Related Issues to Rural-to-Urban Migration in China

By

Sen Xue

A thesis submitted for the degree of

Doctor of Philosophy at

The Australian National University

September 2015

Declaration

This thesis is my original work. It contains no material that has been presented for a degree at this or any other university. To my best knowledge, it contains no copy or paraphrasing of work published by another person, except where explicitly acknowledged.

I am responsible for remaining errors and omissions.

Sen Xue
Research School of Economics
College of Business and Economics
The Australian National University
September, 2015

To my beloved wife Tian Hang and family.

Acknowledgement

First, I am deeply indebted to my principal supervisor, Professor Xin Meng. She provided me with a variety of data sources, constantly encouraged me to expand my research, and offered me invaluable advice when I met difficulties. She also provided me with substantial financial support during my PhD study. Without her support I would not have been able to complete this thesis. I cannot thank her enough. I also would like to express my sincere appreciation to Dr Tue Gorgens and Professor Robert Gregory. My supervisor, Dr Tue Gorgens, spent considerable time on my thesis and provided valuable comments on the econometric methodology, for which I am extremely grateful. I also greatly benefited from discussions with my supervisor, Professor Robert Gregory. He taught me how to discern the important issues in reality. I also thank him for providing comments on Chapter 3.

I would like to express my gratitude to my family, especially my wife Tian Hang. For a young researcher, nothing is better than having a smart and critical wife who is also doing a PhD. Her criticism sometimes made me frustrated, but always pushed me to work harder.

Completing a PhD is a challenging journey. As Chapter 2 shows, social networks are important to those in difficulty and this is true for the PhD student, for whom friendship is particularly valuable. I thank my PhD fellows, Jamie Cross, Minhee Chae, Chenghan Hou, Sanghyeok Lee, Qingyin Ma, Tomohito Okabe, Guanglong Ren, and Jilu Zhang for their social support.

Finally, my thank goes to the seminar participants in the 2013 and 2014 PhD conferences in Economics and Business, 2014 Australian Econometric Society Meeting, the 2014 Australian Health Economist Society Conference, and brownbag seminars at the Australian National University in 2013 and 2014. I am grateful for their constructive comments.

Abstract

Over the past two decades, China has witnessed unprecedented rural-to-urban migration. This thesis includes three self-contained empirical chapters, which are related to rural-to-urban migration in China.

Migrants are vulnerable to mental health problems, so it is important to understand the factors that can mitigate their mental stress. Chapter 2 investigates the relationship between the social networks and mental health of Chinese rural-to-urban migrants. The empirical analysis is based on a unique migrant survey from the Rural-to-Urban Migration in China (RUMIC) project, which includes the most up-to-date information about Chinese rural migrants. Using OLS and fixed effect models, I find that larger networks are correlated with better mental health. I use the instrumental variable approach to mitigate endogeneity bias. Both IV and fixed effect IV estimates indicate that social networks significantly help reduce mental health problems. The heterogeneity analysis suggests that the effect is larger for migrants with smaller social networks or with limited access to social welfare. In addition, females benefit more from their social networks than males.

Migrants are socially segregated and discriminated against in their destination cities in China. Chapter 3 uses a large representative survey to investigate whether interpersonal contact between urban locals and migrants could improve urban locals' attitudes towards migrants. The OLS estimates show that having previous contact experience with migrants is positively and significantly correlated with urban locals' willingness to interact with migrants. I adopt Lewbel (2012)'s heteroskedasticity identification approach to mitigate endogeneity bias between contact and attitudes. The estimates indicate that having previous contact experience with migrants could significantly improve willingness to engage in non-intimate interactions, but has no significant effect on willingness to engage in intimate interactions.

The migrant household survey of the Rural-to-Urban Migration in China Project is the largest longitudinal survey documenting the city life of rural migrants in China. The survey has been widely used in migration studies. However, a large proportion of respondents have left the survey sample, because migrants tend to be very mobile. Chapter 4 studies whether attrition is random, and the extent to which attrition biases estimated results of the first six waves of the survey. The empirical analysis suggests that there are systematic differences between the non-attritors and attritors. The non-attritors tend to be socio-economically better off, enjoy larger income gains from migration, be more willing to stay in cities and are more likely to be self-employed than the attritors. This chapter finds that there is likely to be some attrition bias and that the non-attritor sample is unrepresentative of the general migrant population at the time of follow-up surveys. Nevertheless, the examples shown in this chapter suggest that, in some cases, attrition bias and sample (un)representativeness could have only limited impact on the

regression coefficients of the individual-level variables which are most relevant to research and policy interests.

Contents

List of Figures

List of Tables

Chapter 1 Introduction to the Thesis

Rural-to-urban migration benefits the economic growth of developing countries. Individually, rural people enjoy better income opportunities when they migrate to urban areas and take up employment in more productive sectors. For society as a whole, rural-to-urban migration helps developing economies urbanise and industrialise (Lewis, 1954) and narrows the income disparity between rural and urban areas. However, migration is an afflicting process as well. Cultural differences and competition for scarce resources create conflicts between natives and migrants, which may undermine the well-being of migrants and create social segregation between rural migrants and urban locals.

