still rely heavily on jobs in formal sectors where skill complexity is required and social skills can be very important.

Related Literature This paper mainly relates to two strands of literature. The first one is the empirical analysis on the relationship between ethnic diversity and economic development. A general perspective is that ethnic diversity is negatively associated with economic opportunities at the regional level. It is the case especially in African countries characterised by high ethnic fragmentation (Michalopoulos, 2012; Michalopoulos and Papaioannou, 2013).[5] Ethnic fragmentation harms the economic performance in these countries as it is associated with under-investment in public goods (Michalopoulos and Papaioannou, 2013), conflict (Amodio and Chiovelli, ming) and collective action failures resulting from difficulties in imposing social sanctions in diverse places (Miguel and Gugerty, 2005).

Discussions at the micro level are relatively scarce. There is some firm-level microeconometric evidence on the direct effect of ethnic divisions on workers' productivity in Kenya which documents that upstream workers undersupply downstream workers at the sacrifice of total output if these people come from different ethnic groups (Hjort, 2014). Another strand of literature looks at how entrepreneurs from a specific ethnic group make use of their ethnic networks to develop social capital and mobilise resources (Iyer and Shapiro, 1999), but this is not directly linked to ethnic diversity. Thus, how the level of ethnic fractionalisation affects labour market outcomes remains unclear.

Some papers established a causal relationship between ethnic diversity and economic outcomes. The first approach relies on the exogenous change of ethnic diversity in the time dimension, for example due to the implementation of new jurisdictions (Alesina et al., 2016). The second approach is based on natural or quasi-experiments which directly affect the level of ethnic diversity. For exam-

[5] More research in developed world finds support for the positive side of diversity (Andersson et al., 2005; Niebuhr, 2010; Ottaviano and Peri, 2006). The relationship between diversity and economic performance can also be non-linear. For example, Nikolova et al. (2013) use data from the post-soviet states and show that entrepreneurship is increasing in ethnic heterogeneity at low level of diversity, while it loses its positive impact when diversity reaches a certain threshold.

ple, Algan et al. (2016) explore an exogenous allocation of public housing in France at the apartment block level and Dahlberg et al. (2012) make use of a policy on the compulsory allocation of refugees in Sweden. In South Africa, however, ethnic diversity does not change dramatically over time, which means there is not enough time variation to identify changing levels of diversity. It is also hard to find proper natural or quasi-experiments due to the political sensitivity of ethnic topics in this country. Therefore, the above two commonly established identification strategies in the current literature are not feasible in our setting.

The second strand of literature concerns the theoretical models on social interaction. There are two key differences between our model and several models documenting social interactions in response to diversity in current literature. On the one hand, unlike models relying on the intrinsic ethnic-specific parameters of taste, preference or discrimination (for example, Morgan and Vardy (2009) shows minority candidates produce noisier signals of their ability), we show that ethnic diversity still affects people's decision in investments in social skills without documenting those assumptions. This is in line with the recent finding that ethnic diversity can be independent of cultural diversity (Desmet et al., 2017). On the other hand, unlike Glaeser et al. (1992) which requires that communication is more extensive or the amount of social connection is larger in more diverse places (Alesina and La Ferrara, 2000), in our model the overall level of social interaction does not necessarily increase with ethnic diversity (overall social interaction is the sum of intra- and inter-ethnic connections). Ethnic diversity results in more investments in social skills because inter-ethnic communication is more costly (or requires more skills) than intra-ethnic connection.

The mechanism in our paper is the closest to, yet distinct in important aspects from, two existing papers. In the story in Lazear (1999a), he finds that immigrants to the U.S. have higher English proficiency when there are smaller proportions of people from their native country in the communities in their destination. Our paper also documents that people are incentivised to learn English to have access to more potential communication partners (in our story we generalise "language" to a broader concept of social skill). The key difference is that they focus on the assimilation of the immigrants to the U.S and therefore the majority group (i.e. the U.S. native) do not respond to the diversity level in different communities. However, both the theoretical model and empirical findings in our paper show the opposite - only groups with large size (analogue to the U.S. native in his paper) are affected by ethnic diversity whereas smaller groups (analogue to the minority group of immigrants in the U.S.) behave indifferently between ethnically diverse and homogeneous places.[6] What generates this difference is that his model is featured by unilateral assimilation of the immigrants to the U.S. while in our model social interaction and skill investments are bilateral. This makes more sense especially in ethnically diverse places where no ethnic group has overwhelming group size. Also due to strong ethnic identities, groups with smaller size will invest in a common or official language rather than the language of the large group. In our modelling part, we show further that unilateral assimilation is not consistent with our empirical results.

In another model on social interactions between different groups, Alesina and La Ferrara (2000) assume that individuals prefer to communicate with people with similar income, race or ethnicity

[6] We control for the proportion of the black over the whole population in our analysis and focus on within-black communication.

and conclude that homogeneous communities have higher levels of social capital. Instead of making the direct assumption of group-based preference, we treat this as an implicit implication of the model and argue that people have preference towards groups similar to them because the cost of intra-ethnic communication is lower.

The paper unfolds as follows. In Section 2, we provide a historical overview of the pattern and formation of ethnic diversity as well as summary statistics on labour market in South African context. In Section 3, we describe the data sources and how we construct the variables of our interest. Section 4 details the empirical methodology, focusing on the instrumental variable and its validity. In Section 5, we comment on the results about how ethnic diversity affects labour market outcomes in post-Apartheid South Africa and how this impact differs across sub-groups. Section 6 proposes a theoretical model with numerical simulation and empirical evidence to explain the main empirical results and rule out some alternative explanations. Finally we draw some conclusions and policy implications in Section 7.

2 Institutional Setting

2.1 Ethnic groups in South Africa and the formation of ethnic diversity

None of the black ethnic groups are indigenous in South Africa. All of them migrated from eastern and central Africa to southern Africa starting from centuries ago, as part of the so-called "Bantu migration".

Before explaining the narratives, two concepts should be made clear. The first is "homeland" which refers to the original settlements of those ethnic groups when they first moved to South Africa. The second is "white areas" or "white South Africa"[7] which refers to places in South Africa outside those homelands. Many years after arrival in South Africa, those black people moved out of their original homelands and ended up in these "white areas" due to different reasons, mainly the pressure of conflicts with the British and Dutch colonisers as well as other ethnic groups. Therefore, "white areas" are not areas where only white people reside, but places outside original black homelands (the proportion of the black over the whole population can still be large in those "white areas").

Based on Mwakikagile (2010) and Gradin (2014), we provide historical narratives on the mass migration of ethnic groups from central Africa towards South Africa, the original settlements of these ethnic groups and the migration of these people out of their homelands to "white areas" in South Africa. The timeline about the history of the settlements and migration of the black ethnic groups outside their own settlements up to the time of South Africa's independence can be found in the upper panel of Figure 1.

The indigenous groups in South Africa are San and Khoikhoi (both are "coloured" groups) residing in the southwestern and southeastern coast about 2000 years ago. Around 700s A.D., black Africans

[7] It became an official terminology during the Apartheid regime.

had settled in the northern part of what is South Africa today.[8] They were members of different Bantu ethnic groups who had moved southward from East-Central Africa (the Great Lake district around Congo) and spoke related languages.

Ethnicity-specific information on the Bantu migration from eastern and central Africa towards South Africa and the formation of ethnic diversity in "white areas" is summarised in Appendix A. The table contains information on the timing of their migration into South Africa, geographical location of original homelands, timing of migration outside homelands and the Bantustans assigned to them during Apartheid (which will be explained in constructing our instrumental variable). For example, Zulu are believed to be descended from a leader named Zulu born in the Congo Basin area. In the 16th century, they migrated to the south and eventually settled in the eastern part of South Africa, an area now known as Kwazulu-Natal. The Zulu empire in the 1800s witnessed their vast migration and expansion of territory.

One indication from the narratives is that the black had settled in the country long before Europeans arrived. For example, the diaries of shipwrecked Portuguese sailors attest to a large Bantu-speaking population in present-day Kwazulu-Natal by 1552. In 1652 Jan van Riebeeck and about 90 other people set up a permanent European settlement as a provisioning station for the Dutch East India Company at Table Bay on the Cape of Good Hope, beginning the era of European colonisation.

Due to the pressure from the potential conflicts with white colonisers and the other ethnic group-s, the nine black ethnic groups began to move out of their homelands or change their territories. By the early 1700s, there were already some African groups migrating into the interior of the country to shield themselves from European domination. By 1750 some white farmers, known as Boers, expanded to the region where they encountered the Xhosa and Zulu. Starting from 1789, a series of wars and conflicts over land and cattle ownership broke out between the Boers and the black ethnic groups. In early 1800s the British replaced the Dutch at the Cape as the dominant force. The Boer-s, defeated by the British, migrated eastwards into today's Kwazulu-Natal and Free State where the conflicts between the Boers and Zulu people continued. Many other ethnic groups have encountered similar conflicts.

The destination of their migration is not well-documented. This information, however, can be reflected from today's distribution of ethnic groups across South Africa. This pattern of migration will also affect today's distribution of ethnic diversity. For example, a place would be more diverse potentially if more ethnic groups moved in. Details will be shown in the next section. One thing which needs to be emphasised here is that in most of the cases the key driving force of emigration from ethnic homelands is the conflict either with the white or with other ethnic groups rather than the economic benefits in the destination.

Importantly, further evidence shows that the mass migration both from central to southern Africa and from homelands to "white areas" within South Africa took place mainly before the spur of industrialisation and modern economy. The discovery of mineral resources is a milestone in the economic development and transformation towards modern South Africa. Diamonds were first discovered in

[8] Some argue it is as early as the third century (Gradin, 2014).

1867 along Vaal and Orange rivers, and in Kimberley in 1871. In 1886, gold was first discovered in Witwatersrand, around today's Johannesburg, which stimulated trade and construction in large dimensions. All this took place after the Bantu migration. This means the migration from homelands to "white" areas, although not completely random, may not be purely driven by the economic prospects in the destination.

In 1910 the Union of South Africa was established, which declared the superior socio-economic status of the white politically and created a white-dominated society. Since then racial discrimination has been a prominent feature of South African society even before the official institution of Apartheid, and the mobility of the black was largely restricted.

Summary demographic statistics about the nine ethnic groups are reported in Table 1.1 for 1996 data and Table 1.2 for 2001 data. The distribution of population share among these nine groups and their labour market outcomes are similar in these two years. In both 1996 and 2001 there are three out of nine ethnic groups (Xhosa, Zulu and South Sotho) who have relatively large population size (i.e. their share of the whole population is over 20%). We define them as *large* groups. Another two ethnic groups have smaller size (Tswana and North Sotho), and are therefore defined as *medium* groups. The remaining four ethnic groups have much smaller population share (less than 5%) and are defined as *small* groups.

2.2 The role of Apartheid in shaping inter-ethnic relations and labour market outcomes

Since mid-1900s, inter-ethnic relationships and labour market outcomes have been significantly shaped by the Apartheid regime and related regulations. The regime reinforced the ethnic identity and destroyed much of the path dependence in the opportunities for education and labour market for the black. The timeline of the Apartheid regime can be found in the lower panel of Figure 1.

Starting in 1948, the ruling Afrikaner National Party (NP) implemented a program of *apartness* and formalized a racial classification system, which transformed into official *Apartheid* by the 1951 *Bantu Authorities Act* and 1953 *Bantu Self-Govern Act*. Each individual living in South Africa belonged to one of the four races (White, Indian, Colored, Black), which essentially defined an individual's social and political rights. In addition, the government over-emphasised the differences among the various ethnic groups, in the spirit of the *"divide et impera"* principle. The ethnic segregation, on top of the racial separation, was aimed at guarantying the political and economic supremacy of the white minority. This exacerbated division of ethnic groups served as a tool for the white to control the black in an easier way (Gradin, 2014).

With the introduction of the *Promotion of Black Self-Government Act* in 1959, the government delimited a number of scattered rural areas as "native reserves" for blacks (called "Bantustans"), one for each ethnic group. The designated areas for the reserves amounted to 13 percent of the total South African territory, while the blacks accounted for more than 75 percent of the total population. Blacks' land ownership was restricted, as well as their ability to freely move and settle in the white South Africa. Internal migration was severely regulated until the repeal of the *Pass Laws Act* in 1986.

With the forced removal of the blacks from the "white areas" of South Africa, the Bantustans became over-densely populated territories, where land was overgrazed and afflicted with serious soil erosion. The economic development of these reserves never materialized, leaving their inhabitants in acute poverty (Christopher, 2001). In 1970, the regime promulgated the *National States Citizenship Act*, which provided citizenship to blacks in their homelands. The ultimate aim was to create a number of ethnicity-based independent states.

In conclusion, the Apartheid regime used separation along racial lines and ethnic lines as a fundamental device for the demarcation of physical and social boundaries for all interactions.

One thing which needs to be pointed out is that Apartheid did not shift the big picture of the magnitude and distribution of ethnic diversity in these "white areas", despite the campaign of forced-removal during this time. During the Apartheid period, 3.5 million (equivalent to $\frac{1}{5}$ of the black South African population in 1980) were forcibly removed from their homes and dumped in areas designed for the black by the Apartheid government. However, our data shows that this forced removal did not lead to large changes in the pattern of the distribution of black South Africans across "white" districts. In 1996 census data, 79.61% of the black population in the "white areas" of our interest never moved in their life. 11.82% moved within their birth district and only 6.63% migrated across districts. These inter-district migrants did not dramatically change the ethnic diversity of "white" districts, as we find the high correlation of district-level ethnic diversity between 1996 and 1985 (the correlation is 0.88, calculated from 1985 and 1996 census by the authors). Therefore it is still reasonable to link contemporaneous distribution of ethnic diversity to the location of historical homelands, despite the large campaign of black migration during the Apartheid era.

The Apartheid regime also severely limited the job opportunities and resources among the black (Posel, 2001). The *Bantu Education Act* of 1953 ensured that non-whites received a substandard quality of education, while access to occupation was regulated by the 1956 *Industrial Conciliation Act*. Whites were authorized to determine the racial allocation of jobs (Mariotti, 2012) and to reserve certain professions for themselves, especially in the manufacturing sector. In particular, the black were banned from semi-skilled and skilled occupations. Similarly, blacks were not allowed to run their own businesses in white areas. In fact, only with the advent of the democracy, in 1993, non-whites were able to make their free occupational choices. This, together with the reallocation of industries, changed the industrial and occupational structures in white areas, which partly weakened the path-dependence in regional demand of black labour. Moreover, the intergenerational occupational persistence, which has been shown to be particularly relevant for employment (Sørensen, 2007; Pasquier-Doumer, 2012; Magruder, 2010), does not represent a very important issue in the early post-Apartheid era. In other words, blacks may rely more on resources outside their families in overcoming the entry barriers to jobs (barriers such as information about trade partners and market opportunities, informal credit and insurance arrangement).

2.3 Labour market in post-Apartheid South Africa

High unemployment rates and large proportion of discouraged workers remain important issues in the South African labour market in the post-Apartheid era (Bhorat and Oosthuizen, 2005; Leibbrandt et al., 2009). Based on 1996 census data, over 60 percent of the working-age black population are either unemployed or out of labour force. A large share of the unemployed in 2005 have never worked in their life. To make things worse, skill-biased technological changes lead to an increase in capital-labour ratio in late 1980s and the whole 1990s, further reducing demand for unskilled labour. At the same time, real wage has been stable or decreasing between 1995 and 2005 (Banerjee et al., 2008). The increase in the supply of unskilled labours, together with the shrinkage in labour demand due to skill-biased technical change as well as the exodus of the white (who are the owners of capital and factories) largely leads to this persistent unemployment issues in the contemporary South African labour market (Banerjee et al., 2008). Furthermore, there is very low informal employment rate in South Africa, which is only 7.7% - 9.7% based on various measures of informality in September 2004 Labour Force Survey (Heintz and Posel, 2007), possibly because there are also entry barriers in those informal sectors (Kingdon and Knight, 2004). This means the formal wage-employed sector is still the main force in absorbing increased labour supply.

Summary statistics on labour market outcomes based on 1996 and 2001 census data confirm this pattern. In Table 1.1 and Table 1.2, in the overall sample, less than 40% are employed over the whole working-age black population, among which self-employment rate is particularly low (3.2% in 1996 and 2.3% in 2001). The slight rise in unemployment rate from 1996 to 2001 is consistent with the current finding that unemployment rate peaked between 2001 and 2003 in South Africa (Banerjee et al., 2008).

There is, however, large heterogeneity among different ethnic groups. In general groups with medium and small sizes are more active in the labour market and more likely to be employed, both in self- and wage-employed jobs. This indicates that groups with smaller size are in general more active in the labour market and more competitive in job search, which can be explained by the theoretical model later on in the paper.

3 Data

For our empirical analysis, we make use of different data sources. We rely on census data for main analysis. There are three years of census data in the post-Apartheid era: 1996, 2001 and 2011, all of which are the 10% sample from the original national sample in publicly available sources. We do not use 2011 census as both the classification and boundary of magisterial districts have changed dramatically after 2001, making it less reliable to match the new system of magisterial districts in 2011 to the older ones. More importantly, in publicly available 2011 census data, there is no information on which magisterial district each individual resides in. As respondents in 1996 and 2001 census cannot be matched, we use them as two separate cross-sectional data-sets.

The unit of analysis is the Magisterial District (MD).[9] There are 354 magisterial districts in South Africa, with an average territory size of 3447.5 km^2 and average population size of 0.1 million in 1996. It is particularly convenient to use the MD as a small-scale geographical unit for comparative analysis, given that all other administrative divisions have been revised and re-demarcated repeatedly since the first democratic election in 1994. It also provides a reasonably large geographical unit to define labour market. Our final sample consists of 210 districts in 2001 census (205 in 1996 census), which are the "white" areas outside the historical homelands. Take 2001 census as an example. The excluded districts are either part of the homelands and thus had distinct political status and partially different laws and labour market regulations (124 districts)[10], or districts where the black population in 2001 accounted for less than 1% of the overall population (11 districts[11]), or they cannot be matched with 1985 census data that is explored in the instrumental variable approach (9 districts).[12]

Status in employment. In both 1996 and 2001 census data, we construct an individual-level binary variable for unemployment. The dummy takes value 1 if one is unemployed or economically inactive and 0 if one is employed (either self-employed or an employee). Among workers who are employed, we also consider the allocation of them between self-employment and wage-employment jobs. More in details, an individual is considered to be self-employed if s/he declares to be either self-employed, employer or work in the family business. To do this, we create another dummy variable only for employed people. It equals 1 if one is self-employed and takes value 0 if s/he declares to be an employee. We only consider working-age black population (15-64 years old).

Ethnicity. Following Amodio and Chiovelli (ming), the ethnolinguistic group each individual belongs to is identified using the information on the first language they speak in the 1996 and 2001 census. There are nine black ethnic groups in the country: Xhosa, Zulu, Swazi, Ndebele, North Sotho, South Sotho, Tswana, Tsonga, and Venda. Following Desmet et al. (2012), we rely on Lewis' *Ethnologue* tree of ethnolinguistic groups (Lewis et al., 2009) to build our measures of ethnic diversity.[13] For each magisterial district and census year, we calculate the relative shares of each ethnic group within the black population and combine them into ethnic diversity index: the *fractionalisation index*.[14] Universally used in the empirical literature on ethnic diversity (Desmet et al., 2017; Easterly and Levine, 1997; Alesina et al., 2003; Alesina and La Ferrara, 2005), the ethno-linguistic fractionalisation index (ELF) is

[9] We calculate the ethnic diversity of the magisterial districts where individuals reside in. There are three reasons why we do not use district of work for the main analysis. Firstly, the mechanism we provide in this paper regarding how ethnic diversity affects labour market outcomes is more related to the districts where one resides (i.e. places where one has social interaction even before entering the labour force) than where one works, which we will explain in the theoretical model. Secondly, the correlation between district of work and district of residence is very high so that they provide similar information. Thirdly, more than half of the black population are unemployed or out of labour force. Therefore the information on their district of work is unavailable and has to be replaced by the information on district of residence, making the district-level information among this group and that among the employed people less comparable.

[10] The boundary of the homelands does not coincide with the boundary of contemporary MD. Taking a conservative method, we define district with less than 10 % overlap with homelands as "white" districts.

[11] This figure is 16 in 1996 census data, which is why the total number of districts of our interest is 205 in 1996.

[12] OLS regression results remain unchanged if we include the nine districts which cannot be matched with 1985 census data.

[13] The nine black ethnolinguistic groups of South Africa belong to the Niger-Congo language family and correspond to level 11 in the tree of ethnolinguistic groups.

[14] We consider another index: polarization index in the robustness check. It has been proved that fractionalisation index performs better in explaining economic outcomes than polarisation index (Alesina et al., 2003).

a decreasing transformation of the Hirschmann-Herfindahl concentration index and is defined as

$$ELF = 1 - \sum_{k=1}^{m} s_k^2$$

where s_k is the population share of ethnolinguistic group k and m is the overall number of groups. Intuitively, the index measures the probability that two individuals who are randomly drawn from the population belong to different ethnic groups. Larger value of the fractionalisation index indicates higher diversity in the magisterial district.

Figure 2 shows how ethnic diversity, measured by the ELF index, is distributed in the districts of our interest in 1980, 1985, 1996 and 2001. Districts in darker colours are those with higher ethnic diversity. There is large variation in ethnic diversity levels across South Africa. In general, districts in the northeastern part of the country are more ethnically diverse than those in the southwestern part. In addition, some districts in the middle part of the country are the most ethnically diverse ones. These patterns will be explained when we construct instrumental variables. Districts coloured in white are those inside original homelands, with less than 1% of the black population or that cannot be matched to 1985 census data. A cross-year comparison shows that the degree of ethnic diversity in these districts is very stable. The patterns are extremely similar between year 1996 and 2001. The spatial distribution of ethnic diversity during Apartheid (1980 and 1985) is slightly different but places with higher (lower) degree of diversity remain ethnically diverse (homogenous) over time. This reveals that the formation of ethnic diversity is a historical event and not largely driven by contemporary migration. A comparison between 1980 and 1996 (or 2001) confirms that the Apartheid regime did not drastically shift the spatial distribution of ethnic diversity.

A more detailed investigation of the distribution of ethnic groups in districts with different degrees of ethnic diversity is in the last column of Table 1.1 and Table 1.2 for year 1996 and 2001 respectively. Ethnic groups with relatively larger population size (e.g. Xhosa) are distributed in more homogeneous places while small groups (e.g. Venda) are in more diverse districts. This implies that homogenous places are dominated by groups with large size over the national population while the distribution of population size over different groups is more even in ethnically diverse districts.

Demographic, socio-economic and geographical controls. From the censuses, we also derive a number of controls, which we introduce in our regressions either at the individual level or as aggregated information at the district level. Individual characteristics include gender, age, educational attainment, marital status and whether one's father is alive. Among the district-level controls, we consider population density, proportion of the blacks, proportion of people working in manufacturing and service sectors, whether the district is mainly rural or urban, and whether there is a river and road crossing the district. Additionally, we introduce other geographical factors, which can potentially shape the economic activities of a region. Starting from the Mineral Resources Data System[15], we compute the density of mine for each district. Our geographical unit here, magisterial district, is large enough to capture activities related to the mining sector. Furthermore, the density of mine has two advantages over a simple dummy for the presence of mining activities. Firstly, it takes into account

[15] Mineral Resources Data System, MRDS, is a collection of reports describing metallic and nonmetallic mineral resources throughout the world. Spatial data is available at: https://mrdata.usgs.gov/mrds/.

the number of mineral resources in each district as the magnitude of the effect of mines can increase with the number of mines available at the district level. Secondly, it captures the fact that mineral resources have larger economic effects in more condensed districts either due to higher population density or lower travel cost to the mines. In order to account for the agricultural suitability of land, we use the measure of terrain ruggedness from Nunn and Puga (2012).[16] We also include the measure of soil quality as another proxy for agricultural suitability. Data comes from the Harmonized World Soil Database from the Food and Agricultural Organization of the United Nations. It is a discrete index ranging from 1 to 7, with a descending order of soil quality.[17] As a proxy for the economic development at the local level, we use the National Oceanic and Atmospheric Administration night-time light satellite images data for 1996 and 2001 (Michalopoulos and Papaioannou, 2013).[18] We also include the number of conflicts in each district as it has been proved to be correlated with ethnic diversity (Amodio and Chiovelli, ming) and potentially affects economic prosperities. "Conflicts" here incorporate violence outside the context of a civil war, including violence against civilians, militia interactions, communal conflict, and rioting. A detailed discussion of conflicts in post-Apartheid South Africa can be found in Amodio and Chiovelli (ming).

The rationale of taking into account these control variables is to control for the main drivers of economic development especially employment which are correlated with ethnic diversity. A detailed discussion is in the section about empirical model specification. Details on the sources of data and methods in constructing district-level control variables are presented in the Appendix B.

Before looking into the data, it is worthwhile to point out some differences in information collected in 1996 and 2001 census. Firstly, 1996 census distinguishes between those who are unemployed and out of labour force (i.e. economically inactive) while 2001 census combines these two categories. We thus conduct analysis separately as well as jointly for these two groups in 1996 data, and compare the results based on the joint group with the corresponding results using 2001 census.

Secondly, we also explore labour market outcomes other than employment status to enrich our analysis on South African labour market, including wage, income and working hours. information on working hours is only available in 2001 census data. We thus focus on 2001 census in calculating hourly income. In addition, a drawback of the income information in the census data in both years is that it calculates income from all possible income sources, including labour market income, social grant and other sources like bonus, rent or interest. As a result, another dataset (i.e. Labour Force Survey) is required for a more precise measurement of wage, which will be discussed in the empirical results.

Thirdly, 1996 census data asks information on both first and second language spoken whereas 2001 census only asks people about the first language they speak. Therefore, we only look at 1996 census to test our channel of social skill acquisition using proficiency of a second language as a proxy

[16] We also tried the measure of slope from the same data source. The results are very similar. We do not include ruggedness and slope at the same time as they are highly correlated (the correlation is larger than 0.9), which potentially leads to multicollinearity issues in regressions.

[17] In the soil quality index, 1 = No ro slight limitations; 2 = Moderate limitations; 3 = Sever limitations; 4 = Very severe limitations; 5 = Mainly non-soil; 6 = Permafrost area; 7 = Water bodies.

[18] Night-light data is at 30-second grid level. Here we take the average night-time light density within each magisterial district by summing up the night-light measure over these grids and dividing it by area of the district.

for social skills.

Fourthly, in the robustness check, we reinforce our analysis by looking at natives and migrants separately to see if our results are purely driven by the selection of migrants in each district. For migrants in each district, we have full information on the exact year of their migration to the current magisterial districts only in 1996 census. In 2001 census only migration between 1996 and 2001 is recorded. Therefore, in 1996 census data non-migrants are defined as those who either never moved or moved within magisterial districts and migrants are defined based on cross-district migration. In 2001 census non-migrants are those who did not migrate between 1996 and 2001 or migrated within magisterial districts while migrants are people who moved across districts between 1996 and 2001.

Table 2.1 and 2.2 compare districts whose ethnic diversity is above and below the medium level of ethnic fragmentation in 1996 and 2001, respectively. The last column shows the p-value corresponding to the t-statistics on the difference between districts with high and low ethnic diversity. In both years more diverse places perform significantly better in all indicators of employment, including employment rate, proportion of self-employed people and employees over the whole working-age black population. Among those people who are employed, there is some difference among sectors and occupations. In 1996 census places with higher diversity have larger proportion of people in the manufacturing sector and less in the service sector and this pattern will change once we include our control variables in regressions. Districts with larger ethnic diversity also have less proportion of people in the unskilled occupations among all workers. The similar pattern holds in 2001 census.

