Cai, Shu; Zimmermann, Klaus F.

Working Paper

Social Identity and Labor Market Outcomes of Internal Migrant Workers

GLO Discussion Paper, No. 716 [pre.]

Provided in Cooperation with:
Global Labor Organization (GLO)

Suggested Citation: Cai, Shu; Zimmermann, Klaus F. (2024) : Social Identity and Labor Market Outcomes of Internal Migrant Workers, GLO Discussion Paper, No. 716 [pre.], Global Labor Organization (GLO), Essen

This Version is available at:
https://hdl.handle.net/10419/281118

Social Identity and Labor Market Outcomes of Internal Migrant Workers*

Shu Cai†, Klaus F. Zimmermann‡

This Version: January 9, 2024

Prepublication Version. Forthcoming European Economic Review.

Abstract

Previous research on internal mobility has neglected the role of local identity contrary to studies analyzing international migration. Examining social identity and labor market outcomes in China, the country with the largest internal mobility in the world, closes the gap. Instrumental variable estimation and careful robustness checks suggest that identifying as local associates with higher migrants' hourly wages and lower hours worked, although monthly earnings seem to remain largely unchanged. Migrants with strong local identity are more likely to use local networks in job search, and to obtain jobs with higher average wages and lower average hours worked, suggesting the value of integration policies.

JEL Classification: J22, J31, J61, Z13

Keywords: assimilation, social identity, labor market, migration, internal mobility, *China's Great Migration*.

Declarations of interest: none.

* We thank the editor Robert M. Sauer and two anonymous referees for their helpful comments and suggestions. We are grateful to François Bourguignon, Corrado Giulietti, Lingqing Jiang, Patrick Kline, Xin Meng, Matloob Piracha, Panu Poutvaara, Ronnie Schöb, Yuanwei Xu and participants at the third RUC–GLO Conference, the International Economic Association World Congress, the Eastern Economic Association Conference, the African Productivity Conference, the 2022 Asian Meeting of the Econometric Society, the European Association of Labour Economists Conference 2022, the GLO Global Conference 2022, the 2023 Annual Meeting of the American Economic Association, the 36th Annual Conference of the European Society for Population Economics, the Global Lecture Series on Chinese Economy, SITES 2023 in Naples, the 2023 Annual Meeting of the Chinese Labor Society in Beijing, and seminars at Jinan University, Renmin University of China, Shanghai Lixin University of Accounting and Finance, and Free University Berlin for their valuable comments. We thank the Migrant Population Service Center of the National Health Commission of China for making the Dynamic Monitoring Survey Data of Migrant Population available to us, and Yuyun Liu for providing the digital version of linguistic data. We also thank Xingjian Zhang and Wei Li for excellent research assistance. Shu Cai thanks financial support from China Natural Science Foundation (Project No. 72173056) and the Fundamental Research Funds for the Central Universities (Project No. 23JNQMX29). All remaining errors are our own.

† Jinan University and Global Labor Organization. E-mail: shucai.ccer@gmail.com.

‡ UNU–MERIT, Maastricht University, Center for Economic Policy Research and Global Labor Organization. E-mail: klaus.f.zimmermann@gmail.com; corresponding author.

1. Introduction

Migration is a major topic of our time. Among a world population of 7.4 billion, about one billion people are migrants. Of these, nearly 750 million are internal migrants and some 250 million are international migrants (UN-DESA-PD, 2016). While economists have given considerable attention to the economic assimilation of immigrants in particular with respect to wages and employment (Chiswick, 1978; Borjas, 1985; Lubotsky, 2007; Abramitzky et al., 2012; Kuziemko and Ferrier, 2014; Duleep et al., 2022), research on the social assimilation of migrants in their host places has been rare and focused primarily on international mobility. However, the identity literature pioneered by Akerlof and Kranton (2000) has recognized that the self-image of migrants and their adaptation in identifying with the social context of host areas in the process of migration is an important factor for economic decisions and labor market outcomes (Battu et al., 2007; Constant and Zimmermann, 2008, 2011; Constant et al., 2009).

This study closes the evidence gap on the economic impact of social assimilation of internal labor mobility, which faces challenges that are different to those of international migration. Specifically, unlike international migration, where keeping one's original identity may have positive labor market effects due to the economic benefits derived from diversity, this is often considered to be less relevant in an internal mobility setting (Ottaviano and Peri, 2005; Alesina and La Ferrara, 2005).

Some types of social identities like ethnic or cultural identities might also be considered to be less diverse and less important for internal migration of foreigners and natives (Skeldon, 2006; Wang and Fan, 2012). But the use of social networks may be crucial for an effective and successful social life and on the labor market even independent of migratory settings (Ioannides and Loury, 2004; Granovetter, 2005; Bayer et al., 2008; Cappellari and Tatsiramos, 2015; Campigotto et al., 2022). For instance, it has been found that public institutions are often much less effective and less common to find jobs than relying on family, friends, own ethnicities or other distinct population groups. Social network effects are particularly relevant for developing countries (see Beaman and Magruder, 2012, for India and Nie and Yan, 2021, for China). Therefore, social identities could still have a major impact.

To challenge the conventional view that social identities are less relevant for internal

migrants (Skeldon, 2006; King and Skeldon, 2010; Ellis, 2012; Wang and Fan, 2012), we focus on the impact of adaptation to local identity in host areas on labor market performance among migrant workers with data from China—the country with the largest recent internal migration experience.[1] The data used in the analysis are from the Dynamic Monitoring Survey of the Migrant Population of China, which was conducted among a representative population of migrants in eight prefectures in China in 2013 and collected detailed information on migrants' identity and their labor market outcomes in host areas, thus providing us with a unique opportunity to examine this question.

To address the endogeneity issue in the identification of the impact, we rely on the exogenous variation in migrants' social identity captured by the linguistic distance between the dialect of the host county and that of the original province (i.e., dialect distance). We will argue that labor market outcomes are beyond social identity plausibly independent of dialect distance conditional on individual sociodemographic characteristics and fixed effects of the original province and the host county provided that communication on the workplace can be done by the commonly spoken Mandarin Chinese. Our instrumental variable estimates reveal that adaptation to local identity increases the hourly wage and reduces the average hours worked per day as well as the likelihood of overworking, keeping the monthly wage unaffected. This documents to what extent economic assimilation follows social identity adaptation, thereby bringing the literature on economic assimilation in context with the one on social identity formation.

On assessing the validity of the exclusion restriction of our identification strategy, we empirically confirm that the communication effect of dialect on migrants' labor market outcomes is economically and statistically insignificant. This may follow from the observation that most migrants can communicate in the workplace by using their Mandarin Chinese knowledge. In addition, we show that sorting in the migration choice exists mainly across the original province and/or across the county of destination, whereas the amount of sorting within origin-destination pair is actually small. More importantly, the remaining sorting on

[1] We use the term "social identity" and "local identity" interchangeably in the rest of the paper, although "social identity" is a much broader concept than "local identity". Studies on international migration usually use the term "ethnic identity" to reflect immigrants' self-image of their identity in the host country. However, this term is not suitable in our context, since most internal migrants in China are *Han* Chinese. See Section 4 below for details about the definition and measurement of "local identity".

observables is not significantly correlated with the instrumental variable or the labor markets outcomes, suggesting that sorting on unobservables is unlikely to be a serious threat to identification in the spirit of Altonji et al. (2005). Moreover, we show that the results remain robust even after controlling for region-of-origin-by-destination-county fixed effects and other interconnected factors such as transportation distance, log of the number of migrants from the same province, and wage gap between the place of origin and the destination.

To further validate the exclusion restriction, we conduct a falsification test that examines the reduced-form relationships between the dialect distance and labor market outcomes with the sample of migrants who resided in the destination county for no more than half a year. If dialectal difference affects migrants' labor market outcomes only through their social identity, then there should be no association between the dialect distance and labor market outcomes for these migrants. This is because social integration takes time, and the labor market advantage of identification with the host place is very unlikely to occur among newly arrived migrants. The results suggest that we cannot reject the null hypothesis that the exclusion restriction is satisfied.

Finally, we examine the robustness of our results to possible violation of the exclusion restriction by employing the plausibly exogenous approach developed by Conley et al. (2012). The exercise indicates that our main results remain robust even when we allow for a plausible amount of direct correlation between the instrumental variable and the labor market outcomes. The stability of the instrumental variable estimates should further alleviate concerns regarding the exclusion restriction.

To examine the potential role of social networks through which commitment to the host community may affect migrants' labor market outcomes, we investigate the impacts of social identity on migrants' access to the networks of local people and the use of local networks in their job search. The results suggest that adaptation to local social identity significantly raises the probability of migrants interacting with local residents and having local neighbors. Moreover, socially assimilated migrants are also more likely to find a job with help from local networks. These results highlight the importance of social networks with local people for explaining the advantages of adopting the local identity.

The present study builds on a small but growing pool of literature that investigates the relationship between migrants' identity and their labor market outcomes in host places (the interplay between social and economic adaptation) with a focus on international migrants. Constant and Zimmermann (2008, 2009) found that among immigrants in Germany assimilated men and women are more likely to work, and women who commit to both host and home social identities are more likely to work than women who are assimilated, but this does not hold for men. Furthermore, they found no significant relationship between ethnic identity and the earnings of men or women. Casey and Dustmann (2010), using German panel data, confirmed a positive association between German identity and employment for females but not for males. They also provided evidence for a positive association between home country identity and employment for only the males among second-generation immigrants. Battu and Zenou (2010) presented evidence for an employment penalty associated with oppositional identity among ethnic minorities in the UK. Using Canadian survey data, Islam and Raschky (2015) found that immigrants' home country identity significantly increases the probability of being unemployed, while a strong host country identity reduces the likelihood of unemployment; but neither of the identities has an effect on immigrants' wage. In Gorinas (2014), the employment of immigrants in Denmark was not systematically associated with measures of ethnic identity but was significantly related to openness to majority norms, particularly for first-generation immigrant women.[2]

Our study contributes to the extant literature in several aspects. First, while most previous literature studies identity and labor market performance for international migrants in developed countries, we focus on internal migration within China, the largest developing country in the world. This makes our study unique for understanding the economic impacts of the social identity of internal migrants.[3] Second, extant studies primarily gauge the likelihood of employment or earnings as labor market outcomes in host places. We complement the literature by examining the quality of jobs, including working hours, the likelihood of overworking,

[2] Related studies include immigration in Sweden (Edin et al., 2003; Nekby and Rödin, 2010), Canada (Pendakur and Pendakur, 2006), France (Delaporte, 2019), Australia (Piracha et al., 2023), Italy (Carillo et al., 2023), and Europe (Bisin et al., 2011).

[3] Some studies conduct within-county analysis of the impact of specific features of language on labor market outcomes. Eugster et al. (2017) examine how culture (in particular attitudes towards work) generated by the linguistic border that separates German from Romance language in Switzerland affects the duration of job searches. Campo et al. (2023) investigate the effect of language future time reference on self-employment among long-term first-generation immigrants in Switzerland. However, neither of them examines the impact of social identity of internal migrants on their labor market outcomes.

hourly wages, and monthly earnings. Third, evidence from most previous studies indicates only an association between immigrants' identity and labor market outcomes except for Islam and Raschky (2015) which exploited genetic distance between immigrants' home and host countries as instruments for immigrants' identity. In this study, we provide causal evidence on the relationships by using exogenous variation in migrants' identity caused by cultural difference between the original and host places. Fourth, we highlight the role of native networks in explaining the labor market advantages of social identity committed to the host place. Networks have been emphasized theoretically as an important channel through which ethnic identity may affect immigrants' labor market outcomes (Battu et al., 2007; Verdier and Zenou, 2017), but they are seldom examined in empirical studies of social assimilation.

Our research also ties into the broad literature on examining the impact of group identity (such as gender and ethnic identity) on economic outcomes, including consumption, financial decisions, labor force participation, inequality, trader feedback, engagement in the workplace, cooperation, and competition (Afridi et al., 2015; Bolton et al., 2020; Guadalupe et al., 2020; Martinangeli and Martinsson, 2020; Olivetti et al., 2020; Bricker et al., 2021), and non-economic outcomes, including conflict, norms, values, and preferences (Desmet et al., 2017; Amodio and Chiovelli, 2018).[4] Our study contributes to this literature by examining the impact of social identity of internal migrants on their labor market outcomes. This is of independent interest given the unique context and salient scale of the subject group.

More broadly, this study relates to a rich literature on the economic assimilation of immigrants, which has focused on investigating earnings gaps between immigrants and natives (Chiswick, 1978; Borjas, 1985; LaLonde and Topel, 1991; Hatton, 1997; Minns, 2000; Card, 2005; Lubotsky, 2007; Abramitzky et al., 2012, 2014). Economic assimilation is not enough to explain all the phenomenon during the process of assimilation, such as ethnic segregation. In addition, the labor market disadvantage of minority groups can be reinforced by their ethnic identity (Battu and Zenou, 2010). Our study contributes to this literature by shifting attention toward social assimilation and examining how social identity affects the labor market performance of minority groups, which suggests the potential value of supporting policies.

[4] See Shayo (2020) for a comprehensive review of evidence from applied economics, and Charness and Chen (2020) for evidence from the experimental literature.

The rest of the paper is organized as follows. Section 2 provides a conceptual framework. Section 3 introduces the background of the study. Section 4 describes the data and measurements. Section 5 lays out the empirical specification and estimation strategy. Section 6 presents the empirical results. The final section concludes.

2. Conceptual Framework

The integration of migrants into the host community is a process of migrant adaptation to aspects related to economics, culture, behaviors, and psychology. It includes not only economic assimilation by catching up with the earnings of the natives, for instance, but also involves behavioral adaptation, self-identification, and cultural affinity with the host community. In this section, we present a conceptual framework to illustrate how these different dimensions of integration interact with each other, and, in particular, how social identity affects the economic assimilation of migrants.

In the spirit of Akerlof and Kranton (2000), the standard utility function can be extended to include individuals' sense of self, namely, identity. To achieve a better self-image, individuals may make a seemingly sub-optimal choice, but their overall utility is maximized. For instance, migrants may be willing to "pay" an income penalty in choosing an occupation to reinforce their identity. Self-identification changes the "payoffs" from different actions. Therefore, it affects individuals' behavioral choice. Consequently, identity will affect economic outcomes of these behaviors and interactions.

It is important noting that we consider identity to be endogenous. That is, migrants can decide on whether to identify with the original and host places in responding to potentials and constraints. Individuals "produce" identity in the terminology of Becker's home production approach. An example for a choice-based approach to identity is the model of Battu et al. (2007), in which non-white individuals determine the level of adaptation to white culture by balancing peer pressure from same-race friends and the beneficial effect of high-quality jobs through whites' social networks which do not suffer from discrimination. Endogenous ethnic identity formation was also systematically analyzed by Alesina and La Ferrara (2005) and Constant et al. (2009). This concept can be applied to the different social identities between local

communities in China.

In principle, migrants enter a host place with a stronger identification with the social context of the original place and a lower commitment to the host society. However, the migrant social identity can and will evolve confronted with the social context of the receiving community. Social identity may be affected by factors of the local macro environment, such as culture and institutions. In particular, social identity is influenced by cultural differences between host and original places, for instance with dialects. Migrants may experience greater anxiety and discomfort by affiliating with the culture of a host community that is more different from their original culture.

To guide our empirical analysis, we investigate the following hypotheses:

H1: Differences in culture between the original and host places will hinder migrants' identification with the host community.

H2: Identification with the host society (social identity) has a beneficial effect on migrants' labor market outcomes.

H3: Social identity affects economic outcomes through migrants' behavioral adaptation to the host community.

3. Background

3.1 China's Great Migration

China has witnessed a massive flow of migration from the interior to the coast, or from poor rural areas to more developed urban areas. According to the National Bureau of Statistics of China, 245 million people migrated outside of their home township over six months in 2013, which is about 18% of China's total population that year. This number is about ten times the size of immigration from Europe to the U.S. during the Age of Mass Migration (Sequeira et al., 2020), and about forty times of the size of the Great Migration of Southern-born African Americans to the urban North and West that occurred between 1910 and 1970 (Black et al., 2015; Stuart and Taylor, 2021). *China's Great Migration* has been described as the "greatest development story in human history" (Gardner, 2017). In the great flow, a substantial number of people were engaged in job-related migration caused by large wage differentials across

regions. This is due partly to decreased marginal labor productivity in the agricultural sector as a result of abundant laborers, and partly because of the accelerated development of the manufacturing and construction industries in urban areas, mainly in coastal cities, after China became a member of the World Trade Organization (Erten and Leight, 2021).

The great migration is accompanied by difficulties of social assimilation due to the vast cultural differences between the places of origin and destination and institutional barriers such as the *hukou* (the household registration) system.[5],[6] China has a broad land spanned by many degrees of latitude and longitude with varied climate zones and complicated terrain. It also has a large population with diverse cultures (Talhelm et al., 2014). Thus, migrants may face great challenges because of significant differences in language, customs, attitudes, eating habits, and other lifestyle factors. For institutional barriers, even though China's *hukou* system has been gradually relaxed over time and non-*hukou* migration has been tolerated, the conversion to local *hukou* and related social benefits (such as pension, education, medical insurance, and permission to purchase housing and vehicles) is still quite restrictive for non-*hukou* migrants (Chan, 2009), making migration in China predominately temporary and individualized (Cai et al., 2022).[7] These cultural and institutional barriers hinder migrants from adapting to local identity, which in turn may affect their labor market performance.

3.2 Dialects in China

China is unique in its language, which has a unified writing system, whereas its spoken language varies substantially across regions. The geographic variation of dialect is the result of historical interactions across regions and linguistic evolutionary processes involving mass migration flows, military borders, and political events. Thus, the similarity of dialects between regions may be informative about these historical interactions and indicate similarity in cultural identity (Falck et al., 2012; Suedekum, 2018).[8]

[5] A large number of studies has examined the wage differentials between local and migrant workers in urban China. Using the same data as this study used, and combining data on local residents from a matched survey, Cai and Zhang (2021) show that migrant workers, on average, have lower hourly wage and longer working hours than their local counterparts.

[6] To some extent, these are akin to barriers faced by international migrants. In this study, we focus on cultural barriers of social assimilation (i.e., linguistic distance). However, institutions in the host community are also shown to be pivotal barriers to the integration of migrants. See Freedman et al. (2018) and Bazzi et al. (2021) for examples of international migration.

[7] See Appendix A for more details of China's *hukou* system.