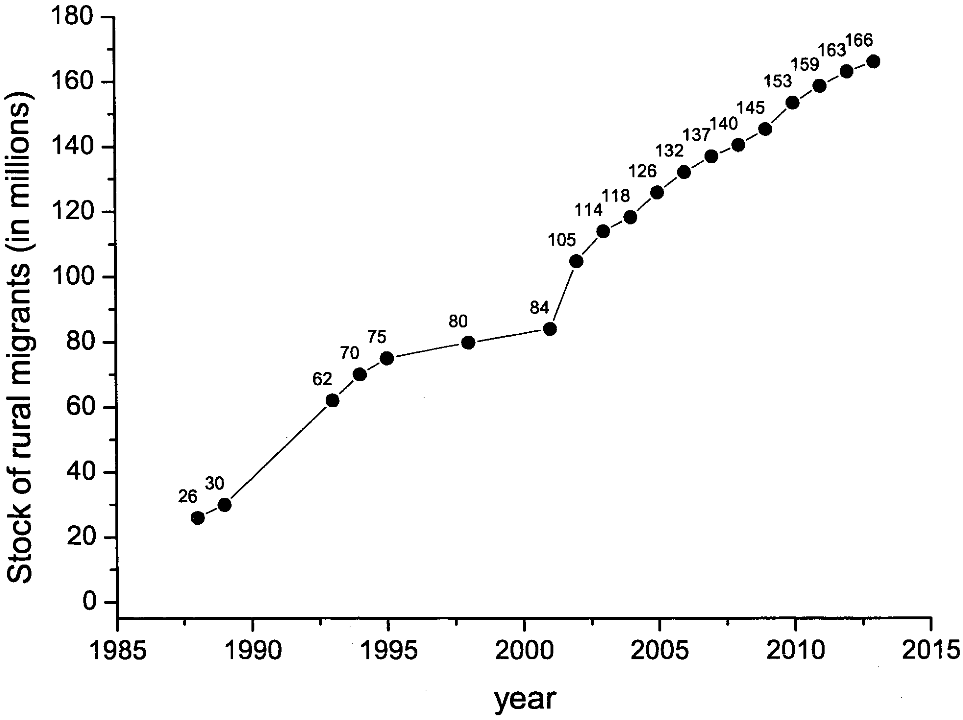
The central issue in policy-making is to reinforce the positive effects of rural-to-urban migration, while suppressing the negative effects. This thesis studies what factors mitigate the adverse impact of rural-to-urban migration and discusses relevant data issues, with a focus on China.

Rural-to-urban migration in China is caused by a series of reforms and structure changes in the economy. On the supply side, the Household Responsibility system implemented in the beginning of 1980s greatly increased the productivity of the agricultural sector, releasing substantial rural labour from the land. On the demand side, the market oriented reforms in urban areas in the late 1980s expanded the non-agricultural sector, increasing the demand of labour. Motivated by the income difference, rural-to-urban migration emerged in the late 1980s and has continued for more than two decades. Rural-to-urban migration has now become a prominent feature of the socio-economic transformation of China.

Figure 1.1 illustrates the growth of rural migrant stock in China.[1] The number of rural migrants was fairly small in the 1980s (around 3% to 4% of rural population), but has since grown rapidly. The average annual growth rate of migrant stock reached 7.7%, from 1988 to 2013. In 2013, there were 166 million rural migrants, which accounts for 18% of the rural population and more than one fifth of full-time employment in urban areas. The weight of numbers suggests that rural migrants are an integral part of the Chinese labour force.

[1] In Figure 1.1, rural migrants are defined as people who have rural household registrations and work outside their home towns for more than six months in a year. The data excludes migrants who have converted their household registration to urban registration, so Figure 1.1 represents the lower bounds of the size of rural migrant stock.

Figure 1.1 Trend in the stock of rural migrants in China



Source: The data from 1988 to 1998 are from Chan (2013). The data from 2001 to 2007 are from World Bank (2009), and the rest are from *The National Monitoring Report of Migrant Workers in 2013 (2013nian quanguo nongmingong jiance diaocha baogao).*

Over the past two decades there has been substantial research examining the impact of this large-scale rural-to-urban migration. Several findings can be summarised as follows. First, rural-to-urban migration contributes remarkably to economic growth in China. Rural-to-urban migration plays an important role in alleviating poverty in rural households (Luo and Yue, 2010) and it is the primary source of urbanisation (Zhang and Song, 2003). Overall, the decomposition exercise conducted by Cai and Wang (1999) suggests that labour reallocation from the agricultural sector to the non-agricultural sector explains 20% of Chinese economic growth from 1982 to 1997, considering all growth factors (i.e., capital formation, labour force growth, human capital formation, labour reallocation, and technology advancement).[2]

Second, contrary to popular opinion, the inflow of rural migrants does not have an adverse impact on urban natives. Meng and Zhang (2010) find that rural migrants have a modestly

[2] Note that rural-to-urban migration may make a stronger contribution to economic growth in later periods, since the migrant stock dramatically increased after 1995.

positive or zero effect on average employment rates and an insignificant effect on the earnings of urban workers. This is contrary to the myth that migrants suppress wages and take away the jobs of natives. Meng and Zhang (2013) also find that rural migrants do not have a significant impact on urban locals' access to education and health services, nor do they affect local crime rates. Rural migrants have only a modest negative effect on access to public transportation. These findings indicate that rural migrants in China consume only limited public resources and do not harm public security.

Third, rural migrants are discriminated against in cities. In the labour market, rural migrants are segregated away from white collar occupations (Meng and Zhang, 2001) and they earn significantly lower wages than their urban counterparts for reasons which cannot be explained by productivity differences between migrants and urban locals (Meng and Zhang, 2001; Lee, 2012). This discrimination may limit the degree to which migrants can assimilate economically (Zhang et al., 2010). Rural migrants also have limited access to social welfare in cities. For example, the migrant and urban household surveys of the RUMIC project indicate that, in 2008, more than 50% of urban workers had social insurance coverage — unemployment (60%), pension (73%), health (72%) and work injury (54%). However, the coverage rates among migrants workers are much lower — unemployment (15%), pension (22%), health (11%) and work injury (22%) (Frijters and Meng, 2010). Similar findings can be found in the Chinese Urban Labour Survey (Cai and Wang, 2008).