The negative correlation between unemployment and ethnic diversity at district level is further confirmed in Figure 3 where we plot the proportion of unemployed (including economically inactive) people over the whole working-age black population against ethnic diversity in each district. The downward-sloping line between these two variables is observed in both 1996 and 2001.

4 Empirical Methodology and Specification

4.1 Baseline model specification and potential bias

We study the relationship between ethnic diversity among the black population living in "white areas" of South Africa and their labour market outcomes. In particular, we examine whether the within-black ethnic diversity affects blacks' employment opportunities. We start by examining the cross-sectional evidence and investigate the relationship separately for year 1996 and 2001. For both of the years we specify our linear probability model as follows:

$$Empl_{ikdp} = \alpha + \beta ELF_{dp} + \gamma \mathbf{X}_{ikdp} + \delta \mathbf{Z}_{dp} + v_{ikdp} \tag{1}$$

where $Empl_{ikdp}$ is a dummy variable for the labour market outcome for individual i of ethnicity k in district d in province p, taking value 1 if one is unemployed or economically inactive, and 0 if employed. We also report the results for wage-employment, self-employment (including self-employed, employer and working in the family business) and the substitution between wage-employment and self-employment within the subsample of the employed people. ELF_{dp} takes the value of the within-black index of ethnic diversity (i.e. fractionalisation index computed in Section 3[19]) in district d in province p. X_{ikdp} is a vector of individual-level characteristics (age, gender, educational attainment, marital status, whether one's father is alive which is a proxy for family financial and non-financial support). Z_{dp} is a set of both time-varying demographic and economic controls as well as time-invariant geographical characteristics at the district level, which will be explained in more detail below.

Unobservables which potentially affect employment rate are included in the term v_{ikdp}. v_{ikdp} can therefore be decomposed into the following items:

$$v_{ikdp} = \theta_p + \lambda_k + \epsilon_{ikdp} \tag{2}$$

ϵ_{ikdp} is the random error term. θ_p is province fixed effect which mainly controls for historical path dependence in job opportunities in each province, as well as province-level fiscal variables including social grant provision and policies on taxation and redistribution. There is also evidence that there is inequality between ethnic groups (Alesina et al., 2016) and that the gaps between different ethnic groups lie in their demographic structure, location, education and labour market outcomes (Gradin, 2014). Therefore we introduce λ_k, ethnic group fixed effects, which allows us to control for mechanical compositional effect and ensures we are comparing individuals from the same ethnic group across districts exposed to different levels of diversity.

Cross-sectional estimates suffer from omitted variable bias originating from ϵ_{ikdp}. For example, the existence of a local economic centre in the district could both create the demand for labour and encourage diversity, in that job opportunities attract individuals from other districts with different ethnic backgrounds. Or more energetic individuals with higher work spirits, who are intrinsically more likely to be employed than the average population, may sort to more diverse districts which have more active atmosphere. In these cases, our results will suffer from upward bias as both ethnic diversity and employment rate are positively correlated with the unobserved district and individual characteristics.

To address the concern that the results are driven by these confounding factors, we first include a rich set of district controls Z_{dp} to limit the information in unobserved items. To account for market size effects, we introduce the population density and urban/rural status of the district. As proxies for local economic development, we use the average night-time light density across 30-second grid areas within each district, and the share of blacks in the district population. For the industrial structure of the district which potentially leads to differences in labour intensity of firms, we control for the proportion of people employed in manufacturing and service sectors. Furthermore, to control for the direct spillover from homelands, we include the distance to homelands which were severely deprived by the Apartheid government. To control for the potential cost of ethnic diversity like conflicts, we add

[19] We use the results about polarization index as a robustness check.

the number of violence in each district in the corresponding years, as conflict has been proved to be associated with ethnic diversity (Amodio and Chiovelli, ming) and potentially job opportunities for the blacks (for example, there might be more closure of factories in more turbulent districts). Finally, to control for agricultural suitability and other geographic factors relevant for the local economic activities we use the terrain ruggedness, the existence of a river and a road crossing the district and the density of mineral resources.

The remaining district-level omitted variables are included in ϵ_{ikdp}. Our results will be biased if they are correlated with employment rates. All this will be dealt with using the instrumental variable discussed later on.

Unobserved information at the individual level in ϵ_{ikdp} might also bias the OLS result. We therefore cluster standard errors at the district level to allow for correlation of the error term across individuals in the same district. Furthermore, as a robustness check, we conduct regressions only on people who are born and remain in the districts (i.e. native people) as well as those who only migrated within districts. If the main results still hold among the native, the potential selection of people moving into places with different levels of diversity based on individual-level criteria will not largely drive the whole story. This will be discussed in more detail in the empirical results.

The relationship between ethnic diversity and labour market outcomes can also be investigated at the district level. Then model (1) would change accordingly. $Empl_{dp}$ would represent the proportion of individuals in unemployment, wage employment and self-employment in district d in province p and the ethnicity fixed effect would be removed. The set of individual characteristics X_{ikdp} should therefore be aggregated at the district level (e.g. average education in each district). The district-level regression becomes:

$$Empl_{dp} = \alpha + \beta ELF_{dp} + \delta \widetilde{\mathbf{Z}_{dp}} + \theta_p + \epsilon_{dp} \tag{3}$$

Here $\widetilde{Z_{dp}}$ include both the individual-level variables in X_{ikdp} aggregated at the district level, and the original district-level variables in Z_{dp}. Similarly, after controlling for province fixed effect θ_p, the remaining items in ϵ_{dp} are still sources of omitted variable bias which will be dealt with using the same instrumental variable approach.

As individual-level regressions contain more information (especially ethnic-specific characteristics captured by ethnicity fixed effects), we mainly report results based on individual-level regressions in our analysis whilst presenting the results of district-level regressions for robustness check.

4.2 Instrumental variable approach

Our instrument for ethnic diversity exploits the historical origins of the location of blacks' homelands. As explained in the institutional setting, the nine black ethnic groups moved long ago from the northern territories of the African continent and settled in different regions of today's South Africa, with one ethnic group occupying one settlements (i.e. defined as "homelands"). Assume the magnitude of migration from the homelands to outside districts decreases with the distance between them and distance is the only determinant in migration. When they moved out of these homelands to the outside districts (i.e. "white" districts which we are focusing on in this paper), the territories that are equally distant to multiple homelands are more likely to be inhabited by individuals with different ethnic origins, and therefore the ethnic diversity will be the highest. On the contrary, places close to only one homeland and far away from the rest become ethnically homogeneous as they have one group dominant in population size migrating from the closest homeland. Visually, this prediction is confirmed by the distribution of ethnic diversity in South Africa in 1996 (Figure 2). As shown before, places with relatively higher diversity are not necessarily places at the border or close to economic centres of the country, but are those in the middle and northeastern part of the territory surrounded by multiple homelands. Furthermore, districts on the far western part of the country present reasonably high level of ethnic diversity although being far away from all homelands. This is because these districts are still equally equidistant from all the homelands.

We therefore need an instrument that captures the equidistance of each district to all the original homelands. Our instrumental variable strategy proceeds in two stages. First, similar to Alesina et al. (2015), we estimate a parsimonious gravity model of migration based on 1985 census data (i.e. pre-1994 distribution of ethnic groups). We aim at predicting the level of within-black ethnic diversity in each white district d, solely as a function of a factor that is plausibly exogenous to labour market outcomes of the blacks: the distance of the district to the homelands. Second, we start from the predicted stocks to construct a diversity index. Specifically, we estimate:

$$N_{dk85} = \alpha + \beta_1 Dis_{dk} + \gamma_k + \epsilon_{dk85} \tag{4}$$

where N_{dk85} is the actual stock of individuals belonging to ethnic group k in (white) district d in 1985; Dis_{dk} is the bilateral Euclidian distance between the centroid of district d and the closest border of homeland for ethnic group k[20]; and γ_k is the homeland fixed effect. The determinants in our model are the ones traditionally employed in the related literature (Mayda, 2010; Beine et al., 2013; Ortega and Peri, 2014; Dumont et al., 2010). In particular, the physical distance between two districts (the homelands and the white areas) accounts for the migration costs, while the homeland fixed effects take into account common shocks in living conditions in the original settlement and the stock of population of each ethnic group in homelands, which can also influence migration decision. Following Santos Silva and Tenreyro (2006), we estimate the model by using the pseudo poisson maximum likelihood (PPML) estimator, which better suits the count data in the dependent variable.[21]

By imposing a universal β_1 to all ethnic groups, we assume that the per-unit migration cost is the

[20] The reason why we use the centroid of the districts instead of capital city is that capital cities are not well-defined at the magisterial district level. We use the border instead of the centroid of the homeland because the shape of the homeland is highly irregular and scattered. Furthermore, the distribution of population within homeland is highly uneven, making the centroid of homeland a less reliable measure in capturing the distance between the destination and the location of potential migrants from homeland.

[21] We do not control for the population size in the destination in the gravity model as it might be endogenously determined by the level of economic development in the destination which potentially affects the flow of migrants into the destination. Here our aim is not to get the most precise estimate of bilateral migration but to construct the counterfactual number of migrants in each district under a hypothetical setting where bilateral migration is only determined by distance between the original homeland and destination.

same for everyone, regardless of their ability and ethnicity. In addition, by ignoring any characteristics of the destination (e.g. population size, economic development and job opportunities) in the gravity model, we impose the condition that the benefit of migration is the same for everyone. Therefore by construction our predicted number of migrants from each homeland is only determined by the distance between homeland and destination.

In principle, the migration stocks could be predicted by 1996 and 2001 data. Nevertheless, we prefer to use the 1985 census data to rule out the selection of migration resulting from the movements of the black population after 1994 (this happened even as early as the repeal of the Pass Law in 1986). In fact, as previously documented (Section 2), the blacks had very limited freedom in choosing their own residential location and were strictly regulated in inter-district migration before 1986. After 1986 these constraints were loosened and the blacks had some freedom to decide where to resettle. Therefore, the distribution of ethnic groups in 1985 is less affected by the simultaneous change of labour market conditions and blacks' selection into "white areas" in the post-Apartheid era. Another reason why we use the 1985 distribution of the black population is that the equidistance to different homelands is a feature which stays stable over time. By sticking to 1985 data we can construct an instrumental variable whose value stays the same between 1996 and 2001 to make the IV regression results in these two years more comparable.[22]

Using the predicted stocks $\widehat{N_{dk}} = \widehat{\alpha} + \widehat{\beta}_1 Dis_{dk} + \widehat{\gamma}_k$, we calculate the predicted share of ethnic group k in the black population of district d and construct the instrument for the fractionalization index ELF:

$$\widehat{ELF} = 1 - \sum_{k=1}^{m} \widehat{s_k}^2 \text{ with } \widehat{s_k} = \frac{\widehat{N_{dk}}}{\sum_{k=1}^{m} \widehat{N_{dk}}} \tag{5}$$

The same instrumental variable approach with the same model specification at the first stage can be applied to district level regressions.

The remaining challenge is to find a proper measure of the original homelands for each ethnic groups. As there is no document about the exact location and boundary of these homelands, we use the territories of Bantustans during Apartheid as proxies for these original homelands. As is discussed in the institutional setting, with the ascent of the apartheid regime, the white-dominated government of South Africa designated specific territories as pseudo-national homelands (i.e. "native reserves", called "Bantustans" in the official documents) for the country's black African population. The Bantustans were organized on the basis of ethnic and linguistic groupings and were a major administrative device for the exclusion of blacks from the "white areas" of South African. The location of the Bantustans is based on the government's knowledge and documents about the historical location of homelands of each ethnic group. Ten Bantustans were created for these nine ethnic groups (there are two Bantustans for Xhosa people - Transkei and Ciskei and other groups each occupies one Bantustan).[23]

[22] We do not find much variation in fragmentation index between 1996 and 2001, which means ethnic diversity stays relatively stable over time.

[23] Therefore we treat Transkei and Ciskei as one homeland in the gravity model. When we calculate the distance between

To verify that the location and territory of Bantustans can be treated as proxies for the original homelands for the black people, we compare the distribution of these Bantustans and the "Murdock map". This map, drawn by an anthropologist George Murdock in 1953[24], provides the information on what the dominant ethnic group is in each geographical unit on the map of the whole African continent at the end of the 19^{th} century. As reflected in the Murdok's map (panel (a) in Figure 4) (each colour represents a certain group dominating the corresponding place in terms of population size), up to the end of the 19^{th} century, each of the nine groups have occupied some specific areas of the country. The Murdock map reveals the distribution of dominant ethnic group in each geographical unit rather than the exact location of original homelands. And the boundary of the geographical units on this map does not coincide with the border of magisterial districts in South Africa. Therefore, the Murdock map roughly implies the spatial distribution of each ethnic group in South Africa resulting from the distribution of original homelands combined with the migration of ethnic groups from these original settlements to other places.

Comparing Murdock's map in panel (a) and the distribution of Bantustans under the Apartheid system in panel (b) in Figure 4, we can find large overlaps of the Bantustans designated to each ethnic group with the region where the same group have dominated historically in Murdock's map. For example, places around the Bantustan designed for Tswana people (the dark green part in panel (b)) are also the places dominated by Tswana people (labeled with the same dark green colour) at the end of the 19^{th} century in Murdock's map in panel (a). Therefore, it is reasonable to use the distribution of Bantustans as proxies for the location of original ethnic homelands.

The map in Figure 5 presents the value of predicted diversity index together with the distribution of Bantustan across the country. The white places with slashes are either places which cannot be plausibly considered as "white" South Africa of our interest as they have more than 10% overlap with Bantustans, or places which cannot be matched with 1985 census data. The spatial pattern of predicted value of ethnic diversity in this figure is similar to the distribution of ethnic diversity in Figure 2 based on the real data. Again, places with the highest predicted ethnic diversity are those amid multiple homelands (mainly in the middle and northeastern part of the country). A more important feature is that the distance to the closest homeland (proxied by Bantustans) does not completely determine the level of predicted ethnic diversity. That is to say, places close to a specific Bantustan (and far from the other ones) may not be highly diverse. It is particularly the case for the districts around the Bantustans of Transkei, Ciskei, Kwazulu and Bophuthatswana. We will discuss this in more detail in the next section.

4.2.1 Test of validity of the instrumental variable

Identification requires the instrument to capture the ethnic diversity pattern observed in 1996 and 2001 and to be uncorrelated with any other determinants of the blacks' labour market outcomes. The

each district and the original homeland of Xhosa people, we measure the distance between each district and Transkei and Ciskei respectively and choose the smaller one.

[24] The map has been digitized by Nathan Nunn, starting from "Tribal Map of Africa" which is a fold out map from the book "Africa: Its peoples and Their Culture History" by George Murdock, 1959.

first condition is satisfied provided that: 1) The historical distribution of ethnic groups within the country varies with and is closely related to the distance of the destination region ("white" district) from multiple Bantustans, and 2) Apartheid did not overturn the historical pattern. As for the second condition, the non-randomness of blacks' homelands could cast doubts on its fulfillment. The proximity to the Bantustans might well be correlated with unobserved factors other than diversity, affecting the blacks' labour market outcomes.

However, the instrument exploits the distance to *multiple* ethnic homelands as a predictor for diversity. As mentioned above, the map in Figure 5 shows that districts with higher predicted diversity are the ones that are "equally" distant to multiple homelands, and not necessarily the ones that are the closest to a specific homeland. For example, although being contiguous to one of the Bantustans - Transkei (identified with the red color in Figure 5), districts in the South-East are among the most ethnically homogeneous areas because they are located at the periphery of other homelands. To further ensure that the instrument only captures the relative distance to multiple homelands and not the proximity to a single Bantustan, in the regression we control for the distance to the closest homeland. As all the homelands are located at the eastern part of the country, controlling for distance to the closest homeland can also deal with the problem that the instrumental variable might purely capture the west-east division of the country.

We argue that, conditional on proximity to a single homeland, the distance to multiple homelands is as good as random. The most direct narrative evidence is that according to the timeline in the institutional setting, the mass migration of the black largely occurred before the discovery of mines, rise of industrial sectors and modern development. This means the migration from homelands to "white" areas is not purely driven by the higher economic prosperity in the destination.

For a more rigorous test of the validity of our instrumental variable, we run regressions to show that the predicted ethnic diversity index is not correlated with potential confounders which determine ethnic diversity and employment simultaneously, conditional on all the control variables in our first stage regressions. Firstly, we test the correlation between the instrumental variable and potential job opportunities. According to agglomeration economics, economic centres, as clusters of economic activities, business and capital inflow, may act as the hub of job creation. Therefore, distance to economic centres may capture the potential job opportunities an individual is exposed to, based on the spillover of economic prosperity from the economic centres. There are five main economic centres in South Africa: Cape Town, Pretoria, Durban, Port Elisabeth and Johannesberg. In the validity test we calculate the distance from the centroid of each magisterial district to the closest economic centre and correlates it with predicted fragmentation index discussed above.

The second potential confounding factor is the economic activity of the white. On the one hand, as the Apartheid regime destroyed the self-employment opportunities, leadership and the training towards skilled occupations of the black in the "white" South Africa, the majority of the employers of wage-employed black people are the whites. Although our main regressions focus on the blacks, the population size and the employment status of the whites are also important in determining black people's employment rate, as they might be the providers of potential jobs to the black workers. On the other hand, the dominance and wealth of the white might potentially affect the migration decision of the early black migrants. Black people from different ethnic groups may move to a district where the white behave relatively better as there are more opportunities (or poorer as there is less stress/competition from the white) and thus the ethnic diversity of the black might be correlated with the behaviours of the white. We then calculate the employment rate of the white among their working-age population for each magisterial district in our sample and see if it relates to ethnic diversity of the black.

Thirdly, path dependence also matters in determining contemporary employment opportunities. As the distribution of black settlements is not completely random, the equidistance to multiple original settlements might reveal some socio-economic characteristics besides the distance itself (i.e. customs, early conflict or the distribution of ancient civilisations) which have long-term impact on contemporary development. This persistence of particular socio-economic features is usually a concern in literature which constructs instrumental variables with geographical variables. However, in our special setting, the Apartheid regime before our sample period compressed the opportunities of education, job opportunities and residential choice nationwide among the black and potentially destroyed part of such historical path dependence. If we can show that the path dependence which potentially correlates with equidistance to homelands was largely destroyed by the Apartheid regime due to the shift in residential patterns and the re-allocation of economic activities both for the black and the white, we will be safer to claim that the historical persistence is not likely to affect contemporary employment opportunities directly. As there is no reliable data to reveal the employment pattern of the black during apartheid, we use the employment pattern of the white in 1980 as a proxy for the remaining path-dependence in employment close to the end of the apartheid and see if it correlates with our instrumental variable measured with 1996 and 2001 data. For the employment status of the white in 1980, we do not consider self-employment as the definition of self-employment is not quite clear under Apartheid regime and therefore has large measurement errors.[25] We also consider the population size of the white in 1980.

The fourth potential confounding factor is the magnitude of contemporary migration. Although historical migration was not mainly driven by economic prospects, it might still be the case that contemporary diversity results from contemporary migrants which are driven by economic opportunities. Therefore, we need to show that our predicted diversity does not relate to the magnitude of contemporary migration which refers to cross-district migration ever happening in one's life in 1996 census and cross-district migration between 1996 and 2001 in 2001 census.

Table 3 shows the results on the validity of the instrumental variable based on 1996 and 2001 census data. We regress a set of variables that potentially affect employment rate on predicted fractionalisation index conditional on all the control variables in the main regressions discussed above. Panel A, B, C and D present the tests on the relationship between predicted ethnic diversity and job opportunities, economic activities of the white, path dependence and contemporary migration, respectively. We obtain the coefficients of the tests by regressing the corresponding dependent variables (as reported in the table) on predicted ethnic diversity conditional on all the control variables in the main

[25] There are four census during Apartheid: 1960, 1970, 1980 and 1985 census. We only consider 1980 census as the data quality is higher than that in 1960 and 1970 census. Publicly available 1985 census data has no information on employment status.

regression. These dependent variables include: distance to the closest economic centre, proportion of white people who are self-employed over the white population in 1996 and 2001, proportion of white people who are employees over the white population in 1996 and 2001, proportion of white people over the whole population in 1996, 2001 and 1980, proportion of white people who are employees over the white population in 1980 and the number of contemporary migrants in each district. We do not find systematic relationships between these potential confounders and our instrumental variable, which means that the predicted ethnic diversity can be considered as a valid instrumental variable.

4.2.2 Other potential threats to the instrumental variable

This section discusses some remaining potential threats to the instrumental variable which are not likely to be measured with available data.

Firstly, one may argue that the original distribution of ethnic homelands is not completely random. The fact that one place is close to multiple homelands at the same time might mean that these homelands are themselves close to each other. Similarly, one possible pre-requisite for a place to be close to only one homeland is that those homelands might be scattered and relatively far away from each other. If the whole region is equipped with better endowments (geography, climate or soil quality) than the others at the time of the Bantu migration from central Africa, this place could attract more than one ethnic groups to establish their homelands, whilst regions with only one ethnic homeland or regions where the distribution of homelands is more scattered might be less attractive in resources and endowments. Therefore, our instrumental variable - the predicted diversity index might just capture the distribution of homelands and the original endowments of the whole surrounding region.

This is not likely to be the case for the following reasons. The first reason is that our instrumental variable captures the equidistance to different homelands conditional on the distance to the closest homeland. By construction places far away from all homelands can still have reasonably high predicted diversity, as long as it is of equidistance to all these homelands. These places are less likely to be affected by the original endowments and resources of ethnic homelands. The second reason is that we have already controlled for geographical endowments (ruggedness, soil quality and river) in each district which are potentially correlated with their initial development by affecting their agricultural production. The third reason is that if our instrumental variable mainly captures the initial economic development and the endowments or resources of the region rather than ethnic diversity, the predicted diversity index should be correlated with the labour market outcomes among both black and white population. However, as is shown in table 3, our instrumental variable is not systematically correlated with the employment rate of white workers. Therefore, it is unlikely that the initial endowments in the regions surrounding ethnic homelands challenge the exclusion condition of the instrumental variable.

Secondly, there is a possibility that districts close to multiple homelands might be the trading centres for people from those homelands whilst trade flows in districts close to only one homeland are less. This might also lead to the difference between these two types of places in the initial economic prosperity and the establishment of cities resulting from trade. Here we show this is unlikely to severely violate the validity of our instrumental variable. Our instrumental variable by construction allows for the case that a place far away from all homelands can be reasonably diverse if it is equidistant to different homelands. And this place is less affected by the initial trade flows among homelands. Furthermore, places with more initial trade flows might become contemporaneous economic centres due to the path dependence in city development and the accumulation of capital and labour. In our validity test we do not find a systematic pattern of the distance to the closest economic centre and predicted diversity index.

Thirdly, one may worry that certain events which attract diverse migrants might happen coincidentally in places close to multiple homelands. For example, the homeland for Tswana group (i.e. the Bantustan of Bophuthatswana) and places in Mpumalanga and Limpopo Province (in the northeastern part of the country) are rich in mineral resources. If our instrumental variable mainly captures the distribution of mineral resources, and if the discovery of mines in a district motivates people of diverse backgrounds to migrate into the district and at the same time boosts economic development, what can be reflected in the predicted ethnic diversity is mainly the effect of mineral resources. In our analysis we have controlled for the density of the mines in each district. More importantly, narrative evidence reveals that the mass migration from central Africa (which can be dated back to the 11^{th} and 12^{th} century) and the emigration from homelands to "white" South Africa happened well before the discovery of mineral resources (mainly starting from the 19^{th} century). Therefore, the discovery of mines and the related events are not likely to violate the validity of our instrumental variable.

4.2.3 First stage results

Table A0 in the Appendix reports the estimated parameters of the gravity model. It suggests that the distance between a white district d and an ethnic group's homeland is strongly negatively correlated with the size of the same ethnic group's population living in district d. Table 4 presents the first-stage regression of the instrument at the individual level. We are interested in both working-age population (age 15-64) and a subsample which have already finished full education (age 25-64). All regressions include province fixed effects and all control variables. Columns 1 and 2 (3 and 4) report the first-stage regression results based on 1996 (2001) census data. In both years the predicted fragmentation index $\widehat{ELF}$ is positively associated with the observed index ELF. The F-statistics is very high in all regressions (i.e. much larger than 10), indicating that the instrument is a very strong predictor of ethnic diversity. Comparing column 1 and 2 reveals that the F-statistics remain stable in both the full sample and the subsample. Comparison between columns 3 and 4 confirms the same pattern in year 2001.

District-level regressions in Appendix Table A1 reveal the same pattern. Predicted ethnic diversity is positively and strongly correlated with the ethnic diversity index in real data. F-statistics of the instrument are still large in all regressions in both year 1996 and 2001. All results consistently show that our predicted ethnic diversity index is strong enough as an instrumental variable.

4.3 Supplementary approach: district-level fixed effect

The fact that we have two-year cross-sectional census data and that the territory of magisterial districts stay stable between 1996 and 2001 motivate us to find a way to construct panel data at district level as a supplementary approach to the instrumental variable specification. From the district-level model specification (3), we realise that the main source of bias comes from the unobserved district-level confounders. Therefore an alternative way to instrumental variable approach to deal with this bias is to control for it directly by including district fixed effect based on a panel of districts. Therefore we construct a balanced panel by matching the magisterial districts between 1996 and 2001[26] and conduct the model (3) by adding magisterial district fixed effect directly. Any time-invariant variables in Z_{dp} and θ_p are dropped automatically. Instead we add time fixed effect u_t in the model.[27]

$$Empl_{dt} = \alpha + \beta ELF_{dt} + \delta \widetilde{\mathbf{Z}_{dt}} + \sigma_d + u_t + \epsilon_{dt} \tag{6}$$

We report the results of this district-level fixed effect model right after the main analysis.

5 Empirical Results

5.1 Ethnic diversity and labour market outcomes

5.1.1 Ethnic diversity on employment

Table 5 summarizes the main results on the effect of ethnic diversity (measured by fractionalisation index) on unemployment rate. The dependent variable is a dummy which equals 1 if one is unemployed or out of labour force and 0 otherwise (including people who are self-employed and employees). In 1996 census data which distinguishes people who are unemployed and out of labour force, we create dummies for unemployment and labour force participation and look at how they respond to ethnic diversity separately. Columns 1-6 report the results in year 1996 while columns 7-8 are for year 2001 when unemployed workers and people out of labour force are combined into one category in the original census data. Furthermore, panel A in Table 5 reports the results based on the cross-sectional OLS regressions at the individual level. Panel B in Table 5 provides the corresponding estimates based on the instrumental variable regressions. We provide results both for the full sample (which gives a

[26] Among 205 magisterial districts in 1996 and 210 districts in 2001, 205 of them can be matched, given that we exclude districts with less than 1% of black people over the whole population.

[27] A potential further specification is to combine the above two approaches and rely on fixed effect-IV approach. The rationale to do this is that some district-level unobservables might change over time which cannot be captured by time-invariant σ_d. In this case, we have the first difference specification:

$$\Delta Empl_{dt} = \alpha + \beta \Delta ELF_{dt} + \delta \Delta \widetilde{\mathbf{Z}_{dt}} + \Delta f_{dt} + \epsilon_{dt}$$

Ideally we can find an instrumental variable for f_{dt}. A similar case to this specification can be found in Dustmann et al. (2017). However, this first-difference specification at district level with instrumental variable is not appropriate here because there is little variation in both the real-world ethnic diversity and the predicted ethnic diversity (i.e. the equidistance to different homelands does not change over time) over time, which is not sufficient for reliable statistical inference.

lower bound of the effect of diversity on employment) and a subgroup of people aged from 25 to 64 who have finished their education (which gives an upper bound of the effect of diversity on employment). All regressions control for the individual and district level characteristics including ethnicity fixed effects discussed above.