[8] Although culture is a broader concept than language, and includes other domains such as traditions, habits, and beliefs, language is well accepted as an important and clear indicator of culture (Herrmann-Pillath et al., 2014). Compared to dialect,

Differences in dialect may affect the local identity of migrants for several reasons. Individuals may bear psychological costs when interacting with people speaking different dialects or may be discriminated against, which may hinder their identification with the host community. In addition to the cultural effect of dialects on identity, they may have a communication effect on their labor market outcomes. However, this should be less of a concern in our study setting because of the popularization of Mandarin Chinese (i.e., *Putonghua*).[9] Using nationally representative data of the labor force from the 2012 China Labor-force Dynamic Survey, Liu et al. (2020) show that 71.6% of internal migrants in China can speak *Putonghua*, and another 12.2% of migrants can understand *Putonghua*, although they cannot speak it. For non-migrants (including rural and urban residents), some 57% can speak *Putonghua* and 16.4% can understand it. Given the high popularization rate of *Putonghua*, people can easily communicate with one another in the workplace. Thus, the difference in dialects between the home and host places mainly affects migrants' identity and is not of labor market relevance because of communication difficulties. We provide more evidence in Section 6.2 to assess the communication effect of dialect.

4. Data and Measurements

The data used in this study are from the Dynamic Monitoring Survey of the Migrant Population of China in 2013. Starting in 2009, the National Health and Family Planning Commission of China conducted an annual nationwide survey of the migrant population. In 2013, the survey included a special module on social integration in eight prefectures. The prefectures were selected to be geographically representative of the main migration destinations in China. Four prefectures were chosen from Eastern China—the Songjiang district in Shanghai, the Suzhou and Wuxi prefectures in Jiangsu Province, and the Quanzhou prefecture from Fujian Province. Two were selected from Middle China—the Wuhan prefecture from Hubei Province and the Changsha prefecture from Hunan Province. Two more

ethnicity and religion are more homogeneous in China.

[9] The *Putonghua* (or the Standard Mandarin) became the official language of People's Republic of China (P. R. C.) in 1956, when the country started to promote it as the common speech nationwide and it also became a mandatory language used in schools and governments (State Council of P. R. C., 1956). According to the Putonghua Popularization Survey conducted by the State Language Commission in 2010, approximately 70% of the Chinese population can speak *Putonghua* compared to about 50% of the population at the end of the last century (see http://www.gov.cn/gzdt/2013-01/04/content_2304386.htm).

were chosen from Western China—the Xi'an and Xianyang prefectures from Shan'xi Province. Each prefecture's migrant population ranks among the top in its region.[10] In total, the survey prefectures cover about one tenth of the total internal migrants in China. Figure 1 illustrates the geographic location of the eight prefectures.

The full population on which the sampling is based includes all migrants aged 15–59 (inclusively). In the survey, a person is considered to be a migrant if he or she lived in a county for at least one month, whereas his or her *hukou* was registered outside the county in which he or she lived at the time of the survey.[11],[12] The survey uses the multi-step Probability Proportionate to Size (PPS) method to conduct the sampling. In the first step, the survey selected the township according to the PPS method within each prefecture. For each of the selected townships, the survey then chose the sampling unit, namely, villages or communities, using the PPS method. In the last step, the survey randomly chose 20 migrants in each sampling unit.

The designed sample size of the eight prefectures is as follows: Songjiang (2,000), Suzhou (4,000), Wuxi (2,000), Quanzhou (2,000), Wuhan (2,000), Changsha (1,880), Xi'an (2,000), and Xianyang (1,000). In the data set, the sample sizes of Suzhou and Wuhan are 3,999 and 1,999, respectively, whereas the sample sizes of the other prefectures are equal to the designed sample size. There is a total of 16,878 migrants from eight prefectures, 68 counties, and 844 villages or communities in the data set.

The survey collected detailed information on migrants' demographic and social characteristics, migration experience, employment status, income, and so on. In particular, related to our main outcome of interest, the survey contains information on the labor market performance of the respondents, including monthly income and work time (average days per

[10] According to the 2015 population census, the migrant population in Shanghai (*shi xia qu*), Suzhou, Wuxi, and Quanzhou rank first, seventh, 12th, and 18th respectively among prefectures in Eastern China (106 prefectures in total); Wuhan and Changsha rank first and second respectively in Middle China (106 prefectures in total); Xi'an and Xianyang rank third and 16th respectively in Western China (133 prefectures in total).

[11] This excludes people who commute between districts within the same city, or people with a separate *hukou* registration place because of temporary business trips, medical treatment, tourism, and family visits, or those serving in the military or studying in secondary school and above.

[12] According to the definition, it excludes those who converted their *hukou* to their destination after migration. Actually, given the very high requirements of *hukou* conversion as described in Appendix A, very few migrants can convert their *hukou* to their host destination. The annual conversion rate is only between 0.15% and 0.2%, even during the period of economic reform (Lu, 2003).

week and average hours per day).[13] It also contains information on other labor market characteristics, including employment type (employee, employer, self-employed, and others), occupation, industry, and types of work unit.[14] Given the difficulties of separating earnings between those from labor inputs and those from capital inputs for employers and the self-employed, we restrict the sample to only employees. To address concerns regarding the sample selection, we account for potential selection bias in a robustness check. The results confirm that sample selection is not a severe threat to the main estimates. See Section 6.4 for details.

For respondents from the eight prefectures, the survey also asked the question "Which of the following types of identity do you think you belong to?" The answers to the question include "local citizen," "new local citizen," "the citizen of your hometown," and "do not know."[15] Only three percent of the respondents said they did not know, which may include individuals who either could not affiliate with both host and home areas or affiliate more or less equally with both. Since the number of those who could not decide is surprisingly small, we drop those respondents from our sample.[16] We therefore can measure migrants' local identity by a dummy, which equals 1 if respondents said they felt they were local citizens or new local citizens, and 0 if they felt they were citizens of their hometown. Some 45% of our sample of migrants who were employees affiliated with the host place.

In addition to the survey data, we also use linguistic data to construct the dialect distance between the original place and the current place of residence of the migrants. The linguistic data on local dialects are from the Chinese Dialect Dictionary (Xu and Ichiro, 1999), which is based on a detailed census conducted by a massive on-site investigation between 1983 and 1987. It identifies the main Chinese dialects and draws a dialect tree constituted of ten dialectal super-groups, 20 dialectal groups, and 105 dialectal sub-groups, according to the similarity of

[13] Specifically, income includes personal employment earnings and operating income, where employment earnings consist of wages, bonuses, overtime pay, allowances, and the equivalent monetary value of food and accommodation provided by the work unit.

[14] The types of work unit include state organizations, state-owned and state-holding enterprises, collective enterprises, individual businesses, private enterprises, Hong Kong, Macao and Taiwan enterprises, Japanese and Korean enterprises, European and American enterprises, Chinese-foreign equity joint ventures, and others.

[15] According to the survey manual, respondents were asked to assess their self-image of their identity without considering their *hukou* status. Therefore, when we mention "identity" in this study, it means the migrants' self-identification with the communities of the original and host places, rather than their assigned categories according to *hukou*. As described earlier, all the respondents of the survey were migrants who did not hold local *hukou*.

[16] This excludes what the identity literature has called "integration" (affiliation with both host and original areas) and "marginalization" (cannot affiliate with both), which was shown to be relevant for international migrants. See Constant and Zimmermann (2008) and Constant et al. (2009), for instance.

phonological and grammatical attributes, such as articulation and pronunciation. The dictionary also classifies every county in China into a dialectal sub-group. Using the linguistic atlas of China, we construct the dialect distance between county-pairs to measure the similarity of their dialects. Specially, following Spolaore and Wacziarg (2009), we coded distance as 0 if the dialects of the two counties belong to the same dialectal sub-group, 1 if they belong to the same dialectal group but different sub-groups, 2 if they belong to the same dialectal super-group but different dialectal groups, and 3 if they belong to different dialectal super-groups. With the county-level matrix of dialect similarity in hand, we then construct a measure of the dialect distance between the residential county of migrants and the province they come from, using the population-weighted dialect distance between the residential county and each county in the original province.[17] The dialect distance is a pair-wised measure of the similarity of linguistic characteristics between dialects. Specifically, it measures the steps required for two dialects to reach a common node in the dialect tree (Liu et al., 2020). Thus, it does not capture specific features of dialects (e.g., grammatical difference in separating future and present events) or the ordinal differences between dialects (e.g., difference in implicit social status of dialect) that may be directly related to labor market outcomes (Chen, 2013). Figure A1 in Appendix E provides an example by demonstrating the bilateral dialect distance between one of the counties in the sample—Chang'an district (the point)—and the potential destination provinces of migration. Although dialect distance and geographic distance are significantly correlated (correlation coefficient=0.43, p=0.000), the geographic distance can only explain 18.5% of the variation of dialect distance. Similarly, the economic differences (measured by wage gap) between the places of origin and destination can explain only 5% of the variation of dialect distance. The statistical results indicate that dialect distance is not directly proportional to geographic distance, or simply reflects economic differences across regions.

Figure 2 plots the probability density distribution of labor market outcomes for respondents who feel they belong to the local citizens and those who feel they do not belong to the local citizens. Panel A illustrates that the distribution of work time (measured by average hours per week) for the group of assimilated migrants is to the left of the distribution for the

[17] We only know the original province of the migrants. The weights are constructed by using data from the population census in 2000.

unassimilated group, whereas Panel B demonstrates that the distribution of hourly wage for the former group is to the right of the distribution for the latter group. Kolmogorov-Smirnov tests indicate that the differences in the distributions of work time and hourly wage between these two groups are statistically significant (see Figure 2). It is worth noting that a substantial proportion of migrants worked over the standard work time (i.e., 40 hours per week) as revealed by Panel A.

Table 1 provides summary statistics for labor market outcomes and demographic characteristics. Column (1) reports the sample mean of migrants who feel they belong to the local citizens, whereas Column (2) reports the sample mean of migrants who do not feel they belong to the local citizens. The last two columns report the difference between the two groups and the *p*-value of the hypothesis that the difference is equal to 0. For labor market outcomes, the monthly income in the assimilated group is 179 *yuan* higher than the unassimilated group, and the difference is statistically significant. The difference in hourly wage between the two groups is 1.59 *yuan* per hour, which is significantly different from 0. The next three rows show that migrants with local identity have significantly less working time than the unassimilated migrants, in terms of average days worked per week, average hours worked per day, and average hours worked per week. For example, on average, migrants who are adapted to local identity work 0.34 hours or about 20 minutes less every day than those who are not adapted to local identity. The next three rows summarize the situations of overwork in both groups. Overwork is common among migrants, as illustrated by the high averages. The unassimilated migrants are more overworked than socially assimilated ones, and the differences are statistically significant. Regarding the demographic characteristics, the age gap between the two groups is not large but statistically significant, whereas difference in gender is not significant. On average, married migrants are more committed to local identity than unmarried ones. The last five rows show that migrants with local identity are generally more likely to have a higher educational level.

Do social identity adaptation and economic assimilation evolve together or independently? We do not have panel data to study this, but know the duration in years migrants in our survey are present in the host area. According to our conceptual framework, the share of migrants with

local social identity and of those economically assimilated should both start low, and rise jointly with a lead of social identity. This analysis brings the literature on economic assimilation in context with the one on social identity formation.

Figure 3 contains the data means for hourly wages, hours worked per week and the share of local social identity for the eleven year cohorts (0, 1, 2…, 10) of the migrants in our sample: social identity against hours worked in panel A, and against wages in panel B. Older cohorts have the expected high local identity rates (about 50-60%), while cohorts 1-3 are below (about 40-45%), and the current cohort (0) is 33.8%. The current cohort (0) has very low wages and very high working hours. For cohorts 1-3, wages are rising fast, and hours worked stabilize at a much lower level. For all older cohorts, wages stabilize at a much higher level not far below the level of the local workers (shown as the dashed line in panel B).[18] This economic assimilation, however, does not take place with respect to hours worked. The mean values for cohorts 4 to 10 are somewhat smaller than the average of cohorts 1-3, but far above the average hours worked by the locals (shown as the dashed line in panel A).

It is understood that cohort data cannot replace panel information to finally judge the dynamic adjustment process, but what we observe is at least consistent with the understanding that identity formation takes place strongly over time, and goes hand in hand with economic assimilation at least for wages but not for hours worked. However, this finding is explorative. Whether social identity drives economic assimilation, we will investigate in the next sections.

5. Empirical Strategy

The following equation for the determinants of migrants' labor market outcomes is estimated:

$$y_{icp} = \alpha_0 + \alpha_1 Identity_{icp} + \alpha X_{icp} + \delta_c + \delta_p + \varepsilon_{icp}, \quad (1)$$

where i represents individuals, c represents the current residential county of individual i, and p represents the home province of the individual. y_{icp} is the labor market outcome of

[18] The average hourly wage and the average hours worked per week (shown below) are calculated by using data from a matched survey of local residents from the same eight prefectures as migrants in our sample. The survey was conducted by the National Health and Family Planning Commission of China as part of the Dynamic Monitoring Survey of the Migrant Population of China in 2013.

individual i, including monthly income, hourly wage, working time, and so on.[19] The key explanatory variable is social identity $Identity_{icp}$, which is a dummy equal to 1 if the individual feels he or she belongs to the group of locals (local citizen or new local citizen), and 0 otherwise. X_{icp} is the vector of control variables, including age, age squared, dummy of male, marital status (including dummies of married once, married two or more times, divorced, and widowed), and education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). δ_c and δ_p are the fixed effects of the county of current residence and the province of the original place of the respondents, respectively. ε_{icp} is the error term, which is clustered by the community of current residence in accordance with the sampling design (Abadie et al., 2017). The parameter α_1 is our main interest. It indicates how migrants' identification with the host community affects their labor market outcomes.

The ordinary least squares (OLS) estimates of α_1 are biased if there is reverse causality or there are omitted variables. For instance, people with better labor market outcomes may feel more integrated with local residents or people who are ambitious to integrate into the host community may intentionally affiliate with host places and work hard for better economic integration.[20] The OLS estimates may also be contaminated by attenuation bias given that answers to the survey question on the identity question may be subject to a measurement error. To address the endogeneity problem, we use dialect distance between the residential and original places of the respondents as an instrument for their social identity. Previous studies have shown that linguistic distance is an important determinant of migrants' social identity (Fouka, 2020; Ginsburgh and Weber, 2020). Thus, we assume that migrants' identification with the host community is a function of the following determinants:

$$Identity_{icp} = \beta_0 + \beta_1 dialect\ distance_{icp} + \beta X_{icp} + \delta_c + \delta_p + \tau_{icp}. \quad (2)$$

The variable $dialect\ distance_{icp}$ is the instrumental variable that measures the distance between the dialect spoken in the residential county c of migrant i at the time of the survey

[19] For an easier interpretation of estimated coefficient, we use the level of earnings as the dependent variable rather than its log transformation. The results are robust if we use the log-transformed earnings.

[20] Panel data may reduce the challenges of tracing the causal relationship between local identity and labor market outcomes. To the best of our knowledge, the only longitudinal survey on China's internal migrants is the Rural-Urban Migration in China (RUMIC). However, the data do not contain a survey question about migrants' local identity.

and the dialect spoken at his or her original province p. We also control for individual characteristics X_{icp} that may affect migrants' identification with the host community. They are the same as in equation (1), which contain age, age squared, dummy of male, marital status, and education categories. Following Constant et al. (2009), we do not include post-migration variables that could be endogenous, although self-identification with the host community may evolve after migration with factors such as time since migration and intention to migrate permanently.[21] By using the dialect distance between the places of origin and destination as the instrumental variable, we exploit the variation in identity caused by cultural differences between the home and host places for identification. Since such variation is determined prior to migration, it is likely to be orthogonal to post-migration factors that may also be correlated with labor market outcomes. In equation (2) we also control for the fixed effects of the current residential county and the fixed effects of the original province, which absorb the determinants of social identity common to all migrants in the same destination county or from the same original province. In particular, we control for economic conditions and public policy of the destination counties via the fixed effects δ_c (e.g., local *hukou*-registration restrictions) and those of the original provinces via the fixed effects δ_p.[22] Therefore, the identification is essentially a within-origin-province and within-destination-county comparison between individuals with varying degrees of similarity between dialects at home and destination places.[23]

The exclusion condition of the instrumental variable estimation is based on the assumption that, conditional on the individual characteristics X_{icp} and the fixed effects of residential county and original province, the dialect distance between the host county and the home province affects migrants' labor market outcomes only through their social identity.

[21] The marital status of migrants was measured at the time of the survey. Admittedly, marital status may change after migrating and it can be endogenous if cultural identity affects marriage formation (Gousse et al., 2023). For example, one concern is that affiliation with the host place may increase the likelihood of marrying local people. However, it is rarely the case that migrants marry locals in the context of urban China. Migrants, especially male migrants, usually return to their hometown for marriage after accumulating enough wealth at the host places (Mu and Yeung, 2020). Migrating for marriage is also not common. In our sample, only 0.31% of the respondents migrated for marriage. Reassuringly, the results of the main estimates are quite similar if we do not control for the variables of marital status (see Table A9 in Appendix E).

[22] The destination-county fixed effects in the specification will account for the differences in local *hukou* policy that arise after the decentralization reforms of *hukou* system (Bosker et al., 2012). It is worth noting that whether one can obtain a local *hukou* does not depend on where the migrants come from (Zhang et al., 2019).

[23] For a robustness check, we also control for the region-of-origin-by-destination-county fixed effects. See details in Section 6.2 below.

One concern regarding the identification assumption is that facility with the local dialect because of less linguistic distance between dialects at the host and home places may have a beneficial effect on labor market outcomes. However, given the popularization of Mandarin Chinese, most migrants should have no difficulty communicating in their workplace. Actually, the results below suggest that the ability to speak or understand a local dialect has no significant effect on migrants' labor market outcomes conditional on variance in social identity caused by the similarity between dialects at the home and destination places. This is consistent with the finding of Liu et al. (2020), who show that the main barrier to China's internal migration caused by dialect distance is due to the difficulties of social integration, whereas the communication effects are small.

Another main concern about the exclusion restriction is that individuals may have a comparative advantage in some destination counties if their home dialect is similar to the local dialect of that county. In case migrants sort across destinations according to their comparative advantage (Bazzi et al., 2016), we would observe a negative correlation between dialect distance and labor market outcomes driven simply by selection on the comparative advantage. To assess such selection, we follow Bayer et al. (2008) and examine the extent of sorting by analyzing the correlation between observable individual characteristics and the average characteristics of other migrants who came from the same home province and lived in the same destination county.[24] The higher the correlation coefficients, the greater the extent of sorting in the choice of destination places.