Given the importance of rural-to-urban migration and the marginalized city life of rural migrants, this thesis examines how to improve the well-being of rural migrants and discusses related data issues which will facilitate future research. In particular, Chapter 2 examines the impact of social networks on the mental health problems of rural migrants in cities. Chapter 3 investigates whether interpersonal contact with migrants improves urban locals' willingness to interact with them. Chapter 4 explores the effect of attrition in the longitudinal migrant household survey of the RUMIC project, to guide future migration research. Chapter 5 concludes with a discussion of the main findings and future research directions.

1.1 Social networks and mental health problems

An important issue of rural-to-urban migration in China is the mental well-being of migrants. Rural migrants in Chinese cities are vulnerable to mental health problems because they experience significant work-related stress, are excluded from local society and have limited access to social welfare (Wong et al., 2008; Qiu et al., 2011; Mou et al., 2011; He and Wong, 2013). Rural migrants have been found to have more mental health problems than non-migrants (Li et al., 2009; Chen, 2011).

Mental health is directly related to individual productivity and social stability; thus, there are tremendous economic and social benefits to be gained from identifying the factors which influence the mental condition of migrants. Motivated by this concern, Chapter 2 investigates the role of social networks in shaping the mental health of rural migrants, using the migrant household survey in the RUMIC Project.

A key challenge in this regard is how to mitigate the endogeneity bias caused by omitted variables and reverse causality. This chapter contributes to the general literature by providing an arguably better way to correct the endogeneity bias between social networks and mental health problems in the observational data. The mental health of migrants who live in cities may not be affected by what has happened in their hometown. This feature of the migrant sample permits us to find plausibly exogenous variation in the characteristics of migrant hometowns, to identify the effect. In particular, I take past rainfall in home county and distance between home village and its closest traffic hub, as the instrumental variables of social networks. Moreover, thanks to the longitudinal nature of the survey used, I am also able to control for individual fixed effects to reduce the omitted variable problem.

Both the IV estimates and fixed effect IV estimates suggest that expanding social networks improves mental health. This finding is robust regardless of whether the instrumental variables are used individually or jointly in the estimations. In exploring the heterogeneous impact, I further find that the protective effect of social networks is stronger for migrants with smaller networks or without access to social welfare. In addition, females benefit more from their social networks than males.

1.2 Interpersonal contact and attitudes

Social segregation between migrants and locals is a pressing problem for many countries. China is no exception. The 2008 urban community survey in the RUMIC project interviewed 786 officers of urban local communities and asked them about the relationship between urban locals and rural migrants. Only 11% of respondents thought that migrants were close to local people, 68% reported that they were not close to each other and 21% said that there was little contact between natives and migrants. This social segregation also affects communication between second generations, and 14% of officers reported that the children of migrants and local children do not play together.

Social segregation can create tension between locals and migrants and thereby hinder economic growth, so it is important to understand how to reduce social segregation and create environments that are welcoming to migrants. Chapter 3 studies whether inter-personal contact improves urban locals' willingness to interact with migrants.

I use the 2005 China General Social Survey (CGSS) to study this issue. This survey uses sophisticated measures of attitudes, in contrast to the crude measures used in many previous studies which simply ask respondents if they like or dislike migrants. The CGSS attitude measures range from willingness to have non-intimate interactions with migrants, such as being colleagues or living in the same community, to willingness to have intimate relationships with migrants, such as having children or relatives marry migrants. Categorizing different "levels" of attitudes allows us to see whether contact works on all the dimensions of attitudes or only some of them.

I adopt the novel heteroskedasticity identification strategy (Lewbel, 2012) to mitigate endogeneity bias and find that contact enhances willingness to have non-intimate interactions with migrants, but does not significantly impact willingness to have intimate interactions. On the one hand, this finding suggests that contact helps reduce some types of social segregation, and thus should be promoted by the government. On the other hand, contact is not a panacea for all discrimination and the government should consider other measures to reduce segregation in intimate relationships and interactions.

1.3 Attrition in the RUMIC Migrant Household Survey

A longitudinal survey enables us to observe behavioural dynamics for the same individuals, and also to control for individual time-invariant variables which could be confounding factors in estimation. Prior to the RUMIC project, most studies of rural-to-urban migration in China relied on cross-sectional datasets (e.g., Rozelle et al., 1999; Meng and Zhang, 2001; Taylor et al., 2003; Lee, 2012). The RUMIC longitudinal migrant household survey has greatly enhanced the study of migration in China.

However, the migrant household survey suffers from an attrition problem, because migrants are inherently mobile and itinerant. The attrition rates between the first two waves reached 64% and Chapter 4 explores the nature and consequence of this large attrition rate.

Specifically, Chapter 4 first examines the predictors of attrition. It finds that the migrants who are more likely to attrite are those who are: salary workers, more socio-economically disadvantaged, less willing to stay in cities and those who gain less income from migration. This finding helps to identify what types of migrants should be more closely tracked in the future migrant survey, if the survey designer would like to reduce attrition. It also helps us to understand the external validity of the results that are obtained from the sample of non-attritors, such as fixed effect estimates.

Chapter 4 then studies attrition bias with an example of the earnings equation. Using characteristics of the baseline waves, I compare the regression coefficients which are obtained from the sample of non-attritors with those derived from the sample of attritors. The comparison suggests that attrition bias possibly exists, but the magnitude of the bias is case-dependent. In some cases, attrition has only limited impact on the individual-level characteristics (e.g., education, gender and years since the first migration) which are relevant to substantive research and policy interest.

Last, Chapter 4 assesses the representativeness of the sample of non-attritors at the time of the follow-up waves. The result indicates that the sample of non-attritors is unrepresentative, using the random refreshments of the follow-up waves as a benchmark. But the regression comparisons on the earnings equation suggest that the impact of sample (un)representativeness on regression coefficients are case-dependent. In some cases, sample (un)representativeness has limited impact on the individual-level characteristics.