In most of the OLS and IV regressions in Table 5 the coefficients of ethnic diversity on unemployment (or labour force participation or these two outcomes altogether) are significantly negative, indicating that within-black diversity increases the rate of employment and labour force participation. Comparing panel A and panel B, the negative and significant coefficients of ethnic diversity remain in IV regressions in many columns. In panel B, comparing columns 2, 4 and 6 reveals that ethnic diversity increases employment mainly by decreasing the number of people who are actively looking for jobs but still unemployed, rather than bringing people into the labour force. Table 5 also shows that the coefficients of ethnic diversity are larger and more significant for the subgroup of people who have finished education than those for the full sample, which confirms that the full-sample analysis gives a lower bound of the effect of ethnic diversity.

We focus on the full sample (i.e. lower bound) to calculate the magnitude of the effects of ethnic diversity on employment based on the results in columns 5 and 7. In panel A in column 5, one standard deviation increase in ethnic diversity index in 1996 is associated with 2.15 percentage point decrease in unemployment (including inactivity), which is 3.51% of the average unemployment (including inactivity) rate.[28] Similarly, in panel A in column 7, one standard deviation increase in ethnic diversity index in 2001 is associated with 3.88 percentage point decrease in unemployment (including inactivity), which is 5.97% of the average unemployment (including inactivity) rate.[29] Correspondingly, in IV regressions, one standard deviation increase in ethnic diversity index in 1996 (2001) decreases unemployment (including inactivity) by 2.61 (4.40) percentage point, which is 4.24% (6.91%) of the average unemployment (including inactivity) in 1996 (2001).

Comparing the magnitude of estimates in OLS and IV regressions in both years shows that the magnitude of the effects of ethnic diversity on unemployment rate increases largely between 1996 and 2001 (from 3.51% of the average unemployment rate to 5.97% in OLS and from 4.24% to 6.91% in IV) and IV estimates are slightly larger than OLS estimates. This can be explained by the fact that IV regressions capture LATE for workers at the margin of being affected by ethnic diversity. They might be the most responsive to ethnic diversity in considering their employment status.

Appendix Table A3 further breaks down employment status into two categories: self-employment and wage-employee. All the independent variables remain the same as those in Table 5. In columns 1 and 3 in Appendix Table A3, the dependent variable is a dummy which equals 1 if one is self-employed and 0 otherwise (including unemployed, inactive and wage employee). The dependent

[28] It can be calculated that the standard deviation of ethnic diversity in 1996 is 0.2659. The coefficient of diversity index in panel A in column 6 is -0.081. Therefore one standard deviation in diversity index decreases unemployment by 0.081 * 0.2659 = 0.0215. From Table 1.1 we know that the average unemployment (including inactivity) rate among the black in "white" districts is 0.613. Therefore this point decrease is 0.0215/0.613= 3.51% of the average unemployment rate.

[29] It can be calculated that the standard deviation of ethnic diversity in 2001 is 0.2586. Therefore in 2001 one standard deviation in diversity index decreases unemployment by 0.146 * 0.2586 = 0.038. From Table 1.2 we know that the average unemployment (including inactivity) rate among the black in "white" districts is 0.636. Therefore this point decrease is 0.038/0.636= 5.97% of the average unemployment rate.

variable in columns 2 and 4 is a similar one which equals 1 if one is an employee and 0 otherwise. Again, panel A (B) reports the results for OLS (IV) regressions.

The results show that in the post-apartheid South African context, within-black ethnic diversity has a positive effect on the labour market outcomes of the blacks, mainly in wage-employment as is shown in columns 2 and 4. Specifically, one standard deviation increase in the fractionalisation index is associated with a 2.31 (3.54) percentage point increase in the wage-employment rate of the working-age black individuals in 1996 (2001), according to the OLS results. This corresponds approximately to a 6.52% (10.39%) increase of the average wage-employment rate among the population of reference in 1996 (2001). In IV regressions, one standard deviation increase in the fractionalisation index increases wage-employment rate by 2.74 (4.60) percentage points in 1996 (2001), which is around 7.71% (13.50%) increase of the average wage-employment rate in 1996 (2001).

Similar to the patterns in Table 5, the effect of ethnic diversity on wage-employment increases from year 1996 to 2001. IV estimators have slightly larger magnitude than OLS estimators for possibly the same reason. We do not find anything significant about self-employment rate. There are two potential reasons to explain why the effect of ethnic diversity on self-employment is not obvious. Firstly, as is described in summary statistics, self-employment rate is only 2-3% among the black South Africans, which means the variation of self-employment rate across districts might be too limited for reasonable statistical inference. In addition, measurement errors in self-employment might be large. If these measurement errors are not random, it will also bias our results. Secondly, it is reasonable that self-employment does not respond as much as wage-employment. Current literature relates self-employment, especially small-scale entrepreneurs, to trust, tolerance or cohesive and homogenous networks (Barr, 1998) and argues that self-employment increases with trust. If ethnic diversity can lower the level of trust, according to some existing papers (Alesina and La Ferrara, 2000, 2002), the overall effect of ethnic diversity on self-employment can be ambiguous.

Table 6 further presents how ethnic diversity affects workers' choice between self-employment and being an employee. As self-employment rate is between 2% - 3% of the whole working-age black population, we drop self-employed people from the whole sample and investigate if ethnic diversity increases the probability of being an employee against unemployed in columns 1 and 5. The magnitude and significance of the coefficients on ethnic diversity index are very similar to those in the corresponding columns (columns 2 and 4) in Appendix Table A3. This shows that most of the effects of ethnic diversity on employment takes place in wage-employed jobs.

Columns 3, 4, 7 and 8 only include employed people and look at the allocation of these workers between self- and wage- employment. The dependent variable equals 1 if one is self-employed and 0 if being an employee. This is to investigate the effect of ethnic diversity on the potential substitution between self- and wage-employment among employed black population. We replicate the results of the main analyses by restricting the sample to people who are either wage-employed or self-employed (i.e. excluding the unemployed and the inactive). Although the self-employment rate might be too low for enough variations to generate significant statistical inference, we find that the coefficients of ethnic diversity are consistently negative in OLS and IV regressions in both years. That is to say, ethnic diversity helps unemployed individuals get into employment; a large fraction of those newly employed people opt for working for others as an employee.

Focusing on the full sample (excluding self-employed people) in column 1 and 5, we find that one standard deviation increase in ethnic diversity index increases the rate of wage employment among the working-age black by 2.29 (3.76) percentage point in 1996 (2001) in OLS regressions, which is around 6.23% (10.67%) increase of the average wage-employment rate in 1996 (2001). In IV regressions, one standard deviation increase in ethnic diversity index increases the rate of wage employment among the working-age black by 2.98 (4.56) percentage point in 1996 (2001), which is around 8.12% (13.04%) increase of the average wage-employment rate in 1996 (2001).[30]

The corresponding district-level regressions based on the model specification 3 are reported in the Appendix Table A4. In these district level regressions, the dependent variables are the proportion of working-age black people who are unemployed or inactive; who are wage-employed; who are self-employed and the proportion of people who are self-employed relative to employees (columns 1-4 and columns 5-8, for year 1996 and 2001 respectively), given the corresponding individual features aggregated at district level and district level controls. OLS (IV) estimators are shown in panel A (B).

The OLS and IV estimates reported in Table A4 confirm the positive impact of diversity on the employment of the blacks. And this positive impact mainly takes place in wage-employment. The effect on employment (and wage-employment) in OLS regressions is slightly smaller than the ones estimated with the individual-level regressions, while the magnitude of the effect in IV regressions is slightly larger than that in individual-level regressions.[31]

5.1.2 Ethnic diversity on wage, income and working hours

In this section we investigate labour market outcomes other than wage to get a more thorough picture of how labour market responds to ethnic diversity in post-Apartheid South Africa. We replicate the above individual-level regressions (both OLS and IV) by replacing the dependent variables with other labour market outcomes, including working hours, hourly wage and monthly earnings. As information on working hours is only available in 2001 census data, we only conduct these analyses based on 2001 data. For data on working hours, if values of self-reported weekly working hour are larger than 80, we treat them as outliers and exclude them from regressions. In addition, we trim the income data by excluding values above 5 standard deviation of the mean income. Hourly wage is constructed by dividing monthly earnings by monthly working hours (i.e. four times weekly working hours).

Data on monthly income in 2001 census includes both labour market earnings and income from other sources such as dividend, rent or social grant. We first report the results based on these rough measures of monthly earnings and replicate the regressions with more precise data on labour market earnings and working hours.

[30] The wage employment rate is 36.69% for 1996 and 34.97% for 2001 after excluding self-employed people.

[31] Columns 4 and 8 report the results on the effect of ethnic diversity on the rate of self-employment relative to wage-employment aggregated at district level by only including black people who are employed. Results in other columns are based on the whole working-age black population.

Panel A in Table 7 reports the OLS and IV regression results on these labour market measures based on 2001 census data. Dependent variables include: log monthly income, log hourly income and weekly working hours. As self-employed workers and employees have very different determinants of working hours and earnings, and that ethnic diversity mainly increases wage-employment rate, we only focus on employees in all regressions.[32] Columns 3 and 6 indicate that ethnic diversity does not affect weekly working hours among the employees. Therefore the increase in employment in response to ethnic diversity comes from the extensive margin by increasing employability of unemployed and inactive people, rather than the intensive margin (measured by weekly working hours). And this extension of the extensive margin of labours is not achieved at the sacrifice of decreased intensive margin.

Columns 1, 2, 4 and 5 show some evidence on the increase in both monthly and hourly income among the black employees in response to higher ethnic diversity. As is stated above, information on income in census data incorporates all potential income sources. Therefore we need another dataset which asks information on labour market earnings in particular. We turn to October Household Survey 1996 to replicate all the results in Panel A.[33] We do not choose year 2001 because starting from year 1998 there is no information on the magisterial districts each individual lives in. The results are in Panel B in Table 7. Columns 3 and 6 confirm that weekly working hours are not responsive to ethnic diversity. In columns 1, 2, 4 and 5 the effects of ethnic diversity on measures of labour market earnings are not significant, possibly because the increase in employment can come from both the supply and demand side of the labour market, or because the measures of nominal earnings are not adjusted for price levels (as there is no price or living cost data at the magisterial district level).

5.2 Supplementary approach: district-level fixed effects

As a supplementary approach to the instrumental variable approach, we provide estimation results on district-level fixed effects models based on the model specification (6) in Table 8. We construct a balanced panel between 1996 and 2001 (205 magisterial districts each). The measures of labour market outcomes (i.e. dependent variables) are: proportion of people who are unemployed or inactive among the whole working-age black population; proportion of employed workers among the whole working-age black population (excluding self-employed people); ratio of the number of self-employed workers versus employees and log monthly income among employees.

Similar to the main IV regression results, higher ethnic diversity is associated with higher employment, mainly in wage-employment but there is no significant correlation between ethnic diversity and monthly income. In particular, in district fixed effect regressions we find some evidence that more diverse districts are associated with higher ratio of wage-employment in relation to self-employment.

The magnitude of coefficients in Table 8 are larger than those in Table 5 and Table 6, which can

[32] There are more observations in columns 3 and 6 than others because there are missing values in income and we trim the income values above 5 standard deviation from the mean.

[33] It is an annual survey staring from 1993 (which was renamed as Labour Force Survey conducted twice a year from 2000 and became a quarterly survey from 2008). In 1996 survey 72890 individuals are covered, among which 16082 have information on work status.

be explained by two possible reasons. Firstly, district-level regressions do not include ethnicity fixed effect which is used to capture ethnicity-specific unobservables which affect the labour market outcomes of each ethnicity such as the attitudes towards work and leisure and ethnic-specific skills. It is however not appropriate to include this fixed effect in the district-level regressions due to the potential multicollinearity problem, as the proportion of each ethnic group in a district is already a component of the ethnic diversity index (i.e. an item in Herfindahl Index).

Secondly, the relatively larger coefficients of panel regressions might reflect some time-varying district-level unobservables. For example, people are more likely to move to ethnically diverse districts as time goes by as a result of increased benefits in the destination (i.e. the economy of the districts with higher ethnic diversity might grow more rapidly than that in more homogeneous districts). In individual-level IV regressions, our instrumental variable is not likely to be correlated with the economic development in the destinations by construction (as the distance between homelands and destination is the only determinant in migration). Therefore the variation of these unobservables over times does not affect our estimates in IV regressions. However, as panel regressions with district-level fixed effects may lead to upward bias of the key estimator as they do not take into account these time-varying unobservables.

5.3 Heterogeneous effects of ethnic diversity on employment

Table 9 split the whole sample into several sub-samples to investigate the heterogeneity in the impact of ethnic diversity on labour market outcomes with individual-level regressions. In particular, we replicate the regressions in the main specification by carrying out the same analysis on these sub-samples. By excluding workers who are self-employed, we use a dummy dependent variable which takes the value 1 if one is an employee and 0 if one is unemployed or inactive.[34] Panel A in Table 9 replicate the same regressions in columns 1 and 5 in Table 6 by splitting the working-age black population into ethnic groups with different population size. Panel B and C look at the allocation of employees among different sectors and occupations in response to ethnic diversity by regressing the probability of working in particular sectors or occupations on ethnic diversity index only among employees.

Panel A split the sample by group size. As is shown in Table 1.1 and 1.2, we have three "large" groups whose population share is above 20%, two "medium" groups whose share is between 10% and 20% and the remaining "small" groups making up less than 5% of the whole black population. We look at these three groups separately and discuss how they are affected by ethnic diversity. The results reveal that only the group with "large" size are positively affected by diversity. None of the columns show that "small" groups response to ethnic diversity of the districts they live while evidence on the "medium" group is mixed. It is not very likely that the results are purely driven by the lack of power of statistical inference due to smaller sample size. In all the regressions for "medium" and "small" groups, the t-statistics is far from being large enough to generate significant inference. Furthermore, in some regressions the coefficients of ethnic diversity are negative, especially for those in the "small"

[34] We also conduct the analysis with a dummy on whether one is unemployed (including inactive people) or not. The results are quite similar.

group in 1996.

In panel B and C, we look at the allocation of industries and occupations among employed workers to show that the improvement in employment rate is not accompanied by the increase in less skilled jobs or the expansion of primary sectors. Otherwise this will lead to a less favourable industrial and occupational structure.

Both 1996 and 2001 census data provides information on the industrial sectors they work, which we classify into agriculture, manufacturing and service sectors. Panel B presents the results on this allocation. There is no evidence to show that ethnic diversity affects the industrial structure of the districts in IV regressions. This further confirms the idea that the employment opportunities generated from ethnic diversity are not purely driven by the expansion of manufacturing sector due to the revolutionary events like the discovery of mines, nor from the expansion of the primary sector.

We study the allocation of employees further by looking into occupations to show if ethnic diversity leads to a less skilled occupational structure. In both 1996 and 2001 census for each worker there is information on the occupation classified into a detailed 3-digit code. We aggregate this 3-digit coding system into types of occupations based on their skill levels: manager, professional, clerk, service worker, craft worker, skilled worker in agricultural sector, machine operator and unskilled worker. The dependent variables in Panel C are dummies on whether one works in one of these occupations. According to the regression results, ethnic diversity decreases people's chance of becoming a machine operator and increases their probability of being a manager, professional employee and clerk. One common feature is that occupations such as manager, professional and clerk require more language and social skills while the demand for social skills is the least among machine operators. This is closely linked to our mechanism through which ethnic diversity influences labour market outcomes, which will be discussed in the modelling part.

5.4 Robustness check

We conduct a series of robustness checks in this section to consolidate the result that ethnic diversity increases employment rate among working-age black population.

Firstly, we use population density and the proportion of black people over the whole population as our control variables. They two altogether capture the information on the total population size of the black in the destination districts. As our census data is a 10% subsample of the original census data and the size of population is calculated with the post-stratification weights in the 10 % sample of census data, these two variables may suffer from measurement errors. Our first robustness check is to replace these two control variables with the total distance from the destination district to all black homelands. The idea is that if all black people in the "white" districts come from historical homelands and the migration from those homelands to the destination decreases with distance, total distance to all homelands will be a proxy for the pool of black population in the destination. In Table 10 we replicate the main analysis in Table 5 and Table 6 by replacing population density and the proportion of black people with total distance to all homelands. Panel A and Panel B show that ethnic diversity still decreases unemployment rate and in particular increases wage employment rate in both 1996 and 2001. The magnitude of the coefficients of ethnic diversity index is larger than that in our main analysis but the magnitude in IV regressions is quite similar. Panel C, D and E report that this effect still holds for (and only for) ethnic groups with larger population size and the magnitude of coefficients is larger than that in main analysis. All this suggests that our results are robust to different measures for our control variables.

Secondly, we provide some further evidence on the argument that our result is not purely driven by the sorting of migrants. That is, we show that the positive correlation between ethnic diversity and labour market outcomes does not purely come from the migrants with higher abilities moving to more diverse places and therefore are performing better in job searching. We divide the whole working-age black population into three sub-samples with different levels of sorting: people who were born and stay in the district or people migrating within districts (i.e. "native" people); people moving across districts (i.e. "migrants"); immigrants moving from other countries ("immigrants").[35] In Table 11 we run the same IV regressions[36] as those in the main analysis separately for these three groups in 1996 and 2001. The dependent variables include a dummy on whether one is unemployed and a dummy on whether one is an employee (excluding self-employed workers).

Columns 1 and 4 show that in both years ethnic diversity positively affects the labour market outcomes for native people who are the least likely to sort to places with higher ethnic diversity, as they were born in these districts and remained there, or moved within districts. The positive effect of ethnic diversity on employment also exists among immigrants in columns 3 and 6, the mostly selected sample based on ability and preference (although the number of immigrants in South Africa belonging to one of the nine ethnic groups is very small compared with the whole black population). Interestingly, there is no effect of ethnic diversity on employment among migrants across districts. As we discussed in the validity of the instrumental variable, there are two potential mechanisms of selection among migrants. Either the selection occurs in the original place, meaning people with higher ability choose to move out; or the selection takes place at the destination, meaning people sort to places with higher economic prosperity or job opportunities or more socially active environment when they decide where to move. The result about cross-district migrants here might suggest that the first selection mechanism is more important - migrants are of higher ability and therefore behave better wherever they end up, which indicates that the relationship between ethnic diversity and employment is not solely driven by the selection of destinations.

Another potential threat to the interpretation of our results as illustrating a positive impact of ethnic diversity on employment is the emigration of the white after the end of Apartheid. It has been observed that there has been a large emigration of the white out of South Africa after 1994 and that white people moved out of the country for the fear of the worsening economic conditions, weaker government capacity, or the revenge from the black after the nightmare of Apartheid. A place has higher within-black diversity might just indicate that the power of the white is weaker in these places (so that the black community can grow and attract people with a diverse background). If this is the

[35] Note that "migrants" and "immigrants" in 2001 census data are those who move across districts or countries between 1996 and 2001, whereas in 1996 census they are the people whose last migration was across districts or countries.

[36] OLS regressions have very similar results. We only show the results about IV regressions here.

case, there would be more white people emigrating from South Africa in a district with larger ethnic diversity index. The mass emigration of the white may lead to many job vacancies to be filled by black workers, consequentially improves the job opportunities of the black. If this story is true, the correlation between ethnic diversity index and employment rate in a district cannot reflect the impact of ethnic diversity as ethnic diversity index here is just a proxy for the power of the white in the district.

We therefore regress the number of the white in 1996 and 2001 respectively and the difference in the number of white residence between 1985 and 1996 (or 1985 and 2001) on ethnic diversity index for each district, using the same set of control variables. We find in Table 12 that the ethnic diversity index is associated with neither the absolute number of the white population nor the difference in the white population before and after the end of Apartheid (which captures the emigration of the white). This confirms that ethnic diversity is positively related to employment not simply because these places have more job vacancies left by the white people who emigrated from the country.

Thirdly, related literature suggest using Conley's standard errors in regressions to account for the spatial correlation in error terms (Michalopoulos and Papaioannou, 2013). Following the analysis in the Appendix Table A4, we use Conley's standard errors to replicate the analysis. It is required to set up a cutoff distance above which there is no spatial correlation. Current cross-country analysis in Africa uses 2000km as the cutoff value (Michalopoulos and Papaioannou, 2013). In our paper, we reduce the cutoff value to 1000km for within-country regressions. Spatially correlated error terms are both implemented in OLS and GMM (using the same IV as that is in the main analysis) regressions. Appendix Table A5 reports all these results. Ethnic diversity still has a positive impact on employment opportunities in all columns in both OLS and IV regressions. Comparing GMM estimates with the previous main IV analysis, we find that the magnitude of the effect of ethnic diversity is in general larger in regressions with spatially correlated errors.

Fourthly, we replace some or include more control variables to see if the results are still robust, as is shown in the Appendix Table A6. In the first column, we use population size to replace density of population to capture the magnitude of market size without scaling for the area of territory. In the validity test of IV, our instrumental variable is positively correlated with the proportion of white people who are self-employed in 1996. Therefore we include proportion of self-employed people over the white population in 1996 (or 2001, depending on the census year) as an additional control in column 2. In column 3 we add the proportion of contemporary migrants to control for the potential sorting of more capable migrants to more diverse districts. The results, especially those in 2001 census are robust to these changes in control variables.

Last but not least, we use non-linear econometric methods to estimate the main regressions. Given that our outcomes are measured by binary variables, we replicate our results by estimating a logit model, a probit model and a probit model with the instrumental variable in both 1996 and 2001. Results are summarized in the Appendix Table A7.1. Marginal effects at average ethnic diversity index are reported in all columns. The positive effect of ethnic diversity on both employment as a whole and wage-employment in particular (excluding self-employed people in columns 4 - 6) is robust to these specifications. The magnitude of the marginal effects is very similar to those in Table 5 and Table 6 in baseline regressions. For example in logit regressions in 2001, the coefficient of ethnic diversity on wage employment is 0.145, which is roughly the same as the corresponding coefficient in OLS regressions in Table 6 (0.144 in column 5). In IV regressions the magnitude in non-linear models is smaller than that in linear IV models but the significance remains the same. For example, in probit regressions with our instrumental variable based on 2001 census data, the coefficient of ethnic diversity on employment is 0.140 while in the corresponding IV regression it is 0.176 (column 5 in Table 6).

In Appendix Table A7.2 we also implement multinomial regressions to take into account the decisions of both self-employment and wage-employment. We construct a variable of employment status which equals 0 if one is unemployed, 1 if one is self-employed and 2 if one is wage employed. Columns in Appendix Table A7.2 captures the marginal effects of ethnic diversity on the decision of self-employment and wage employment, relative to the outcome of unemployment. Columns 1 and 2 report the cross-sectional multinomial logit regression results while columns 3 and 4 report the multinomial probit regressions with our instrumental variable. The results indicate that wage employment rate responds positively to ethnic diversity while there is no robust evidence on self-employment rate. The magnitude of the effect of ethnic diversity in multinomial logit regressions is similar to that in our main analysis while the magnitude in IV regressions is much larger in multinomial models than in linear probability models. For example, the coefficient is 0.135 in 2001 in multinomial logit regressions and 0.125 in OLS regressions with linear probability models. The corresponding coefficient is 0.981 in multinomial probit regressions with instrumental variables and 0.182 in IV regressions with linear probability models.

5.5 Decomposing ethnic diversity index

We can decompose ethnic diversity index into two components: the number of groups and the dispersion of group size. Suppose there are m ethnic groups in a district. Group k has a population share s_k over the whole population. Our ethnic diversity index is thus decomposed in the following way:

$$ELF = 1 - \sum_{i=1}^{m} s_k^2 = 1 - \sum_{k=1}^{m}((s_k - \frac{1}{m}) + \frac{1}{m})^2$$

It is obvious to get the following decomposition:

$$ELF = 1 - \underbrace{\frac{1}{m}}_{1/\text{No. of groups}} - \underbrace{\sum_{k=1}^{m}(s_k - \frac{1}{m})^2}_{\text{Dispersion of group size}} \tag{7}$$

Leaving aside the constant term (i.e. 1), the first item in the ethnic diversity index is an inverse of number of ethnic groups and the second item captures the dispersion of group size. We therefore replicate the main analysis by using these two items to replace ethnic diversity index for both the whole sample and the subsamples with different population size and see how they are correlated with wage employment rate.

The corresponding results are in columns 1-4 in Table 13.1. We find that in 1996 and 2001, both of the two terms are negatively correlated with wage employment rate, which means that employment rate increases with the number of ethnic groups in a district and decreases if the distribution of group size becomes more uneven (column 1). Again, a detailed investigation of subgroups shows that these two components only affect ethnic groups with relatively larger size (column 2 - 4).

As both components are significantly correlated with wage employment rate, we need to disentangle these two factors for a further investigation of the mechanism through which ethnic diversity improves labour market opportunities. Here our instrumental variable not only provides an exogenous variation on ethnic diversity but also helps disentangle these two components by fixing one and exploring the variation in the other. By construction the instrumental variable is calculated based on the distance from each district to all the historical black homelands. That is to say, the number of groups in the predicted ethnic diversity index is fixed, which is the same as the total number of homelands. Therefore the only variation in the instrumental variable comes from the uniformity in the distance to different homelands (which refers to the distribution of population size among all these ethnic groups). By applying this instrumental variable, ethnic diversity has a clear meaning here: a more diverse place implies the distribution of group size is more even, which is independent of the number of ethnic groups.

To verify this argument, we run the IV regressions similar to our main analysis in Table 13.2. In Panel A, we control for the number of groups (i.e. the corresponding variable is an inverse of the number of groups) and use the predicted ethnic diversity index as an instrumental variable for the dispersion of group size for both 1996 and 2001. IV regressions in column 1 - 6 imply that given the number of groups, a decrease in the dispersion of group size (i.e. group size is distributed in a more even way, which is the case in a more diverse place) will significantly increase the wage employment rate. Again, this only works for ethnic groups with relatively larger population size. Furthermore, the instrumental variable remains strong in all these columns as the F statistics is close to or larger than 10, especially for large groups.

In Panel B we control for the dispersion of group size and instrument the number of groups (i.e. the corresponding variable is an inverse of the number of groups). In all columns the instrumental variable is weak as F-statistics is around or below 3. This confirms that the instrumental variable does not capture the number of different groups.

As a further check, we change the measure of ethnic diversity from the fragmentation index to the polarization index which, given the number of groups, cannot be monotonically mapped to the dispersion of groups size[37] and see if it still significantly affects employment chance. The polarization index (P) such as Montalvo and Reynal-Querol (2005) has also been widely used in literature. The index captures the deviation of the distribution of the ethnic groups from the bipolar distribution (which represents the highest level of polarization). Following the notations in defining fractionali-

[37] Current literature shows the fractionalisation and polarisation index are highly correlated at low levels, while being uncorrelated and negatively correlated at intermediate and high levels, respectively (Montalvo and Reynal-Querol, 2005).

sation index, the index is computed as:

$$P = 1 - \sum_{k=1}^{m} \left(\frac{1/2 - s_k}{1/2} \right)^2 s_k$$

We use the same "equidistance" measure as an instrumental variable for ethnic diversity here. Following the same approach as that for fractionalisation index, we use predicted polarisation index obtained from the predicted stock of ethnic groups in each district as an instrumental variable for real polarisation index. After getting predicted population share of each ethnic group $\widehat{s_k}$ in each district based on the gravity model (4), we get the predicted polarisation index:

$$\widehat{P} = 1 - \sum_{k=1}^{K} \left(\frac{1/2 - \widehat{s_k}}{1/2} \right)^2 \widehat{s_k}$$

We use this predicted polarisation index as an instrumental variable for P from real data and conduct both OLS and IV regressions. We report the first-stage outcomes in Appendix Table A8.1 and the individual-level regressions in Appendix Table A8.2. First-stage regressions show that the predicted polarisation index is a good indicator of the real polarisation index, although the instrumental variable is not very strong.