The results are presented in Table A1 of Appendix E. Columns (1) and (2) report the unconditional correlation coefficients and their significance levels. As shown, the individual and average group characteristics are positively and highly correlated in terms of age, sex, marital status, levels of education, number of children, ethnicity, and *hukou* status, indicating there is indeed a significant amount of sorting in the choice of destination places among migrants. However, Column (3) demonstrates that the associations reduce substantially when

[24] Like Bayer et al. (2008), we randomly choose a respondent in each group indicated by original province and destination county to avoid a negative correlation mechanically if all individuals were used in the estimation. The average characteristics of the group are calculated by excluding the chosen individual. Similar to Bayer et al. (2008), we drop the groups with less than six respondents in the analyses to reduce measurement error, although the results are not sensitive to this restriction. See Appendix B for more details of the method.

we account for the fixed effects of the original province and destination county separately. For most observables, inclusion of the fixed effects reduces the associations by more than 50 percent. Some correlations even turn to be negative. Column (4) shows that many of the correlations turn out to be statistically insignificant. The results imply that the amount of sorting on observables is driven primarily by factors of the common places of origin or destination[25], whereas the amount of sorting due to factors related to the same pair of origin and destination places is actually small, albeit not exactly equal to zero. Columns (5) and (6) suggest that the amount of sorting reduces even further when we also control for the region-of-origin-by-destination-county fixed effects.

To assess the importance of sorting within the same destination-origin pair in explaining the relationship between language similarity and labor market outcomes, Table A2 in Appendix E examines the associations between the dialect distance and the average characteristics of other migrants with the same places of origin and destination. Panel A shows that the correlations are not significantly different from 0 for most observables in regressions separately controlling for the destination-county fixed effects and the original-province fixed effects, except that migrants are more likely to be surrounded by better educated fellow townsmen in destination places with a dialect more similar to their home dialect. Panel B shows similar results when we further control for the region-of-origin-by-destination-county fixed effects. The association between dialect distance and average education turns out to be only marginally significant at the level of 10%.

To investigate whether the remaining within-destination-origin-pair sorting on observables has any significant impacts on migrants' labor market outcomes, we further control for the average characteristics of migrants with the same destination county and home province based on equation (1). As seen at the bottom of Table A3 in Appendix E, the *p*-values of the joint significance test reveal that the average group characteristics do not significantly predict any of the labor market outcomes except for the propensity of working over eight hours per day. These results strongly support our identification assumption of the exclusion restriction under the situation of a small amount of sorting within the pair of origin and destination places

[25] For instance, people from some areas may have the comparative advantage required in a certain industry and thus are more likely to migrate; or some cities may be more attractive to migrants because of less restrictions in obtaining a local *hukou*.

that exists in the data. In the analysis below, we provide additional examinations on the validity of the instrument variable.

6. Results

6.1 Main Results

Table 2 reports the OLS regression results of a variety of labor market outcomes on the dummy variable of identification with the host community, as specified in equation (1). Column (1) shows that the correlation between monthly income and feeling local is positive but not statistically significant. The magnitude of the estimate is small compared to the mean of monthly income. This is consistent with Constant and Zimmermann (2009), who also found no significant correlation between ethnic identity and earnings of immigrant workers in Germany. Column (2) shows that assimilated migrants have higher hourly wages. On average, the hourly wage of the socially assimilated group is 0.66 *yuan* higher than that of the unassimilated group given other factors fixed. Columns (3) to (5) report a negative and significant relationship between working time and commitment to the host community. On average, adaptation to the local identity is associated with a decrease of 1.48 hours in the working time every week. Columns (6) to (8) show a negative and significant association between the likelihood of overworking and identification with the host community. For instance, the feeling of belonging to local citizens is associated with a four percentage point decrease in the probability of working over 40 hours per week.

We proceed with the two-stage least squares (2SLS) estimation of the impact of social identity on the labor market outcomes from equations (1) and (2) and report the results in Panel A of Table 3. Column (1) displays a strong first-stage relationship between commitment to the host place and dialect distance in our sample. The greater the dialect distance between the original place and the destination, the lower the likelihood of holding identity committed to the host place. The point estimate indicates that if a migrant who originally moved within the same dialect sub-group chooses to move outside the dialect sub-group, the probability of affiliating with the host place would be 10 percentage points lower. The first-stage Kleibergen-Paap *F*-statistic of the instrumental variable is 52.3, which is by far greater than the conventional

critical value (i.e., 10), suggesting no weak instrumental variable problem.[26] Column (2) shows that migrants' monthly income increases as they adapt to local identity. However, the coefficient is economically and statistically insignificant. Column (3) shows that identification with the host community increases migrants' hourly wages, whereas the coefficient is marginally significant (t=1.56). The magnitude of the estimate is economically sizeable. On average, feeling assimilated can increase migrants' hourly wages by 3.25 *yuan*, which represents about a 24-percentage point increase above the 13.8-*yuan* baseline hourly wage.

In Columns (4) to (6), we examine whether social identity would change migrants' working time. As shown, commitment to the host community significantly reduces average working time per week by about nine hours. The results in Columns (7) to (9) show that affinity with the host place significantly reduces the probability of overworking for migrants as well. Specifically, identification with the host community reduces the probability of overworking beyond the regular eight hours a day by 44 percentage points. This is large, compared with the average rate of overworking (i.e., 48 percentage points). In other words, identifying with the host place almost solely eliminates the likelihood of migrants' overworking on a daily basis. The other estimates indicate that social identity reduces the probability of working for over five days per week by 22 percentage points, and reduces the probability of working for more than 40 hours a week by 26 percentage points.

A comparison of the OLS and IV estimates suggests that the magnitude of the IV estimates of the impact of social identity on working time is smaller (i.e., more negative) than the magnitude of the OLS estimates. One explanation is that the OLS estimates are biased due to omitted variables such as migrants' ambition of integrating into the host community and the hostility among locals. As evidenced by Jaschke et al. (2022), a more hostile environment of local community may induce faster cultural convergence of migrants, but may also hinder their economic assimilation. Moreover, a possible measurement error of local identity may also bias the OLS estimates towards zero. Lastly, the IV estimator identifies the average treatment effect for the compliers, namely, migrants whose local identity was affected by the similarity between

[26] The inferences of the second-stage estimates remain robust to the use of the adjusted critical value from Lee et al. (2022). Specifically, given the first-stage F-statistic is 52.3, the corresponding critical value $\sqrt{c_{0.05}(F)}$ is between 2.099 and 2.147 according to Table 3A of Lee et al. (2022). It is well below $|t|$ for all outcome variables with significant IV estimates, except for the indicator of working more than five days per week ($|t|$=2.118).

dialects at the home and destination places. That might be different from the average treatment effect among non-compliers.

Panel B of Table 3 reports the reduced-form estimates of the relationship between dialect distance and labor market outcomes. As expected, larger distance between the dialects of the home province and that of the destination county is associated with lower hourly wage and more hours worked per day (or per week), although the former is statistically insignificant. The results also indicate large dialect distance is associated with a higher likelihood of being overworked, whereas its association with monthly earnings is economically and statistically insignificant. Overall, these results are consistent with the first- and second-stage results of the IV estimates reported in Panel A.

6.2 Validity of Identification

Our IV strategy rests on the assumption that the dialect distance between the host and home places affects migrants' labor market outcomes only through self-identification with the host community, conditional on the set of control variables and fixed effects. One main concern regarding the assumption is that the instrumental variable may affect labor market outcomes through the communication effect of dialect. To address this concern, in equations (1) and (2), we further control for two dummies that indicate whether the migrants can speak the local dialect and whether they can understand it.[27] The results are reported in Panel A of Table 4. As shown, the coefficients of the two dummies are not significantly different from 0 in regressions of all labor market outcomes, except that migrants who speak the local dialect are more likely to work over 40 hours a week. This may not be surprising given that most migrants may have no difficulty to communicate in the workplace using Mandarin Chinese. By controlling for the communication effect of dialect, the impacts of social identity are actually similar to those in Panel A of Table 3. Identification with the host community reduces the hours worked per day or the hours worked per week. It also reduces the likelihood of overworking. If anything, the results are even stronger than the benchmark estimates.

[27] The ability to speak or understand the local dialect was measured when the survey was conducted. In Appendix D.1, we assess the potential importance of knowledge of the local dialect on arriving in the host region, and provide evidence consistent with the hypothesis that knowledge of the local dialect upon migration does not play a significant role in the relationship between dialect distance and migrants' labor market outcomes.

In the above examination, we assume the individual skills using the local dialect are exogenous by treating them as control variables. However, people may intentionally acquire skills of local dialect to achieve better labor market outcomes. Although we do not observe a significant association between language skills and labor market outcomes in most cases, the possibility of endogenous controls may still remain.[28] To further assess the role of the communication effect, we examine the heterogeneous relationships between dialect distance and labor market outcomes by proximity to *Putonghua* of the dialect at the destination place. The idea is that, if the communication effect is indeed important in affecting labor market outcomes, we would expect the effect to be more salient in places where the dialect is more different from *Putonghua*. In the sample, the dialects at Xi'an, Xianyang, and Wuhan prefectures belong to the same dialectal super-group as *Putonghua* (i.e., *Guanhua* or Mandarin), whereas the dialects at the other five prefectures belong to different dialectal super-groups.[29] Thus, we define a dummy which equals one if migrants were at one of the other five prefectures with larger dialect distance to *Putonghua* and add the interaction term of the dummy with the linguistic distance between dialects at the home province and destination county in the reduced-form regressions.

Panel B of Table 4 reports the OLS estimation results. For migrants residing in prefectures where their dialects are similar to *Putonghua*, the communication effect should be small. However, we still observe significant associations between the dialect distance and labor market outcomes on working time. Actually, the estimates are quite similar to those reported in Panel B of Table 3. Furthermore, the estimated coefficients of the interaction term indicate no significant heterogeneity along the proximity of dialect at the destination city to *Putonghua*, suggesting that the communication effect of dialect is unlikely to be an important channel through which the dialect distance affects migrants' labor market outcomes. Overall, the results in Table 4 should reduce concerns about possible violation of the exclusion restriction due to the commutation effect.

[28] For this consideration, we do not control for the variables of language skills in the baseline specification.

[29] The *Putonghua* was established as Standard Chinese on the basis of a dialect spoken at Luanping county of Chengde prefecture near Beijing. It belongs to Beijing *Guanhua*, one particular dialectal group of the dialectal super-group of *Guanhua*. The dialects spoken at Xi'an, Xianyang, and Wuhan all belong to the same dialectal super-group (i.e., *Guanhua*) as *Putonghua*. In contrast, the dialects spoken at Songjiang, Suzhou, and Wuxi belong to the dialectal super-group of *Wu*, whereas the dialects spoken at Quanzhou and Changsha belong to the dialectal super-group of *Min* and *Xiang*, respectively.

Another concern of identification is that the dialect distance is correlated with some bilateral factors that may affect the labor market performance of migrants. To address such concern, Table 5 further conducts a battery of robustness checks on the IV estimates based on alternative specifications of equation (1). Panel A controls for region-of-origin-by-destination-county fixed effects, where provinces are classified into six regions (i.e., North China, Northeastern China, East China, Central China, Southwest China, and Northwest China) according to the National Bureau of Statistics.[30] The fixed effects may absorb attitude biases (e.g., trust or discrimination) of local residents in some county toward migrants from a specific region (Guiso et al., 2009). As shown, the estimates do not change much compared with the benchmark results.[31]

To account for geographic distance, which may relate to both dialect distance and the labor market outcomes of migrants, Panel B controls for the log of transportation distance from the administrative center of the home province to the destination county. The results show that the estimates are actually quite similar to the basic results.

Migrants from linguistically less distant provinces may have a larger number of migrant peers from the same province in a particular destination county. Consequently, they may have a better chance of success in the local labor market. To account for such a possibility, Panel C controls for the log of the number of migrants from the same province in the destination county by using data from the population census in 2010. The results suggest that the impacts of social identity on work time are even stronger and that the estimate of the impact on the hourly wage turns out to be significant at the level of 10%, although the estimated impact on the likelihood of working over five days per week is marginally significant (t=1.55). Since the stock of migrants may reflect bilateral connections in a broader sense, which include across-region links caused by political events such as the send-down movement (Kinnan et al., 2018), the results above should also reduce concerns of such bilateral connections.[32]

Relatedly, people may worry that our instrumental variable may affect not only the

[30] See https://data.stats.gov.cn/english/easyquery.htm?cn=E0101 for details of the classification.
[31] To further address the concern about dialect-based discrimination, we control for the type of dialects spoken in the province of origin interacted with the destination-county dummies. The results are also quite similar to the benchmark estimates. See Appendix D.2 for details.
[32] Appendix D.3 presents more evidence which suggest that social connections in the host region prior to migration is unlikely to threaten the exclusion restriction.

migrants' performance on the local labor market, but also their sorting into the local labor market, which may go through factors independent of social identity. As discussed in Section 5, sorting within the pair of destination county and original province is much less extensive than sorting into a certain destination county or from a certain original province in the data. Meanwhile, an examination of the remaining sorting on observable attributes within origin-destination pairs indicates that they are neither significantly correlated with the instrumental variable nor are important determinants of labor market outcomes. Furthermore, the IV estimates controlling for the average characteristics of fellow townsmen residing in the same destination county are quite similar to the benchmark estimates (see Table A4 in Appendix E). Although we can only examine sorting on the basis of observables, it can be informative of the potential sorting of unobservables (Altonji et al., 2005; Oster, 2019). The above results imply that the exclusion restriction of the instrumental variable with respect to sorting on unobservables is likely to be a reasonable assumption.

To further address the concern of potential within-origin-destination-pair sorting on unobservable factors, in Panel D of Table 5, we control for the wage differentials between the places of origin and destination, the primary determinants of sorting across locations in migration choices. Specifically, we construct prefecture-level average wages by using data from the population census in 2005, and then weight them by population of the original prefecture to get the measure of the gap between wages at the original province and destination prefecture. As shown, the results are quite similar to the benchmark estimates, indicating that our instrumental variable estimates are unlikely confounded by wage differentials across regions.[33]

Finally, to further validate the exclusion restriction, we perform a falsification test. Specifically, we examine the possible direct effect of the dialect distance on migrants' labor market outcomes by estimating reduced-form regressions in the sample of new migrants who resided in the destination county for less than, or equal to, six months, exploiting the fact that integration with local networks (the main mechanism evidenced below) takes time and the

[33] The estimates remain robust if we also control for region-of-origin-by-destination-county fixed effects, the log of transportation distance, the log of the number of migrants from the same province in the destination county in 2010, and the average characteristics of fellow townsmen residing in the same destination county. The results are available upon request.

beneficial effect of self-identification with the host community on labor market outcomes is very unlikely to occur among newly arrived migrants. Panel A of Table 6 reports the results. As shown, none of the labor market outcomes is significantly associated with dialect distance. Further, the magnitude of the estimates is generally quite small. In contrast, for migrants who resided in the destination county for more than six months, the reduced-form associations between dialect distance and labor market outcomes are economically and statistically significant in most cases.

These results strongly support the identification assumption. In particular, the results should further reduce the concerns related to the communication effect of dialect and sorting in the destination choice of migrants. Both effects should appear among the new migrants if they are indeed salient. However, we do not find any evidence of these effects from the falsification test. Table 6 also shows that for both groups of migrants commitment to the local place is significantly negatively correlated with linguistic distance between dialects of the home and host places. The association is stronger for new migrants than those who have resided in the host county for more than half a year. These results are consistent with the understanding that our IV estimation explores variation in pre-determined social identity that is caused by the cultural difference between the home and host places and that the association can be attenuated when social identity evolves after the migrants arrive in the host place.[34]

6.3 Plausible Exogeneity of the Instrumental Variable

While the above results show no clear evidence of the violation of the exclusion restriction, we examine the sensitivity of our results when the instrumental variable is only plausibly exogenous by using the method developed by Conley et al. (2012), which provides unbiased IV estimates in situations where the exclusion restriction of the instrumental variable does not hold precisely. Specifically, consider a generalization of our second-stage equation

$$y_{icp} = \alpha_0 + \alpha_1 Identity_{icp} + \gamma Dialect\ distance_{icp} + \alpha X_{icp} + \delta_c + \delta_p + \varepsilon_{icp},$$

where γ captures the direct effect of dialect distance on migrants' labor market outcomes other

[34] In Appendix C, we use surname distance (a measure of genealogical relatedness) between host and home provinces as an alternative instrumental variable. The point estimates of the IV regressions are largely comparable to the benchmark results in sign and magnitude, although they are less statistically significant.

than effects through the channel of their commitment to the host place. Given γ, we can obtain an unbiased IV estimate of α_1 from the modified equation

$$\widehat{y_{icp}} = \alpha_0 + \alpha_1 Identity_{icp} + \alpha X_{icp} + \delta_c + \delta_p + \varepsilon_{icp},$$

where $\widehat{y_{icp}} \equiv y_{icp} - \gamma Dialect\ distance_{icp}$.

Following the idea presented in the falsification test above, we estimate γ by conducting a reduced-form regression in the sample of new migrants who resided in the destination county for no more than six months. As a practical manner, we construct a summary index consisting of the average Z-score of all outcomes variables with significant benchmark results as shown in Columns (6) to (9) in Panel A of Table 3 in the main text.

The coefficient of γ in the regression on the summary index is estimated to be negative (i.e., -0.037) and statistically insignificant (the 90% confidence interval is [-0.111, 0.036]). These results imply that the true effect of social identity on the summary index of working time is actually more negative (i.e., a stronger effect) than the benchmark IV estimate which is -0.696 (p=0.002) if $\gamma = -0.037$. Applying the method of Conley et al. (2012), Figure A2 in Appendix E illustrates the 90 percent confidence interval boundaries for IV estimates of the effect of social identity on the summary index when we assume the value of γ varies on the interval [-0.111, 0.036]. As shown, we are still able to confirm a significantly beneficial effect of identification with the host community on labor market outcomes (i.e., less working time) even when we allow for a plausible amount of imperfect exogeneity of the instrumental variable. Actually, for the 90 percent confidence interval for the IV estimate to include 0, γ must be greater than 0.036. This possibility is only 0.05 according to the estimates of γ reported above. In other words, the probability of a violation of the exclusion restriction that would make the results insignificant at the 10%-level is only 5 percent. Overall, the exercise suggests our conclusion from the main results is robust to possible deviations from the perfect exogeneity assumption.