Chapter 2 Social networks and mental health problems: Evidence from rural-to-urban migrants in China

2.1 Introduction

Mental health is important human capital. From an individual's point of view, mental health not only reflects current well-being, but also influences future productivity and future well-being. Mental health problems can trap a person in a disadvantageous position in the labour and marriage markets (e.g., Bartel and Taubman, 1979, 1986; Ettner et al., 1997; Frijters et al., 2010). In extreme cases, mental health problems can make someone vulnerable to suicide. At the macro level, improving mental health can help governments allocate resources more efficiently. Mental health problems impose a huge burden on society, via health care costs and potential productivity losses. The societal cost of mental health problems was estimated to be 1.8% of GDP (193.2 billion US dollars) in the US in 2002 (Kessler et al., 2008) and 6% of GDP (798 billion euros) in Europe in 2010 (Olesen et al., 2012).[3] Given the negative consequences of mental health problems, it is important to know which factors affect mental health.

This chapter examines the role of social networks in shaping mental health. The relationship between social networks and health has long been studied. In theory, social networks have both beneficial and harmful effects on mental health. On the one hand, social networks may improve a person's mental health by enhancing their sense of social integration and by buffering stress. On the other hand, participating in social networks can have a psychological cost in terms of indebtedness and obligation if the person finds it difficult to respond to the needs of others (Kawachi and Berkman, 2001). A large body of research has empirically examined the correlation between social networks and mental health in western societies, revealing both positive and negative correlations (see review by Kawachi and Berkman, 2001; Smith and Christakis, 2008; Cohen and Janicki-Deverts, 2009). In contrast to the abundant literature on developed countries, there are very few studies which focus on developing countries.

This chapter extends the literature to a developing country - China. In particular, I investigate this issue in the context of internal migration in China. In the last 20 years, China has experienced unprecedented rural-to-urban migration. In 2013, 166 million rural workers worked

[3] Please note that in Olesen et al. (2012) the estimate includes the cost of neurologic disorders.

outside their hometown for more than half a year (NBS, 2014). Since migration is a stressful process (Bhugra, 2004), and mental health of migrants tends to deteriorate during migration (e.g., Wu and Schimmele, 2005; Rivera et al., 2015), understanding how to maintain mental health of such a large group of migrants can provide tremendous economic and social benefits to China.

China's unique cultural and institutional systems also make it an interesting case study. A crucial aspect of the net effect of social networks is the psychological cost of supporting others. For people with limited resources, this cost might be large. On the one hand, most Chinese rural migrants possess only limited economic and social resources. They are usually lower paid and less likely to hold white-collar jobs than urban locals (e.g., Meng and Zhang, 2001; Deng and Li, 2010). On the other hand, the demand for social networks among these migrants is high, because Chinese society has a long tradition of relying on social networks and the unique "guest worker" system prevents rural migrants from accessing social welfare when they experience difficulty (Meng, 2012). Investigation under this context helps to understand to what extent social networks are protective for the low socio-economic status people.

In addition to investigating the case of Chinese migrants, this chapter also contributes to the literature by enriching the methodology for addressing endogeneity issue. Because of the problems of reverse causality and omitted variable, the correlation between mental health and social networks cannot reveal causation. In the literature of psychology and public health, psychologists and epidemiologists usually examine causality through randomised controlled trials. However, as noted by Cohen (2004), Cohen and Janicki-Deverts (2009) and Ertel et al. (2009), evidence on the causal effect of social networks is quite scarce and has two gaps. First, the existing experimental studies cannot estimate the effect of the natural social networks (e.g., acquaintances, friends or family members), because network interventions in most experimental studies are based on support provided by strangers, such as nurses, social workers and psychologists (Cohen, 2004) rather than natural networks. To my knowledge there is no experiment-based study examining the effect of natural networks in the literature. Second, these experiments are mainly clinical trials, which may suffer from substitution bias. As Heckman and Smith (1995) stated, if "human subjects recognize that they have been denied treatment and attempt to obtain it elsewhere", then the control group would be contaminated. In this case, substitution bias would cause the effect of social networks to be underestimated. Because of these two issues, this chapter employs a different strategy - the instrumental variable approach - to mitigate endogeneity bias, and I apply this approach to the observational data which allows me to look at the effect of natural social networks.

I use the migrant household survey from the Rural-to-Urban Migration in China project (RUMIC) to study this issue. This survey is the only large-scale dataset on internal migration in China. It records detailed information on respondents' mental health, social networks and other socioeconomic aspects. The data have two advantages for estimating the effect of social networks. First, the survey has a large longitudinal component which extends across five years. The key challenge to identifying the effect is the endogeneity issue between social networks and mental health. The longitudinal data enable me to use the within-individual variation to estimate the effect, which removes the endogeneity bias caused by individual time-invariant factors. Second, in contrast to previous surveys which have been confined to a specific region or sector with limited samples (e.g., Wong et al., 2008; Mou et al., 2011; Qiu et al., 2011), the RUMIC survey provides large representative samples in 15 cities which received migrants from more than half the rural counties and districts in China. Such a large-scale representative sample offers rich variation in the hometown conditions of migrants, which allows me to find plausibly exogenous variation to mitigate the endogeneity bias.

In this chapter, I use the General Health Questionnaire 12 to measure mental health problems. I measure network size in cities as the number of contacts made during the last Chinese Lunar New Year while living in urban areas at the survey time. The empirical results generally support the hypothesis that social networks reduce mental health problems. In particular, the OLS and fixed effect estimates show that larger social networks are significantly associated with fewer mental health problems. Evidences from the instrumental variable estimations further indicate that social networks improve the mental health of migrants, even if taking endogeneity issue into account. In addition, the analysis of the heterogeneous effect suggests that social networks have greater beneficial effects for migrants with smaller networks or without access to social welfare, and females benefit more from their social networks than males.