Appendix Table A8.2 reports the results of the polarisation index on both employment rate in general and wage-employment rate in particular (again excluding self-employed people). The effect of polarisation index is not significant in most regressions. As the polarisation index not only reflects the diversity of ethnic groups but is also weighted by the relative group size, it cannot be monotonically mapped to the dispersion of group size. Therefore the insignificance of the coefficients of polarisation index in these regressions indicates that the dispersion of group size is the main driving force in labour market outcomes in our setting.

Summary of empirical results. The above empirical section consolidates the following results which are the basis for the theoretical model in the next section:

1. Ethnic diversity increases employment among the working-age black population and this mainly takes place in wage-employed jobs.

2. The positive effect of ethnic diversity on employment only works for the ethnic groups with relatively large size.

3. Ethnic diversity affects employment opportunities through the change in the dispersion of population size among different ethnic groups.

6 How Does Ethnic Diversity Affect Employment: A Theoretical Model and Mechanism

We propose a theoretical framework consistent with our empirical findings above to explain the positive effects of ethnic diversity on employment and the heterogeneity of the effects across sub-groups. More specifically, we focus on social skill investment which increases with ethnic diversity. The model can be verified by both numerical simulation and empirical evidence from real data.

6.1 A theoretical framework

The story is as follows. Assume that inter-ethnic communication requires more skills than intra-ethnic interaction. In a more diverse place, the necessity to communicate with individuals from different ethnic groups motivates people to learn and practise more social skills. The acquisition of this extra skill, which is helpful in reducing coordination costs or increasing labour productivity (which we will discuss later on), can make them more competitive in the labour market and increase their chances of finding jobs.

In more detail, people obtain utility from interacting with those both inside and outside their own ethnic group. Establishing a relationship with someone from a different ethnic group requires more skills than with those from the same ethnic group (this may be due to barriers like language). In a more ethnically diverse place people have to communicate with a larger proportion of individuals outside their own ethnic group to maintain a certain level of social connection. Therefore they put in more efforts in developing social skills, as long as the benefit of interacting with a different ethnic group outweighs the cost of learning efforts. Social skills here can be of many types, including both cognitive skills like language and non-cognitive skills like pro-social traits. When people are in the labour force, these skills are beneficial to their labour market performance, in addition to their human capital investment.

What needs to be emphasised here is that more ethnically diverse places do not necessarily have more overall social interaction in general but the investment in social skills should be higher because a larger proportion of social interaction comes from inter-group connection and inter-ethnic interaction requires more skills than intra-ethnic communication.

The distinction between social connection and investment in social skills is analogue to the literature which differentiates social connectedness and network formation (Chay and Munshi, 2015). Their story implies that there exists a threshold above which social connectedness and network-based outcomes are positively correlated. Similarly, in our story, the level of social connection can be high in both ethnically homogeneous and diverse places, but investment in social skills is only high when a large proportion of this social connection takes place between ethnic groups as intra-ethnic communication is relatively costless.

6.1.1 Model setup

We provide a model of a coordination game to explain the mechanism. We assume that individuals gain utility from social interaction at the cost of investing in social skills. As the cost of communicating with a different ethnic group is larger than that with the same group, we assume communication within each ethnic group is costless. The cost of investment in social skills for inter-ethnic interaction is c per unit. We also assume that the amount of investment in social skills x_{ik} equals the output of the investment (i.e. the amount of skills acquired) for individual i in ethnic group k. We have the following setup of a coordination game:

Players. Each group only differs in terms of their population size. Suppose there are m ethnic groups in total. We denote these different groups as m different sets $N_1, N_2, \dots N_m$, each with a group size n_k and $k = 1, 2, \dots m$. The overall population in each district is N, so that $\sum_{k=1}^{m} n_k = N$. Each ethnic group then has the share s_k and $k = 1, 2, \dots m$ over the whole population in the district. Here $s_k = \frac{n_k}{N}$.

Strategies. Each individual i in group k invests x_{ik} in social skills. For simplicity we assume x_{ik} is a binary variable which equals 1 (0) if i invests (does not invest).[38] One can only participate in inter-ethnic social interactions if he invests in social skills. The total amount of people each individual i in ethnic group N_k with a group size n_k has access to in the inter-ethnic communication is calculated as $x_{ik} \sum_{j \neq k} \sum_{q \in N_j} x_{jq}$. There is complementarity between i's own investment in social skills and the overall investment level of people outside group k. Therefore the total number of people interacting with i (both inside and outside his own group) can be calculated as $n_k + x_{ik} \sum_{j \neq k} \sum_{q \in N_j} x_{jq}$.

One important feature of the strategy is that by construction we assume skill investment is bilateral. If $x_{ik} = 0$, i cannot benefit from social interaction even if everyone outside his group invests in social skills. As is discussed in the literature review part, papers discussing the social behaviour of ethnic minorities in the American society argues that investment in social skills is unilateral as ethnic minorities try to assimilate to the American society by learning the language spoken by the American majority while the Americans do not put in any efforts in learning additional language. However, in our case, if we assume that skill investment is unilateral such that the "small" ethnic group assimilates to the "large" group by learning their language while the "large" group do not need to make any investment, we should observe the pattern that only the "small" ethnic group responds to ethnic diversity, which directly contradicts our empirical finding (where we find only "large" ethnic group responds to ethnic diversity). Therefore, a more reliable assumption in our setting is that both groups put efforts in learning additional social skills. A reasonable example is that both groups learn a common language. This is also consistent with our proxy of social skill later on, which is the proficiency of English/Afrikaans as the second language. In this case one can communicate with people from another ethnic group only if both learn a second official language.

Utility. Individual i belonging to group k obtains utility from social interaction which depends

[38] One can potentially treat x_{ik} as a continuous variable or make x_{ik} heterogeneous in communicating with different ethnic groups. For example, similar to Akerlof (1997), we can introduce the investment of x_{ikj} if individual i is interacting with group j, and x_{ikj} is a decreasing function of social distance between groups k and j. However, this binary setting of x_{ik} is already enough to explain the key empirical findings about ethnic diversity discussed above.

on the size of his own groups n_k and the number of people he can reach in other ethnic groups, the latter relying on both his own investment in social skills and the efforts from other ethnic groups. The utility from overall social interaction is written as $f(n_k + x_{ik} \sum_{j \neq k} \sum_{q \in N_j} x_{jq})$, which is assumed to be increasing at a diminishing rate. That is, $f' > 0$ and $f'' < 0$. The implication is that utility from social interaction increases as more people participate in communication, but this has a diminishing return as people get tired from social life when the number of contacts increases. We can thus write the net utility U_{ik} from overall social interaction for individual i in group k as follows:

$$U_{ik} = f(n_k + x_{ik} \cdot \sum_{j \neq k} \sum_{q \in N_j} x_{jq}) - cx_{ik}$$

We then normalise the amount of social interaction by the overall population size N in the corresponding district. By doing this we control for the whole population size and what matters in social interaction is the share of each group over the whole population rather than the absolute level of group size. There are three reasons to do this normalisation. Firstly, each group's share of population size, instead of the absolute level of group size, is directly linked to our measure of ethnic diversity index. Secondly, controlling for the magnitude of overall population size in each district is consistent with our empirical analysis where we control for the population density in each district and investigate the remaining variation in ethnic diversity. Thirdly, the interpretation of the utility function becomes more intuitive. As the total amount of time for social interaction is limited for each individual, what matters more in social connection is not the total amount of people one has access to, but the probability of establishing connection to a person one randomly meets in the district per unit time. After this normalisation, the utility function becomes:

$$U_{ik} = f(s_k + \frac{x_{ik} \cdot \sum_{j \neq k} \sum_{q \in N_j} x_{jq}}{N}) - cx_{ik} \tag{8}$$

Here we also assume that the per unit cost of social interaction is the same in different districts. In principle one can extend the model by allowing the cost c to vary across districts. If this is the case, how the investment in social skills responds to ethnic diversity might be ambiguous. On the one hand, it can increase with the level of ethnic diversity, which will be explained by our mechanism. On the other hand, it may decrease with ethnic diversity as more ethnically diverse districts might have more conflicts, which discourage people from social interaction. However, as we find the positive effect of ethnic diversity on employment in the empirical part, we argue that the positive side of ethnic diversity is more important than its negative side. Furthermore, how conflict responds to ethnic diversity has been discussed in other literature already and is not the central focus of this paper. Therefore, we only focus on the explaining the positive effect of ethnic diversity by simplifying other potential factors at the negative side. In our numerical simulation we also set different values for c to see how this affects our results.

Equilibrium. In this paper we focus on pure strategy Nash equilibrium. In this game, player i from group k chooses either $x_{ik} = 1$ or $x_{ik} = 0$ to maximise his total utility from social interaction, given the population share of each ethnic group as well as the investment of x among people outside group k. In the pure strategy Nash equilibrium, no one has the incentive to deviate from his current decision.

Clearly the coordination game has multiple equilibria. For example, $x_{ik} = 0, \forall i, k$ is a Nash equilibrium. This is because starting with this initial condition, no one has the incentive to deviate. In more detail, for an individual i in group k, his utility from social interaction is:

$$U_{ik} = \begin{cases} f(s_k) - c, & \text{if } x_{ik} = 1 \\ f(s_k), & \text{if } x_{ik} = 0 \end{cases}$$

Therefore individual i always gets higher utility by not investing in social skills. That is to say, in order for the social interaction to happen, there might be some initial efforts to stimulate communication.

6.1.2 Key features of the Nash equilibrium with the maximal level of skill investment

Each Nash equilibrium is characterised by its own level of skill investment so that it does not make sense to conduct comparative statics across different Nash equilibria in this setting. Moreover, the ultimate Nash equilibrium in a given district with a given distribution of group size purely depends on its initial condition (i.e. how many groups choose $x = 1$ and how many choose $x = 0$ in this district originally). As there is no particular selection criterion of the initial condition in each district, it is reasonable to assume that the initial conditions are assigned randomly for each district. In this case, each district falls in any of its own possible Nash equilibria with equal probability. Therefore, the expected level of skill investment in each district at the equilibrium is determined by the range of its possible Nash equilibria (i.e. the Nash equilibria with the maximal and minimal level of investment in social skills). That is to say, to capture the expected level of skill investment in each district, we can just focus on the range of possible Nash equilibria in each district instead of discussing each Nash equilibrium individually.

As discussed before, the Nash equilibrium with the minimal level of skill investment is the same for all districts (i.e. everyone chooses $x = 0$). In this case, the range of possible Nash equilibria in each district is only determined by the Nash equilibrium with the maximal level of skill investment. Therefore in the following discussion we only focus on the equilibrium where the number of individuals investing in social skills is as large as possible, and see how this equilibrium state changes in response to group size. By doing this, we can indirectly demonstrate how the expected level of skill investment changes with ethnic diversity in each district.

One important feature of this particular equilibrium is that to guarantee the maximum participation in inter-ethnic communication, individuals always choose to invest in social skills unless the net utility from doing so is strictly smaller than that from deviating. In other words, even if the individual is indifferent between investing and not investing, he will always choose to invest in social skills.

In addition, we derive the following two lemmas which capture the key characteristics of the Nash equilibrium in this coordination game with the maximal level of skill investment.

Lemma 1. *In each district, people from the same ethnic group choose the same amount of investment.*

Proof. Suppose player 1 and player 2 both come from ethnic group k with group size n_k. Without loss of generality we assume $x_{1k} = 1$ and $x_{2k} = 0$. We focus on the pure strategy equilibrium with the maximum number of skill investment. As both 1 and 2 maximise their utility from social interaction, we have:

$$\begin{cases} f(s_k + \frac{\sum_{j \neq k} \sum_{q \in N_j} x_{jq}}{N}) - c \geq f(s_k), & \text{for player 1} \\ f(s_k + \frac{\sum_{j \neq k} \sum_{q \in N_j} x_{jq}}{N}) - c < f(s_k), & \text{for player 2} \end{cases}$$

Clearly these two inequalities contradict each other. Therefore we must have $x_{1k} = x_{2k} = 1$ or $x_{1k} = x_{2k} = 0$. □

Based on this, we have lemma 2:

Lemma 2. *In each district, people from different groups choose the same amount of investment as long as the population size of these groups is the same.*

Proof. Suppose player i and player j come from ethnic group k and l, and $s_k = s_l$. Without loss of generality we assume $x_{ik} = 1$ and $x_{jl} = 0$. According to lemma 1, everyone from group k (l) chooses $x_{ik} = 1$ ($x_{jl} = 0$). As both i and j maximise their utility from social interaction, we have:

$$\begin{cases} f(s_k + s_l \cdot 0 + \frac{\sum_{p \neq k, p \neq l} \sum_{q \in N_p} x_{pq}}{N}) - c \geq f(s_k), & \text{for player } i \\ f(s_l + s_k \cdot 1 + \frac{\sum_{p \neq k, p \neq l} \sum_{q \in N_p} x_{pq}}{N}) - c < f(s_l), & \text{for player } j \end{cases}$$

When $s_k = s_l$, these two inequalities hold altogether if and only if $f(s_k + \frac{\sum_{p \neq k, p \neq l} \sum_{q \in N_p} x_{pq}}{N}) - c > f(s_k + s_k + \frac{\sum_{p \neq k, p \neq l} \sum_{q \in N_p} x_{pq}}{N})$ for each possible x_{pq} in group p. As $f' > 0$, $n_k \geq 0$, $c > 0$, this inequality cannot hold.

Therefore we must have $x_{ik} = x_{jl} = 1$ or $x_{ik} = x_{jl} = 0$. □

6.2 Social interaction, skill acquisition and distribution of group size

6.2.1 Analytical predictions: staring from a symmetric case

Combining lemma 1 and lemma 2, we can link the size distribution of ethnic groups to social skill investments. To guarantee the maximal level of skill investment in equilibrium, we start with the initial condition where $x_{ik} = 1, \forall i, k$ and study people's incentive to deviate from this condition.

We derive an analytical proposition based on a symmetric case where all groups in a district have the same population size (i.e. group size is distributed evenly). We later on show that it is not feasible to prove the proposition with an arbitrary distribution of group share. But we will provide a numerical simulation based on the generalised density function of group share to verify the proposition.

Consistent with the empirical strategy, we fix the total number of ethnic groups in a district and see how a more even (uneven) distribution of these groups affect social skill acquisition. Suppose the number of groups m is fixed but groups are not distributed evenly. Starting from the point where each group has the same population size and compare it with the case of asymmetric size distribution among all these groups, we have the following proposition:

Proposition 1. *Suppose the total number of different groups is given. Compared with the symmetric case where each group has the same population size in the district and everyone invests in social skills, social skill investment decreases when the dispersion of group size in a district increases (i.e. the distribution of population size among different groups becomes more uneven).*

Proof. Given the total number of different groups m, the dispersion of group size can be captured by $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$.[39] Starting from the symmetric case where every group has $s = \frac{1}{m}$, when the distribution of group size is more uneven, the gap in the population share among all these groups becomes larger. One implication is that if $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ becomes larger, either there exists one s_k which is extremely large or there exist several s_k which are larger than the fixed mean value of group share $\frac{1}{m}$ (Otherwise the overall population size is smaller than N). As a result, to show that a higher proportion of people will deviate from the initial condition when the distribution of group size is more dispersed, we just need to show that larger groups are more likely to deviate.

Starting from $x_{ik} = 1, \forall i, k$ as the Nash Equilibrium. The utility of social interaction for individual j in group k is:

$$U_{jk} = \begin{cases} f(s_k + (1 - s_k)) - c, & \text{if } x_{jk} = 1 \\ f(s_k), & \text{if } x_{jk} = 0 \end{cases}$$

Individual j in this group will deviate if:

$$f(1) - c < f(s_k) \Rightarrow s_k > s^* \tag{9}$$

Suppose in the symmetric case no one deviates, which means $f(1) - c \geq f(\frac{1}{m})$. When group sizes are more unevenly distributed in a district, the population share of the largest group(s) becomes larger than $\frac{1}{m}$. Suppose in a district, s_k is the largest group in the distribution of group size ($s_k >$

[39] In principle the dispersion of group size can also be captured by the variance or standard deviation of group share. Here $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2 = m * Var(s_k)$. We do not use the $Var(s_k)$ to measure the dispersion here is that to prove the level of investment changes with number of different groups and the dispersion of group size, we must hold one fixed and get the other to vary. In statistics $Var(s_k)$ decreases intrinsically with m. Therefore, it is very hard to find the same $Var(s_k)$ for different m. Thus it is better to scale $Var(s_k)$ up by a scalar m. In another case, if we want to hold m constant and see the changes in the dispersion, it does not matter whether we use $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ or $Var(s_k)$ as $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2 = m * Var(s_k)$.

$\frac{1}{m}$), it is straightforward that it is more likely to have $s_k > s^*$ when group sizes are more unevenly distributed in the district. In this case the largest group k will deviate and choose $x_{jk} = 0$. For the remaining groups, suppose group l is the second largest group. Given the largest group deviates from the equilibrium $x_{jk} = 1, \forall j$, the same logic shows that for group l to deviate as well, we must have:

$$f(1 - s_k) - c < f(s_l) \tag{10}$$

Since $f' > 0$, we find that the motivation for deviating increases with group size. In particular, when the dispersion of group sizes is larger in a district, more groups will have large sizes (at least larger than the mean value of group size) so that they will deviate from the initial condition where everyone chooses to invest in social skills.

□

Although not the story in our paper, this model can also be generalised to the case where the dispersion of group size is fixed and the number of ethnic groups varies. The same model can explain how skill investment improves with the increase in the number of different groups. Details are in the Appendix C.

6.2.2 Numerical simulation with a convoluted density function of group size: algorithm and results

In this section we give a more generalised verification of proposition 1 with numerical simulation by allowing for a convoluted density function of group size in a district. The logic is similar to that behind the equation 9.

Suppose there are m groups in total in a district. Without loss of generality we rank them by an ascending order of group size. We have:

$$s_1 \leq s_2 \leq \ldots \leq s_k \leq s_{k+1} \leq s_m$$

We start from the initial condition where everyone invests in social skills. The largest group will deviate if:

$$f(1) - c < f(s_m)$$

After the largest group deviate, the second largest group will deviate if:

$$f(1 - s_m) - c < f(s_{m-1})$$

In general, suppose s_k is the last group which deviates from the initial condition. We have:

$$\begin{cases} f(1 - s_m - s_{m-1} - \ldots - s_{k+1} - s_k) - c = f(s^*) \\ s_{k-1} \leq s^* < s_k \end{cases} \tag{11}$$

Given the population share of each group over the total population in the district, the largest s^*, which gaurantees the maximal level of social interaction at the equilibrium, is unique.

The total proportion of people deviating from the initial condition is $Y = \sum_{s_k > s^*}^{m} s_k$. And the overall level of skill investment is $1 - Y = 1 - \sum_{s_k > s^*}^{m} s_k$.

It is not feasible to get an analytical analysis on how Y changes with m or $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ with an arbitrary distribution of group size. This is mainly because Y depends on both the number of groups which deviate from the initial condition and the population share of these groups, which is not easily captured simultaneously by a density function of group share. Furthermore, with different parameter values c, how Y reacts to the number of groups and the dispersion of group size is not always unambiguous. For example, in some particular distribution, we may find that Y decreases with the dispersion of group share at some point, which we have encountered in our numerical simulation. However, if the amount of tests in the numerical simulation is large enough, we can get overwhelming results that support our propositions.

Algorithm. We need to numerically show that for a convoluted distribution of group share in a district, the proportion of people who deviate from investing in social skills (i.e. Y) increases with the dispersion of group size $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ when the number of groups is fixed. Suppose the utility function is $f(x) = \sqrt{x}$. c in principle can take any positive values. We conduct our simulation based on three of them: $c = 0.1$, $c = 0.2$ and $c = 0.5$. For each c, the steps of simulation are as follows.

1. Draw s_k, $k = 1, 2, \ldots, m$ from a convoluted distribution of s, but make sure that $\sum_{k=1}^{m} s_k = 1$ (Appendix D explains how to make the constrained draws in more detail).
2. Choose a particular m as we want to fix the number of groups.
3. Rank each s_k in an ascending order.
4. Suppose the largest group share is s_m. $Y = 0$ if $\sqrt{s_m} \leq \sqrt{1} - c$. Otherwise move to the next step.
5. Suppose the second largest share is s_{m-1}. $Y = s_m$ if $\sqrt{s_{m-1}} \leq \sqrt{1 - s_m} - c$. Otherwise move to the next step.
6. Continue until we find s^* which satisfies Equation 11. $Y = \sum_{s_k > s^*}^{m} s_k$.
7. $Y = 1$ if we search till the smallest s_1 but still could not find such s^*.
8. Operate another draw of s_k. Repeat the steps 3-7 to get different Y. In our simulation we conduct 100000 tests.

9. Fix m, we calculate the standard deviation of group share in each test (i.e. $SD(s)$) and finally draw a figure of mean value of Y over $SD(s)$.

The simulation results are in Figure 6. Here we fixed the number of groups m and see how the proportion of people who deviate from investment changes with the dispersion of group size in a district. For each value labeled in the x axis, we conduct 100000 tests to get the mean value of Y. For each test, we also do the same simulation for different c. Consistent to proposition 1, the probability of deviating increases with the dispersion of group size in a district. This is robust to different numbers of groups we set in our simulation. The intuition is that when the distribution of group size becomes more uneven, there is a larger chance that we can have groups with very large size and these are the groups which are the most likely to deviate.

One interesting finding is that when c is relatively large and the number of groups is small (which means each group is important), proportion of people who deviate from the initial condition can decrease with the dispersion of group size (panel a in Figure 6). This is because in this case we can have two districts, one having only one very large group and the other having several large groups. The relative magnitude of their overall population share in the corresponding district can be ambiguous. This result can actually tell how our paper is reconciled with the current empirical finding that ethnic diversity is negatively correlated with economic outcomes. This means, if the conflict level in a district is too high (i.e. cost of investment is too large), a more diverse district (i.e. less dispersion of group size) can potentially have less investment in social skills, which might be harmful to economic outcomes.

6.2.3 Social interaction, skill acquisition and ethnic diversity

We prove from the above propositions that skill investment is higher when the group size is more evenly distributed. And how does these relate to ethnic diversity?

Equation 7 indicates that ethnic diversity decreases with the dispersion of group size. Based on proposition 1, we have the following proposition 2:

Proposition 2. *Social skill investment increases with ethnic diversity (which means a more even distribution of group size given the number of groups).*

Following proposition 1, we also have proposition 3:

Proposition 3. *Ethnic groups with relatively smaller group size are not affected by the ethnic diversity.*

This is because when the initial condition is $x_{ik} = 1, \forall i, k$, in the Nash equilibrium with the maximal level of social interaction, the small group will not deviate as long as their group size is below a certain level (regardless of the strategies of the large group). In other words, they always choose to participate in inter-ethnic communication and invest in social skills regardless of ethnic diversity levels. Therefore the small group will in general have more social skill investment than the large group but their social skill investment is not affected by ethnic diversity of the district. The intuition is that as the small groups get relatively less utility from intra-group communication, they rely more on inter-group connection and therefore are less sensitive to the incentive to deviate caused by changes in the level of ethnic diversity.

One thing to notice is that in our data "large", "medium" and "small" groups are defined by the group size in the national population while in the model "small" and "large" groups are defined at district level. However, definitions at these two levels are compatible in our data. A detailed investigation of the population share in each district in both 1996 and 2001 shows that in general groups with large population size at the national level are also the dominant group in ethnically homogeneous districts, while groups with small population share at the national level also makes up a very small part of the population in those districts. In diverse places the population size of these groups becomes more balanced.

6.2.4 Social skills and labour market outcomes

The social skills acquired through inter-group interactions in a diverse place might potentially improve workers' employment opportunities in several ways.

Less search cost in job hunting. Social skill lowers the cost of searching for potential jobs, therefore increasing labour supply. More social skills help individuals build closer and stronger intra-group contacts. For example, people with higher social skills are better at making use of networks and other methods in gaining job information or asking for referrals. Current literature shows that social network is an important factor in providing more job opportunities for low-educated labours both in South Africa (Magruder, 2010) and in other developing countries (Munshi, 2003).

Increased productivity of certain skills. Recent literature which incorporates different tasks in the production function (Acemoglu and Autor, 2011) and highlights the importance of social skills (Deming, 2017). Under the framework that low and high-skill workers have their own comparative advantages in dealing with different tasks and the range of tasks performed by low-skill workers is determined by where their comparative advantages are, Deming (2017) explains that social skill increases the productivity of certain tasks by allowing workers with comparative advantages to trade their tasks, which leads to more efficient production. In our story, acquiring additional social skills may also potentially increase the productivity of certain tasks and increase the employment chances for low-skilled workers by allowing them to perform a wider range of tasks.

Overcoming skill deficit. A simple explanation on why social skill stimulates employment is that it works as a substitute for other skills required by employers. In particular, low-educated workers may lack skills necessary for certain occupations, which prevents them from getting the position. For example, if the candidate for the position of a salesman lacks necessary skills of communication, proficiency of additional language may compensate for this communication skills and guarantees him for the position. As the substitutability between social skill and skills acquired through formal education helps more people qualified for the positions they apply for in a more diverse place, the employment rate will increase accordingly. Skill acquisition from inter-group interaction here functions in a way similar to what is emphasised in related literature that community-based network can work as a substitute for endowments by helping individuals from disadvantaged families get out of low-skill occupational traps (Munshi, 2011).

6.3 Ethnic diversity, social skill acquisition and employment: empirical evidence

In this section we provide some evidence to show that social skill acquisition increases with ethnic diversity. There is no straightforward information in census data on social skills. The closest one we can approach is the information on second language at home, including whether or not one speaks a second language and which language they speak. A black person is considered to have some proficiency in a second language if he speaks either one of the nine ethnic languages or a common language (English or Afrikaans). Language is often considered as a cognitive skill which can be learnt from school. In this setting, however, controlling for educational background and investigating into the heterogeneity in the acquisition of language skills among sub-groups, we hope the proficiency of the second language can capture some information on the skills one acquires from inter-group interactions.

More importantly, whether one speaks a second language (and which language he speaks) reflects more of his investment in social skills than the inheritance of language skills from his parents. This results from a series of laws and regulations during the Apartheid regime. Firstly, inter-racial marriage was prohibited during Apartheid starting from 1949 when the Prohibition of Mixed Marriage Act came into effect. The act was repealed in 1985 by the Immorality and Prohibition of Mixed Marriages Amendment Act. In 1996 and 2001 census, parents and spouse of the working-age black people of our interest either lived through Apartheid when marriage between black and white (or black and coloured) was abandoned, or they got married before the independence of South Africa from the British colonisation when there was already informal racial segregation. Thus it is not very likely that the proficiency of English or Afrikaans among the current generation was purely obtained from their parents in the inter-racial marriage. Even among the black population, inter-ethnic marriage is also rare. As is discussed at the beginning of the paper, inter-ethnic relationship was deteriorated during Apartheid so that marrying someone from another ethnic group is not a common case. Appendix table A9 shows that in 1996 census, the contemporary inter-ethnic marriage rate is less than 4%. This phenomenon is even more rare in the parental generation as their inter-ethnic marriage rate is only 1%. Although the sample is selected as only spouse and parents cohabiting with the household head are included in the census, this statistics can still reflect the low inter-ethnic marriage rate.