6.4 Selection Bias

One main concern of the sample construction is the post-migration selection. That is, if the unassimilated migrants are more likely to leave the host places and are also more likely to

be unsuccessful in the labor market, then our estimates based on the sample of migrants who remain staying at the host places can be biased.[35]

To assess the extent to which post-migration selection can bias the estimates, we utilize the time level variation in return migration and examine the heterogeneity of the impact of local identity on migrants' labor market outcomes. Specifically, we exploit the fact that the financial crisis, which occurred around 2008 and 2009, substantially increased the return of internal migrants in China (Giles et al., 2013). We expect the problem of sample selection to be more severe for migrants in our sample who arrived in the host places before the financial crisis, whereas the issue of post-migration selection is less severe for migrants who arrived in the host places after the crisis. We conduct subsample analyses based on whether the migrants arrived in the host places before or after the crisis.

Table A4 in Appendix E shows that the beneficial effects of local identity on migrants' labor market outcomes exist mainly among the sample of migrants who arrived in the host places after the financial crisis, whereas for migrants who arrived in the host places before the financial crisis, the impacts are statistically insignificant for most outcomes. The results imply that our benchmark findings of the beneficial effects of local identity on migrants' labor market outcomes are unlikely to be driven largely by post-migration selection. If anything, this kind of selection may lead to an under-estimation of the effects.

Another concern is that, in the main analyses, we exclude employers and self-employed migrants due to the difficulty of separating their labor earnings from capital returns. To address potential bias caused by sample selection, we use Heckman's selection model to take into account of this. Specifically, we use the indicators of participating in social insurance programs in hometowns as instrumental variables for the selection indicator, namely, being an employee in destination places.[36] The estimation results suggest that the indicators of participating in social insurance programs in hometowns are negatively correlated with being an employee (*F*-statistic=45.9), whereas they are unlikely to directly impact labor earnings and work time at

[35] Using survey data from sending areas could identify return migrants, but the problem of such data is that they typically do not contain identity questions about the host and home areas.

[36] The indicators include whether the migrants were participating in the following social insurance programs in their hometowns: the New Rural Cooperative Medical Scheme, the Medical Insurance Scheme for Urban Workers, the Medical Insurance Scheme for Urban Residents, the Urban Pension Insurance Scheme, and the Rural Pension Insurance Scheme.

the migration destinations. We then control for the inverse Mills ratio predicted from the selection model to account for potential bias caused by the sample selection. Table A5 in Appendix E reports the results. As shown, the estimates are nearly the same as the benchmark results, except that the impact of social identity on hourly wage turns out to be significant at the level of 5% after correcting for selection bias.

One related issue is that we exclude unemployed migrants in the analyses because we cannot observe their wage and working hours. The unemployment rate is very low among migrants—only 1% in our sample. Meanwhile, Table A6 in Appendix E shows that there is no significant impact of affiliating with the host community on the likelihood of being unemployed. These findings should alleviate the concern of a possible selection bias caused by excluding unemployed migrants in the analyses.

Overall, the results are robust in a battery of alternative specifications. This should reduce concerns related to the potential violation of the exclusion condition of our estimation strategy and sample selection in the main analyses.

6.5 Mechanisms

To investigate the mechanisms of how social identity affects labor market outcomes, following the conceptual framework, we consider the effects of identity on migrants' social network and choice of residence. In addition, we explore the connections migrants use during job search, which are important for obtaining higher-quality jobs.

We first investigate the effects of social identity on migrants' network and neighborhood choice. Columns (1) and (2) in Table 7 show that identification with the host community significantly increases the probability of interacting with locals, whereas it reduces the probability of interacting with people from migrants' place of origin. Identification with the host community also lowers the chance of participating in the activities of ethnic organizations, although this result is not statistically significant. Columns (4) to (6) report the estimated effects of social identity on migrants' choice of residence. We find that commitment to the host community significantly increases the probability of having local neighbors by 21 percentage points and reduces the probability of having non-local neighbors by 37 percentage points.

These results indicate that socially assimilated migrants are more likely to interact with local citizens and are less likely to interact with people from their hometown. They are also more likely to live in a community with mostly local citizens and less likely to live with non-local citizens.

Social interactions and residence choice may play an important role in information diffusion and labor market outcomes (Bayer et al., 2008; Bollinger et al., 2020). To examine such a channel, Table 8 examines the impacts of social identity on migrants' job search. We find that identification with the host community significantly increases the probability of finding jobs through local people by 17 percentage points. Migrants are also more likely to find a job self-dependently. However, socially assimilated migrants are less likely to find a job through family members, relatives, or friends. These results indicate that networks with local citizens are an important channel through which migrants can obtain high-quality jobs for reasons such as alleviating information friction in the job-searching process (Abel et al., 2020).[37]

To explore the extent to which the benefits of identification with the host community for labor market outcomes are through job attainment, in equation (1) we further control for a vector of dummies indicating types of occupation, industry, and work unit of the migrants to account for potential labor market segmentation (Wang and Conesa, 2022). Table A8 in Appendix E reports the IV estimates. The coefficients on work time and propensity to overwork are still negative and significant, although the magnitudes are smaller in absolute values than those in Panel A of Table 3. These results indicate that occupation, industry, and type of work unit can partially explain the impact of identification with the host community on reducing work time, likely through the beneficial effects of networking with local citizens. The significantly higher hourly wage and lower likelihood of overwork of the assimilated migrants conditional on the job characteristics indicate the possibility of having substantial differences

[37] Table A7 in Appendix E assesses the exclusion condition of the IV regressions on the variables of migrants' behaviors by conducting falsification tests similar to that in Table 6. The results suggest that the reduced-form association between dialect distance and the behavioral variables is not significantly different from 0 for new migrants in most cases. One exception is that they were significantly more likely to interact with ethnic people if they came from a province with a larger linguistic distance from the dialect spoken at the destination county. These results are consistent with the conjecture that establishing relations with local people took much more time than making connections with ethnic networks. Therefore, for new migrants we may not observe their behavioral adaptation even if they identify with the host community. Overall, the results support the assumption of exclusion restriction for identification in the mechanism analysis.

in the quality of jobs even within the same occupation, industry, and type of work unit.

6.6 Heterogeneity Analysis

In this section, we examine the heterogeneity of the impacts of social identity on the labor market outcomes of migrants concerning gender, age, and the education level of the migrants. Specifically, we conduct subsample analysis by defining migrants are young if they are 30 years old or younger and defining migrants are highly educated if they have completed college education or above.

Table 9 presents the results of the heterogeneity analysis by estimating IV regressions separately for each subgroup. Column (1) reports the first-stage results. As shown, among every subgroup, the dialect distance significantly reduces the likelihood of holding an identity committed to the host place. Similar to the baseline results, Columns (2) to (9) show that identification with the host community significantly reduces the work hours of male migrants and increases their hourly wage. For females, the signs of the estimates are the same as those of males, but the estimates are economically and statistically insignificant. The younger migrants obtain significant beneficial effects of reducing work hours from affiliating with the host community. Identification with the host community also decreases work hours for older migrants, but the magnitude of the impact is about half of that on younger migrants, and is statistically insignificant. Older migrants benefit from affiliating with the host community mainly in terms of increasing their hourly wage by 5.26 *yuan*, which is statistically different from 0 at the 10% significance level. Finally, the results show that migrants with a low-level of education benefit more in terms of reducing work time from their commitment to the host community, whereas the effects on labor market outcomes are insignificantly different from 0 among migrants with a college education or above. This echoes the finding of Carillo et al. (2023) which show that the labor market effect of integration is higher for the lower educated immigrants in Italy.

Overall, the results of the heterogeneity analysis indicate that the beneficial effects of identification with the host community on labor market outcomes appear mainly among migrants who are male, 30 years old or younger, and have a high-school education or below.

In contrast, for migrants who are female, older than 30 years old, or have a college education or above, the effects are economically and statistically insignificant in most cases. The results imply that the former group should be the target of a possible integration policy.

7. Conclusions

This study examines the impact of the social identity of China's internal migrants on their labor market outcomes by exploring plausibly exogenous variation in identification with the host community captured by the dialect distance between the original and current place of residence. To deal with concerns related to violation of the exclusion restriction, we take into account of a possible communication effect of dialect on labor market outcomes and sorting in the choice of migration destinations. We also check the sensitivity of our results by relaxing the strict exogeneity assumption of the instrumental variable, using the method developed by Conley et al. (2012).

We find consistent evidence that identification with the host community increases migrants' hourly wages and reduces the average number of working hours and the likelihood of overworking. Specifically, the benchmark estimates suggest that the hourly wage increases by 3.25 *yuan*, or 24% of its mean, as migrants socially assimilate into the local place. Commitment to the host community also significantly reduces the average work time per day by 1.17 hours, and eliminates the likelihood of overwork on a daily basis.

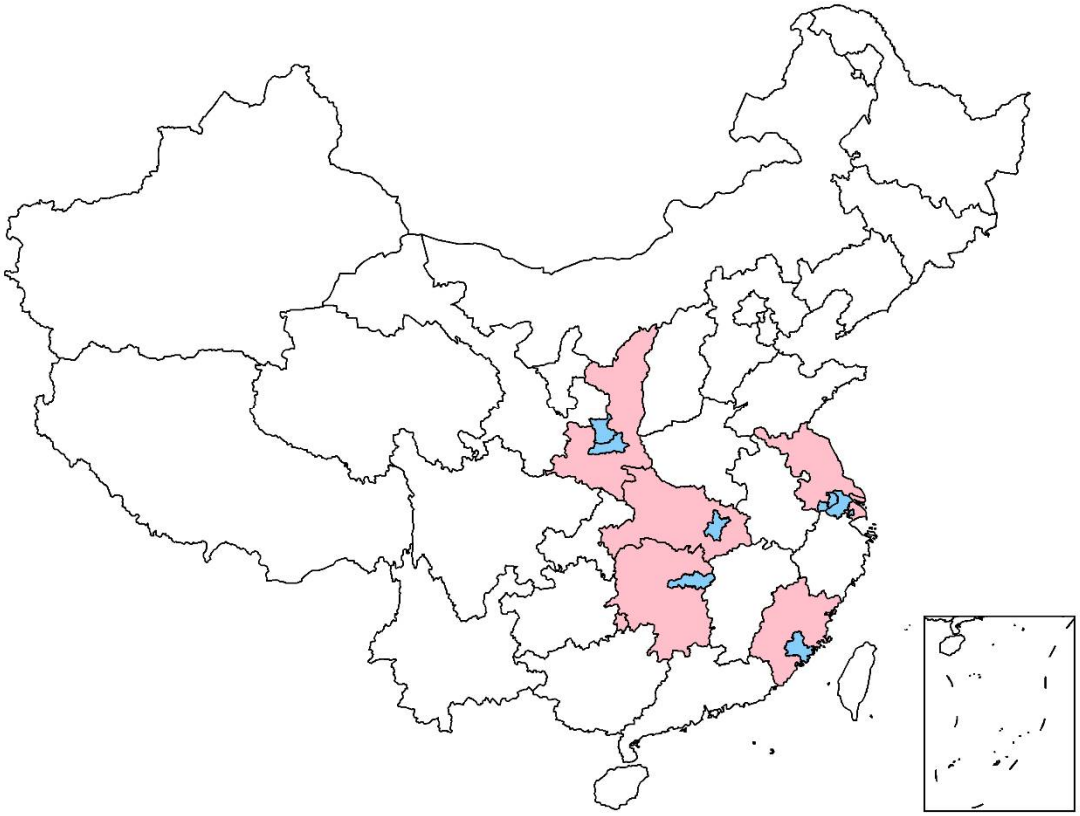
Further analyses of the mechanisms suggest that identification with the host community significantly raises the likelihood of interacting with locals and living in local neighborhoods. It also significantly increases the propensity of finding a job with help from the locals, likely through extended access to local networks. In line with these results, we find a reduction in the negative effect of affinity with the host community on working hours and the propensity of overwork by controlling for job characteristics, including types of occupation, industry, and work unit. However, the benefits of identification with the host community on labor market outcomes among migrants are still significant after accounting for these job characteristics, indicating possible differences in the quality of jobs even within the same occupation, industry, and type of work unit.

Our findings of the beneficial labor market impacts of identifying with the host community among internal migrants in China qualitatively align with other studies on international migrants. These studies found that ethnic identity is significantly associated with the likelihood of getting work and wages, but has no significant relationship with earnings among immigrants in developed countries, including Germany, Canada, Australia, and Italy (Constant and Zimmermann, 2009; Islam and Raschky, 2015; Piracha et al., 2023; Carillo et al., 2023). The similarity of the findings indicates that social identity is relevant for both international and internal migrants, although they face different challenges in the process of their migration. We highlight the role of social networks in explaining the beneficial impacts among internal migrants in China. This can be particularly relevant for developing countries, where public institutions are often much less effective. It is worthwhile for future research to examine whether the results also hold for internal migrants in other countries.

The findings of the study suggest that integration policies promoting migrants' identification with the host communities may benefit their economic assimilation as well, implying these policies can be valuable on a large scale. Furthermore, our heterogeneity analyses suggest that the integration policy should target vulnerable migrants such as those who have lower human capital and face larger barriers in integration.