The structure of this chapter is as follows. Section 2.2 reviews the literature on social networks and mental health and provides related background information on the mental health of migrants in China. Section 2.3 introduces the data used. Section 2.4 describes the methodology. Section 2.5 discusses the results, and Section 2.6 concludes with discussion.

2.2 Literature review

2.2.1 Social networks and mental health

The role of social networks in determining mental health has been widely discussed in psychology and public health.[4] Several mechanisms have been proposed to describe how social networks affect mental health. Psychologists Cohen and Wills (1985) developed two of the most prominent models - the main effect model and the stress-buffering model - to explain the beneficial effects of social networks. The main effect model predicts that the social interaction provided by an individual's networks can generate positive psychological states by increasing sense of security, social belonging and recognition of self-worth, regardless of whether an individual is actually experiencing difficulty. The stress-buffering model focuses on situations when an individual is in a crisis. This model posits that, before a crisis, a person's expectation that his/her networks will provide necessary help can mitigate his/her stress about a future crisis. During a crisis, a person's stress can also be reduced by the material and emotional support provided by his/her networks.[5]

Kawachi and Berkman (2001) point out the negative aspects of social networks. The reciprocal nature of social networks means that they may inflict a psychological cost if an individual finds it difficult to respond to the needs of his/her network members. This negative effect can be particularly great for people with limited social and economic resources. Thus, given these mechanisms, whether social networks are beneficial or harmful to mental health is an empirical question.

There are countless empirical studies in psychology and public health testing the association between social networks and mental health, and most of the existing literature finds a positive association, although some negative correlations have also been found (e.g., Rose, 2000; Sapp et al., 2003; Cohen, 2004; Cohen and Janicki-Deverts, 2009; Ertel et al., 2009). Nevertheless, compared to the correlation studies, there are very few studies investigating causality. These studies exclusively employ randomized experiments to identify causality, by comparing the mental health related outcomes of the treatment group that receives social support from nurses, social workers or psychologists, with those of the control group. There is no consensus about the conclusions drawn from these experimental studies. Mittelman et al. (1995), Harris et al.

[4] To my knowledge there is little economic literature on this topic.

[5] Please refer to Thoits (2011) for the detailed channels through which social network improves mental health in the main effect model and stress-buffering model.

(1999) and Goodwin et al. (2001) find evidence that the support from social networks may improve the mental health of carers of patients with Alzheimer's disease, chronically depressed women and patients suffering from breast cancer. On the other hand, studies of Heller et al. (1991), Brand et al. (1995) and Frasure-Smith et al. (1997) suggest that social networks have no significant effect on the mental health of people with low perceived support or patients recovering from myocardial infarction. These mixed results call for new evidence on the effect of social networks on mental health, especially on the impact of natural networks rather than assigned strangers. This chapter extends the literature to the natural networks of migrants.

2.2.2 Mental health situation of rural migrants in China

Over the past two decades, the prevalence of mental health problems in medium and large-sized cities in China has almost tripled (Ministry of Health of the People's Republic of China, 2010). Rural migrants are no exception and several studies document high rates of mental health problems among Chinese rural migrants. Qiu et al. (2011) found that 23.7% of migrant workers in Chengdu had clinically relevant depression symptoms, and 12.8% of migrants were consistent with a clinical diagnosis of depression. In Shenzhen, 21.4% of migrant workers were found to have clinically relevant depression symptoms (Mou et al., 2011). He and Wong (2013) surveyed female migrants in Shenzhen, Kunshan, Dongguan and Shanghai - four important destination cities for migrants. They show that 24% of female migrants had poor mental health. Similarly, Wong et al. (2008) found that 25% of male migrants and 6% of female migrants in Shanghai were mentally distressed. The existing literature also indicates that the mental health of migrants is worse than their urban and rural counterparts. Li et al. (2009) compared migrants with urban residents in Beijing and rural residents in emigrating regions and found that both the urban and rural residents were mentally healthier than the migrants. Chen (2011) showed that the psychological distress of migrants in Beijing was greater than that of urban locals, using multivariate regression techniques. As an exception, Li et al. (2007) found that migrants in Hangzhou were mentally healthier than urban locals, but their mental health was still worse than that of rural people in Western Zhejiang, their main place of origin. These studies suggest that a significant proportion of Chinese migrants have poor mental health, a finding that deserves our attention.

In the literature, a set of factors, such as physical health (or self-rated health), working conditions (e.g., working hour, relationship with colleagues, job satisfaction), economic status (e.g., salary, employment difficulties), life behaviour (e.g., smoking, usage of internet), social support, city adaption and discrimination, are shown to be strongly correlated with mental health of the Chinese rural migrants (e.g., Wong et al., 2008; Mou et al., 2011; Qiu et al., 2011;

Li et al., 2007). However, these studies are almost all based on small surveys, and none of them discuss the endogeneity issue. This chapter complements the literature by more carefully dealing with endogeneity bias using a large-scale survey data.

2.3 Data

2.3.1 Description of Data

The RUMIC migrant household survey

The main data used in this chapter are from the Rural-to-Urban Migration in China (RUMIC) project. The RUMIC project aims to provide a longitudinal dataset to document the socio-economic impact of internal migration in China. It comprises three independent surveys: the migrant household survey, the urban household survey and the rural household survey. This chapter uses the migrant household survey.[6]

The RUMIC migrant household survey is currently the largest longitudinal survey of rural migrants in China. It covers 15 cities in 9 provinces or municipalities: Guangzhou, Dongguan, Shenzhen, Zhengzhou, Luoyang, Hefei, Bengbu, Chongqing, Shanghai, Nanjing, Wuxi, Hangzhou, Ningbo, Wuhan and Chengdu. These 15 cities represent both the largest migration sending places and destinations (Gong et al., 2008), and also cover the coastal, central and western regions of China, providing the survey with rich geographic and economic variations. Each wave contains around 5000 households.