Furthermore, whether one speaks a second language is not very likely to capture the language proficiency of individuals before they decided to move out of the homelands. As is discussed in the institutional setting, there were almost no indigenous black people in the "white" areas in South Africa and the contemporary population in these districts are mainly the decedents of the migrants from different homelands before the arrival of white colonisers. Therefore it is unlikely that those ancestors learnt English or Afrikaans before migration. Furthermore, in the institutional setting, we have already shown that over 90% of the black population surveyed in 1996 census either never moved or moved within districts up till the time when they were surveyed. Even among recent migrants in 2001, intra-district migration is much larger than inter-district migration. This further shows that the distribution of ethnic diversity in our census data is largely inherited from the historical pattern rather than driven by contemporaneous migration who potentially acquired language skills before migrating.

To prove the channel in our theoretical model, we first show that ethnic diversity improves social skill acquisition (i.e. measured by second language proficiency) and then we demonstrate that higher social skill is correlated with higher employment rate conditional on ethnic diversity. As the information on second language proficiency is only reported in 1996 census data, we only show the results in 1996 census in this section.

Appendix table A9 also reveals that the proportion of people who speak a second language is not too small. Among the whole black population, around 22.5% speaks a second language, 8.7% (13.8%) of which speaks a common language (ethnic language). In the regression analysis we focus on the common language (English or Afrikaans) instead of ethnic language as the former one is more related to labour market performance in wage-employment and less likely to reflect family inheritance as the ban on inter-racial marriage was more strict than inter-ethnic marriage during Apartheid.[40]

We introduce a dummy variable on whether one can speak English or Afrikaans as a second language and regress it on ethnic diversity in 1996, conditional on the same set of control variables in the main analysis. Simple OLS regressions may suffer from the same problem as is discussed before. For example, there are two potential types of selection of migrants related to their language proficiency. Firstly, migrants with higher ability are able to move out of the homelands and these people might have already mastered a second language prior to migration. Secondly, migrants with better language proficiency choose to move to a more diverse area where there are more job opportunities. If the first type of selection is the case, people with higher ability than their counterparts in the original homelands can potentially move to both ethnically homogenous and diverse places. Thus we should not see any correlation between ethnic diversity and proficiency of second language if language skills are purely captured by the selection of migrants at the time of moving out of homelands. The second selection of migration comes from the fact that migrants with higher ability (including language efficiency) move to more diverse places as migrants who cannot speak a common language may find it difficult to communicate with people outside their ethnic groups. To deal with this selection, we use the same instrumental variable approach as is implemented in the main analysis (using predicted value of ethnic diversity in 1996 as an instrument for real ethnic diversity).

Table 14.1 shows both OLS and IV regression results about how ethnic diversity affects individuals' second language proficiency. Panel A and B investigate the results for the whole black population and the heterogeneity of the effects of ethnic diversity by group size. The coefficients in Panel A in both OLS and IV regressions are significantly positive, indicating that ethnic diversity increases the probability of learning a second language (English or Afrikaans). In Panel B, a comparison between groups with large, medium and small population size indicates that ethnic diversity has a strong and positive effect on language skills only among the ethnic groups with relatively large population size,

[40] But in regressions the proficiency of both common language and ethnic language can respond to ethnic diversity.

which is consistent with proposition 3 in the model. In addition, the instrumental variable remains strong in both whole-sample and sub-sample regressions.

One concern in interpreting the positive impacts of ethnic diversity on language proficiency as a result of social interaction is that in a more diverse place the importance of English or Afrikaans is more highlighted. For example, firms will have a more favourable environment for employees to learn an official language, as employees have to serve customers or talk to colleagues whose first language is different from these employees' own ethnic language. In this case language proficiency is developed after one has found a job rather than before, and the probability of finding a job is affected by other factors. To show that the improvement in language proficiency is driven by the need for social interaction instead of a skill purely developed in the workplace, we split the sample into subgroups with different ages: young people below the age of 15, working-age people (15-64) and people who are retired ($\geq$65) in Table 14.2. We find the positive and significant effect of ethnic diversity on language proficiency in IV regressions among all the three age groups, which indicates that the language skill can be achieved even before people enter the labour force, therefore is not purely driven by the requirement from the workforce.

As a further check of our mechanism, we decompose the ethnic diversity index into the inverse of number of groups and the dispersion of group size and conduct the same OLS and IV regressions as described in Table 13.1 and 13.2 by replacing the dependent variables with a dummy on whether one can speak English or Afrikaans as the second language. The results are in columns 5 - 8 in Table 13.1 and 7 - 9 in Table 13.2. All the results imply that language proficiency increases when the distribution of group size becomes more even, and this only works for groups with large population size.

We then look at whether acquisition of social skills improves labour market outcomes by regressing employment probabilities on the proficiency of a second language (English or Afrikaans) conditional on ethnic diversity, as is presented in Table 15. The dependent variable in Panel A is a dummy on whether one is employed or not (including unemployed and inactive) while in Panel B the dependent variable equals 1 if one is an employee and 0 if one is unemployed or inactive. The independent variable in all these OLS regressions is a dummy on whether one can speak English or Afrikaans as a second language. Again we look at the whole sample and the difference among groups with large, medium and small population size. In all regressions learning a second common language is positively and significantly associated with higher employment rate (both overall employment rate and wage-employment rate).

6.4 Summary of the theoretical model and mechanism

In summary, diversity along ethnic lines could provide individuals with social skills, which improves their employability. That is to say, even if ethnic diversity does not necessarily increase the amount of overall social interactions within a district, it may still motivate people in more diverse areas to learn and practise more skills such as a common/official language. This is because communication with individuals from different ethnic groups requires more efforts and skills than intra-ethnic interaction. The acquisition of this extra skill, which is helpful in reducing coordination costs or increasing productivity of certain skills, could increase individuals' chances of finding a job.

The key point of the story is that it is not the overall amount of social interaction that drives the whole story, but the composition of the social interaction. That is to say, more diverse districts do not necessarily have larger amount of social interaction but a larger proportion of the interaction comes from inter-ethnic communication, which is more challenging than intra-ethnic connection and therefore gives people more motivation to invest in social skills.

In our model, without imposing any intrinsic difference in taste, skills or attitudes between different ethnic groups, the tradeoff between the cost of and benefit from developing social skills leads to the conclusion that inter-ethnic social interaction and investment in social skills are the most likely to occur in a place where the distribution of group size is relatively even, which implies a larger ethnic diversity. It is because starting from an initial condition where everyone invests in social skills, less people deviate from this investment decision in the equilibrium state in a more ethnically diverse place. This effect occurs mainly among the ethnic groups with relatively large group size. In the labour market, the acquisition of these extra social skills is helpful in lowering the barrier to formal jobs by reducing coordination and search costs, by increasing productivity of certain skills or by substituting for some necessary skills which are otherwise not available.

6.5 Ruling out some alternative explanations

Ethnic diversity might positively affect the labour market outcomes of the blacks through several channels. Here we rule out some alternative explanations through which ethnic diversity improves labour market outcomes based on our data and narratives.

Labour supply: skill complementarity. There might be some skill complementarities among different ethnic groups, as each may have their own comparative advantages in skills. For example, South Sotho are believed to have special skills as shaft-sinkers on the mines (Guy and Thabane, 1988). Therefore, diversity generates creativity and innovative environment by combining people with different skills. In this case, we can also expect diversity to affect differently individuals with different level of education. A priori, we would expect to find a stronger effect for the higher educated whose activities would benefit more from knowledge-sharing and problem solving.

In Appendix Table A10 we replicate the main results by splitting the sample into people with high and low educational levels. According to the compulsory schooling law in the post-Apartheid South Africa, one has to go to school when reaching the age of seven and stays at school until the age of fifteen or the ninth grade. We therefore use 9 years of schooling as a cutting point between high and low-educated people. "High education" refers to people with more than 9 years of schooling (i.e. high school, college and postgraduate) while "low education" means no education, primary and junior high school education. We present both OLS and IV results in both years.[41] In 1996 the positive and significant effect of ethnic diversity on wage-employment rate only exists among low-

[41] The results are robust to other definitions of "high" and "low" educational categories. For example, we also split the sample into people with more and less than 7 years of schooling, and people whose years of schooling are above and below the mean value in the district where they live.

educated working-age black population. The magnitude of the coefficients of ethnic diversity index is also larger among the low-educated group. In IV regressions in 2001 the positive effect of ethnic diversity still only holds for low-educated people. However, there is some difference in its magnitude between 1996 and 2001. From 1996 to 2001 the magnitude of the coefficient of ethnic diversity index increases largely from 0.05 to 0.12 for high-educated people while for low-educated people the increase is smaller (from 0.14 to 0.19). A more detailed split of the sample reveals that the increase in the magnitude of the effect of ethnic diversity on wage-employment rate takes place only among college graduates while for high-school graduates the coefficient is insignificant and the magnitude is still around 0.05. All this indicates the substitutability rather than the complementarity between education and ethnic diversity.

Furthermore, if ethnic diversity generates skill complementarity, it might also give birth to new occupations as new skills can be learnt from other ethnic groups and this creates opportunities for occupations which rely on otherwise infeasible tasks. Therefore, if ethnic diversity stimulates new ideas and skills, we may observe a larger range of occupations in a more diverse place. We regress the range of occupations in each district.[42] on ethnic diversity. The results from 1996 and 2001 census are in Appendix Table A11.1 and Appendix Table A11.2. We do not find any positive relationship between diversity and potential new occupations in either OLS or IV regressions .

Labour supply: social grant. Social grant, such as Old Age pension, potentially dis-incentivise labour force participation in South Africa (Banerjee et al., 2008). At the same time there is a possibility that a more ethnically homogenous place is associated with higher level of public goods provision, which might include social grants. In particular, governments in a more ethnically homogeneous place might be willing to offer more social grants due to the nepotism towards the dominant group in that place or less coordination cost among ethnic groups. If the receipt of social grants dis-incentivise working-age people to enter labour force, this could also explain the association between higher ethnic diversity and higher employment rate. However, this is not the case in our setting for two reasons. Firstly, provision of social grants is mainly designed at the national level, which does not vary across magisterial districts. Secondly, we include province fixed effects to account for potential discrepancy of social grants at province level.

Labour demand: discrimination. Discrimination in the labour market is a potential reason why homogeneous places discourage employment, as employers deliberately prevent the minority groups from gaining job opportunities and therefore the demand for minority labours is declined (Goldberg, 1982). It has been proved that the disutility from discrimination against minority groups in the production network harms the productivity of co-workers (Hjort, 2014; Borjas and Bronars, 1989).

A more diverse place can reduce the discrimination against minority groups by encouraging higher level of tolerance and openness. As the chance of interacting and communicating with other ethnic groups increases in a more ethnically diverse place, discrimination in the labour market becomes less of an issue, either because employers have access to more information about the productivity and behaviours of ethnic minorities, or because they are more open to people from different backgrounds.

[42] We measure the range of tasks by counting the total number of different occupations observed in each district. Occupations are counted in 3-digit code level.

If this story is the case, we would expect that ethnic groups with smaller size benefit more from increased ethnic diversity than those with relatively larger size, which contradicts our empirical evidence.

Labour demand: diversity of taste. Another potential driving force of labour demand might be the diversity of taste. As people from different ethnic groups have diversified tastes for consumption goods, the variety of consumption increases when a place becomes more ethnically diversified. This induces the diversity of production as well, resulting in higher variety of labour inputs in the production process. When different labour inputs are complementary in the production function, this love for variety of labours increases the total demand for labour, therefore improving workers' chance in the labour market. However, if this is the case, we should see the positive effect of ethnic diversity among both large and small ethnic groups, which also contradicts the empirical findings. There is also related literature about how greater diversification of sectoral demands reduces unemployment (Neumann and Topel, 1991). However this works under the condition that workers are mobile enough, which is not likely to be a prevalent case in South Africa where many black people locate far away from economic centres and the transportation cost is very high to them.

7 Conclusion and Discussion

This paper provides empirical support for the positive role played by within-black ethnic diversity and blacks' labour market outcomes in post-Apartheid South Africa based on an instrumental variable approach. We also propose a theoretical model to explain how the need for inter-ethnic social interaction stimulates investment in social skills in more ethnically diverse places, making black workers better equipped for the labour market.

The finding reveals that ethnic identity, together with inter-ethnic relationship, is still a distinctive feature shaping people's social life and labour market in modern South African society. The distinction between ethnic groups does not fade away after years of integration, resulting from the Apartheid regime which reinforced ethnic identity. In addition, although the climate of hatred and mistrust generated by the Apartheid system had substantial repercussions on the social fabric, inter-ethnic connections still occur within the black population.

Our result is different from, yet can be reconciled with the association between ethnic diversity and inter-ethnic cleavages or the erosion of social cohesion. Firstly, most of those literature highlights the under provision of public goods and social capital in ethnically fragmented communities in developing countries (Alesina et al., 2016), or the conflict between different ethnic groups (Amodio and Chiovelli, ming). Our story takes a different angle by focusing on skill investment motivated by social interaction. This can just be another side of inter-personal relations which can co-exist with conflicts or coordination problems. Secondly, in our numerical simulation, we show that the level of investment in social skills can decrease with ethnic diversity (i.e. increase with the dispersion of group size) when per unit cost of investment c is large enough. This means, if the conflict level in a district is too high (i.e. cost of investment is too large), ethnic diversity can potentially decrease investment and the consequential economic growth. Thirdly, we have shown in our model that the initial condition in skill investment is important in shaping the ultimate equilibrium. If the society starts from the situation where no one actively participates in inter-ethnic communication, benefits from inter-ethnic connection will stay at the low level forever. Therefore, societies where ethnic diversity is negatively associated with socio-economic indicators might have worse initial conditions in inter-ethnic interaction.

We also find the heterogeneous effects of ethnic diversity on labour market outcomes for different sub-groups. In particular, labour market outcomes of the ethnic groups with larger size are more responsive to ethnic diversity. This indicates that our story is not likely to be the case where the minority assimilates to the majority by integrating into their culture and language, nor is it the story that diversity alleviates discrimination against minority groups (in both cases only the small group will respond to diversity level). Rather, in our story groups with both large and small sizes participate in social interaction and invest in social skills in response to ethnic diversity.

More importantly, what drives our story is not the number of groups but the distribution of group size. Different from many other papers, we do not impose any intrinsic difference in taste or preference among different ethnic groups. We show that people respond differently in places with low and high levels of ethnic diversity not because ethnically diverse districts bring about more groups which contribute to something unique in these diverse places, but because the relative size of their group results in different motivations to invest in social skills.

Could any interventions be designed to increase employment opportunities for the black South African? As is presented in the theoretical framework, a successful intervention must encourage more inter-ethnic connection which can motivate people to invest in more social skills. It can be an efficient policy as we show that the initial investment in social skills is important to the ultimate equilibrium. Therefore, an attempt at fostering inter-ethnic communication in a more diverse society will have long-lasting effects on overall skill investments. Policies which directly improve black people's social skills may also be effective in preparing them for better employment opportunities.

These interventions to improve people's labour market performance have far-reaching implications not only in different aspects of South African society but also in dealing with ethnic issues all over the world. On the one hand, reducing unemployment can have other important consequences on South African society. For example, it has been estimated that in contemporary South Africa a 10 percentage point reduction in unemployment lowers the Gini coefficient by 3 percent (Anand et al., 2016).

On the other hand, this paper can also shed light on dealing with inter-ethnic relations in other African countries or even developed countries. In recent decades, Western societies have also become considerably more ethnically diverse due to the net immigration flows and the growing presence of ethnic communities (Putnam, 2007), which gives rise to more social problems. For example, there is some negative evidence of ethnic diversity on the support for redistribution which in particular harms low-income earners (Dahlberg et al., 2012). Furthermore, current immigration policies in the US and the European refugee crisis also require urgent modification in policy interventions to improve inter-ethnic relationships and explore the positive impact of ethnic diversity on economic outcomes, to which our mechanism about inter-ethnic interactions can be generalised. Our identification strategy can also be generalised to studies on other types of diversity or migration. For example, replacing homelands with individuals' countries of origin, one can instrument the ethnic composition of immigrants in Europe or the U.S. with a measure of equidistance to multiple home countries (Alesina et al. (2015) implements an approach similar to this).

References

Acemoglu, D. and D. Autor (2011). *Skills, tasks and technologies: Implications for employment and earnings*, Volume 4. Elsevier Inc.

Akerlof, G. A. (1997). Social distance and social decisions. *Econometrica 65*(5), 1005--1027.

Alesina, A., R. Baqir, and W. Easterly (1997). Public goods and ethnic divisions. *The Quarterly Journal of Economics 114*(4), 1243--1284.

Alesina, A., A. Devleeschauwer, S. Kurlat, and R. Wacziarg (2003). Fractionalization. *Journal of Economic Growth 8*, 155--194.

Alesina, A., C. Gennaioli, and S. Lovo (2016). Public goods and ethnic diversity: evidence from deforestation in Indonesia.

Alesina, A., J. Harnoss, and H. Rapoport (2015). Birthplace diversity and economic prosperity.

Alesina, A. and E. La Ferrara (2000). Participation in heterogeneous communities. *The Quarterly Journal of Economics 115*(3), 847--904.

Alesina, A. and E. La Ferrara (2002). Who trusts others? *Journal of Public Economics 85*(2), 207--234.

Alesina, A. and E. La Ferrara (2005). Ethnic diversity and economic performance. *Journal of Economic Literature XLIII*(September), 762--800.

Alesina, A., S. Michalopoulos, and E. Papaioannou (2016). Ethnic inequality. *Journal of Political Economy 124*(2), 428--488.

Algan, Y., C. Hemet, and D. D. Laitin (2016). The social effects of ethnic diversity at the local level: a natural experiment with exogenous residential allocation. *Journal of Political Economy 124*(3), 696--732.

Amodio, F. and G. Chiovelli (forthcoming). Ethnicity and violence during democratic transitions: evidence from South Africa. *Journal of the European Economic Association*.

Anand, R., S. Kothari, and N. Kumar (2016). South Africa: labor market dynamics and inequality.

Andersson, R., J. M. Quigley, and M. Wilhelmsson (2005). Agglomeration and the spatial distribution of creativity. *Papers in Regional Science 84*(3), 445--464.

Banerjee, A., S. Galiani, J. Levinsohn, Z. McLaren, and I. Woolard (2008). Why has unemployment risen in the New South Africa? *Economics of Transition 16*(4), 715--740.

Barr, A. (1998). Enterprise performance and the functional diversity of social capital. Technical report, Centre for the Study of African Economies, University of Oxford.

Beine, M., F. Docquier, and M. Schiff (2013). International migration, transfer of norms and home country fertility. *Canadian Journal of Economics/Revue canadienne d'économique 46*(4), 1406--1430.

Bhorat, H. and M. Oosthuizen (2005). The post-Apartheid South African labour market.

Borjas, G. J. and S. G. Bronars (1989). Consumer discrimination and self-employment. *Journal of Political Economy 97*(3), 581--605.

Chay, K. and K. Munshi (2015). Black networks after emancipation: evidence from reconstruction and the Great Migration.

Christopher, A. J. (2001). *The atlas of changing South Africa*. London and New York: Routledge.

Dahlberg, M., K. Edmark, and H. Lundqvist (2012). Ethnic diversity and preferences for redistribution. *Journal of Political Economy 120*(1), 41--76.

Deming, D. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, 1593--1640.

Desmet, K., I. Ortuño-Ortín, and R. Wacziarg (2012). The political economy of linguistic cleavages. *Journal of Development Economics 97*(2), 322--338.

Desmet, K., I. Ortuno-Ortin, and R. Wacziarg (2017). Culture, ethnicity, and diversity. *American Economic Review 107*(9), 2479--2513.

Dumont, J.-C., G. Spielvogel, and S. Widmaier (2010). International migrants in developed, emerging and developing countries. *OECD Social, Employment and Migration Working Paper No.114*.

Dustmann, C., U. Schoenberg, and J. Stuhler (2017). Labor Supply Shocks, Native Wages, and the Adjustment of Local Employment. *The Quarterly Journal of Economics 132*(1), 435--483.

Easterly, W. and R. Levine (1997). Africa's growth tragedy: policies and ethnic divisions. *The Quarterly Journal of Economics 112*(4), 1203--1250.

Eraydin, A., T. Tasan-Kok, and J. Vranken (2010). Diversity matters: immigrant entrepreneurship and contribution of different forms of social integration in economic performance of cities. *European Planning Studies 18*(4), 521--543.

Fainstein, S. S. (2005). Cities and diversity: should we want it? can we plan for it? *Urban Affairs Review 41*(1), 3--19.

Glaeser, E. L., H. D. Kallal, J. A. Scheinkman, and A. Shleifer (1992). Growth in cities. *Journal of Political Economy 100*(6), 1126--1152.

Glaeser, E. L., J. Scheinkman, and A. Shleifer (1995). Economic growth in a cross-section of cities. *Journal of Monetary Economics 36*(1), 117--143.

Goldberg, M. S. (1982). Discrimination, nepotism, and long-run wage differentials. *The Quarterly Journal of Economics*, 307--319.

Gomes, J., K. Desmet, and O.-O. Ignacio (2016). The Geography of Linguistic Diversity and the Provision of Public Goods.

Gradin, C. (2014). Poverty and ethnicity among black South Africans.

Guy, J. and M. Thabane (1988). Technology, ethnicity and ideology: basotho miners and shaft-sinking on the South African gold mines. *Journal of Southern African Studies 14*(2), 257--278.

Heintz, J. and D. Posel (2007). Revisiting informal employment and segmentation in the South African labor market.

Hjort, J. (2014). Ethnic divisions and production in firms. *The Quarterly Journal of Economics*, 1899--1946.

Iyer, G. R. and J. M. Shapiro (1999). Ethnic entrepreneurial and marketing systems: Implications for the global economy. *Journal of International Marketing*, 83--110.

Kingdon, G. G. and J. Knight (2004). Unemployment in South Africa: the nature of the beast. *World Development 32*(3), 391--408.

Lazear, E. P. (1999a). Culture and language. *Journal of Political Economy 107*(6), 95--124.

Lazear, E. P. (1999b). Globalisation and the market for team-mates. *Economic Journal 109*, 15--40.

Leibbrandt, M., I. Woolard, H. McEwen, and C. Koep (2009). Employment and inequality outcomes in South Africa.

Lewis, M. P., G. F. Simons, and C. D. Fennig (2009). *Ethnologue: languages of the world*, Volume 16. SIL international Dallas, TX.

Magruder, J. R. (2010). Intergenerational networks, unemployment, and persistent inequality in South Africa. *American Economic Journal: Applied Economics 2*(1), 62--85.

Mariotti, M. (2012). Labour markets during apartheid in south africa. *The Economic History Review 65*(3), 1100--1122.

Mayda, A. M. (2010). International migration: a panel data analysis of the determinants of bilateral flows. *Journal of Population Economics 23*(4), 1249--1274.

Michalopoulos, S. (2012). The origins of ethnolinguistic diversity. *American Economic Review 102*(4), 1508--1539.

Michalopoulos, S. and E. Papaioannou (2013). Pre-colonial ethnic institutions and contemporary african development. *Econometrica 81*(1), 113--152.

Miguel, E. and M. K. Gugerty (2005). Ethnic diversity, social sanctions, and public goods in Kenya. *Journal of Public Economics 89*, 2325--2368.

Montalvo, J. G. and M. Reynal-Querol (2005). Ethnic polarization, potential conflict, and civil wars. *The American Economic Review 95*(3), 796--816.

Morgan, J. and F. Vardy (2009). Diversity in the workplace. *American Economic Review 99*(1), 472--485.

Munshi, K. (2003). Networks in the modern economy: Mexican migrants in the U.S. labor market. *The Quarterly Journal of Economics 118*(2), 549--599.

Munshi, K. (2011). Strength in numbers: networks as a solution to occupational traps. *Review of Economic Studies 78*(3), 1069--1101.

Mwakikagile, G. (2010). *South Africa as a multi-ethnic society*. Dar es Salaam, Tanzania: Continental Press.

Neumann, G. and R. Topel (1991). Employment risk, diversification, and unemployment. *The Quarterly Journal of Economics*, 1341--1365.

Niebuhr, A. (2010). Migration and innovation: Does cultural diversity matter for regional r&d activity? *Papers in Regional Science 89*(3), 563--585.

Nikolova, E., D. Simroth, et al. (2013). Does cultural diversity help or hinder entrepreneurs? evidence from eastern europe and central asia. Technical report, European Bank for Reconstruction and Development.

Nunn, N. and D. Puga (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics 94*(1), 20--36.

Ortega, F. and G. Peri (2014). Openness and income: The roles of trade and migration. *Journal of International Economics 92*(2), 231--251.

Ottaviano, G. I. and G. Peri (2006). The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography 6*(1), 9--44.

Pasquier-Doumer, L. (2012). Intergenerational transmission of self-employed status in the informal sector: a constrained choice or better income prospects? evidence from seven west african countries. *Journal of African Economies 22*(1), 73--111.

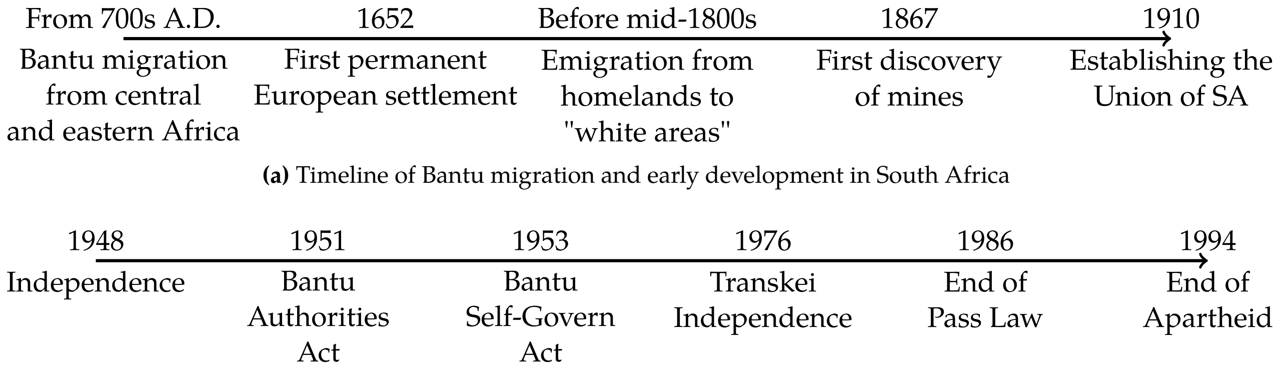
Posel, D. (2001). What's in a name? racial categorisations under apartheid and their afterlife. *Transformation 47*, 50--74.

Putnam, R. D. (2007). E pluribus unum: diversity and community in the twenty-first century the 2006 johan skytte prize lecture. *Scandinavian Political Studies 30*(2), 137--174.

Santos Silva, J. and S. Tenreyro (2006). The log of gravity. *The Review of Economics and statistics 88*(4), 641--658.

Sobel, R. S., N. Dutta, and S. Roy (2010). Does cultural diversity increase the rate of entrepreneurship? *The Review of Austrian Economics 23*(3), 269--286.

Sørensen, J. B. (2007). Closure and exposure: Mechanisms in the intergenerational transmission of self-employment. In *The Sociology of Entrepreneurship*, pp. 83--124. Emerald Group Publishing Limited.