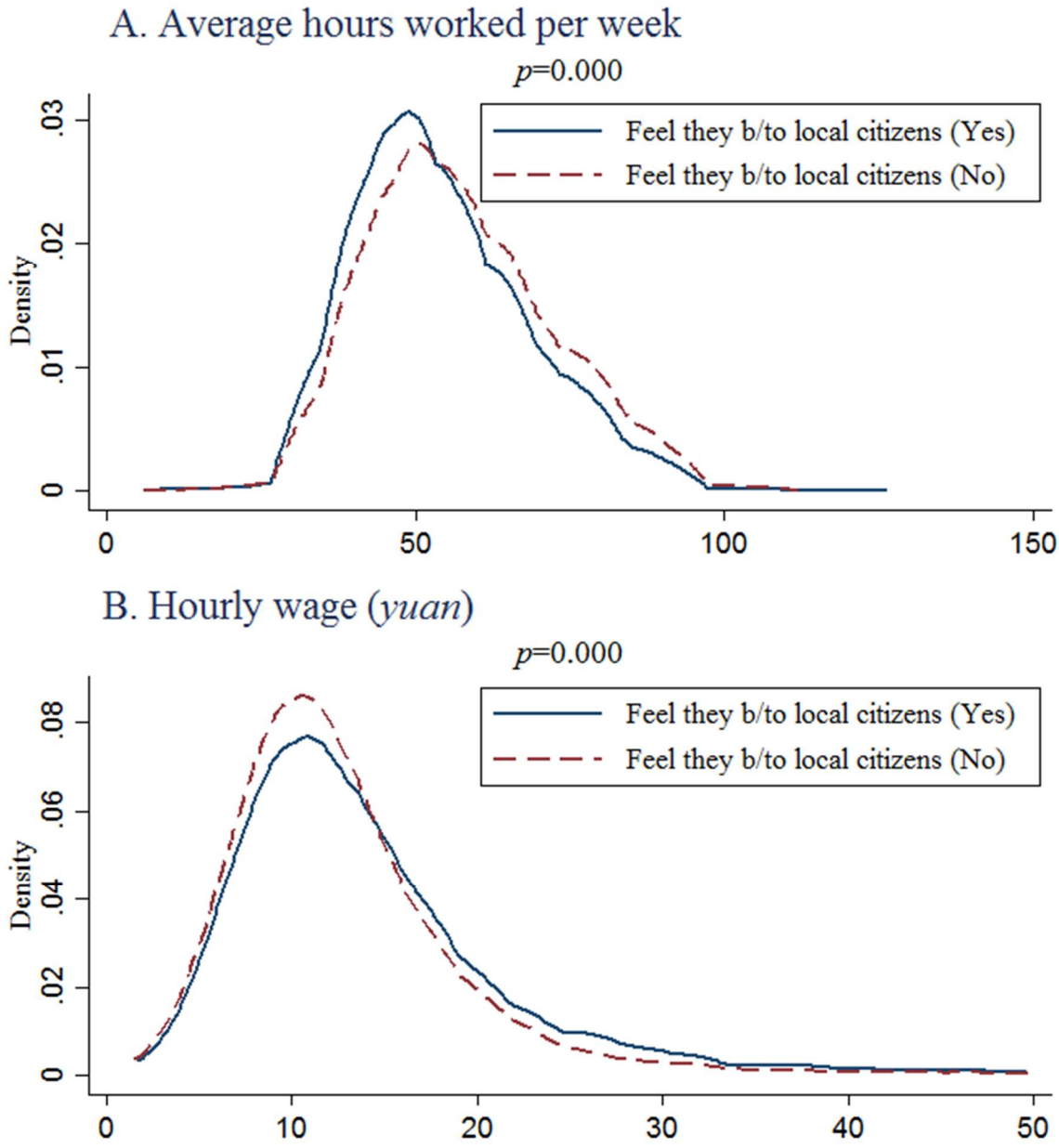
Figures and Tables

Figure 1. Sampled prefectures



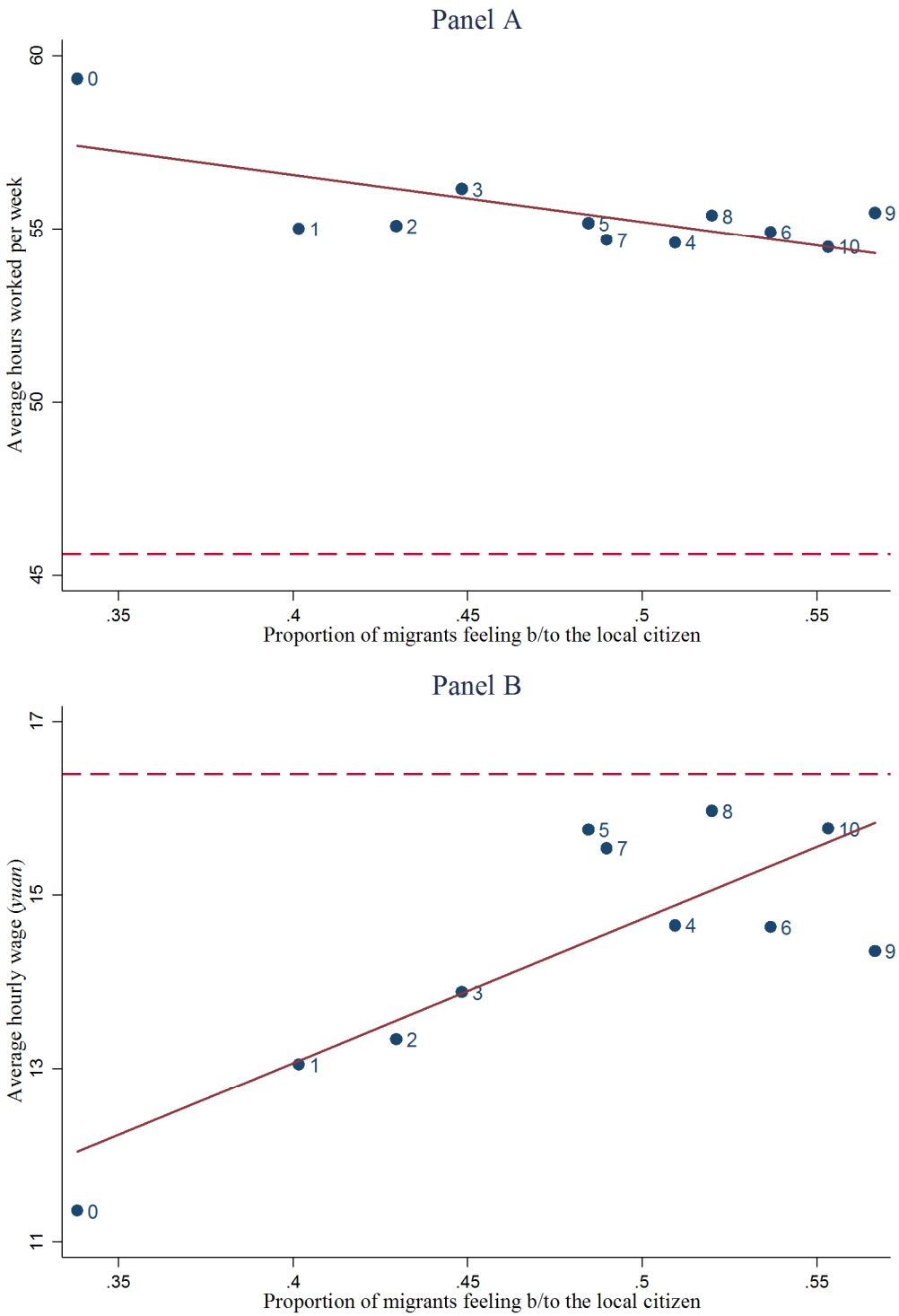
Notes: The figure illustrates the location of provinces (in pink) and prefectures (in blue) in the sample. These include the Songjiang district in Shanghai, the Suzhou and Wuxi prefectures in Jiangsu Province, the Quanzhou prefecture in Fujian Province, the Wuhan prefecture in Hubei Province, the Changsha prefecture in Hunan Province, and the Xi'an and Xianyang prefectures in Shan'xi Province.
Source: Own construction.

Figure 2. Distribution of labor market outcomes by identity



Notes: The figure plots the kernel density distribution of average hours worked per day (Panel A) and hourly wage (Panel B) for respondents who feel they belong to local citizens (solid line) and those who feel they do not belong to local citizens (dash line). The numbers above the charters are the corrected *p*-values of the two-sample Kolmogorov-Smirnov test for equality of distribution functions.
Source: Dynamic Monitoring Survey of the Migrant Population of China 2013.

Figure 3. Migrant social-economic assimilation line



Notes: The figure contains the data means for hourly wages, hours worked per week and the share of local social identity for the eleven year cohorts (0, 1, 2…, 10) of the migrants in our sample: social identity against hours worked in panel A, and against wages in panel B. The dash line in Panels A and B indicates the average hours worked per week and average hourly wage by the locals, respectively.
Source: Own construction based on data from the Dynamic Monitoring Survey of the Migrant Population of China 2013.

Table 1. Sample characteristics by identity

	Feel they belong to local citizens?		Difference	
	Yes	No	(1) - (2)	*p*-value
	(1)	(2)	(3)	(4)
Labor market outcomes				
Monthly income	3195.448	3016.731	178.717	0.000
Hourly wage	14.675	13.089	1.586	0.000
Average days worked per week	5.983	6.078	-0.095	0.000
Average hours worked per day	9.002	9.338	-0.336	0.000
Average hours worked per week	54.188	57.094	-2.906	0.000
Overwork (days per week > 5)	0.735	0.800	-0.064	0.000
Overwork (hours per day > 8)	0.417	0.526	-0.109	0.000
Overwork (hours per week > 40)	0.769	0.837	-0.068	0.000
Demographic characteristics				
Age	33.190	32.633	0.557	0.002
Male (yes=1)	0.546	0.551	-0.005	0.612
Never married (yes=1)	0.242	0.315	-0.073	0.000
Married one time (yes=1)	0.742	0.663	0.079	0.000
Married two or more times (yes=1)	0.007	0.007	-0.001	0.613
Divorced (yes=1)	0.007	0.012	-0.005	0.016
Widowed (yes=1)	0.002	0.003	-0.001	0.612
Education level below middle school	0.089	0.126	-0.037	0.000
Education level of middle school	0.552	0.613	-0.061	0.000
Education level of high school	0.186	0.163	0.023	0.003
Education level of college	0.169	0.096	0.073	0.000
Education level above college	0.004	0.002	0.002	0.040

Notes: The number of observations is 9,790, and 45% of the respondents feel they belong to local citizens. Column (1) reports the sample mean of migrants who feel they belong to local citizens, whereas Column (2) describes the sample mean of migrants who do not feel they belong to local citizens. Column (3) reports the difference in means between the two groups. The last column reports the *p*-value on testing the hypothesis that the difference is equal to 0.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013.

Table 2. OLS estimates of association between identity and labor market outcomes

			work time			overwork		
	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Feel they b/to local citizens (yes=1)	59.20	0.66***	-0.06***	-0.16***	-1.48***	-0.04***	-0.06***	-0.04***
	(40.80)	(0.23)	(0.02)	(0.05)	(0.39)	(0.01)	(0.01)	(0.01)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	3,097.18	13.80	6.04	9.19	55.78	0.77	0.48	0.81
Observations	9,761	9,761	9,790	9,790	9,790	9,790	9,790	9,790

Notes: The table reports the results from OLS regressions as specified in equation (1) in the text. The other control variables include age, age squared, dummy of male, marital status (including dummies of married once, married two or more times, divorced, and widowed), and education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013.

Table 3. IV estimation on the impact of identity on labor market outcomes

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: 2SLS estimates									
Feel they b/to local citizens (yes=1)		39.58	3.25	-0.23	-1.17***	-9.21**	-0.22**	-0.44***	-0.26**
		(400.30)	(2.09)	(0.21)	(0.44)	(3.76)	(0.11)	(0.15)	(0.11)
Dialect distance (0-1-2-3)	-0.10***								
	(0.01)								
KP *F*-statistic	52.31								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel B: Reduced-form estimates									
Dialect distance (0-1-2-3)		-4.06	-0.33	0.02	0.12***	0.94**	0.02**	0.04***	0.03**
		(41.27)	(0.21)	(0.02)	(0.04)	(0.36)	(0.01)	(0.01)	(0.01)
Control VARs		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome		3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations		9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780

Notes: Panel A reports the results from IV regressions as specified in equations (1) and (2) in the text. Panel B reports the OLS estimates of the reduced-form relationships between dialect distance and labor market outcomes. The other control variables are the same as those in Table 2. "KP *F*-statistic" denotes the cluster-robust Kleibergen-Paap (KP) *F*-statistic on testing weak instruments. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table 4. Examination of communication effect

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: 2SLS estimates									
Feel they b/to local citizens (yes=1)		-31.37	4.29	-0.29	-1.54**	-12.01*	-0.31*	-0.58**	-0.37**
		(618.83)	(3.31)	(0.34)	(0.74)	(6.28)	(0.18)	(0.25)	(0.18)
Can speak local dialect (yes=1)	0.15***	28.33	-0.55	0.04	0.22	1.77	0.04	0.07	0.06*
	(0.02)	(106.75)	(0.58)	(0.06)	(0.14)	(1.14)	(0.03)	(0.05)	(0.03)
Can understand local dialect (yes=1)	0.11***	23.40	-0.12	-0.02	0.01	-0.17	0.02	0.02	0.01
	(0.02)	(81.64)	(0.46)	(0.05)	(0.11)	(0.93)	(0.03)	(0.04)	(0.03)
Dialect distance (0-1-2-3)	-0.06***								
	(0.01)								
KP *F*-statistic	21.08								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel B: Reduced-form estimates									
Dialect distance (0-1-2-3)		20.16	-0.41	0.05	0.10	1.02*	0.04**	0.05**	0.04**
		(59.32)	(0.31)	(0.03)	(0.07)	(0.58)	(0.02)	(0.02)	(0.02)
Dialect distance (0-1-2-3) × City with		-40.70	0.13	-0.04	0.02	-0.14	-0.02	-0.01	-0.01
larger dialect distance to *Putonghua*		(82.09)	(0.45)	(0.05)	(0.09)	(0.80)	(0.02)	(0.03)	(0.02)
Control VARs		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE		Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome		3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations		9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780

Notes: Panel A reports the results of IV estimation of regressions which further controls for dummies indicating whether the migrants can speak the local dialect and whether they can understand the local dialect on the basis of specification in Panel A of Table 3. Panel B reports the OLS estimates of the heterogeneous relationships between dialect distance and labor market outcomes by proximity to *Putonghua* of dialect at destination city. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table 5. Robustness checks

			work time			overwork		
	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Control for region-of-origin-by-destination-county FE								
Feel they b/to local citizens (yes=1)	623.66	5.26	-0.42	-1.12**	-11.19**	-0.30**	-0.48**	-0.34**
	(779.30)	(3.70)	(0.27)	(0.57)	(4.97)	(0.15)	(0.20)	(0.15)
Observations	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel B: Control for log of transportation distance								
Feel they b/to local citizens (yes=1)	227.22	3.75	-0.35	-1.14**	-10.69**	-0.34**	-0.43**	-0.39***
	(574.41)	(3.07)	(0.27)	(0.57)	(4.90)	(0.14)	(0.19)	(0.15)
Observations	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel C: Control for log of migrants from the same province								
Feel they b/to local citizens (yes=1)	241.73	4.40*	-0.23	-1.49***	-11.65***	-0.17	-0.47***	-0.19*
	(455.17)	(2.46)	(0.22)	(0.51)	(4.15)	(0.11)	(0.16)	(0.11)
Observations	9,615	9,615	9,643	9,643	9,643	9,643	9,643	9,643
Panel D: Control for wage gap								
Feel they b/to local citizens (yes=1)	331.14	4.62*	-0.30	-1.05**	-9.21**	-0.31**	-0.37**	-0.31**
	(451.13)	(2.53)	(0.26)	(0.50)	(4.49)	(0.13)	(0.17)	(0.13)
Observations	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports the IV estimates of alternative specifications for equation (1) in the text. Panel A controls for the region-of-origin-by-destination-county fixed effects. Panel B controls for the log of transportation distance from the administrative center of the home province to the destination county. Panel C controls for the log of the number of migrants from the same province in the destination county. Panel D controls for the wage gap between the original province and the destination prefecture. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, population census 2000, population census 2010 (number of migrants in Panel C), population census in 2005 (wage in Panel D), and own construction (transportation distance in Panel B).

Table 6. Falsification test

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Subsample—Years since arrival less than or equal to half a year									
Dialect distance (0-1-2-3)	-0.16***	-45.53	-0.08	0.01	-0.07	-0.26	-0.01	-0.02	-0.02
	(0.03)	(91.25)	(0.42)	(0.05)	(0.10)	(0.87)	(0.02)	(0.03)	(0.02)
KP *F*-statistic	32.35								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.33	2,746.92	11.47	6.15	9.52	58.94	0.85	0.59	0.88
Observations	1,426	1,421	1,421	1,426	1,426	1,426	1,426	1,426	1,426
Panel B: Subsample—Years since arrival more than half a year									
Dialect distance (0-1-2-3)	-0.09***	-10.79	-0.37*	0.02	0.13***	0.98**	0.03**	0.05***	0.03***
	(0.02)	(42.67)	(0.23)	(0.02)	(0.05)	(0.39)	(0.01)	(0.02)	(0.01)
KP *F*-statistic	34.99								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.47	3,156.92	14.2	6.02	9.13	55.25	0.76	0.46	0.79
Observations	8,354	8,330	8,330	8,354	8,354	8,354	8,354	8,354	8,354

Notes: Panel A reports the OLS estimates of the first-stage (reduced-form) relationship(s) between dialect distance and identity (labor market outcomes) using a subsample of migrants who stayed in the city for less than, or equal to, half a year, whereas Panel B reports the OLS estimates using a subsample of migrants who stayed in the city for more than half a year. The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table 7. IV estimation on the influence of identity on network and neighborhood choice

Outcome variables	Networks			Neighborhood		
	interact with ethnic people (yes=1)	interact with local people (yes=1)	member of ethnic organization (yes=1)	neighbors are mostly local citizens (yes=1)	neighbors are mostly non-local citizens (yes=1)	the number of local and non-local neighbors is similar (yes=1)
	(1)	(2)	(3)	(4)	(5)	(6)
Feel they b/to local citizens (yes=1)	-0.12*	0.52***	-0.08	0.21*	-0.37***	0.16
	(0.06)	(0.14)	(0.12)	(0.12)	(0.13)	(0.13)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes
Original Province FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.93	0.42	0.12	0.23	0.49	0.28
Observations	9,780	9,780	9,780	9,405	9,405	9,405

Notes: The table reports the IV estimates on migrants' network and neighborhood choice, using the dialect distance between the original and destination places as an instrumental variable for the sense of belonging to the local citizens. The other control variables include age, age squared, dummy of male, marital status (including dummies of married once, married two or more times, divorced, and widowed), and education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table 8. IV estimation on the influence of identity on job search

Outcome variables	find the job through family/relatives or friends/classmates (yes=1)	find the job on their own, or start a business on their own (yes=1)	find the job through local people (yes=1)	find the job through government, social agency, internet, job fair, and others (yes=1)
	(1)	(2)	(3)	(4)
Feel they b/to local citizens (yes=1)	-0.44***	0.21*	0.17**	0.06
	(0.15)	(0.12)	(0.07)	(0.10)
Control VARs	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes
Original Province FE	Yes	Yes	Yes	Yes
Mean of outcome	0.47	0.31	0.06	0.16
Observations	9,774	9,774	9,774	9,774

Notes: The table reports the IV estimates on migrants' job search, using the dialect distance between the original and destination places as an instrumental variable for the sense of belonging to the local citizens. The other control variables include age, age squared, dummy of male, marital status (including dummies of married once, married two or more times, divorced, and widowed), and education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table 9. IV estimation of the heterogeneous impact of identity on labor market outcomes

	first-stage	seconod-stage							
				work time			overwork		
	feel they b/to local citizens (yes=1)	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Male sample	-0.10***	-14.56	4.67*	-0.35	-1.97***	-14.30***	-0.28**	-0.62***	-0.33**
	(0.02)	(524.50)	(2.75)	(0.28)	(0.63)	(5.36)	(0.14)	(0.22)	(0.14)
KP *F*-statistic	33.93	34.56	34.56	33.93	33.93	33.93	33.93	33.93	33.93
Observations	5,364	5,348	5,348	5,364	5,364	5,364	5,364	5,364	5,364
Female sample	-0.10***	64.50	1.25	-0.14	-0.04	-2.53	-0.15	-0.16	-0.16
	(0.02)	(464.05)	(2.51)	(0.29)	(0.54)	(4.64)	(0.16)	(0.17)	(0.16)
KP *F*-statistic	30.55	30.80	30.80	30.55	30.55	30.55	30.55	30.55	30.55
Observations	4,416	4,403	4,403	4,416	4,416	4,416	4,416	4,416	4,416
Younger sample	-0.12***	-333.81	1.71	-0.30	-1.63***	-12.78***	-0.28**	-0.63***	-0.25**
	(0.02)	(453.62)	(2.27)	(0.27)	(0.57)	(4.90)	(0.13)	(0.19)	(0.12)
KP *F*-statistic	33.98	34.69	34.69	33.98	33.98	33.98	33.98	33.98	33.98
Observations	4,603	4,592	4,592	4,603	4,603	4,603	4,603	4,603	4,603
Older sample	-0.09***	476.62	5.26*	-0.20	-0.76	-6.17	-0.18	-0.21	-0.28*
	(0.02)	(572.89)	(2.98)	(0.31)	(0.61)	(5.41)	(0.16)	(0.20)	(0.17)
KP *F*-statistic	23.40	23.40	23.40	23.40	23.40	23.40	23.40	23.40	23.40
Observations	5,177	5,159	5,159	5,177	5,177	5,177	5,177	5,177	5,177
Low-education sample	-0.09***	90.20	3.36	-0.27	-1.46***	-11.79**	-0.24*	-0.51***	-0.25**
	(0.02)	(449.68)	(2.27)	(0.24)	(0.55)	(4.72)	(0.12)	(0.19)	(0.12)
KP *F*-statistic	33.38	33.74	33.74	33.38	33.38	33.38	33.38	33.38	33.38
Observations	8,494	8,471	8,471	8,494	8,494	8,494	8,494	8,494	8,494
High-education sample	-0.15***	394.03	5.85	-0.14	-0.40	-1.70	-0.17	-0.18	-0.31
	(0.03)	(649.65)	(3.71)	(0.42)	(0.49)	(4.32)	(0.20)	(0.17)	(0.21)
KP *F*-statistic	26.11	26.11	26.11	26.11	26.11	26.11	26.11	26.11	26.11
Observations	1,286	1,280	1,280	1,286	1,286	1,286	1,286	1,286	1,286

Notes: The table reports the results from IV regressions for each sub-sample. Migrants are defined as young if they are 30 years old or younger, and are highly educated if they have completed college education or above. All the regressions also include the same control variables as those in Table 2, as well as destination-county fixed effects and original-province fixed effects. "KP *F*-statistic" denotes the cluster-robust Kleibergen-Paap (KP) *F*-statistic on testing weak instruments. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Appendix

A. *Hukou* System in China

China's *hukou* (household registration) system registers the household address and *hukou* type of every citizen. According to the system, every person is classified as having an agricultural *hukou* or a non-agricultural *hukou*. Moreover, they are categorized according to their registered (or official) place of residence. The crucial part of the *hukou* institution is that it determines one's entitlements to social benefits, including access to a pension, public education, medical insurance, and permission to purchase housing and vehicles. A non-agricultural *hukou* is entitled to more social benefits than an agricultural *hukou*, whereas residents of large cities are usually entitled to more social benefits than those residing in small cities. The distinction between an agricultural *hukou* and a non-agricultural *hukou* has become less important over time, and the main barrier to accessing social benefits of a certain city is whether having a local *hukou* in that city (Chan, 2009).