This survey includes both longitudinal and repeated cross-sectional components. The baseline survey was conducted in 2008, consisting of randomly sampled migrant households from 15 cities.[7] From 2009, every year the survey tracks the migrant households which were interviewed in the previous year, and the households which were tracked successfully are included in the present survey as the longitudinal component. However, due to the high mobility of migrants, the survey has a high attrition rate.[8] To maintain a sufficient sample size, from 2009 the survey replaces the attrited households with random refreshments in each follow-up survey. Thus, the

[6] It is important to note that this survey is conducted in in-migration areas, and only includes migrants who were living in the city at the time of interview. Hence, in this chapter I focus on the effect among the migrants who stay in cities. I am not going to generalise the effect to the whole population of rural people who have had migration experience or who may potentially migrate in the future.

[7] See the discussion on the sample representativeness in (Gong et al., 2008).

[8] For more discussion on attrition see Chapter 5.

sample in the baseline wave and the random refreshments in the follow-up waves constitute a repeated cross-sectional sample. Given this special design, this survey offers me two opportunities. First, I can use the longitudinal component to control for individual fixed effect to provide more internally valid estimates. Second, I can use the repeated cross-sectional sample to provide estimates which are free of attrition bias and also are relatively representative of the general migrant population.

Measure of mental health problems

This chapter mainly uses the 2008, 2009, 2011 and 2012 waves of the migrant household survey to obtain information on mental health, social networks and other individual characteristics.[9] Mental health information is sourced from the survey's General Health Questionnaire (GHQ) 12. In psychological studies, GHQ is a widely used screening instrument for detecting psychiatric disorders. The abbreviated version, GHQ 12, is frequently used to measure mental health conditions or subjective well-being in the economic literature (e.g., Clark and Oswald, 1994; Gardner and Oswald, 2007; Frijters et al., 2009; Cornaglia et al., 2012). GHQ 12 consists of 12 questions, which focus on "two main classes of phenomena: inability to carry out one's normal 'healthy' functions, and emergence of new phenomena that are distressing" (Graetz, 1991). The answer to each question has a 4-point scoring system, generally denoting not stressed (1), slightly stressed (2), fairly stressed (3) and highly stressed (4). Respondents who were more than 16 years old and present at the time of the interview were asked to answer these questions.

There are several ways to measure mental health problems using GHQ 12. I use the Likert score to measure mental health problems in the main analysis. This measure is widely used in the literature (e.g., Gardner and Oswald, 2006, 2007; Akay et al., 2013). To construct the Likert score, I first sum the answers of all the questions in GHQ 12 and then minus 12, so the Likert score ranges from 0 to 36. The larger the Likert score, the worse the mental health condition . In the robustness check, I also consider the GHQ score, another measure of mental health problems used in the literature (e.g., Clark and Oswald, 1994). The GHQ score counts the number of items for which respondents reported "fairly" or "highly" stressed. It ranges from 0 to 12. Similar to the Likert score, a larger GHQ score indicates worse mental health condition. I choose the Likert Score in the main analysis, because it has better distributional property (i.e., with less skewness and kurtosis, Graetz, 1991).[10]

[9] The 2010 wave does not include information on migrants' mental health, so it is not included in my analysis.

[10] Another way to measure mental health is to use factor analysis to measure different aspects of mental health (Cornaglia et al., 2012). However, although several studies have found the 12 items in GHQ 12 could be attributed to different factors, the factor structures (especially for the number of factors) is different across populations. For

Measure of social networks

The RUMIC migrant household survey contains a module to collect information on social networks. Theoretically speaking, an ideal measure of social networks would reflect both the size of the network, i.e., the number of network members, and the quality of the network, i.e., the help which the network can offer. However, due to data constraints, this chapter only discusses the impact of the size of social networks. The measure of network size comes from the question "During the period of the recent Chinese Lunar New Year, how many people in total did you send your greetings to in various ways (including visiting/phone call/mail/e-mail, etc..)? Among them, approximately how many person(s) is (are) relative(s); how many person(s) is (are) friend(s) and acquaintance? Among them, approximately how many person(s) is (are) currently living in the city; how many person(s) have city Hukou"?[11][12] As Chinese people have a tradition of greeting friends to maintain their social networks during the Lunar New Year, this question provides me with an opportunity to measure the approximate size of a person's networks. In particular, this chapter uses the number of greeted people who lived in cities during the survey time as the variable to measure network size in urban areas.[13]

Other control variables

All the variables used in this chapter are extracted from the RUMIC migrant household survey except for two economic variables at the destination cities and rainfall related to variables.

example, Graetz (1991) finds that GHQ 12 can be modelled as three factors using Australian youth samples, but Kih et al. (1997) find only two factors in the Turkish sample. In addition, the study of Doi and Minowa (2003) find that there are three factors for Japanese male adults, but only two factors for Japanese women. Given that there is no agreement on the factor structure of Chinese rural migrants and I do not find a robust factor structure in this dataset, I do not use this method.

[11] Please note that the period of the Chinese New Year usually lasts 15 days in rural areas.

[12] "Hukou" refers to the household registration system in China. There are two types of Hukou: city (non-agricultural) Hukou and rural (agricultural) Hukou. Usually people born in urban areas have city Hukou.

[13] Note that this dataset also allows us to construct network size in rural areas. However, in this chapter only the effect of urban network size is chosen to be investigated, because of two reasons. First, since this survey records mental health information when migrants stayed in the city, networks in urban areas are presumably more relevant to the recorded mental health information. Second, due to unavailability of a convincing instrumental variable, analysing the causal effect of network size in rural area is practically infeasible. In the following main analysis, I do not control for the network size in rural areas in the regressions, as it may be endogenous and bias the coefficients of other variables. But in the robustness check I check whether the results are sensitive to controlling for it, and the results turn out to be robust (see Panel 3 in Table 2.10).