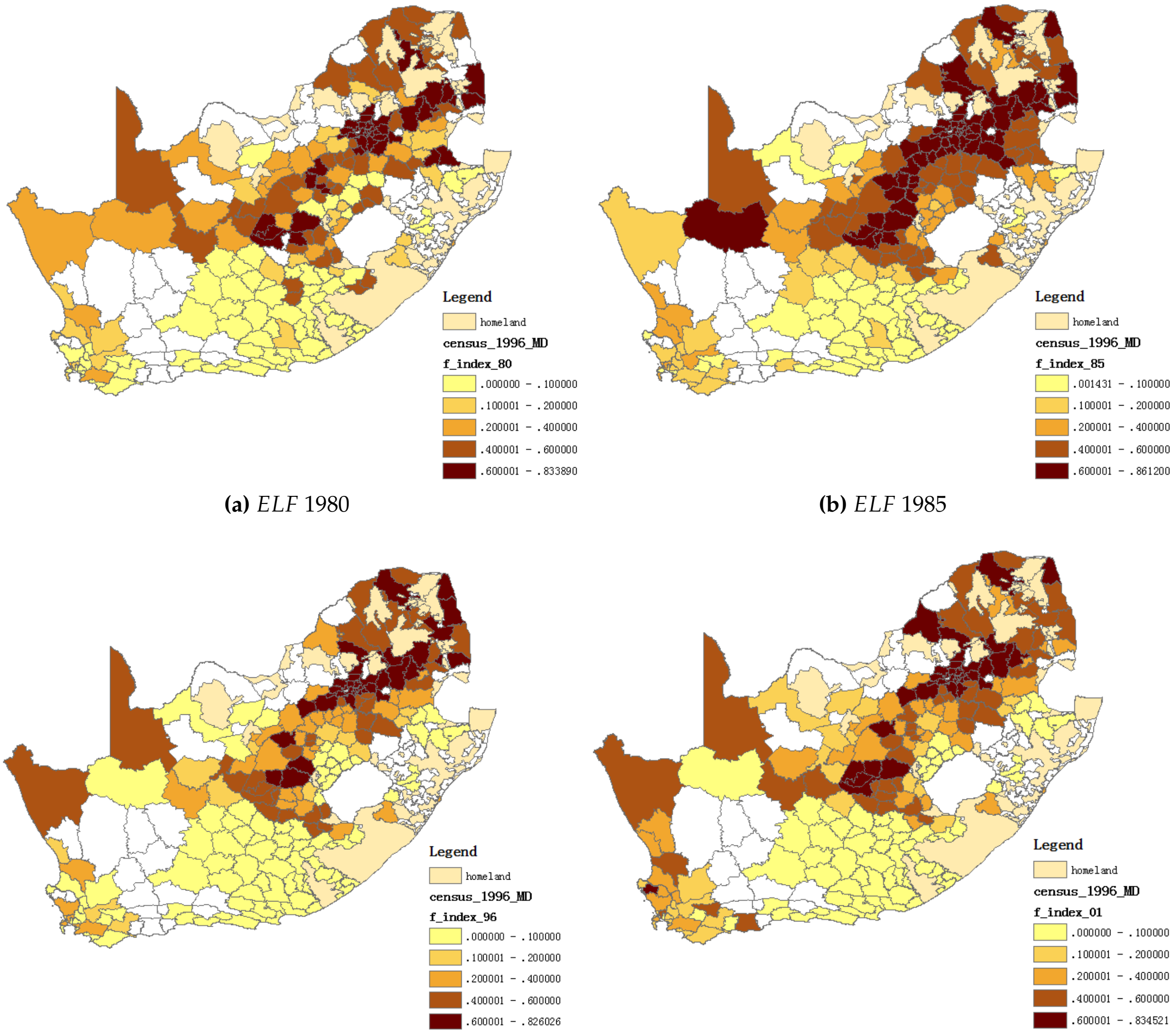


(a) Timeline of Bantu migration and early development in South Africa

(b) Timeline of modern South Africa starting from Apartheid

Notes: The figures presents the timeline of important nodes in South African history: Bantu migration from central and eastern Africa, emigration of ethnic groups from original homelands, the White colonisation, the discovery of mines and Apartheid regime. Sources of narratives: Mwakikagile (2010) and Gradin (2014).

Figure 1. Timeline of Bantu migration, historical development and Apartheid regime in South Africa



(a) *ELF* 1980

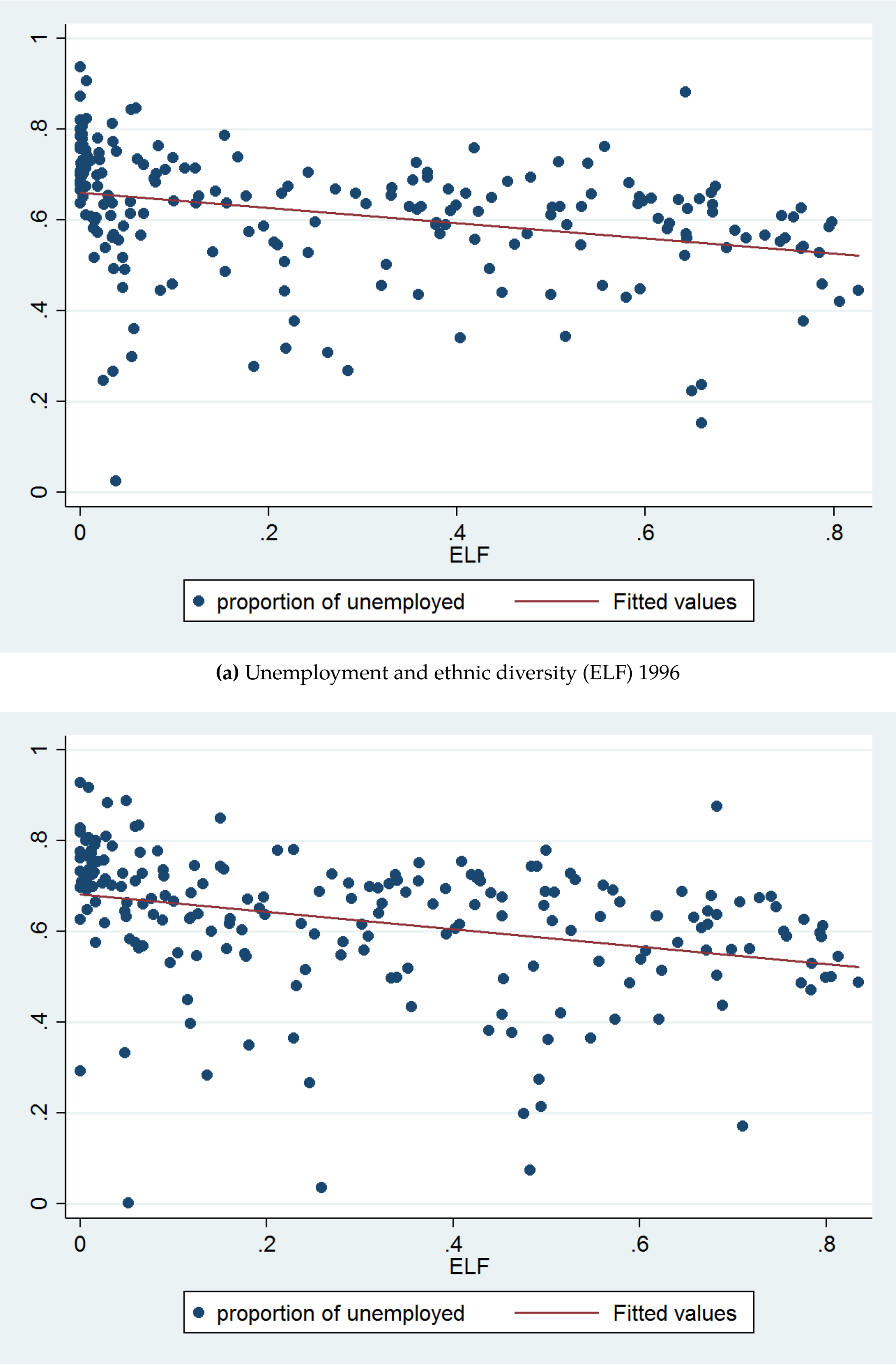
(b) *ELF* 1985

(c) *ELF* 1996

(d) *ELF* 2001

Notes: The figure presents the geographical pattern of ethnic diversity across South African districts in 1980, 1985, 1996 and 2001 for the "white areas". Within-black ethnic diversity is measured with Fractionalisation Index analogue to Herfindahl Index. The results are calculated by the authors based on the corresponding census data.

Figure 2. Distribution of ethnic fractionalization index: 1980. 1985, 1996, 2001

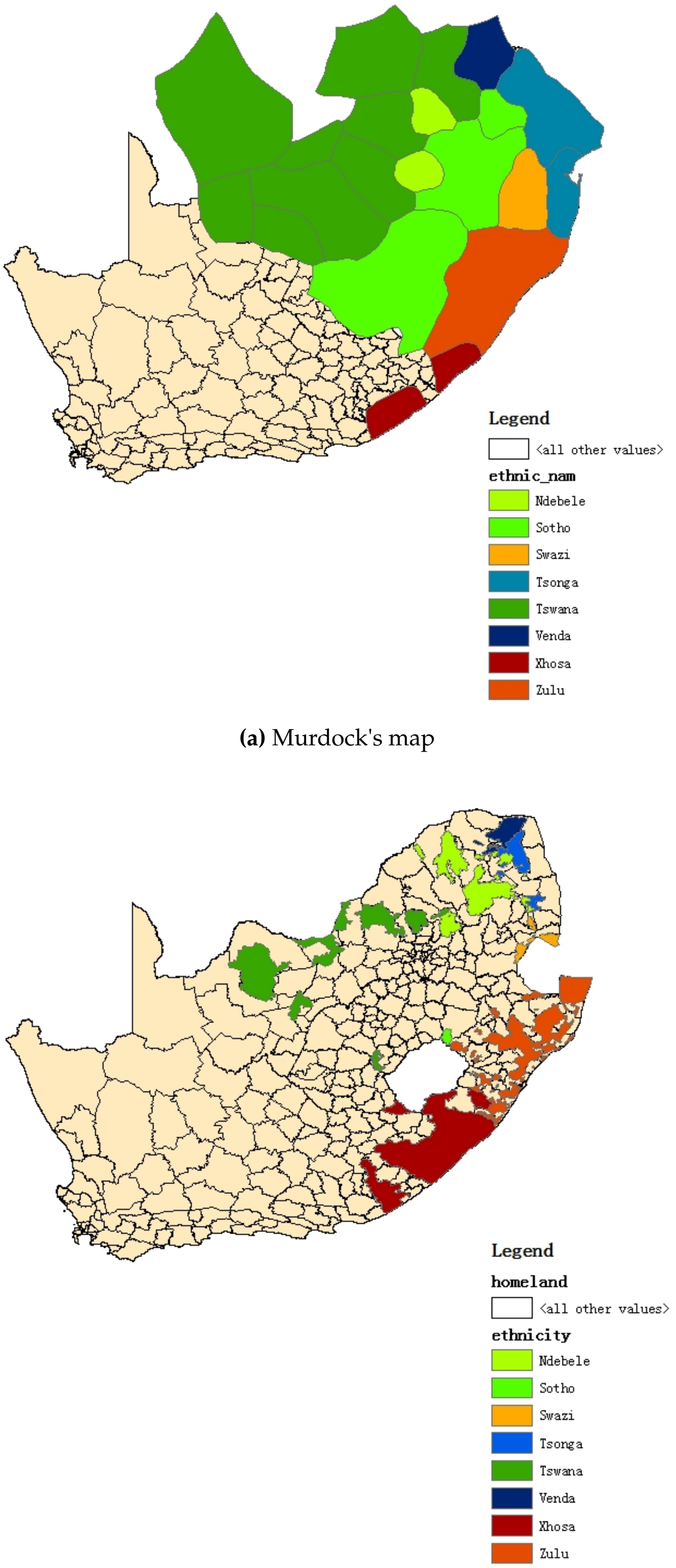


(a) Unemployment and ethnic diversity (ELF) 1996

(b) Unemployment and ethnic diversity (ELF) 2001

Notes: The figures present the results on the correlation between ethnic diversity and unemployment rate. Both are measured at the magisterial district level (therefore unemployment rate is calculated as the proportion of unemployed people over the whole working-age black population in a district). The results are calculated by the authors based on 1996 and 2001 census data.

Figure 3. The relationship between ethnic diversity and unemployment in South Africa in 1996 and 2001

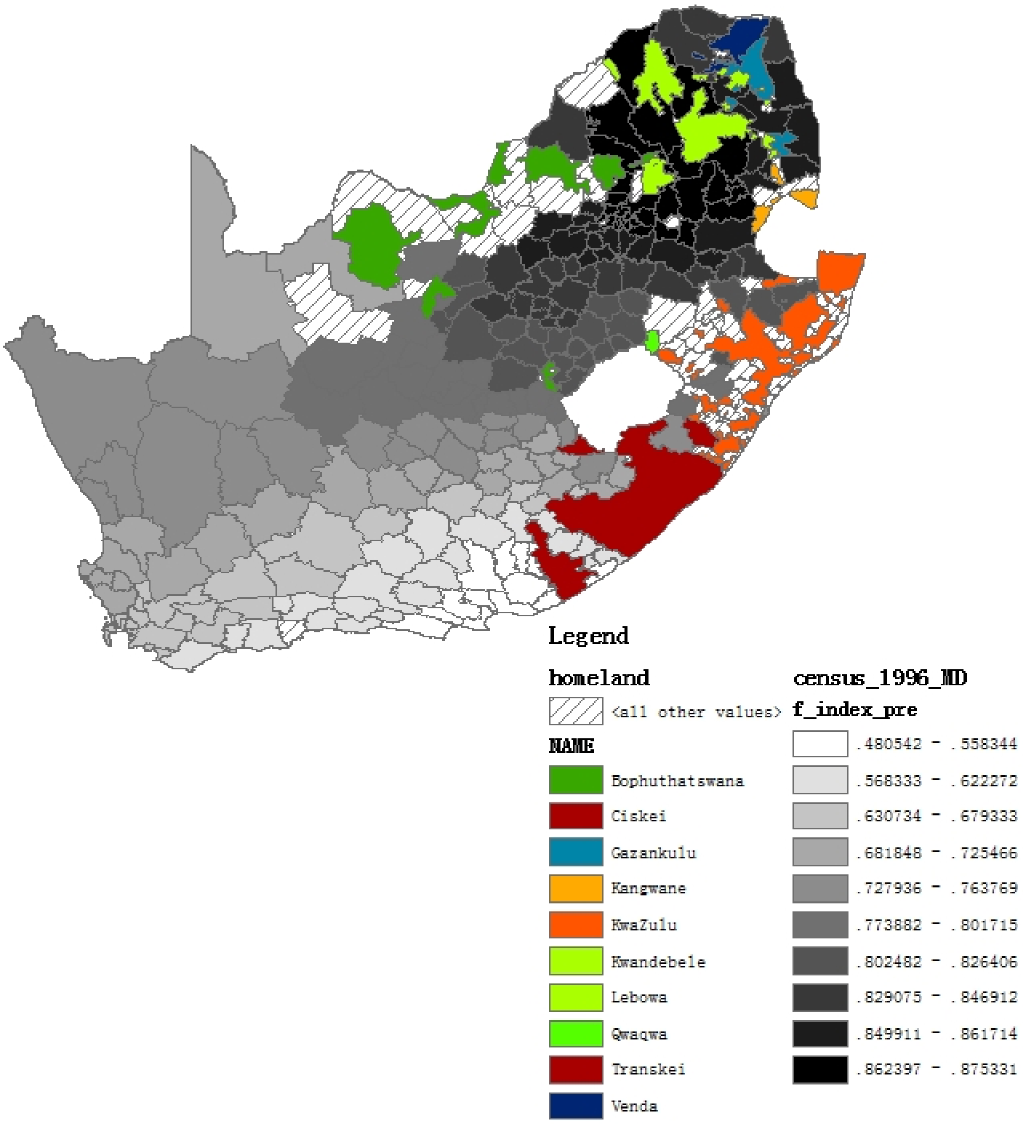


(a) Murdock's map

(b) Bantustan

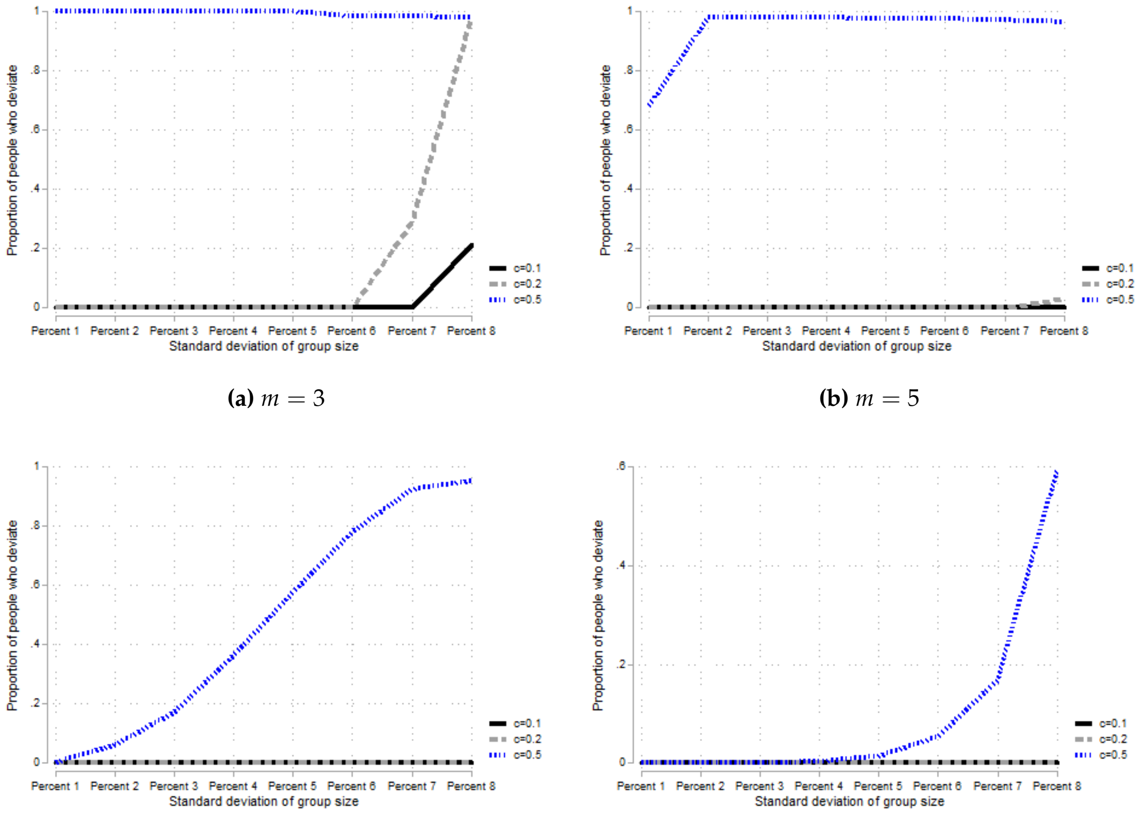
Notes: The figures compare the distribution of ethnic groups in South Africa in Murdock map and the location of Bantustans as proxies for ethnic homelands. Murdock map comes from George Murdock's 1959 work which illustrates the dominant ethnic group in each geographical unit, which is highly consistent with the Bantustans for these ethnic groups assigned by the Apartheid government. This confirms that the location of these Bantustans can well reflect the spatial distribution of original homelands for those ethnic groups.

Figure 4. Comparison between the historical settlements of the black ethnic groups and Bantustans



Notes: The figure shows the spatial distribution of our instrumental variable for ethnic diversity - the predicted ethnic fractionalisation index. Following the idea that districts more (less) equidistant to multiple homelands are more (less) diverse, we first calculate the stock of each ethnic group in each district based on the distance between the district and the corresponding homeland with a gravity model. The instrumental variable is a predicted fractionalisation index calculated based on the predicted stock of black migrants.

Figure 5. Distribution of predicted ethnic fractionalization index



(a) $m = 3$

(b) $m = 5$

(c) $m = 7$

(d) $m = 9$

Notes: The figures show the results on the numerical simulation of the proportion of people who deviate from investing in social skills in response to the dispersion of group size in the district. Each group has population share s_i and there are m groups in total. The dispersion of group size is measured by $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2$. In each graph we hold the number of groups m constant. We also consider different per unit cost of investment c.

Figure 6. Numerical simulation results on how the level of investments in social skills responds to dispersion of group size

Table 1.1. Summary statistics of demographics and employment among black ethnic groups in 1996

	Population size	Share of the	Self employed	Wage employee	Unemployed	Inactive	Unemployed	ELF
		black population					+inactive	
Xhosa	2229452	0.252	0.027	0.312	0.273	0.388	0.661	0.230
			[0.162]	[0.463]	[0.445]	[0.487]	[0.474]	[0.301]
Zulu	2073036	0.234	0.037	0.349	0.264	0.350	0.614	0.521
			[0.189]	[0.477]	[0.441]	[0.477]	[0.487]	[0.273]
South Sotho	2009582	0.227	0.026	0.349	0.242	0.383	0.625	0.467
			[0.159]	[0.477]	[0.428]	[0.486]	[0.484]	[0.23]
Tswana	1039138	0.117	0.026	0.384	0.224	0.367	0.590	0.525
			[0.158]	[0.486]	[0.417]	[0.482]	[0.492]	[0.243]
North Sotho	770110.7	0.087	0.038	0.412	0.236	0.314	0.549	0.700
			[0.192]	[0.492]	[0.424]	[0.464]	[0.498]	[0.116]
Tsonga	295688.6	0.033	0.075	0.434	0.247	0.244	0.491	0.715
			[0.263]	[0.496]	[0.431]	[0.430]	[0.500]	[0.109]
Ndebele	185065.8	0.021	0.040	0.358	0.227	0.375	0.602	0.689
			[0.195]	[0.480]	[0.419]	[0.484]	[0.490]	[0.111]
Swazi	181467.1	0.020	0.041	0.405	0.224	0.330	0.554	0.604
			[0.198]	[0.491]	[0.417]	[0.470]	[0.497]	[0.166]
Venda	80189.34	0.009	0.048	0.497	0.202	0.252	0.455	0.714
			[0.214]	[0.500]	[0.402]	[0.434]	[0.498]	[0.113]
Overall	8863729.5	1.000	0.032	0.355	0.251	0.361	0.613	0.274
			[0.177]	[0.479]	[0.434]	[0.48]	[0.487]	[0.266]

Note: The number and proportion of each ethnic group in the whole black population are calculated in the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Population size is not always an integer because it is weighted by each person's weight in the census data. Employment outcomes are calculated from individual-level 1996 census data among the working-age blacks. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group. Other labour market outcomes are calculated in similar ways. The mean degree of ethnic diversity index is calculated at the district level. All other statistics are calculated at the individual level.

Table 1.2. Summary statistics of demographics and employment among black ethnic groups in 2001

	Population size	Share of the black population	Self employed	Wage employee	Unemployed+inactive	ELF
Xhosa	3105625	0.249	0.017	0.299	0.684	0.251
			[0.130]	[0.458]	[0.465]	[0.298]
Zulu	2798132	0.224	0.025	0.331	0.643	0.558
			[0.156]	[0.471]	[0.479]	[0.264]
South Sotho	2531013	0.203	0.020	0.324	0.657	0.500
			[0.139]	[0.468]	[0.475]	[0.221]
Tswana	1373413	0.110	0.018	0.373	0.610	0.578
			[0.132]	[0.484]	[0.488]	[0.225]
North Sotho	1341608	0.107	0.027	0.396	0.577	0.689
			[0.163]	[0.490]	[0.494]	[0.148]
Tsonga	552403.3	0.044	0.048	0.421	0.531	0.714
			[0.214]	[0.494]	[0.50]	[0.128]
Ndebele	292188.3	0.023	0.029	0.370	0.601	0.708
			[0.168]	[0.483]	[0.490]	[0.131]
Swazi	324071.7	0.026	0.028	0.376	0.597	0.579
			[0.164]	[0.484]	[0.491]	[0.189]
Venda	172927.4	0.014	0.034	0.457	0.509	0.724
			[0.183]	[0.498]	[0.500]	[0.119]
Overall	12491382	1.000	0.023	0.341	0.636	0.302
			[0.149]	[0.474]	[0.481]	[0.259]

Note: The number and proportion of each ethnic group in the whole black population are calculated in the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Population size is not always an integer because it is weighted by each person's weight in the census data. Employment outcomes are calculated from individual-level 2001 census data among the working-age blacks."Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group. Other labour market outcomes are calculated in similar ways. The 2001 census data does not distinguish unemployed and economically inactive people. The mean degree of ethnic diversity index is calculated at the district level. All other statistics are calculated at the individual level.

Table 2.1. Summary statistics of ethnic fragmentation and labour market outcomes in 1996

		High ELF			Low ELF		
	Mean	S.d	Obs	Mean	S.d.	Obs	ttest
ELF	0.507	0.134	102	0.044	0.071	103	***
Self employment	0.028	0.044	102	0.021	0.044	103	***
Wage employee	0.400	0.110	102	0.320	0.120	103	***
Unemployed	0.570	0.110	102	0.658	0.118	103	***
Agriculture	0.466	0.138	102	0.454	0.134	103	
Manufacture	0.115	0.105	102	0.090	0.105	103	*
Service	0.419	0.114	102	0.455	0.130	103	**
Manager	0.014	0.032	102	0.012	0.045	103	
Profession	0.070	0.063	102	0.082	0.081	103	*
Clerk	0.032	0.056	102	0.020	0.049	103	***
Serve	0.073	0.059	102	0.063	0.069	103	*
Craft	0.107	0.092	102	0.125	0.108	103	
Skilld agriculture	0.121	0.079	102	0.107	0.079	103	*
Operator	0.088	0.071	102	0.062	0.063	103	***
Unskill	0.495	0.118	102	0.529	0.111	103	**

Note: This table compares labour market outcomes in districts with relatively high (i.e. above the median value) and low levels of ethnic diversity. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Employment outcomes are calculated from individual-level 1996 census data among the working-age population. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group."Wage employee" and "unemployed" are calculated in similar ways. We only focus on people who are employed when comparing the allocation of workers across industrial sectors and occupations.

Table 2.2. Summary statistics of ethnic fragmentation and labour market outcomes in 2001

		High ELF			Low ELF		
	Mean	S.d	Obs	Mean	S.d.	Obs	ttest
ELF	0.527	0.126	105	0.077	0.084	105	***
Self employment	0.022	0.055	105	0.019	0.045	105	
Wage employee	0.396	0.114	105	0.315	0.118	105	***
Unemployed	0.582	0.118	105	0.667	0.118	105	***
Agriculture	0.338	0.155	105	0.376	0.152	105	
Manufacture	0.183	0.130	105	0.096	0.089	105	***
Service	0.478	0.138	105	0.527	0.152	105	*
Manager	0.017	0.051	105	0.017	0.068	105	
Profession	0.082	0.075	105	0.080	0.076	105	
Clerk	0.056	0.055	105	0.054	0.084	105	
Serve	0.081	0.071	105	0.076	0.069	105	
Craft	0.059	0.071	105	0.084	0.093	105	***
Skilld agriculture	0.117	0.084	105	0.074	0.071	105	***
Operator	0.108	0.075	105	0.088	0.071	105	***
Unskill	0.480	0.120	105	0.527	0.121	105	**

Note: This table compares labour market outcomes in districts with relatively high (i.e. above the median value) and low levels of ethnic diversity. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. Employment outcomes are calculated from individual-level 2001 census data among the working-age black population. "Self-employed" refers to the proportion of self-employed people in each ethnic group over the whole working-age population of the corresponding ethnic group."Wage employee" and "unemployed" are calculated in similar ways. We only focus on people who are employed when comparing the allocation of workers across industrial sectors and occupations.