Conversion to a local *hukou* is quite restrictive even nowadays. The requirements for obtaining a local *hukou* vary across cities due to recent decentralization reforms of the *hukou* system (Bosker et al., 2012). Specifically, local governments set the conditions for obtaining a local *hukou*, which usually include purchasing high-end apartments, making large business investments, or holding an advanced degree or professional qualifications. Given such demanding conditions, only very few people can meet the requirements and convert their *hukou* (Chan, 2009).

In the era of planned economy, the *hukou* system severely restricted internal migration. After the reform and opening-up, the system continued but was less restrictive on migration, partly due to the increase in labor demand in the coastal cities. People can move to, and work in, a place different from their *hukou* registration location, although they cannot enjoy the same social benefits as the native people with local *hukou*. The unequal access to social benefits strongly prohibits migrants from permanently settling in the receiving places, making migration in China mostly temporary and individualized. This further hinders the assimilation of migrants into the host places.

B. Assess Sorting in the Choice of Migration Destinations

One concern about the exclusion restriction of our instrumental variable estimation is that the instrumental variable, i.e., the dialect distance between the residential county and province of origin, may affect migrants' labor market outcomes due to sorting in the choice of destination places. For example, if migrants are sorted across destinations according to their comparative advantage, we would also observe a negative correlation between dialect distance and labor market outcomes, even without changing social identity.

To assess the extent of sorting in the choice of destination places, we examine the correlation between observable individual characteristics and the average characteristics of other migrants who came from the same home province and lived in the same destination county. If there are origin-destination-specific factors affecting migrants' choice of destination places, we would expect positive correlations between the characteristics of the individual and those of other migrants with the same places of destination and origin. In the spirit of Altonji et al. (2005), the analyses can also indicate a possible correlation in unobservable factors.

Like Bayer et al. (2008), for each pair of province-of-origin-by-destination-county, we randomly select only one migrant and calculate the correlation between characteristics of the individual and the average value of corresponding characteristics among other respondents who came from the same province and worked in the same destination county. Sampling only one respondent for each origin-destination pair is done to avoid a negative correlation mechanically if all individuals were used in the estimation, which arises because the characteristics of each individual are used in calculating the average characteristics of all the other respondents in the same origin-destination pair but not for those of the individual. Since the sampling is random, the estimates of the correlation coefficients are unbiased. To reduce measurement error, we drop the observation from origin-destination pairs when the number of respondents is less than six, but the results remain robust if we use other criteria, such as five, four, or three.

Table A1 in Appendix E reports the estimates of the correlation coefficients between the characteristics of the chosen migrants and the average characteristics of other migrants who came from the same province and worked in the same destination county. Hence, the number of observations for estimating the correlation coefficients equals the number of pairs of the

province of the origin and the destination county.

As the benchmark, Columns (1) and (2) report the unconditional correlation coefficients and their significance levels. The higher the correlation coefficients, the greater the extent of overall sorting in the choice of destination places. Columns (3) and (4) report the statistics of correlations conditional on destination county fixed effects and original province fixed effects, whereas Columns (5) and (6) report the statistics of correlations further conditional on the region-of-origin-by-destination-county fixed effects. For each estimate of the conditional correlation, we first regress both the individual and average characteristics on the corresponding fixed effects, and then report the correlation coefficients between the residuals. Therefore, the conditional correlations reflect the extent of sorting by partialling out factors of the destination county and the province of origin that are relevant in the choice of destination places.

The results in Table A1 show that the unconditional correlation coefficients are positive and significant, indicating there is indeed a significant amount of sorting in the choice of destination places among migrants. However, the correlation coefficients reduce substantially when they are estimated conditional on destination county fixed effects and original province fixed effects, and many of the correlations turn out to be statistically insignificant. The results imply that the amount of sorting on observables is driven primarily by factors of the place of origin or those of the destination, whereas the amount of sorting due to factors related to the same pair of origin and destination places is small. Columns (5) and (6) indicate that the amount of sorting reduces even further when we account for region-of-origin-by-destination-county fixed effects.

Overall, the results imply that the amount of sorting in migrants' choice of destination places due to original-province-by-destination-county factors can be quite small, if any. This should alleviate the concern of possible violation of the exclusion restriction due to sorting in the choice of destination places, when we have controlled for individual characteristics and the fixed effects of residential county and original province.

C. Use of Surname Distance as the Instrumental Variable

In this section, we use the surname distance as an instrumental variable to estimate the impact of identity on the labor market outcomes of internal migrants in China.

In the Chinese population, surnames are transmitted via the male line. It is like the transmission of Y-chromosome genes, except that surnames are also passed on to females (Du et al., 1992). Surnames have been widely investigated by geneticists, anthropologists, and scientists in many other fields, given the considerable similarity of geographic distribution of surnames and genes (Chen et al., 2019).[41] Therefore, we use the surname distance between populations as a measure of genealogical relatedness between populations.

Following the literature, we define isonymy within a region i as $I_i = \sum_{k=1}^{S} p_{ki}^2$, where p_{ki} is the proportion of the population with surname k among the entire population in region i, and S is the total number of surnames. The isonymy between region i and j is defined as $I_{ij} = \sum_{k=1}^{S} p_{ki}p_{kj}$, which captures the similarity of surname distribution between populations in the two regions. Accordingly, the surname distance between region i and j is measured by the Nei's index (Nei, 1972), which is a normalization of isonymy between the two regions, namely, $N_{ij} = -\log(I_{ij}/\sqrt{I_i I_j})$.[42] The Nei's index equals 0 when the surname distribution of two populations are identical, and is positive when the distributions differ. Similar to the dialect distance, a higher Nei's index is associated with larger cultural differences between two populations.

We use data constructed by Du et al. (1992) to measure the surname distance between provinces in China and match them with the survey data used in the main analysis. By doing so, we can get a measure for every migrant of the distance in surname distribution of populations in the home province and the province of the destination place. The correlation coefficient between the measures of surname distance and linguistic distance is 0.554. For reference, Spolaore and Wacziarg (2009) reported that the correlation coefficient between genetic distance and linguistic distance is 0.227 across countries in the world. The relatively

[41] Many Chinese surnames appeared around 4,000 years ago, which is at least 3,000 years earlier than those in Europe or Japan. Meanwhile, the Chinese population uses fewer surnames and has much larger isonymous groups than Europe or Japan (Du et al., 1992).

[42] We achieve similar results if we measure surname distance by the relative isonymy of two regions (i.e., $I_{ij}/\sqrt{I_i I_j}$) or we construct the indexes based on the 19 most common surnames or 1,035 less common surnames, according to Du et al. (1992). This should reduce concerns that the information contained in the less popular surnames cannot be adequately revealed in the measure of isonymy (Chen et al., 2019).

high pairwise correlation confirms the validity of our measure of linguistic distance.

Table A10 in Section E of this Appendix reports the results of the 2SLS by using surname distance between the home and host provinces as the instrument variable for self-identification with the host community. As shown, the IV estimates of the impact of identity on migrants' labor market outcomes are largely comparable to the benchmark results in sign and magnitude, although the estimates are less statistically significant. We only observe significant impacts of identity on the likelihood of overworking using the alternative instrument. This is likely due to weaker power in the first stage, as suggested by the *F*-statistic, although it is also larger than 10.

D. Additional Examination on Exclusion Restriction

We had conducted many analyses to assess the validity of the exclusion restriction in the text. While we note it is (in general) impossible to directly test the exclusion restriction, we conduct further analyses and add more discussion on the exclusion restriction in this section to confirm robustness and reliability of our results.

D.1 Knowledge of Local Dialect upon Migration

One concern about the exclusion restriction is that knowledge of local dialect upon migration, which is determined by dialect distance, can impact migrant's labor market outcomes. Since the survey did not measure the ability to speak or understand the local dialect upon migration, we cannot directly account for migrants' knowledge of the local dialect when they arrived in the host region.

To assess the potential importance of knowledge of the local dialect on arriving in the host region, we conduct analyses by separating those who had migrated recently from those who had migrated a longer time ago. For the newly arrived migrants, the measurement of the knowledge of the local dialect almost captures their knowledge of the local dialect on arriving in the host region. Table A11 in Appendix E shows that even for this subgroup of migrants, there is no significant association between knowledge of the local dialect and labor market outcomes. The results are consistent with the fact that people can easily communicate with one

another in the workplace given the high popularization rate of *Putonghua*, and communication with the local dialect does not play a significant role in migrants' labor market outcomes.

To further address this concern, we examine the heterogeneity in the reduced-form association between dialect distance and labor market outcomes by proximity to *Putonghua* of the dialect at the destination place, separately for those who had migrated recently and those who had migrated a longer time ago. If the knowledge of local dialect upon migration is indeed important in affecting labor market outcomes, we would expect the effect to be more salient in places where the dialect is more different from *Putonghua*, particularly for those who arrived in the host region recently. However, Table A12 in Appendix E shows that the coefficients of the interaction terms between dialect distance and proximity to *Putonghua* of the dialect at the destination place are not significantly different from 0 for the subgroup of newly arrived migrants. The results imply that knowledge of local dialects upon arrival does not play a significant role in the relationship between dialect distance and migrants' labor market outcomes.

In sum, these results should further reduce the concern that the communication effect may threaten the exclusion restriction.

D.2 Dialect-based Discrimination

Dialects may connect to certain stereotypes. It is possible that dialect-based discrimination directly affects migrants' labor market outcomes. However, such discrimination may not necessarily correlate with the linguistic distance between the dialect of a specific place and that of the host place.

As was already done in the text, we controlled for region-of-origin-by-destination-county fixed effects as a robustness check (see Panel A of Table 5). The fixed effects may absorb discrimination by local residents in some counties toward migrants from a specific region. The results are robust compared to the benchmark estimates.

As another robustness check, we control for the type of dialects spoken in the province of origin interacted with the destination-county dummies. Reassuringly, Table A13 in Appendix E shows that the results are also quite similar to the benchmark estimates. The results should

further reduce concerns about violation of the exclusion restriction due to dialect-based discrimination.

Another piece of evidence comes from a falsification test. If the dialect distance affects migrants' labor market outcomes through dialect-based discrimination, we would expect significant correlations between the dialect distance and labor market outcomes among the new migrants. However, as shown by the results in Table 6, none of the labor market outcomes is significantly associated with dialect distance, and the magnitude of the estimates is generally quite small when using the subsample of newly arrived migrants.

Overall, the above results suggest that dialect-based discrimination is unlikely to threaten the validity of the exclusion restriction.

D.3 Social Contacts in the Host Region prior to Migration

Dialect distance may correlate with social connections in the host region prior to migration, which in turn could affect labor market outcomes. To examine the extent to which this may threaten our identification assumption, we provide several pieces of empirical evidence.

First, we control for the log of the number of migrants from the same province in the destination county in 2010, which is used as a proxy to measure the social connections prior to migration. As shown by the results in Panel C of Table 5, the estimates are quite similar to the benchmark estimates.

In addition, if social connection prior to migration indeed plays an important role in the association between dialect distance and migrants' labor market outcomes, we would expect it exists especially among newly arrived migrants. However, as shown by the results of the falsification test in Panel A of Table 6, we find that the associations between dialect distance and migrants' labor market outcomes are economically and statistically insignificant among new migrants. The results suggest that social connection prior to migration does not play an important role in the relationship between dialect distance and migrants' labor market outcomes. This should further reduce the concern that social connection prior to migration may threaten the identification assumption of our instrumental variable estimation.

Taking account of all the evidence, we believe that the exclusion restriction is justified

and robust.

E. Supplementary Figures and Tables

See the online supplementary materials for additional figures and tables.

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, Jeffrey Wooldridge. 2017. "When Should You Adjust Standard Errors for Clustering?" NBER Working Papers 24003.

Abel, Martin, Rulof Burger, Patrizio Piraino. 2020. "The Value of Reference Letters: Experimental Evidence from South Africa," *American Economic Journal: Applied Economics* 12 (3): 40-71.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review* 102 (5): 1832-1856.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration," *Journal of Political Economy* 122(3): 467-717.

Afridi, Farzana, Sherry Xin Li, Yufei Ren. 2015. "Social Identity and Inequality: The Impact of China's *hukou* System," *Journal of Public Economics* 123: 17-29.

Akerlof, George A., Rachel E. Kranton. 2000. "Economics and Identity," *The Quarterly Journal of Economics* 115: 715-753.

Alesina, Alberto, Eliana La Ferrara. 2005. "Ethnic Diversity and Economic Performance," *Journal of Economic Literature* 43(3): 762-800.

Altonji, Joseph G., Todd E. Elder, Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy* 113(1): 151-184.

Amodio, Francesco, Giorgio Chiovelli. 2018. "Ethnicity and Violence during Democratic Transitions: Evidence from South Africa," *Journal of the European Economic Association* 16(4): 1234-1280.

Battu, Harminder, McDonald Mwale, Yves Zenou. 2007. "Oppositional Identities and the Labor Market," *Journal of Population Economics* 20(3): 643-667.

Battu, Harminder, Yves Zenou. 2010. "Oppositional Identities and Employment for Ethnic Minorities: Evidence from England," *The Economic Journal* 120: F52-F71.

Bazzi, Samuel, Sarah Burns, Gordon Hanson, Bryan Roberts, John Whitley. 2021. "Deterring

Illegal Entry: Migrant Sanctions and Recidivism in Border Apprehensions," *American Economic Journal: Economic Policy* 13(3): 1-27.

Bazzi, Samuel, Arya Gaduh, Alexander Rothenberg, Maisy Wong. 2016. "Skill Transferability, Migration, and Development: Evidence from Population Resettlement in Indonesia," *American Economic Review* 106(9): 2658-2698.

Bayer, Patrick, Stephen L. Ross, Giorgio Topa. 2008. "Place of Work and Place of Residence: Informal Hiring Networks and Labor Market Outcomes," *Journal of Political Economy* 116: 1150-1196.

Beaman, Lori, Jeremy Magruder. 2012. "Who Gets the Job Referral? Evidence from a Social Networks Experiment," *American Economic Review* 102(7): 3574-3593.

Bisin, Alberto, Eleonora Patacchini, Thierry Verdier, Yves Zenou, Andrea Ichino, Etienne Wasmer. 2011. "Ethnic Identity and Labour Market Outcomes of Immigrants in Europe," *Economic Policy* 26(65): 59-92.

Black, Dan A., Seth G. Sanders, Evan J. Taylor, Lowell J. Taylor. 2015. "The Impact of the Great Migration on Mortality of African Americans: Evidence from the Deep South," *American Economic Review* 105 (2): 477-503.

Bollinger, Bryan, Jesse Burkhardt, Kenneth T. Gillingham. 2020. "Peer Effects in Residential Water Conservation: Evidence from Migration," *American Economic Journal: Economic Policy* 12(3) 107-33.

Bolton, Gary E., Johannes Mans, Axel Ockenfels. 2020. "Norm Enforcement in Markets: Group Identity and the Volunteering of Feedback," *The Economic Journal* 130: 1248-1261.

Borjas, George. 1985. "Assimilation, Changes in Cohort Quality, and Earnings of Immigrants," *Journal of Labor Economics* 3(4): 463-489.

Bosker, Maarten, Steven Brakman, Harry Garretsen, Marc Schramm. 2012. "Relaxing Hukou: Increased Labor Mobility and China's Economic Geography," *Journal of Urban Economics* 72: 252-266.

Bricker, Jesse, Jacob Krimmel, Rodney Ramcharan. 2021. "Signaling Status: The Impact of Relative Income on Household Consumption and Financial Decisions," *Management*

Science 67(4): 1993-2009.

Cai, Shu, Albert Park, Winnie Yip. 2022. "Migration and Experienced Utility of Left-behind Parents: Evidence from Rural China," *Journal of Population Economics* 35(3): 1225-1259.

Cai, Shu, Xingjian Zhang. 2021. "Anatomy of the Wage Gap between Local and Migrant Workers in Urban China: New Evidence from Matched Data," Working paper. http://dx.doi.org/10.2139/ssrn.3933758

Campigotto, Nicola, Chiara Rapallini, Aldo Rustichini. 2022. "School Friendship Networks, Homophily and Multiculturalism: Evidence from European Countries," *Journal of Population Economics* 35(4): 1687–1722.

Campo, Francesco, Luca Nunziata, Lorenzo Rocco. 2023. "Business is Tense: New Evidence on How Language Affects Economic Activity," *Journal of Economic Growth*, forthcoming.

Cappellari, Lorenzo, Konstantinos Tatsiramos. 2015. "With a Little Help from My Friends? Quality of Social Networks, Job Finding and Job Match Quality," *European Economic Review* 78: 55-75.

Card, David. 2005. "Is the New Immigration Really so Bad?" *The Economic Journal* 115(507): F300-F323.

Carillo, Maria Rosaria, Vincenzo Lombardo, Tiziana Venittelli. 2023. "Social Identity and Labor Market Outcomes of Immigrants," *Journal of Population Economics* 36: 69-113.

Casey, Teresa, Christian Dustmann. 2010. "Immigrants' Identity, Economic Outcomes and the Transmission of Identity across Generations," *The Economic Journal* 120: F31-F51.

Chan, Kam Wing. 2009. "The Chinese Hukou System at 50," *Eurasian Geography and Economics* 50(2): 197-221.

Charness, Gary, Yan Chen. 2020. "Social Identity, Group Behavior, and Teams," *Annual Review of Economics* 12: 691-713.

Chen, Jiawei, Liujun Chen, Yan Liu, Xiaomeng Li, Yida Yuan. 2019. "An Index of Chinese Surname Distribution and Its Implications for Population Dynamics," *American Journal of Physical Anthropology* 169: 608-618.

Chen, M. Keith. 2013. "The Effect of Language on Economic Behavior: Evidence from

Savings Rates, Health Behaviors, and Retirement Assets," *American Economic Review* 103(2): 690-731.

Chiswick, Barry R. 1978. "The Effect of Americanization on the Earnings of Foreign-born Men," *Journal of Political Economy* 86(5): 897-922.