I include growth rates of GDP and real minimum wages relative to the previous year in the destination cities as the control variables. The GDP information is from various *City Statistical Yearbooks of China*, and the minimum wage information is provided by Ministry of Human Resources and Social Security and the China Academy of Labour and Social Security.

The rainfall data is obtained from the Meteorological Information Centre, which collects daily rainfall information from 824 national climatological base stations. These base stations aim to provide accurate and representative weather information for analysing climate change in China. I match the home counties of the respondents with the nearest weather stations and take the information from the closest weather station to proxy the rainfall information in the home county of the migrant. Based on this matching, I generate rainfall-related variables.[14] [15]

2.3.2 Sample construction and general picture of key variables

In the RUMIC migrant survey, only the household head or spouse answered the questions about social networks and home village. In the following analysis, I restrict attention to respondents who answered these questions to avoid the potential problem of measurement error. In the robustness check, I include all household members in the analysis.

In the survey, there are 19873 household heads or spouses providing information on mental health problems, social networks, and home village. I exclude respondents aged below 16 or above 65 or those with city Hukou to focus on rural migrants who are in the labour market. This leaves 19462 observations. Of these 19462 observations, 17533 observations provide necessary data in the covariates. To reduce the measurement error, I further apply two sample restrictions. First, I remove 58 observations who contacted more than 150 people in urban areas during the Chinese Lunar New Year.[16] Second, I remove 221 observations with a monthly wage above

[14] Please note that the home county is identified by the Hukou information in the RUMIC migrant household survey. Around 12.2% of respondents did not provide accurate information on home counties. For these respondents, I match the weather station with the location of the local government in their home prefecture. Also, 0.2% of respondents did not provide precise information on home prefecture. These observations are excluded from the analysis.

[15] The average distance between the location of the local county/prefecture government and the nearest weather station is around 35 kilometres. For one observation (a migrant who came from Huolinguole, a remote county in Inner Mongolia), the distance between the local county government and the closest weather station exceeded 100 kilometres. In the robustness check I found that including or excluding this observation has no effect on the results.

[16] Dunbar (1993) and Hill and Dunbar (2003) find that a person's maximum network size is approximately 150 people due to cognitive limits, so I apply this sample restriction.

10000 Chinese yuan (approximately 1630 US dollars). These two restrictions result in the final sample of 17254 observations for the main analysis.[17]

In the following analysis, I construct three samples: the repeated cross-sectional sample which consists of the initial wave and the random refreshments in each follow-up wave, the longitudinal sample which contains the observations which appeared in at least two waves, and the pooled cross-sectional sample which includes all the 17254 observations across waves. I use the repeated cross-sectional sample to provide estimates free of attrition bias, the longitudinal sample to provide fixed effect estimates, and the pooled cross-section sample to provide the efficient estimates.

Table 2.1 presents summary statistics for the three samples, respectively. Panel A shows that, in the repeated cross-sectional sample, the average Likert score of GHQ 12 is 7.64, and the average network size is 13 contacted people in cities. 64% of respondents are male. The majority of the sample is migrants with an education level below senior high school level. The average age of migrants is 30 years, and the average period since their first migration to the city is 7.8 years, which suggests that a large proportion of migrants first migrated to urban areas when they were relatively young. The average height of migrants is around 167 cm, and the mean level of monthly income is 1901 yuan.

Across the three samples, there is no notable difference in Likert score and network size, but there are some differences on several other characteristics. Comparing Panels A and B, the differences generally indicate that the migrants in the longitudinal sample are older, wealthier, more likely to be self-employed and married. They also stayed in cities longer than the migrants in the repeated cross-sectional sample. These differences could also be observed between the repeated cross-sectional sample and the pooled cross-sectional sample. Given these differences, I conduct the regression analysis on each sample separately.

Figure 2.1 presents the unconditional relationship between mental health problems and size of social networks. The Likert score decreases as the network size increases, and the variance

[17] I also checked whether the missing values and sample restrictions resulted in serious sample selection problem. In particular, I replace all the missing values in control variables with zero and include the dummy variables indicating the missing values in the control variables. I use all the observations which do not have missing values in the dependent variable, endogenous variable and instrumental variables to replicate all the following analysis. I find that the magnitudes of the estimates are similar to those in the main analysis, but the significance level drops slightly perhaps due to the measurement error.

becomes larger when networks expand as the sample size shrinks.[18] It is interesting to see that the downward trend of the curve mainly concentrates in the region where the network size does not exceed 50, which indicates that the effect of networks is possible to be heterogeneous for different network sizes. In the analysis of heterogeneous effect (Section 2.5.3), I will explore the potential non-linear effect by restricting the sample to respondents whose contacts are not more than 50, and then compare the estimates from this sub-sample with the estimates from the whole sample to see whether there are differences in estimates.[19] In the main analysis, I use the whole sample to give a complete picture, avoiding potential sample selection problem.[20]

2.4 Empirical strategy

In this chapter, I first use OLS regression to estimate the effect of social networks on mental health problems. The formation of mental health problems can be presented as the follows:

$$\mathrm{MHP}_{\mathrm{ijt}} = \beta_1 * \mathrm{SN}_{\mathrm{ijt}} + \mathrm{X}_{\mathrm{ijt}} * \beta_2 + \tau_t + c_j + \varepsilon_{\mathrm{ijt}}, \qquad (2.1)$$

where subscripts i, t and j denote individual, year and city of destination, $\mathrm{MHP}_{\mathrm{ijt}}$ is the measure of mental health problems, $\mathrm{SN}_{\mathrm{ijt}}$ is the measure of migrants' social networks in cities, $\mathrm{X}_{\mathrm{ijt}}$ represents the vector of covariates, τ_t is the year fixed effect, c_j is the destination city fixed effect, and $\varepsilon_{\mathrm{ijt}}$ is the unobserved factor.