Table 3. Validity of the instrumental variable

	[1]	[2]
Dependent variable	1996	2001
Panel A: Job opportunities		
Distance to the closest economic centre	-274959.5	-245255.3
	(301774.502)	(260543.026)
Panel B: Economic activities of the white		
Share of white who are self employed contemporarily	0.230*	0.031
	(0.139)	(0.136)
Share of white who are wage employed contemporarily	0.095	0.185
	(0.170)	(0.158)
Proportion of white	0.335	0.149
	(0.221)	(0.140)
Panel C: Path dependence		
Share of white who are wage employed in 1980	-0.227	-0.245
	(0.217)	(0.219)
Proportion of white in 1980	-0.118	-0.707***
	(0.255)	(0.238)
Panel D: Contemporary migration		
Number of migrants	-44201.67	9845.154
	(37163.296)	(24771.553)
District controls	YES	YES
Individual controls (district average)	YES	YES
Province fixed effect	YES	YES
Obs	205	210

Note: This table conducts validity test of the instrumental variable based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. All regressions are at the district level. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

Table 4. First-stage regression results: individual level regressions

	[1]	[2]	[3]	[4]
	1996		2001	
	Age 15-64	Age 25-64	Age 15-64	Age 25-64
Predicted ELF	1.515***	1.488***	1.653***	1.623***
	(0.320)	(0.326)	(0.292)	(0.293)
Edu	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.000	0.000	0.001	0.001*
	(0.001)	(0.001)	(0.001)	(0.001)
Age	-0.000	-0.000*	-0.000	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
Married	0.003*	0.002	0.004***	0.003**
	(0.002)	(0.001)	(0.001)	(0.001)
Father alive	0.000	0.001	0.002**	0.002**
	(0.001)	(0.001)	(0.001)	(0.001)
Pop density	0.000***	0.000***	0.000**	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Urban	0.012	0.012	0.003	0.002
	(0.011)	(0.011)	(0.010)	(0.010)
River	0.084***	0.080***	0.062**	0.060**
	(0.029)	(0.029)	(0.027)	(0.028)
Density mine	0.434	0.337	0.781	0.697
	(0.709)	(0.698)	(0.717)	(0.695)
Prop black	-0.290***	-0.289***	-0.421***	-0.421***
	(0.056)	(0.056)	(0.077)	(0.077)
Distance closest	-0.000**	-0.000**	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)
Ruggedness	0.005	0.006	-0.005	-0.004
	(0.008)	(0.008)	(0.007)	(0.007)
Soil quality	0.051*	0.052*	0.054**	0.053*
	(0.029)	(0.029)	(0.027)	(0.027)
Per capita light	0.340	0.353	0.583	0.562
	(0.232)	(0.231)	(0.369)	(0.361)
Road	0.009	0.013	0.011	0.015
	(0.030)	(0.030)	(0.029)	(0.028)
Conflict	0.018*	0.017*	-0.004**	-0.004**
	(0.009)	(0.009)	(0.002)	(0.002)
Proportion manu	0.305**	0.313***	0.261***	0.272***
	(0.118)	(0.117)	(0.076)	(0.074)
Proportion service	0.377***	0.373***	0.198**	0.191**
	(0.129)	(0.129)	(0.086)	(0.084)
Ethnicity fixed effect	YES	YES	YES	YES
Province fixed effect	YES	YES	YES	YES
F-statistics of the instrument	22.36	20.87	32.04	30.60
R-squared	0.875	0.879	0.887	0.890
Observations	464,130	318,610	697,369	484,639

Note: This table reports the first-stage results of the instrumental variable based on 1996 and 2001 census data and report the F-statistics of the instrumental variable. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. All regressions are at the individual level. We report all the control variables, both district-level variables especially geographical features and individual-level controls for socio-economic status. We control for ethnicity and province fixed effects. *** p<0.01, ** p<0.05, * p<0.1.

Table 5. Ethnic diversity, unemployment and labour force participation: individual level regressions

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Unemployed		Inactive		Unemployed + inactive		Unemployed + inactive	
	1996		1996		1996		2001	
	Age 15-64	Age 25-64	Age 15-64	Age 25-64	Age 15-64	Age 25-64	Age 15-64	Age 25-64
Panel A: OLS estimates								
Ethnic fractionalisation ELF	-0.024	-0.031	-0.057**	-0.051**	-0.081**	-0.082**	-0.146***	-0.148***
	(0.018)	(0.027)	(0.023)	(0.022)	(0.032)	(0.039)	(0.037)	(0.042)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	0.033	0.075	0.153	0.094	0.195	0.123	0.175	0.107
Observations	464,130	318,610	464,130	318,610	464,130	318,610	697,368	484,639
Panel B: IV estimates								
Ethnic fractionalisation ELF	-0.142***	-0.126**	0.043	-0.036	-0.098	-0.163*	-0.170*	-0.206**
	(0.046)	(0.064)	(0.061)	(0.061)	(0.076)	(0.087)	(0.088)	(0.098)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
F statistics of the instrument	22.36	20.87	22.36	20.87	22.36	20.87	32.04	30.60
R-squared	0.032	0.074	0.153	0.094	0.195	0.123	0.175	0.106
Observations	464,130	318,610	464,130	318,610	464,130	318,610	697,369	484,639

Note: This table reports results about the effect of ethnic diversity on unemployment rate at individual-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We separate unemployed and economically inactive groups only for 1996 results as these two categories are combined in 2001 census. "Unemployed + inactive" is a dummy variable which equals 1 if one is unemployed or inactive and 0 if one is employed. *** p<0.01, ** p<0.05, * p<0.1.

Table 6. Ethnic diversity and employment status: individual level regresions

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Wage employment		Self/wage		Wage employment		Self/wage	
	1996		1996		2001		2001	
	Age 15-64	Age 25-64	Age 15-64	Age 25-64	Age 15-64	Age 25-64	Age 15-64	Age 25-64
Panel A: OLS estimates								
Ethnic fractionalisation ELF	0.086***	0.087**	-0.024	-0.020	0.144***	0.147***	0.013	0.012
	(0.033)	(0.041)	(0.017)	(0.017)	(0.037)	(0.043)	(0.013)	(0.013)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	0.194	0.126	0.011	0.01	0.173	0.106	0.008	0.008
Observations	449,200	305,099	180,535	162,333	681,529	470,552	253,809	228,519
Panel B: IV estimates								
Ethnic fractionalisation ELF	0.112	0.173*	-0.055	-0.030	0.176**	0.215**	-0.043	-0.041
	(0.077)	(0.091)	(0.043)	(0.040)	(0.087)	(0.098)	(0.037)	(0.034)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
F statistics of the instrument	22.46	21	20.24	20.31	32.33	30.88	27.93	28.61
R-squared	0.194	0.126	0.011	0.01	0.173	0.106	0.008	0.007
Observations	449,200	305,099	180,535	162,333	681,529	470,552	253,809	228,519

Note: This table reports results about the effect of ethnic diversity on employment and the allocation between self- and wage-employment at individual-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In column 1, 2, 5, 6 we drop self-employed people as they are a very small proportion of the whole working-age population. Column 3, 4, 7, 8 are based only on the employed black people. *** p<0.01, ** p<0.05, * p<0.1.

Table 7. Ethnic diversity, intensive margin and wage: individual level regressions

	[1]	[2]	[3]	[4]	[5]	[6]
	OLS	OLS	OLS	IV	IV	IV
	Log monthly income	Log hourly wage	Hour	Log monthly income	Log hourly wage	Hour
Panel A: Individual level, census data						
Ethnic fractionalisation ELF	0.326***	0.362***	-1.279	0.497***	0.414	2.482
	(0.071)	(0.090)	(1.286)	(0.190)	(0.268)	(3.809)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES
F statistics of the instrument				28.28	28.28	28.12
R-squared	0.345	0.314	0.053	0.345	0.314	0.052
Observations	228,256	228,256	232,533	228,256	228,256	232,533
Panel B: Individual level, LFS data						
Ethnic fractionalisation ELF	-0.0772	-0.0286	0.107	0.439	0.0361	23.23
	(0.253)	(0.246)	(2.943)	(0.996)	(0.870)	(16.51)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES
F statistics of the instrument				5.497	5.322	5.591
R-squared	0.482	0.474	0.054	0.480	0.474	0.018
Observations	3,478	3,615	3,660	3,478	3,615	3,660

Note: This table reports results about the effect of ethnic diversity on other labour market outcomes at individual-level regressions, including working hour, hourly wage and monthly earnings. We only report the result in 2001 as there is no information on hours of working in 1996 census. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We control for province fixed effects. Ethnic diversity is measured with fractionalisation index. All the columns only focus on employees. *** p<0.01, ** p<0.05, * p<0.1.

Table 8. Ethnic diversity and employment: district fixed effects models

	[1]	[2]	[3]	[4]
Dependent variable	unemploy + inactive	wage employ	self/wage	log monthly income
Ethnic fractionalisation ELF	-0.291***	0.341***	-0.133*	-0.382
	(0.0709)	(0.072)	(0.0696)	(0.365)
Individual controls (district average)	YES	YES	YES	YES
District controls	YES	YES	YES	YES
R-squared	0.493	0.488	0.244	0.730
Observations	410	410	410	410

Note: This table reports results about the effect of ethnic diversity on employment based on the district-level balanced panel. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables which vary over time and individual-level controls aggregated at district level and province fixed effects. Ethnic diversity is measured with fractionalisation index. The dependent variable in column 1 is the proportion of unemployed over the whole working-age black population. Column 2 is defined in a similar way but we exclude those who are self-employed. Column 3 has the dependent variable which is the ratio of the number of self-employed to that of employees at district level. Column 4 only focuses on back people who are employed. *** p<0.01, ** p<0.05, * p<0.1.

Table 9. Heterogeneous effects of ethnic diversity on wage-employment: individual level regressions

	[1]	[2]	[3]	[4]
	OLS	IV	OLS	IV
Dependent variable	1996	1996	2001	2001
Panel A: By ethnicity				
Large	0.111***	0.125	0.148***	0.182**
	(0.035)	(0.081)	(0.032)	(0.086)
Obs	320,901	320,901	459,108	459,108
Medium	0.018	-0.150	0.203***	-0.029
	(0.075)	(0.210)	(0.061)	(0.207)
Obs	91,373	91,373	149,632	149,632
Small	-0.016	-0.606*	0.035	0.210
	(0.104)	(0.326)	(0.083)	(0.415)
Obs	36,926	36,926	72,789	72,789
Panel B: By industrial sector				
Agriculture	0.068**	0.034	0.016	0.067
	(0.033)	(0.070)	(0.031)	(0.073)
Obs	165,605	165,605	180,227	180,227
Manufacturing	-0.024***	-0.008	-0.008	-0.012
	(0.009)	(0.019)	(0.010)	(0.020)
Obs	165,605	165,605	180,227	180,227
Service	-0.044	-0.026	-0.008	-0.055
	(0.027)	(0.061)	(0.028)	(0.073)
Obs	165,605	165,605	180,227	180,227
Panel C: By occupation				
Manager	0.004	0.016	0.006*	0.023***
	(0.004)	(0.010)	(0.003)	(0.008)
Obs	153,294	153,294	224,942	224,942
Profession	-0.016	0.107**	-0.011	0.092*
	(0.016)	(0.046)	(0.013)	(0.052)
Obs	153,294	153,294	224,942	224,942
Clerk	0.015**	-0.004	0.022**	0.033
	(0.006)	(0.014)	(0.010)	(0.024)
Obs	153,294	153,294	224,942	224,942
Serve	-0.022*	0.037	0.015	-0.044
	(0.011)	(0.033)	(0.016)	(0.038)
Obs	153,294	153,294	224,942	224,942
Craft	-0.015	-0.033	-0.043	0.024
	(0.029)	(0.067)	(0.027)	(0.048)
Obs	153,294	153,294	224,942	224,942
Skilled agriculture	0.002	-0.073	0.014	-0.059*

Continued on next page

Table 9 -- continued from previous page

	[1]	[2]	[3]	[4]
	OLS	IV	OLS	IV
Dependent variable	1996	1996	2001	2001
	(0.021)	(0.046)	(0.015)	(0.033)
Obs	153,294	153,294	224,942	224,942
Operator	-0.036**	-0.143***	-0.038**	-0.058
	(0.016)	(0.047)	(0.018)	(0.038)
Obs	153,294	153,294	224,942	224,942
Unskilled	0.068*	0.094	0.035	-0.010
	(0.035)	(0.072)	(0.035)	(0.064)
Obs	153,294	153,294	224,942	224,942
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES

Note: This table reports the main results about the heterogeneous effects of ethnic diversity on the probability of being an employee at individual-level regressions by sub-groups in both 1996 and 2001 census. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. All the columns in Panel B and Panel C only focus on employees to illustrate the allocation of employed workers across different industrial sectors and occupations. *** p<0.01, ** p<0.05, * p<0.1.

Table 10. Using total distance to all homelands as a proxy for black population size

	[1]	[2]	[3]	[4]
	1996		2001	
	Unemployed + active	Wage employee	Unemployed + active	Wage employee
Panel A: OLS estimates whole sample				
Ethnic fractionalisation ELF	-0.210***	0.212***	-0.218***	0.216***
	(0.045)	(0.045)	(0.045)	(0.045)
R-squared	0.188	0.186	0.172	0.170
Observations	464,130	449,200	697,369	681,529
Panel B: IV estimates whole sample				
Ethnic fractionalisation ELF	-0.158*	0.174**	-0.159**	0.161**
	(0.087)	(0.088)	(0.080)	(0.081)
F statistics of the instrument	35.45	35.24	39.74	39.71
R-squared	0.182	0.186	0.172	0.170
Observations	464,130	449,200	697,369	681,529
Panel C: IV estimates large group				
Ethnic fractionalisation ELF	-0.235***	0.244***	-0.222***	0.220***
	(0.074)	(0.074)	(0.058)	(0.059)
F statistics of the instrument	47.53	47.14	48.29	48.19
R-squared	0.182	0.181	0.165	0.162
Observations	330,792	320,901	468,704	459,108
Panel D: IV estimates medium group				
Ethnic fractionalisation ELF	0.474	-0.367	-0.388	0.440
	(1.018)	(0.989)	(0.412)	(0.393)
F statistics of the instrument	0.907	0.904	2.009	2.238
R-squared	0.187	0.188	0.179	0.176
Observations	94,256	91,373	153,041	149,632
Panel E: IV estimates small group				
Ethnic fractionalisation ELF	0.284	-0.259	0.219	-0.187
	(0.240)	(0.256)	(0.187)	(0.194)
F statistics of the instrument	19.48	18.90	15.20	15.40
R-squared	0.215	0.221	0.190	0.194
Observations	39,082	36,926	75,624	72,789
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES

Note: This table reports results about the effects of ethnic diversity on the probability of being unemployed and being a wage employee at individual-level regressions by sub-groups in both 1996 and 2001 census. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. Compared with the main analysis, we replace population density and the proportion of the black with total distance to all homelands. *** p<0.01, ** p<0.05, * p<0.1.

Table 11. Ethnic diversity and employment: separating native, migrants and immigrants

	[1]	[2]	[3]	[4]	[5]	[6]
	Unemployed + inactive			Wage employment		
	Native	Migrants	Immigrants	Native	Migrants	Immigrants
Panel A: IV estimates, 1996 census						
Ethnic fractionalisation ELF	-0.161**	0.142	-0.506	0.171**	-0.131	0.530*
	(0.079)	(0.133)	(0.321)	(0.081)	(0.134)	(0.316)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES
F statistics of the instrument	24.05	15.52	8.432	24.12	15.51	8.826
R-squared	0.191	0.193	0.299	0.188	0.196	0.330
Observations	305,458	128,215	4,657	296,864	122,956	4,283
Panel B: IV estimates, 2001 census						
Ethnic fractionalisation ELF	-0.153*	-0.247	-0.991*	0.158**	1.048*	0.252
	(0.080)	(0.211)	(0.592)	(0.080)	(0.624)	(0.209)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES
F statistics of the instrument	33.73	16.65	6.618	33.90	17.14	6.666
R-squared	0.171	0.196	0.289	0.168	0.198	0.312
Observations	568,260	119,696	20,390	556,296	116,089	19,250

Note: This table reports results about the effect of ethnic diversity on employment separately for native, migrants and immigrants at individual-level regressions based on 1996 and 2001 census data. "Native" is defined as people who were born in the district and never move out or within-district migrants. "Migrants" are cross-district migrants while "immigrants" are those who migrated from another country. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In columns 4-6 we drop self-employed people as they are a very small proportion of the whole working-age population. *** p<0.01, ** p<0.05, * p<0.1.

Table 12. Ethnic diversity and the emigration of the white

	[1]	[2]	[3]	[4]	[5]	[6]
		1996			2001	
	Number of white in 1996	Number of white in 1985	Diffrence: 96 - 85	Number of white in 2001	Number of white in 1985	Diffrence: 01 - 85
Panel A: OLS estimates						
Ethnic fractionalisation ELF	12758.975	765.232	11993.742	-9966.072	4529.285	-14495.357
	(25151.926)	(31380.028)	(13971.953)	(29215.642)	(25530.141)	(15013.818)
R-squared	0.766	0.781	0.452	0.767	0.895	0.782
Observations	205	205	205	210	210	210
Panel B: IV estimates						
Ethnic fractionalisation ELF	145935.526	226342.130	-80406.604	33390.776	15067.044	18323.731
	(185457.367)	(204618.284)	(59332.187)	(106236.562)	(84187.211)	(33223.976)
F statistics of instruments	10.19	10.19	10.19	29.85	29.85	29.85
R-squared	0.735	0.720	0.349	0.765	0.894	0.776
Observations	205	205	205	210	210	210
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES

Note: This table looks at whether ethnic diversity is correlated with the number of white population in 1996 and 2001 and the emigration of the white from the district after the end of Apartheid at district-level regressions. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

Table 13.1. Decomposing ethnic diversity into number of groups and dispersion of population share

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
	Wage employee				Common language (only 1996)			
	Whole sample	Large	Medium	Small	Whole sample	Large	Medium	Small
Panel A: 1996 census								
Dispersion of size	-0.087***	-0.110***	-0.015	0.014	-0.035*	-0.028	-0.054	0.015
	(0.033)	(0.034)	(0.073)	(0.102)	(0.018)	(0.020)	(0.033)	(0.032)
1/No. of groups	-0.170***	-0.194***	-0.207	0.405	0.016	0.020	0.127	-0.097
	(0.044)	(0.045)	(0.311)	(0.555)	(0.028)	(0.031)	(0.366)	(0.177)
R-squared	0.194	0.187	0.201	0.230	0.068	0.068	0.072	0.061
Observations	449,200	320,901	91,373	36,926	654,116	469,737	131,601	52,778
Panel B: 2001 census								
Dispersion of size	-0.148***	-0.152***	-0.207***	-0.033				
	(0.037)	(0.031)	(0.057)	(0.083)				
1/No. of groups	-0.284***	-0.266***	0.419	-0.555				
	(0.065)	(0.057)	(0.777)	(1.126)				
R-squared	0.173	0.164	0.177	0.200				
Observations	681,529	459,108	149,632	72,789				
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES

Note: This table reports results based on the decomposition of ethnic diversity index into items relating to number of ethnic groups and group share, and how these two items are associated with employment rate at individual-level OLS regressions with 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We look at both the whole sample and sub-samples split by sample size. *** p<0.01, ** p<0.05, * p<0.1.

Table 13.2. Verifying that the instrument variable captures the dispersion of group size

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
	Wage employee 1996			Wage employee 2001			Common language 1996		
	Large	Medium	Small	Large	Medium	Small	Large	Medium	Small
Panel A: Instrument dispersion of size									
Dispersion of size	-0.166**	0.166	0.570*	-0.213**	0.030	-0.202	-0.136*	0.068	-0.012
	(0.077)	(0.197)	(0.319)	(0.084)	(0.203)	(0.429)	(0.073)	(0.168)	(0.097)
1/No. of groups	-0.249***	-0.073	0.864	-0.335***	0.531	-0.603	-0.086	0.224	-0.116
	(0.080)	(0.374)	(0.604)	(0.107)	(0.721)	(1.059)	(0.079)	(0.434)	(0.190)
F statistics of the instrument	17.98	7.662	9.421	23.68	12.11	4.009	18.31	8.424	10.03
R-squared	0.187	0.200	0.223	0.164	0.175	0.199	0.066	0.070	0.061
Observations	320,901	91,373	36,926	459,108	149,632	72,789	469,737	131,601	52,778
Panel B: Instrument number of groups									
Dispersion of size	-0.270	0.057	0.113	-0.505	0.065	-0.044	-0.313	-0.006	0.011
	(0.248)	(0.156)	(0.135)	(0.751)	(1.802)	(0.101)	(0.322)	(0.060)	(0.036)
1/No. of groups	-0.566	1.514	4.145	-1.781	56.536	-8.462	-0.619	1.256	-0.289
	(0.573)	(2.867)	(2.830)	(3.415)	(377.818)	(26.742)	(0.716)	(1.169)	(0.630)
F statistics of the instrument	0.897	1.128	3.477	0.204	0.0212	0.331	1.112	1.393	3.392
R-squared	0.184	0.196	0.220	0.142	-0.825	0.193	0.038	0.066	0.061
Observations	320,901	91,373	36,926	459,108	149,632	72,789	469,737	131,601	52,778
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES	YES

Note: This table reports results based on the decomposition of ethnic diversity index into items relating to number of ethnic groups and group share, and how these two items are associated with employment rate at individual-level IV regressions with 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We look at the sub-samples split by group size. *** p<0.01, ** p<0.05, * p<0.1.

Table 14.1. Ethnic diversity and skill acquisition: second language

	[1] OLS	[2] IV
Panel A: Whole sample		
Overall	0.036**	0.109*
	(0.018)	(0.065)
F statistics of the instrument		23.07
Obs	654,116	654,116
Panel B: By ethnicity		
Large	0.028	0.161**
	(0.020)	(0.078)
F statistics of the instrument		17.30
Obs	469,737	469,737
Medium	0.052	-0.078
	(0.034)	(0.188)
F statistics of the instrument		9.052
Obs	131,601	131,601
Small	-0.014	0.024
	(0.031)	(0.093)
F statistics of the instrument		10.76
Obs	52,778	52,778
Individual controls	YES	YES
District controls	YES	YES
Province FE	YES	YES

Note: This table reports results about the effect of ethnic diversity on the acquisition of social skills (proficiency of second language as a proxy) at individual-level regressions based on 1996 census data (as there is no information on the second language in 2001 census). The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. We look at both the whole sample and sub-samples split by group size. *** p<0.01, ** p<0.05, * p<0.1.

Table 14.2. Ethnic diversity and skill acquisition across age groups: second language

	[1] <15	[2] 15-64	[3] >=65
Panel A: Large group			
ELF	0.121**	0.174**	0.236**
	(0.057)	(0.088)	(0.110)
F statistics of the instrument	17.34	17.22	14.89
Obs	108,939	342,890	17,908
Panel B: Medium group			
ELF	-0.079	-0.064	-0.458
	(0.115)	(0.209)	(0.350)
F statistics of the instrument	11.10	8.097	12.96
Obs	28,848	97,754	4,999
Panel C: Small group			
ELF	0.093	-0.005	-0.033
	(0.073)	(0.106)	(0.122)
F statistics of the instrument	11.96	9.896	15.56
Obs	10,238	40,690	1,850
Individual controls	YES	YES	YES
District controls	YES	YES	YES
Province FE	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on the acquisition of social skills (proficiency of second language as a proxy) at individual-level regressions based on 1996 census data (as there is no information on the second language in 2001 census). The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In particular, we split the sample by age groups. *** p<0.01, ** p<0.05, * p<0.1.

Table 15. Skill acquisition and labour market outcomes: second language

	[1] Overall	[2] Large	[3] Medium	[4] Small
Panel A: Unemployed as dependent variable, conditional on diversity				
Second official	-0.133***	-0.123***	-0.144***	-0.176***
	(0.013)	(0.010)	(0.022)	(0.036)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
R-squared	0.201	0.193	0.212	0.229
Obs	461,942	329,416	93,673	38,853
Panel B: Wage employ as dependent variable, conditional on diversity				
Second official	0.132***	0.121***	0.143***	0.179***
	(0.013)	(0.011)	(0.023)	(0.036)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
R-squared	0.199	0.191	0.209	0.236
Obs	447,103	319,580	90,817	36,706

Note: This table reports results about the relationship between social skill acquisition (proficiency of second language as a proxy) and employment at individual-level regressions based on 1996 census data. We control for ethnic diversity and investigate whether this language skill is positively correlated with employment chances. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In Panel A we keep the whole working-age black sample while in Panel B we drop self-employed people as they are a very small proportion of the whole working-age population. *** p<0.01, ** p<0.05, * p<0.1.

A Appendix. Bantu migration and the formation of ethnic diversity from historical narratives

Below we provide a summary of the history of the Bantu migration from central and eastern Africa and the settlement of these groups in South Africa for each ethnic groups in details. Narrative evidence is summarised from Mwakikagile (2010) and Gradin (2014).

Ethnicity	Time of migration into SA	Homelands	Time of moving into white areas	Bantustan
Xhosa	Before 1400s	Today's Eastern Cape	After conflicts with the native Khoisan	Ciskei and Transkei
Zulu	16th century	Eastern part, today's Kwazulu-Natal	Early 1800s	KwaZulu
Swazi	15th and 16th centuries	Southern part of Tongaland in what is now Mozambique	17th and 18th centuries into the Pongola River	KaNgwane
Ndebele	Before 1835	Today's Northern Province, Mpumalanga and Gauteng	By 1835 towards Swaziland and Northern Transvaal	KwaNdebele, Lebowa
North Sotho	1500s	Today's Limpopo and Northwest	After the war with Boers and Ndebele	Qwawa
South Sotho	1500s	Today's Limpopo and Northwest	After the war with Boers and Ndebele	Qwawa
Tswana	1500s	Today's Limpopo and Northwest	After the war with Boers and Ndebele	Bophuthatswana
Tsonga	Before the early 1500s	Close to today's Mozambique	After conflicts with Zulu	Gazankulu
Venda	Before 800s A.D.	A mountainous area in the northern part close to Limpopo River	800s A.D. to Matopo Hills	Venda

B Appendix. Data source and construction of district-level variables

In this section we present data sources and the construction of our district-level control variables in detail. Emphasis has been given on those geographical measures.

Variable	Data source	Construction of variable
Panel A: From census		
Area of the district	Census 1996 and 2001 district-level shape file.	Calculated from the shape file directly in ArcGIS.
Population density	Census 1996 and 2001.	Calcualte the total number of black in each district in census data and divide it by area.
Proportion of the black	Census 1996 and 2001.	Calculate the number of black over the whole population.
Proportion of manufacturing	Census 1996 and 2001.	Calculate the number of people working in manufacturing sector over the whole employed black people.
Proportion of service	Census 1996 and 2001.	Calculate the number of people working in service sector over the whole employed black people.
Urban/rural	Census 1996 and 2001.	Information on whether one lives in an urban or rural settlement is explicit in census data.