Conley, Timothy, Christian Hansen, Peter Rossi. 2012. "Plausibly Exogenous," *Review of Economics and Statistics* 94(1): 260-272.

Constant, Amelie F., Klaus F. Zimmermann. 2008. "Measuring Ethnic Identity and Its Impact on Economic Behavior," *Journal of the European Economic Association* 6: 424-433.

Constant, Amelie F., Klaus F. Zimmermann. 2009. "Work and Money: Payoffs by Ethnic Identity and Gender," *Research in Labor Economics* 29: 3-30.

Constant, Amelie F., Klaus F. Zimmermann. 2011. "Migration Ethnicity and Economic Integration," In: Jovanovic M. N. (eds.) *International Handbook of Economic Integration*. Edward Elgar Publishing, Cheltenham, pp 145-168.

Constant, Amelie F., Liliya Gataullina, Klaus F. Zimmermann. 2009. "Ethnosizing Immigrants," *Journal of Economic Behavior & Organization* 69: 274-287.

Delaporte, Isaure. 2019. "Ethnic Identity and the Employment Outcomes of Immigrants: Evidence from France," GLO Discussion Paper No. 345. https://ideas.repec.org/p/zbw/glodps/345.html

Desmet, Klaus, Ignacio Ortuño-Ortín, Romain Wacziarg. 2017. "Culture, Ethnicity, and Diversity," *American Economic Review* 107 (9): 2479-2513.

Du, Ruofu, Yida Yuan, Juliana Hwang, Joanna Mountain, L. Luca Cavalli-Sforza. 1992. "Chinese Surnames and the Genetic Differences between North and South China," *Journal of Chinese Linguistics* 5: 1-66, 68-93.

Duleep, Harriet, Xingfei Liu, Mark Regets. 2022. "How the Earnings Growth of US Immigrants was Underestimated," *Journal of Population Economics* 35: 381–407.

Edin, Per-Anders, Peter Fredriksson, Olof Åslund. 2003. "Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment," *The Quarterly Journal of Economics* 118(1): 329-357.

Ellis, Mark. 2012. "Reinventing US Internal Migration Studies in the Age of International

Migration," *Population, Space and Place* 18: 196-208.

Erten, Bilge, Jessica Leight. 2021. "Exporting Out of Agriculture: The Impact of WTO Accession on Structural Transformation in China," *The Review of Economics and Statistics* 103(2): 364-380.

Eugster, Beatrix, Rafael Lalive, Andreas Steinhauer, Josef Zweimüller. 2017. "Culture, Work Attitudes, and Job Search: Evidence from the Swiss Language Border," *Journal of European Economic Association* 15(5): 1056-1100.

Falck, Oliver, Stephan Heblich, Alfred Lameli, Jens Sudekum. 2012. "Dialects, Cultural Identity, and Economic Exchange," *Journal of Urban Economics* 72:225-239.

Fouka, Vasiliki. 2020. "Backlash: The Unintended Effects of Language Prohibition in US Schools after World War I," *The Review of Economic Studies* 87(1): 204-239.

Freedman, Matthew, Emily Owens, Sarah Bohn. 2018. "Immigration, Employment Opportunities, and Criminal Behavior," *American Economic Journal: Economic Policy* 10(2) 117-51.

Gardner, Bradley M. 2017. *China's Great Migration: How the Poor Built a Prosperous Nation*. The Independent Institute, Oakland, California.

Giles, John, Albert Park, Fang Cai, Yang Du. 2013. "Weathering a Storm: Survey-based Perspectives on Employment in China in the aftermath of the Global Financial Crisis." In: Banerji A, Newhouse D, Robalino D, Paci P (eds.) *Working through the Crisis: Jobs and Policies in Developing Countries during the Great Recession*. The World Bank, Washington, D.C.

Ginsburgh, Victor, Shlomo Weber. 2020. "The Economics of Language," *Journal of Economic Literature* 58(2): 348-404.

Gorinas, Cedric. 2014. "Ethnic Identity, Majority Norms, and the Native–Immigrant Employment Gap," *Journal of Population Economics* 27: 225-250.

Gousse, Marion, Nicolas Jacquemet, Jean-Marc Robin. 2023. "Culture and Marriage: Lessons from German Reunification," Working paper.

Granovetter, Mark. 2005. "The Impact of Social Structure on Economic Outcomes," *Journal of Economic Perspectives* 19(1): 33-50.

Guadalupe, Maria, Zoe Kinias, Florian Schloderer. 2020. "Individual Identity and Organizational Identification: Evidence from a Field Experiment," *American Economic Review* 110: 193-198.

Guiso, Luigi, Paola Sapienza, Luigi Zingales. 2009. "Cultural Biases in Economic Exchange?" *The Quarterly Journal of Economics* 124(3): 1095-1131.

Hatton, Timothy J. 1997. "The Immigrant Assimilation Puzzle in Late Nineteenth Century America," *The Journal of Economic History* 57(01):34-62.

Herrmann-Pillath, Carsten, Alexander Libman, Xiaofan Yu. 2014. "Economic Integration in China: Politics and Culture," *Journal of Comparative Economics* 42(2): 470-492.

Islam, Asadul, Paul Raschky. 2015. "Genetic Distance, Immigrants' Identity, and Labor Market Outcomes," *Journal of Population Economics* 28(3): 845-868**.**

Ioannides, Yannis M., Linda Datcher Loury. 2004. "Job Information Networks, Neighborhood Effects, and Inequality," *Journal of Economic Literature* 42(4): 1056-1093.

Jaschke, Philipp, Sulin Sardoschau, Marco Tabellini. 2022. "Scared Straight? Threat and Assimilation of Refugees in Germany," NBER Working Paper 30381.

King, Russell, Ronald Skeldon. 2010. "'Mind the Gap!' Integrating Approaches to Internal and International Migration," *Journal of Ethnic and Migration Studies* 36(10): 1619-1646.

Kinnan, Cynthia, Shing-Yi Wang, Yongxiang Wang. 2018. "Access to Migration for Rural Households," *American Economic Journal: Applied Economics* 10(4): 79–119.

Kuziemko, Ilyana, Joseph Ferrier. 2014. "The Role of Immigrant Children in Their Parents' Assimilation in the U.S., 1850-2010," *Human Capital in History: The American Record*. University of Chicago Press.

LaLonde, Robert J., Robert H. Topel. 1991. "Immigrants in the American Labor Market: Quality, Assimilation, and Distributional Effects," *American Economic Review* 81(2): 297-302.

Lee, David S., Justin McCrary, Marcelo J. Moreira, Jack Porter. 2022. "Valid *t*-Ratio Inference for IV," *American Economic Review* 112(10): 3260-3290.

Liu, Yuyun, Yang Jiao, Xianxiang Xu. 2020. "Promoting or Preventing Labor Migration? Revisiting the Role of Language," *China Economic Review* 60, 101407.

Lu, Yilong. 2003. *The Household Registration System: Control and Social Differentiation*. Beijing: The Commercial Press.

Lubotsky, Darren. 2007. "Chutes or Ladders? A Longitudinal Analysis of Immigrant Earnings," *Journal of Political Economy* 115(5): 820-867.

Martinangeli, Andrea F.M., Peter Martinsson. 2020. "We, the Rich: Inequality, Identity and Cooperation," *Journal of Economic Behavior and Organization* 178: 249-266.

Minns, Chris. 2000. "Income, Cohort Effects and Occupational Mobility: A New Look at Immigration to the United States at the Turn of the 20th Century," *Explorations in Economic History* 37: 326-350.

Mu, Zheng, Wei-Jun Jean Yeung. 2020. "Internal Migration, Marriage Timing and Assortative Mating: A Mixed-method Study in China," *Journal of Ethnic and Migration Studies* 46(14), 2914–2936.

Nei, Masatoshi. 1972. "Genetic distance between populations," *The American Naturalist* 106: 283-292.

Nekby, Lena, Magnus Rödin. 2010. "Acculturation Identity and Employment among Second and Middle Generation Immigrants," *Journal of Economic Psychology* 31(1): 35-50.

Nie, Peng, Weibo Yan. 2021. "Effects of Social Networks on Job Attainment and Match Quality: Evidence from the China Labor-Force Dynamics Survey," IZA Discussion Paper No. 14504.

Olivetti, Claudia, Eleonora Patacchini, Yves Zenou. 2020. "Mothers, Peers, and Gender-Role Identity," *Journal of the European Economic Association* 18(1): 266-301.

Oster, Emily. 2019. "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics* 37(2): 187-204.

Ottaviano, Gianmarco I.P., Giovanni Peri. 2005. "Cities and Cultures," *Journal of Urban Economics* 58: 304-337.

Pendakur, Krishna, Ravi Pendakur. 2005. "Ethnic Identity and the Labour Market," Working paper. http://www.sfu.ca/~pendakur/pendakur_and_pendakur_ethnic_identity.pdf

Piracha, Matloob, Massimiliano Tani, Zhiming Cheng, Ben Zhe Wang. 2023. "Social Assimilation and Immigrant' Labour Market Outcomes," *Journal of Population*

Economics 36: 37-67.

Sequeira, Sandra, Nathan Nunn, Nancy Qian. 2020. "Immigrants and the Making of America," *Review of Economic Studies* 87(1): 382-419.

Shayo, Moses. 2020. "Social Identity and Economic Policy," *Annual Review of Economics* 12: 355-389.

Skeldon, Ronald. 2006. "Interlinkages between Internal and International Migration and Development in the Asian Region," *Population, Space and Place* 12: 15-30.

Spolaore, Enrico, Romain Wacziarg. 2009. "The Diffusion of Development," *The Quarterly Journal of Economics* 124(2): 469-529.

State Council of P. R. C.. 1956. *Instructions on the Promotion of Putonghua*. Beijing: Science Press.

Stuart, Bryan A., Evan J. Taylor. 2021. "The Effect of Social Connectedness on Crime: Evidence from the Great Migration," *The Review of Economics and Statistics* 103(1): 18-33.

Suedekum, Jens. 2018. "Economic Effects of Differences in Dialect," *IZA World of Labor* 414 doi: 10.15185/izawol.414

Talhelm, T., X. Zhang, S. Oishi, C. Shimin, D. Duan, X. Lan, S. Kitayama. 2014. "Large-scale Psychological Differences within China Explained by Rice versus Wheat Agriculture," *Science* 344 (6184): 603-608.

United Nations, Department of Economic and Social Affairs, Population Division. 2016. *International Migration Report 2015: Highlights*. ST/ESA/SER.A/375.

Verdier, Thierry, Yves Zenou. 2017. "The Role of Social Networks in Cultural Assimilation," *Journal of Urban Economics* 97: 15-39.

Wang, Wenfei Winnie, C. Cindy Fan. 2012. "Migrant Workers' Integration in Urban China: Experiences in Employment, Social Adaptation, and Self-identity," *Eurasian Geography and Economics* 53(6): 731-749.

Wang, Yan, Juan Carlos Conesa. 2022. "The Role of Demographics and Migration for the Future of Economic Growth in China," *European Economic Review* 144: 104076.

Xu, Baohua, Miyata Ichiro. 1999. *Chinese Dialect Dictionary*. Zhonghua shuju, Beijing.

Zhang, Jipeng, Ru Wang, Chong Lu. 2019. “A Quantitative Analysis of Hukou Reform in Chinese Cities: 2000–2016,” *Growth and Change* 50:201-221.

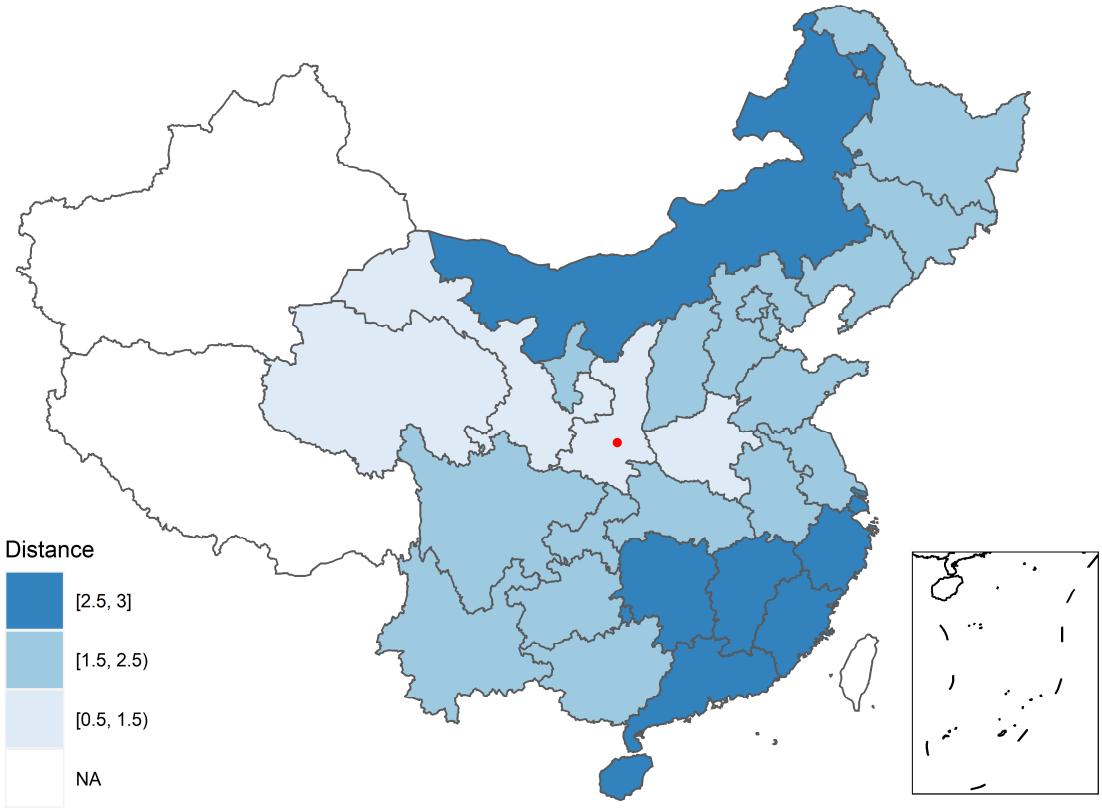
Online Supplementary Materials

for "Social Identity and Labor Market Outcomes of Internal Migrant Workers"

by Shu Cai, Klaus F. Zimmermann

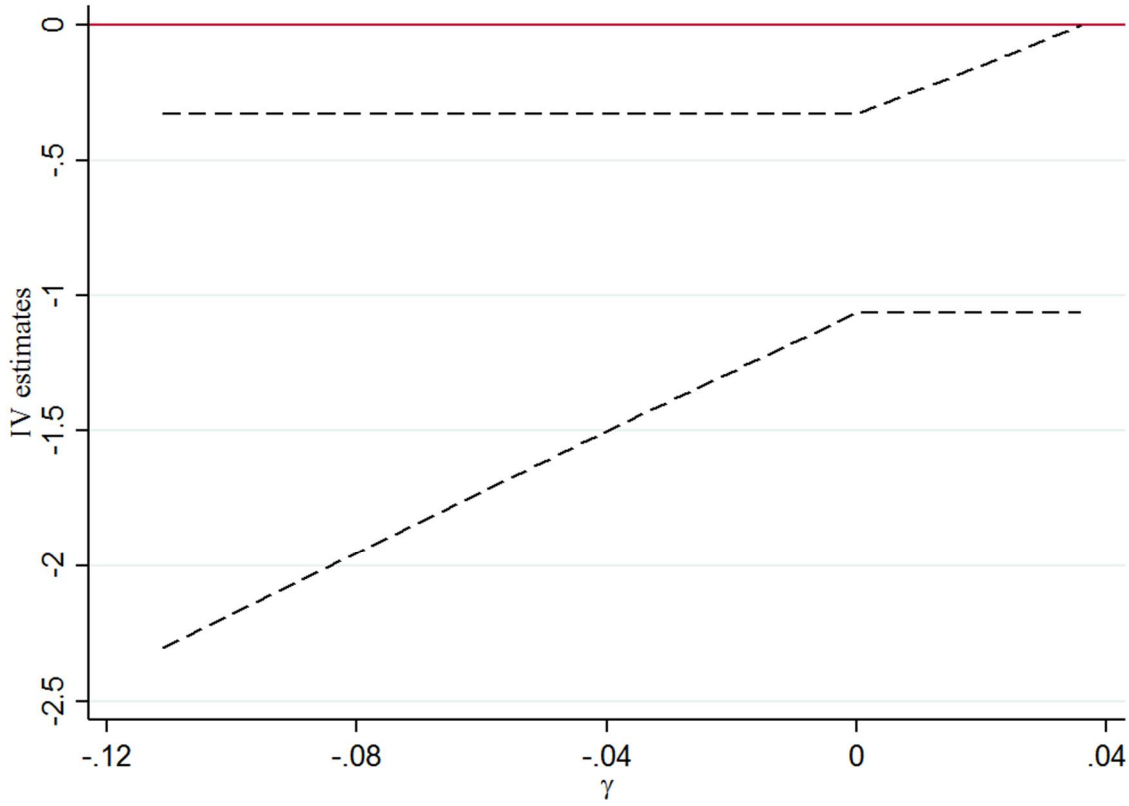
E. Supplementary Figures and Tables

Figure A1. Bilateral dialect distance—An example



Notes: The figure demonstrates the bilateral dialect distance between one of the counties in the sample—the Chang'an district (the red point on the map) and the potential destination provinces of migration. It is measured by the population-weighted dialect distance between Chang'an district and each county in the province of destination.
Source: Own construction.

Figure A2. Bounds of IV estimates under plausible exogeneity



Notes: The figure illustrates the upper and lower bounds of the 90 percent confidence interval for the IV estimates of the effect of identity on the summary index when γ takes the values on the interval [-0.111, 0.036].
Source: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A1. Correlation between individual and average characteristics of respondents with the same residential county and original province

	Unconditional		Conditional on County FE+Original province FE		Conditional on County FE+Original province FE+Region of origin by destination county FE	
	correlation coefficient	*p* -values	correlation coefficient	*p* -values	correlation coefficient	*p* -values
	(1)	(2)	(3)	(4)	(5)	(6)
Age	.207	.001	.000	.996	-.050	.418
Age 15-24	.196	.001	.101	.104	.006	.919
Age 25-34	.241	.000	.107	.083	.114	.065
Age 35-59	.209	.001	-.009	.880	-.013	.836
Male	.126	.042	.052	.398	.019	.761
Married	.220	.000	.050	.416	-.028	.650
Education: Middle school or below	.436	.000	.126	.041	.078	.208
Education: High school	.204	.001	.098	.111	.162	.008
Education: College or above	.478	.000	.235	.000	.063	.307
Number of children	.240	.000	.011	.855	-.081	.191
Number of children=0	.232	.000	.029	.641	-.061	.327
Number of children=1	.098	.115	-.076	.220	-.236	.000
Number of children⩾2	.184	.003	-.015	.813	-.144	.020
Han ethnicity	.259	.000	.148	.016	.141	.023
Nonagriculture hukou	.327	.000	.122	.048	.097	.115

Notes: The table reports the correlation between individual characteristics and the average value of corresponding characteristics among other respondents who came from the same province and worked at the same destination county. The observation unit of the estimation is the category defined by residential county and original province, and an individual is randomly chosen from each category. The categories with less than six respondents were dropped in the analysis. The number of observations is 263. Column (1) reports the raw correlation, Column (3) reports the correlation conditional on residential county fixed effect and original province fixed effect, and Column (5) reports the correlation by further isolating the region-of-origin-by-destination-county fixed effect. Columns (2), (4), (6) report the corresponding *p*-values for each correlation coefficient.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013.

Table A2. Dialect distance and average characteristics of respondents with the same residential county and original province

Dependent variables	Average characteristics												
	Age 15-24	Age 25-34	Age 35-59	Male	Married	Middle school or below	High school	College or above	Num of children =0	Num of children =1	Num of children ⩾2	Han ethnicity	Non-agricultur e hukou
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
Panel A													
Dialect distance (0-1-2-3)	-0.00	-0.01	0.01	0.00	-0.01	0.04**	-0.03**	-0.01	-0.00	-0.00	0.01	-0.00	0.00
	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.04)	(0.03)	(0.03)	(0.00)	(0.01)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	482	482	482	482	482	482	482	482	482	482	482	482	482
R-squared	0.319	0.342	0.411	0.340	0.350	0.585	0.433	0.536	0.345	0.275	0.406	0.420	0.538
Panel B													
Dialect distance (0-1-2-3)	-0.02	0.00	0.02	-0.00	-0.01	0.05*	-0.03	-0.02	-0.01	-0.00	0.02	0.00	0.02
	(0.02)	(0.02)	(0.02)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)	(0.04)	(0.03)	(0.03)	(0.00)	(0.02)
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original Region by County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	413	413	413	413	413	413	413	413	413	413	413	413	413
R-squared	0.523	0.457	0.565	0.490	0.540	0.718	0.618	0.646	0.559	0.468	0.524	0.570	0.657

Notes: The table reports the OLS estimates of the association between the dialect distance and the average of a series of characteristics of respondents with the same residential county and original province. The observation unit is the original province by residential county. Panel A includes residential county fixed effect and original province fixed effect, whereas Panel B further controls for original region by residential county fixed effects. The number of observations is different in the two panels due to dropping of singleton observations in the estimation. The standard errors in parentheses are clustered at the levels of original province and residential county. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A3. Examination of potential bias caused by sorting within destination county and original province

			work time			overwork		
	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: IV benchmark estimates								
Feel they b/to local citizens (yes=1)	39.58	3.25	-0.23	-1.17***	-9.21**	-0.22**	-0.44***	-0.26**
	(400.30)	(2.09)	(0.21)	(0.44)	(3.76)	(0.11)	(0.15)	(0.11)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel B: IV estimates controlling for average characteristics								
Feel they b/to local citizens (yes=1)	-347.97	1.58	-0.41	-1.06**	-9.89**	-0.26**	-0.51***	-0.29**
	(483.47)	(2.45)	(0.26)	(0.48)	(4.19)	(0.12)	(0.18)	(0.13)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Control for average characteristics	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	9,503	9,503	9,530	9,530	9,530	9,530	9,530	9,530
p -value of joint test on average characteristics	0.814	0.488	0.151	0.157	0.174	0.178	0.024	0.209

Notes: Panel A reports the benchmark IV estimates. Panel B reports the results of IV regressions by further controlling for average characteristics (including dummies of age group, male, married, educational categories, and number of children) of fellow townsmen residing in the same destination county on the basis of benchmark regressions. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A4. Examination on post-migration selection

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Subsample—Arrival year earlier than the 2009 financial crisis									
Feel they b/to local citizens (yes=1)		660.44	7.27	0.13	-1.17	-5.57	-0.19	-0.50*	-0.36
		(962.68)	(5.36)	(0.45)	(0.78)	(6.91)	(0.23)	(0.29)	(0.25)
Dialect distance (0-1-2-3)	-0.09***								
	(0.02)								
KP *F*-statistic	14.86								
Mean of outcome	0.52	3332.21	15.06	6.01	9.12	55.17	0.74	0.45	0.78
Observations	4,081	4,070	4,070	4,081	4,081	4,081	4,081	4,081	4,081
Panel B: Subsample—Arrival year later than the 2009 financial crisis									
Feel they b/to local citizens (yes=1)		-158.41	1.62	-0.29	-1.10**	-9.53**	-0.19*	-0.38**	-0.18
		(385.45)	(1.97)	(0.23)	(0.51)	(4.24)	(0.11)	(0.17)	(0.11)
Dialect distance (0-1-2-3)	-0.11***								
	(0.02)								
KP *F*-statistic	38.94								
Mean of outcome	0.40	2928.08	12.9	6.05	9.23	56.22	0.79	0.50	0.82
Observations	5,699	5,681	5,681	5,699	5,699	5,699	5,699	5,699	5,699

Notes: Panel A reports the IV estimates using a subsample of migrants who arrived at the destination place earlier than the 2009 financial crisis (i.e., arrival year was less than or equal to 2009), and Panel B reports the IV estimates using a subsample of migrants who arrived at the destination place after the 2009 financial crisis (i.e., arrival year was larger than 2009). The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A5. IV estimation by correcting for sample selection bias

			work time			overwork		
	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Feel they b/to local citizens (yes=1)	120.64	3.72*	-0.21	-1.23***	-9.37***	-0.21**	-0.44***	-0.25**
	(363.00)	(1.96)	(0.21)	(0.44)	(3.57)	(0.11)	(0.14)	(0.10)
Inverse Mills ratio	380.07*	1.80	0.11	-0.01	0.92	0.05	0.08	0.05
	(217.59)	(1.12)	(0.10)	(0.24)	(1.84)	(0.05)	(0.07)	(0.05)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	3,094.18	13.80	6.03	9.19	55.76	0.77	0.48	0.80
Observations	9,631	9,631	9,660	9,660	9,660	9,660	9,660	9,660

Notes: The table reports the IV estimates by further controlling for the inverse Mills ratio to adjust for potential sample selection of only including the employee, which is instrumented by indicators of participation in social insurance programs in hometowns. The standard errors are computed using bootstrap with 500 replications. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A6. IV estimation on the impact of identity on unemployment

	OLS		IV	
	unemployment	unemployment	feel they b/to local citizens (yes=1)	unemployment
	(1)	(2)	(3)	(4)
Dialect distance (0-1-2-3)		-0.00	-0.09***	
		(0.00)	(0.01)	
Feel they b/to local citizens (yes=1)	0.00			0.03
	(0.00)			(0.02)
KP *F*-statistic			83.25	
Observations	15,435	15,420	15,420	15,420
Mean of outcome	0.01	0.01	0.48	0.01
Control VARs	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes

Notes: Columns (1) and (2) report the estimates from OLS regressions on unemployment, whereas Columns (3) and (4) report the estimates from IV regressions on unemployment. The other control variables include age, age squared, dummy of male, marital status (including dummies of married first time, married second or more times, divorced, widowed), education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A7. Reduced-form association between dialect distance and mediator variables

	Networks			Neighborhood			Job searching			
	interact with ethnic people (yes=1)	interact with local people (yes=1)	member of ethnic organization (yes=1)	neighbors are mostly local citizens (yes=1)	neighbors are mostly non-local citizens (yes=1)	the number of local and non-local neighbors is similar (yes=1)	find the job through family/relatives or friends/classmates (yes=1)	find the job on their own, or start a business on their own (yes=1)	find the job through local people (yes=1)	find the job through government, social agency, internet, job fair, and others (yes=1)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: Subsample—Years since arrival less than or equal to half a year										
Dialect distance (0-1-2-3)	0.03**	-0.05	0.01	-0.01	0.05	-0.04	0.01	-0.03	0.01	0.01
	(0.01)	(0.03)	(0.02)	(0.03)	(0.04)	(0.03)	(0.03)	(0.02)	(0.02)	(0.02)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.93	0.37	0.12	0.21	0.52	0.27	0.55	0.24	0.06	0.14
Observations	1,426	1,426	1,426	1,365	1,365	1,365	1,426	1,426	1,426	1,426
Panel B: Subsample—Years since arrival more than half a year										
Dialect distance (0-1-2-3)	0.01	-0.05***	0.01	-0.02*	0.03**	-0.01	0.05***	-0.02	-0.02***	-0.01
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.93	0.43	0.12	0.24	0.48	0.28	0.45	0.32	0.06	0.17
Observations	8,354	8,354	8,354	8,040	8,040	8,040	8,348	8,348	8,348	8,348

Notes: The table reports the OLS estimates of the reduced-form relationships between dialect distance and mediator variables for the subsample of migrants who stayed in the city for less than, or equal to, half a year (Panel A) and the subsample of migrants who stayed in the city for more than half a year (Panel B). The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A8. IV estimation on the impact of identity on labor market outcomes conditional on job characteristics

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Feel they b/to local citizens (yes=1)		311.04	4.07**	-0.19	-0.98**	-7.74**	-0.19*	-0.36**	-0.22**
		(381.43)	(2.04)	(0.21)	(0.44)	(3.80)	(0.11)	(0.15)	(0.11)
Dialect distance (0-1-2-3)	-0.10***								
	(0.01)								
KP *F*-statistic	50.26								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Occupation, Industry, Unit type	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780

Notes: The table reports the results of IV regressions as specified in equations (1) and (2) in the text. The other control variables include age, age squared, dummy of male, marital status (including dummies of married once, married two or more times, divorced, and widowed), and education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college), and dummies indicating types of occupation, industry, and work unit. "KP *F*-statistic" denotes the cluster-robust Kleibergen-Paap (KP) *F*-statistic on testing weak instruments. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A9. Robustness checks: excluding marital status as the control variables

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Baseline results									
Dialect distance (0-1-2-3)	-0.10***								
	(0.01)								
Feel they b/to local citizens (yes=1)		39.58	3.25	-0.23	-1.17***	-9.21**	-0.22**	-0.44***	-0.26**
		(400.30)	(2.09)	(0.21)	(0.44)	(3.76)	(0.11)	(0.15)	(0.11)
KP *F*-statistic	52.31								
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Panel B: Excluding marital status as controls									
Dialect distance (0-1-2-3)	-0.10***								
	(0.01)								
Feel they b/to local citizens (yes=1)		52.63	3.29	-0.24	-1.16***	-9.17**	-0.22**	-0.44***	-0.26**
		(398.24)	(2.07)	(0.21)	(0.44)	(3.74)	(0.10)	(0.15)	(0.11)
KP *F*-statistic	52.78								
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3,096.76	13.80	6.04	9.19	55.79	0.77	0.48	0.81

Notes: Panel A reports the baseline estimates from IV regressions as specified in equations (1) and (2) in text. The other control variables include age, age squared, dummy of male, marital status (including dummies of married first time, married second or more times, divorced, widowed), education categories (including dummies of education level of middle school, education level of high school, education level of college, and education level above college). Panel B reports IV estimates by excluding marital status as the control variables. "KP *F*-statistic" denotes the cluster-robust Kleibergen-Paap (KP) *F*-statistic on testing weak instruments. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A10. IV estimation on the impact of identity on labor market outcomes using surname distance as instrumental variable

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Feel they b/to local citizens (yes=1)		-362.98	-0.42	-0.18	-0.99	-7.28	-0.28	-0.43*	-0.29*
		(594.55)	(2.81)	(0.33)	(0.71)	(5.95)	(0.17)	(0.23)	(0.17)
Surname distance	-0.30***								
	(0.07)								
KP *F*-statistic	20.98								
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3,097.06	13.80	6.04	9.19	55.79	0.77	0.48	0.81
Observations	9,785	9,756	9,756	9,785	9,785	9,785	9,785	9,785	9,785

Notes: The table reports the IV estimates using the surname distance as the instrumental variable for identity. The other control variables are the same as those in Table 2. "KP *F*-statistic" denotes the cluster-robust Kleibergen-Paap (KP) *F*-statistic on testing weak instruments. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013 and Du et al. (1992).

Table A11. Examination on exclusion restriction: knowledge of local dialect upon migration

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Subsample—Years since arrival less than or equal to half a year									
Feel they b/to local citizens (yes=1)		464.78	0.86	-0.09	0.83	3.62	0.12	0.16	0.20
		(662.86)	(3.21)	(0.38)	(0.82)	(6.72)	(0.15)	(0.24)	(0.15)
Can speak local dialect (yes=1)	0.16***	-130.44	-0.02	0.04	-0.28	-1.14	-0.03	-0.07	-0.04
	(0.04)	(126.23)	(0.64)	(0.08)	(0.18)	(1.47)	(0.04)	(0.06)	(0.04)
Can understand local dialect (yes=1)	0.11***	-80.61	-0.40	-0.03	-0.18	-1.24	-0.01	-0.00	-0.03
	(0.03)	(120.92)	(0.65)	(0.06)	(0.17)	(1.36)	(0.03)	(0.05)	(0.03)
Dialect distance (0-1-2-3)	-0.12***								
	(0.03)								
KP *F*-statistic	15.7								
Mean of outcome	0.33	2746.08	11.47	6.15	9.52	58.93	0.85	0.59	0.88
Observations	1,426	1,421	1,421	1,426	1,426	1,426	1,426	1,426	1,426
Panel B: Subsample—Years since arrival more than half a year									
Feel they b/to local citizens (yes=1)		63.34	5.79	-0.32	-2.11**	-15.68*	-0.46*	-0.72**	-0.57**
		(757.42)	(4.24)	(0.45)	(1.02)	(8.48)	(0.25)	(0.35)	(0.26)
Can speak local dialect (yes=1)	0.15***	30.92	-0.73	0.05	0.32*	2.39*	0.06	0.10	0.09**
	(0.02)	(128.41)	(0.71)	(0.07)	(0.18)	(1.45)	(0.04)	(0.06)	(0.04)
Can understand local dialect (yes=1)	0.11***	-4.16	-0.31	-0.01	0.08	0.37	0.04	0.03	0.03
	(0.02)	(93.70)	(0.54)	(0.06)	(0.14)	(1.13)	(0.03)	(0.04)	(0.04)
Dialect distance (0-1-2-3)	-0.06***								
	(0.02)								
KP *F*-statistic	12.93								
Mean of outcome	0.47	3156.58	14.2	6.02	9.13	55.25	0.76	0.46	0.79
Observations	8,354	8,330	8,330	8,354	8,354	8,354	8,354	8,354	8,354

Notes: Panel A reports the IV estimates using a subsample of migrants who stayed in the city for less than, or equal to, half a year, whereas Panel B reports the IV estimates using a subsample of migrants who stayed in the city for more than half a year. The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A12. Examination on exclusion restriction: knowledge of local dialect upon migration

Reduced-form estimates			work time			overwork		
	monthly income	hourly wage	aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Subsample—Years since arrival less than or equal to half a year								
Dialect distance (0-1-2-3)	44.32	-0.04	0.02	0.02	0.34	-0.00	0.02	-0.02
	(133.85)	(0.59)	(0.08)	(0.17)	(1.40)	(0.03)	(0.05)	(0.03)
Dialect distance (0-1-2-3) × City with larger dialect distance to *Putonghua*	-159.68	-0.08	-0.02	-0.16	-1.07	-0.02	-0.07	-0.00
	(178.39)	(0.83)	(0.11)	(0.26)	(2.08)	(0.05)	(0.08)	(0.05)
Mean of outcome	2746.08	11.47	6.15	9.52	58.93	0.85	0.59	0.88
Observations	1,421	1,421	1,426	1,426	1,426	1,426	1,426	1,426
Panel B: Subsample—Years since arrival more than half a year								
Dialect distance (0-1-2-3)	-12.02	-0.55	0.05	0.10	1.00*	0.04**	0.04	0.05**
	(62.54)	(0.34)	(0.03)	(0.07)	(0.61)	(0.02)	(0.03)	(0.02)
Dialect distance (0-1-2-3) × City with larger dialect distance to *Putonghua*	2.03	0.29	-0.04	0.05	-0.03	-0.03	0.01	-0.02
	(90.49)	(0.49)	(0.05)	(0.10)	(0.86)	(0.03)	(0.03)	(0.03)
Mean of outcome	3156.58	14.2	6.02	9.13	55.25	0.76	0.46	0.79
Observations	8,330	8,330	8,354	8,354	8,354	8,354	8,354	8,354

Notes: Panel A reports the OLS estimates using a subsample of migrants who stayed in the city for less than, or equal to, half a year, whereas Panel B reports the OLS estimates using a subsample of migrants who stayed in the city for more than half a year. The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.

Table A13. Examination on exclusion restriction: dialect-based discrimination

	feel they b/to local citizens (yes=1)	monthly income	hourly wage	work time			overwork		
				aver. days worked per week	aver. hours worked per day	aver. hours worked per week	overwork (days per week>5)	overwork (hours per day>8)	overwork (hours per week>40)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dialect distance (0-1-2-3)	-0.15***								
	(0.02)								
Feel they b/to local citizens (yes=1)		27.01	4.19	-0.42**	-1.03**	-10.60***	-0.19*	-0.35**	-0.29**
		(449.92)	(2.57)	(0.21)	(0.47)	(3.96)	(0.12)	(0.16)	(0.12)
KP *F*-statistic	40.12								
Observations	9,780	9,751	9,751	9,780	9,780	9,780	9,780	9,780	9,780
Control VARs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
County FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Original province FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Mean of outcome	0.45	3097.18	13.8	6.04	9.19	55.78	0.77	0.48	0.81

Notes: The table reports the results from IV regressions as specified in equations (1) and (2) in text, by further controlling for original province dialects interacted with destination county dummies. The original province dialects are measured by dummies indicating whether people of the province speak certain super dialect. The other control variables are the same as those in Table 2. The standard errors in parentheses are clustered by residential community. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.
Data: Dynamic Monitoring Survey of the Migrant Population of China 2013, Chinese Dialect Dictionary, and population census 2000.