Variable	Data source	Construction of variable
Panel B: Sources on geography		
Overlap of district and homeland	A map (shape file) of homeland provided by Tim Brophy and Adrian Frith.	Intersecting the boundary of districts with that of homelands and seeing the overlap in ArcGIS.
River	Census 2001 river shape file.	Overlapping shape file of districts and river and directly calculating in ArcGIS.
Road	Census 2001 major road shape file.	Overlapping shape file of districts and road and directly calculating in ArcGIS.
Ruggedness	From Nunn and Puga (2012). We also tried the measure of slope from the same data source with similar results.	Same as Nunn and Puga (2012).
Soil quality	Harmonized World Soil Database. http://www.fao.org/soils-portal/soil-survey/soil-maps-and-databases/harmonized-world-soil-database-v12/en/.	Calculating average soil quality measures in a district (average of the index over grids in a district).
Density of mine	Mineral Resources Data System (MRDS) https://mrdata.usgs.gov/mrds/.	Overlapping shape file of districts and mines. Calculating number of mines in each district and dividing it by area.
Nighlight per capita	The National Oceanic and Atmospheric Administration night-time light satellite images. www.noaa.gov/stories/our-earth-at-night.	Calculating nightlight measures in a district (summation of the index over grids). Dividing it by the whole population in the district obtained from census data.
Distance from district to homeland	A map (shape file) of homeland provided by Tim Brophy and Adrian Frith.	Calculating Euclidean between centroid of districts and the border of homelands.
Distance to closest homeland	A map (shape file) of homeland provided by Tim Brophy and Adrian Frith.	Choosing the mininum value of the distance to all homelands.
Conflict	The Geo-referenced Event Dataset of the Uppsala Conflict Data Program (UCDP-GED v1.5) for 1996. The Armed Conflict Location and Event Data Project (ACLED) database for 2001.	Same as Amodio and Chiovelli (ming).

C Appendix. Extending the model to the case where the number of ethnic groups varies

Although in our story we fix the number of ethnic groups, here we show that our model can also be applied to the case where the dispersion of group size is fixed and the number of ethnic groups changes. It can also explain the result that an increase in the number of ethnic groups will improve skill investment. We have the following proposition:

Proposition 4. *In a symmetric setting, suppose each group has the same group size in the district. Social skill investment increases with the number of different ethnic groups in a district.*

Proof. Consider the symmetric case where each group has the same group size. In this case for any ethnic group k, we have $s_k = \frac{1}{m}, \forall k = 1, 2, \ldots m$. According to lemma 1 and lemma 2, everyone has the same social skill investment, regardless of his ethnic group.

We can re-write the utility function of social interaction for an individual i in any ethnic group k in the following way:

$$U_{ik} = \begin{cases} f(\frac{1}{m} + (1 - \frac{1}{m})) - c, & \text{if } x_{ik} = 1 \\ f(\frac{1}{m}), & \text{if } x_{ik} = 0 \end{cases}$$

For $x_{ik} = 1$, $\forall i, k$ to be a Nash Equilibrium, no player is going to deviate by choosing $x = 0$ instead. Suppose c satisfies $c < f(1)$, we have:

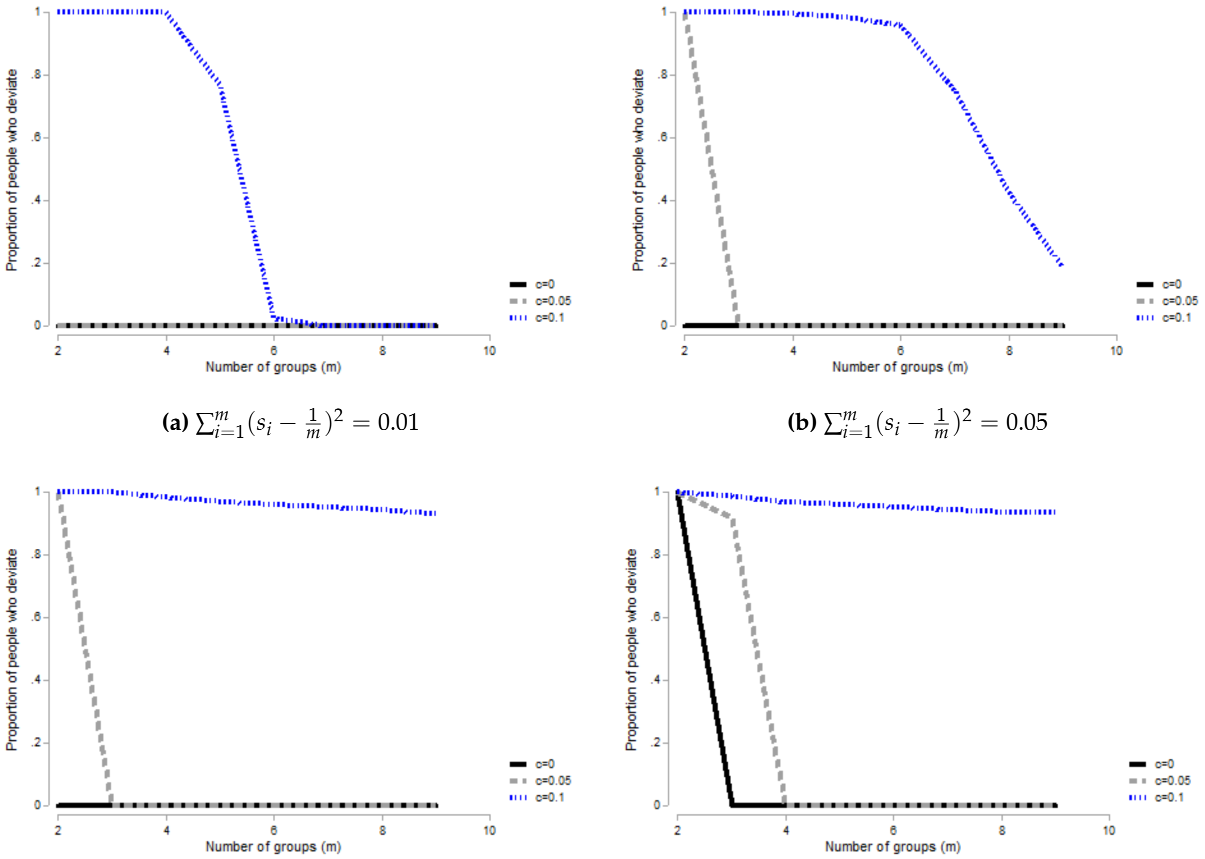
$$f(1) - c \geq f(\frac{1}{m})$$

Since $0 < f(1) - c < f(1)$ and $f' > 0$, there exists a fixed s^* such that $f(1) - c = f(s^*)$. Given $f' > 0$ and $f(s^*) \geq f(\frac{1}{m})$, we have:

$$m \geq \frac{1}{s^*} \tag{12}$$

Therefore, the larger the m is, the more like the Nash Equilibrium $x_{ik} = 1, \forall i, k$ will be maintained. □

This can also be verified via numerical simulation in Figure A1. Here we hold the dispersion of group size $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ constant and see how the proportion of people who deviate from investment changes with the number of groups in a district. Similar to what the above proposition shows, proportion of people who deviate decreases with the number of groups in a district. This is robust to different levels of dispersion of group size and different per unit cost of investment (c). The intuition is that when the number of groups increases, each group becomes less important in group size, indicating that people cannot get enough utility from intra-ethnic interaction and therefore have more motivation to invest in social skills to be able to communicate with those outside their group. We also find that for each certain number of groups and level of dispersion of size distribution, the probability of deviating increases with per unit cost of investment.



(a) $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2 = 0.01$

(b) $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2 = 0.05$

(c) $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2 = 0.1$

(d) $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2 = 0.2$

Notes: The figures show the results on the numerical simulation of the proportion of people who deviate from investing in social skills in response to the number of groups in the district. Each group has population share s_i and there are m groups in total. The dispersion of group size is measured by $\sum_{i=1}^{m}(s_i - \frac{1}{m})^2$. In each graph we hold the dispersion of group size constant. We also consider different per unit cost of investment c.

Figure A1. Numerical simulation results on how the level of investments in social skills responds to number of groups

D Appendix. Explanation on how to draw data for simulation

In this Appendix we explain in more detail how to draw a series of $s_k, k = 1, 2, \ldots, m$ from a convoluted distribution of s under certain constraints in our simulation.

D.1 Hold the number of groups constant

Given a particular value of m, we just need to make sure $\sum_{k=1}^{m} s_k = 1$ when we draw these m different values of s_k.

Choose a particular value of m and hold it as a constant, we start by drawing $d_k, k = 1, 2, \ldots, m$ from a uniform distribution at any positive interval (here we use the interval [0,1]). Set $s_k = \frac{d_k}{\sum_{k=1}^{m} d_k}$. It is straightforward to prove that by this definition $\sum_{k=1}^{m} s_k = 1$. Also by this transformation s_k no longer follows uniform distribution, which satisfies our requirement that in the numerical simulation we need a convoluted density function of s. Therefore $s_k, k = 1, 2, \ldots, m$ is the series of our simulated data which represents each group's share over the whole population in the real data.

D.2 Hold the dispersion of group size constant

In this cases the value of $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ has to be fixed. That is to say, we need to make sure $\sum_{k=1}^{m} s_k = 1$ and $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2 = T$ (T is a constant) when we draw these m different values of s_k.

Similarly, we start by drawing $d_k, k = 1, 2, \ldots, m$ from a uniform distribution at any positive interval (here we use the interval [0,1] again). Set $y_k = \frac{y_k}{\sum_{k=1}^{m} y_k}$.

Choose a particular value of $\sum_{k=1}^{m}(s_k - \frac{1}{m})^2$ (suppose it equals T) and hold it as a constant, we define $s_k = \frac{1}{m} + \frac{y_k - \frac{1}{m}}{\sqrt{\frac{\sum_{k=1}^{m}(y_k - \frac{1}{m})^2}{T}}}$. We can prove that $s_k, k = 1, 2, \ldots, m$ is the series of our simulated data which represents each group's share over the whole population in the real data.[43]

This is because:

$$\sum_{k=1}^{m} s_k = \sum_{k=1}^{m} \frac{1}{m} + \sum_{k=1}^{m}(y_k - \frac{1}{m})\frac{1}{\sqrt{\frac{\sum_{k=1}^{m}(y_k - \frac{1}{m})^2}{T}}} = 1$$

$$\sum_{k=1}^{m}(s_k - \frac{1}{m})^2 = \sum_{k=1}^{m}(y_k - \frac{1}{m})^2 \frac{1}{\frac{\sum_{k=1}^{m}(y_k - \frac{1}{m})^2}{T}} = T$$

Again, although we start from the uniform distribution, after the transformation, the distribution of s_k becomes convoluted.

After making the draws, we conduct 100000 tests for each possible value of m and calculate the mean

[43] One potential problem is that by this transformation y_k might be negative. In our simulation, with relatively large numbers of T this can occur in few occasions. As the number of tests is large enough, we just drop those test with at least one negative y_k in our simulated sample.

value of Y for each m (following the institutional setting, we choose $m = 2, 3, \ldots, 9$). Then we draw a figure of the mean value of Y over the corresponding m.

E Appendix. Tables and figures

Table A0. Gravity model predicting the stock of black population in white districts: PPML estimator

	Dependent variable: ethnic population N_{kd}		
	Coef.	Std. Err.	t-stat
Distance Dis_{kd}	-.0039	(.0007)	-5.17
Ethnic group fixed effects:			
Group 1	.9750	(.2139)	4.56
Group 2	.6133	(.1769)	3.47
Group 3	.1778	(.2248)	0.79
Group 4	-.4604	(.2311)	-1.99
Group 5	.2220	(.2259)	0.98
Group 6	.8940	(.1803)	4.96
Group 8	.0469	(.1833)	0.26
Group 9	-.8184	(.2776)	-2.95
Constant	9.157	(.2176)	42.08
R-squared	.092		
Observations	1989		

Note: This table reports results about the gravity model which helps estimate the stock of each ethnic group in each "white" district based on 1985 census data. The sample is for all the "white" magisterial districts which can be matched to 1996 and 2001 census. We control for homeland fixed effects and run a regression of the stock of ethnic groups on the distance between their corresponding homelands and each district using PPML models. *** p<0.01, ** p<0.05, * p<0.1.

Table A1. First-stage regression results: district level regressions

	[1] 1996	[2] 2001
Predicted ELF	1.122***	1.517***
	(0.351)	(0.278)
Mean edu	0.039**	0.024
	(0.017)	(0.019)
Prop 15-64	0.161	0.901
	(0.878)	(1.006)
Prop <15	-0.257	-0.146
	(0.960)	(1.113)
Prop male	0.444	0.685
	(0.431)	(0.438)
Mean urban	0.025	0.003
	(0.121)	(0.069)
Pop density	0.000	0.000
	(0.000)	(0.000)
River	0.039	0.044
	(0.029)	(0.027)
Density mine	0.630	0.665
	(0.877)	(0.847)
Prop black	-0.154**	-0.202***
	(0.068)	(0.072)
Distance closest	-0.000	-0.000**
	(0.000)	(0.000)
Ruggedness	-0.001	-0.004
	(0.006)	(0.005)
Soil quality	0.055*	0.024
	(0.030)	(0.029)
Per capital light	0.156	0.055
	(0.241)	(0.364)
Road	0.018	0.001
	(0.031)	(0.028)
Conflict	0.022*	-0.007***
	(0.012)	(0.002)
Proportion manu	-0.028	0.016
	(0.137)	(0.103)
Proportion service	0.005	0.129
	(0.174)	(0.106)
Province fixed effect	YES	YES
F-statistics of the instrument	10.19	29.85
R-squared	0.874	0.898
Observations	205	210

Note: This table reports first-stage regression results for our instrumental variable at the district-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

Table A2. Using total distance as an instrumental variable

	[1]	[2]	[3]	[4]
	1996		2001	
	Individual	District	Individual	District
Total distance	-0.000	0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
District controls	YES	YES	YES	YES
Individual controls (district average)	YES	YES	YES	YES
Province fixed effect	YES	YES	YES	YES
F-statistics of the instrument	1.844	0.0145	1.640	0.355
R-squared	0.863	0.866	0.874	0.884
Observations	464,130	205	697,369	210

Note: This table reports results about the effect of ethnic diversity on employment rate at individual-level and district-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In particular we use total distance to all homelands to replace our instrumental variable in main analysis. *** p<0.01, ** p<0.05, * p<0.1.

Table A3. Ethnic diversity and different employment status: individual level regressions

	[1]	[2]	[3]	[4]
	Self employment	Wage employment	Self employment	Wage employment
	1996	1996	2001	2001
Panel A: OLS estimates				
Ethnic fractionalisation ELF	-0.005	0.087***	0.009**	0.137***
	(0.006)	(0.033)	(0.005)	(0.036)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
R-squared	0.014	0.173	0.01	0.159
Observations	464,130	464,130	697,369	697,369
Panel B: IV estimates				
Ethnic fractionalisation ELF	-0.004	0.103	-0.008	0.178**
	(0.014)	(0.075)	(0.012)	(0.084)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
F statistics of the instrument	22.36	22.36	32.04	32.04
R-squared	0.014	0.173	0.01	0.159
Observations	464,130	464,130	697,369	697,369

Note: This table reports results about the effect of ethnic diversity on self- and wage-employment rate at individual-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. "Self-employment" is a dummy variable which equals 1 if one is self-employed and 0 for all other working-age black population. *** p<0.01, ** p<0.05, * p<0.1.

Table A4. Ethnic diversity and employment: district level regressions

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
		1996				2001		
	Unemployed + inactive	Wage employee	Self employment	Self/wage	Unemployed + inactive	Wage employee	Self employment	Self/wage
Panel A: OLS estimates								
Ethnic fractionalisation ELF	-0.079***	0.076***	0.003	0.017	-0.121***	0.109***	0.012	0.026
	(0.028)	(0.028)	(0.008)	(0.020)	(0.036)	(0.036)	(0.007)	(0.019)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
R-squared	0.885	0.865	0.559	0.458	0.876	0.861	0.362	0.281
Observations	205	205	205	205	210	210	210	210
Panel B: IV estimates								
Ethnic fractionalisation ELF	-0.203**	0.190**	0.013	-0.038	-0.147	0.197**	-0.050*	-0.092*
	(0.090)	(0.096)	(0.023)	(0.067)	(0.090)	(0.080)	(0.030)	(0.049)
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES
F statistics of the instrument	10.19	10.19	10.19	10.19	29.85	29.85	29.85	29.85
R-squared	0.871	0.853	0.554	0.429	0.875	0.854	0.182	0.102
Observations	205	205	205	205	210	210	210	210

Note: This table reports results about the effect of ethnic diversity on employment and the allocation between self- and wage-employment at district-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. Dependent variables are the proportion of people in each employment status over the whole working-age black population. *** p<0.01, ** p<0.05, * p<0.1.

Table A5. Ethnic diversity and employment: district level regressions using spatially correlated standard errors

	[1]	[2]	[3]	[4]	[5]	[6]
		1996			2001	
	Uunemployed + inactive	Wage employee	Self/wage	Unemployed + inactive	Wage employee	Self/wage
Panel A: OLS estimates						
Ethnic fractionalisation ELF	-0.072***	0.073***	-0.002	-0.125***	0.127***	0.004
	(0.018)	(0.018)	(0.020)	(0.037)	(0.035)	(0.023)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES
Panel B: IV (GMM) estimates						
Ethnic fractionalisation ELF	-0.385**	0.378**	0.024	-0.193*	0.220*	-0.147
	(0.156)	(0.159)	(0.076)	(0.114)	(0.122)	(0.076)
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on employment and the allocation between self- and wage-employment at district-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. Dependent variables are the proportion of people in each employment status over the whole working-age black population. We use Conley's standard errors with spatial correlations for both OLS and GMM analysis. We use 1000km as the cutoff value above which there is no spatial correlation. *** p<0.01, ** p<0.05, * p<0.1.

Table A6. Robustness check with different control variables

	[1]	[2]	[3]
Panel A: 1996 census, IV estimates			
ELF	0.174**	0.053	0.115
	(0.083)	(0.080)	(0.074)
Population size	0.000**		
	(0.000)		
Proportion self white 96		0.352***	
		(0.125)	
Proportion of migration			0.000**
			(0.000)
F statistics of the instrument	24.35	19.83	22.47
R-squared	0.193	0.194	0.194
Observations	449,200	449,200	449,200
Panel B: 2001 census, IV estimates			
ELF	0.217***	0.162*	0.170*
	(0.084)	(0.083)	(0.091)
Population size	0.000**		
	(0.000)		
Proportion self white 01		0.216*	
		(0.112)	
Proportion of migration			0.000**
			(0.000)
F statistics of the instrument	31.77	31.87	32.25
R-squared	0.173	0.173	0.173
Observations	681,529	681,529	681,529
Individual controls	YES	YES	YES
District controls	YES	YES	YES
Province FE	YES	YES	YES

Note: This table reports the main results about the effects of ethnic diversity on the probability of being an employee at individual-level regressions with different control variables in both 1996 and 2001 census. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index.. *** p<0.01, ** p<0.05, * p<0.1.

Table A7.1. Estimations based on non-linear econometric models

	[1]	[2]	[3]	[4]	[5]	[6]
	Unemployed + inactive			Wage employment		
	Logit	Probit	IV Probit	Logit	Probit	IV Probit
Panel A: 1996 census						
Ethnic fractionalisation ELF	-0.080**	-0.078**	-0.078	0.085**	0.083**	0.082
	(0.033)	(0.033)	(0.080)	(0.034)	(0.034)	(0.081)
Observations	464,130	464,130	464,130	449,200	449,200	449, 200
Panel B: 2001 census						
Ethnic fractionalisation ELF	-0.148***	-0.145***	-0.144	0.145***	0.143***	0.140
	(0.038)	(0.038)	(0.091)	(0.039)	(0.038)	(0.089)
Observations	697,369	697,369	697,369	681,529	681,529	681,529
Individual controls	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on employment based on non-linear econometric models in 1996 and 2001. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. In column 4, 5 and 6 we drop self-employed people as they are a very small proportion of the whole working-age population. *** p<0.01, ** p<0.05, * p<0.1.

Table A7.2. Estimations based on multinomial econometric models

	[1]	[2]	[3]	[4]
	Mlogit		IV Mprobit	
	Self employment	Wage employee	Self employment	Wage employee
Panel A: 1996 census				
Ethnic fractionalisation ELF	-0.008	0.086**	0.106	0.586
	(0.007)	(0.034)	(0.381)	(0.418)
Observations	464,130	464,130	464,130	464,130
Panel B: 2001 census				
Ethnic fractionalisation ELF	0.012**	0.135***	0.408	0.981***
	(0.005)	(0.038)	(0.427)	(0.369)
Observations	697,369	697,369	697,369	697,369
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on employment based on multinomial econometric models (both with and without instrumental variables) in 1996 and 2001. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

Table A8.1. RQ index as a measure of ethnic diversity: first stage regressions

	[1]	[2]	[3]	[4]
	Individual level		District level	
	1996	2001	1996	2001
	RQ	RQ	RQ	RQ
Predicted RQ	-1.957***	-1.693***	-1.686**	-1.592**
	(0.662)	(0.644)	(0.818)	(0.721)
District controls	YES	YES	YES	YES
Individual controls	YES	YES	YES	YES
Province fixed effect	YES	YES	YES	YES
F-statistics of the instrument	8.754	6.903	4.252	4.876
R-squared	0.769	0.780	0.751	0.776
Observations	464,130	697,369	205	210

Note: This table reports first-stage results about our instrumental variable for polarisation index based on 1996 and 2001 census data, at both district- and individual-level regressions. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with polarisation index. *** p<0.01, ** p<0.05, * p<0.1.

Table A8.2. RQ index as a measure of ethnic diversity: individual level regressions

	[1]	[2]	[3]	[4]
	1996	1996	2001	2001
	Unemploy	Wage employment	Unemploy	Wage employment
Panel A: OLS estimates				
RQ	-0.008	0.011	-0.050*	0.049
	(0.026)	(0.027)	(0.030)	(0.030)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
R-squared	0.195	0.193	0.175	0.172
Observations	464,130	449,200	697,369	681,529
Panel B: IV estimates				
RQ	0.032	-0.016	0.183	-0.172
	(0.085)	(0.086)	(0.134)	(0.132)
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES
F statistics of the instrument	8.754	8.795	6.903	6.959
R-squared	0.195	0.193	0.169	0.167
Observations	464,130	449,200	697,369	681,529

Note: This table reports results about the effect of ethnic diversity on employment at individual-level regressions based on 1996 and 2001 census data. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with polarisation index. In column 2 and 4 we drop self-employed people as they are a very small proportion of the whole working-age population. *** p<0.01, ** p<0.05, * p<0.1.

Table A9. Inter-ethnic marriage rate and ethnic diversity: 1996 census

	Mean	Std. Dev.	Obs
Inter-ethnic marriage			
Own generation	0.966	0.18	96,031
Parental generation	0.99	0.0966	532
Second language among married people			
Any second language	0.2356	0.424	95,580
Second English/Afrikaans	0.0888	0.284	95,580
Second ethnic language	0.147	0.354	95,580
Second language among whole sample			
Any second language	0.225	0.418	203,327
Second English/Afrikaans	0.087	0.283	203,327
Second ethnic language	0.138	0.345	203,327

Note: This table reports inter-ethnic marriage rate (i.e. marriage between different ethnic groups within the black population). Ethnicity is identified from the first language spoken by both household head and spouse for the current generation, and household head's parents for the parental generation. We also report the proportion of the black population who can speak a second language.

Table A10. Ethnic diversity and wage employment rate: by education level

	[1]	[2]	[3]	[4]
	Low edu		High edu	
	OLS	IV	OLS	IV
Panel A: 1996 census				
Ethnic fragmentation index	0.100***	0.142*	0.037	0.051
	(0.035)	(0.078)	(0.029)	(0.089)
F statistics of the instrument		23.89		14.69
R-squared	0.184	0.184	0.275	0.275
Obs	297,206	297,206	151,994	151,994
Panel B: 2001 census				
Ethnic fragmentation index	0.136***	0.191**	0.154***	0.115
	(0.041)	(0.090)	(0.029)	(0,101)
F statistics of the instrument		33.91		25.78
R-squared	0.172	0.171	0.221	0.221
Obs	390,222	390,222	291,307	291,307
Individual controls	YES	YES	YES	YES
District controls	YES	YES	YES	YES
Province FE	YES	YES	YES	YES

Note: This table reports the main results about the heterogeneous effects of ethnic diversity on the probability of being an employee at individual-level regressions by educational levels in both 1996 and 2001 census. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features, individual-level controls aggregated at district average and ethnicity fixed effects. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. "High" ("Low") education is defined as years of schooling above (below) 9. *** p<0.01, ** p<0.05, * p<0.1.

Table A11.1. Ethnic diversity and the range of occupations: 1996

	[1] Var manager	[2] Var profession	[3] Var clerk	[4] Var serve	[5] Var craft	[6] Var skill agri	[7] Var operator	[8] Var unskill
Panel A: OLS estimates								
Ethnic fragmentation index	0.611	0.361	-1.323	-1.228	-0.320	-1.499	-2.040	-3.813***
	(0.965)	(3.601)	(0.911)	(1.119)	(0.589)	(1.807)	(1.704)	(1.223)
R-squared	0.793	0.865	0.825	0.762	0.538	0.818	0.827	0.803
Obs	205	205	205	205	205	205	205	205
Panel B: IV estimates								
Ethnic fragmentation index	-4.534	-10.272	-8.039***	-10.269*	-4.614*	-22.605**	-23.246***	-9.949*
	(4.595)	(15.507)	(3.034)	(5.651)	(2.405)	(9.911)	(7.629)	(5.247)
F statistics of the instrument	10.19	10.19	10.19	10.19	10.19	10.19	10.19	10.19
R-squared	0.761	0.856	0.760	0.653	0.363	0.660	0.612	0.769
Obs	205	205	205	205	205	205	205	205
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on the variety of occupations among employees in 1996. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.

Table A11.2. Ethnic diversity and the range of occupations: 2001

	[1] Var manager	[2] Var profession	[3] Var clerk	[4] Var serve	[5] Var craft	[6] Var skill agri	[7] Var operator	[8] Var unskill
Panel A: OLS estimates								
Ethnic fragmentation index	0.319	-0.960	-1.252**	0.459	-0.982*	-0.898	-0.713	-1.494
	(1.081)	(3.594)	(0.619)	(0.583)	(0.570)	(1.679)	(1.744)	(0.924)
R-squared	0.843	0.888	0.829	0.824	0.571	0.870	0.860	0.809
Obs	210	210	210	210	210	210	210	210
Panel B: IV estimates								
Ethnic fragmentation index	-0.589	-4.808	-5.316***	0.432	-3.628***	-11.541***	-6.318	-4.933*
	(2.093)	(10.362)	(1.856)	(1.600)	(1.349)	(3.753)	(5.270)	(2.554)
F statistics of the instrument	29.85	29.85	29.85	29.85	29.85	29.85	29.85	29.85
R-squared	0.842	0.887	0.787	0.824	0.515	0.825	0.849	0.791
Obs	210	210	210	210	210	210	210	210
Individual controls	YES	YES	YES	YES	YES	YES	YES	YES
District controls	YES	YES	YES	YES	YES	YES	YES	YES
Province FE	YES	YES	YES	YES	YES	YES	YES	YES

Note: This table reports results about the effect of ethnic diversity on the variety of occupations among employees in 2001. The sample is only for the "white" magisterial districts which can be matched to 1985 census and whose black population accounts for more than 1% of the overall population. We control for district-level variables especially geographical features and individual-level controls aggregated at district average. We also control for province fixed effects. Ethnic diversity is measured with fractionalisation index. *** p<0.01, ** p<0.05, * p<0.1.