In the baseline model, I include a set of individual characteristics in the covariates $\mathrm{X}_{\mathrm{ijt}}$. These characteristics are age, gender, years since the first migration, education, marital status, number of children, height, self-reported health, monthly wage, self-employment, and number of people over 16 years old in the household. These characteristics have also been used in other mental health studies (e.g., Frijters et al., 2009; Stillman et al., 2009; Bjrklund, 1985; Akay et al., 2012). Since the death of a family member can greatly affect one's mental health, I also include whether any family members had passed away in the previous 12 months in the covariates. To control for the impact of local economy on mental health, I include the growth rates of GDP and

[18] Figure 2.A.1 suggests that the unconditional relationships are similar among the three samples.

[19] An alternative way to explore non-linearity in the effect is to add square or inverse terms of social networks in the regression specification. However, I choose not to do so because of the weak instrumental variable problem when multiple endogenous variables are used.

[20] I have three other reasons for using the full sample in the main analysis. First, the full sample can strengthen the instrumental variable, making the estimations more precise. Second, even if the effect is non-linear as Figure 2.1 shows, the full sample could provide the lower bound of the effect, making the estimates more conservative. Third, I use the BACON method proposed by Billor et al. (2000) to test the existence of outliers, but found no outlier. Thus, I do not have reason to remove observations.

real minimum wages relative to the previous year in the destination cities in X_{ijt}. The covariates of the extended model also include characteristics of the migrants' hometown and dummy variables of occupation and industry, to reduce the omitted variable problem. The hometown characteristics are average long-term rainfall in the home county, average daily wages of unskilled labour in the home village, the village's distance from its closest county seat and whether it is located in a mountainous area, and presence of a medical centre in the home village.

My goal is to identify β_1 in Equation (2.1). If the estimated β_1 is negative, then social networks help reduce mental health problems; otherwise, social networks do not have a beneficial effect. If social networks are exogenous, the OLS estimate of β_1 can be interpreted as the causal effect of social networks on mental health problems. However, there are three reasons why this may not be the case. First, there may be reverse causality between social networks and mental health. A person's mental health may affect his/her relationship with others and thereby his/her social networks. Second, certain unobserved personal attributes can be correlated with both social networks and mental health. For example, introverted people may have fewer friends and be more likely to have mental health problems (Kawachi and Berkman, 2001; McKenzie et al., 2002). Third, since social networks are measured by self-reported retrospective information, the data may contain large measurement error. An indication of the measurement error is that 52% of respondents rounded their answers on social networks to a multiple of five. All of these factors could create a substantial bias in the OLS estimator; hence, the OLS estimator cannot consistently estimate the effect of social networks.[21]

One way to reduce the endogeneity bias is to use the fixed effect (FE) model as follows:

$$MHP_{ijt} = \beta_1 * SN_{ijt} + X_{ijt} * \beta_2 + \tau_t + n_i + \varepsilon_{ijt}, \qquad (2.2)$$

where n_i is the individual fixed effect and X_{ijt} includes only the time-variant characteristics. The main advantage of the FE model is that it explicitly controls for the individual fixed effect n_i, which removes the bias caused by the time-invariant omitted variables. However, FE model still has two limitations. First, it cannot resolve the endogeneity bias which is associated with unobserved time-variant factors. Second, measurement error can induce large attenuation bias in the FE estimator. Pischke (2007) summarized that if the correctly measured variables are persistent and the measurement errors are uncorrelated with each other across waves, then the attenuation bias would be particularly large in the FE estimator. Considering the nature of misreporting in the social networks measure, this is a real possibility.

[21] Please note that the direction of the OLS bias is undetermined here. On the one hand, it is intuitive to think that reverse causality might cause negative bias, since people with fewer mental problems are more attractive and thereby tend to have a larger networks; on the other hand, people with more mental health problems could choose a destination with larger existing networks to relieve their stress. Hence, the direction of the OLS bias is unknown.

An alternative way to mitigate the endogeneity bias and circumvent the disadvantages of the FE estimator is to adopt the instrumental variable approach. The instrumental variable approach jointly estimates Equation (2.1) (or Equation 2.2) and an equation of social networks:

$$SN_{ijt} = \gamma_1 * Z_{ijt} + X_{ijt} * \gamma_2 + \tau_t + c_j(n_i) + \epsilon_{ijt}, \qquad (2.3)$$

where Z_{ijt} is the instrumental variable which identifies the effect of social networks. A valid instrumental variable should satisfy two conditions. First, the instrument(s) must be correlated with the endogenous variable SN_{ijt} (relevance condition); and second, the instrument(s) cannot be correlated with the disturbance ε_{ijt} in Equation (2.1) or (2.2) (exclusion restriction). In this chapter, I use the previous spring and summer rainfall (e.g., from April to August) in the home county and the distance between the home village and its closest traffic station as the instrumental variables. In the following I discuss the validity and construction for each instrumental variable.

Previous spring and summer rainfall in the home county

In the literature, it is normal to use rainfall to instrument for migrants' networks (e.g., Munshi, 2003; Giles and Yoo, 2007). The correlation between rainfall and migrants' social networks comes from the fact that rainfall is a push factor for migration which could affect agricultural income and subsequently influence the migration motivation of rural people. Based on the 2008 and 2009 waves of the RUMIC rural household survey, Table 2.2 confirms this argument. It shows that, in China, the previous spring and summer rainfall increases agricultural income and consequently reduces rural people's migration intention (see data details in Appendix A). As migrants tend to move to destinations where there are existing networks (Bao et al., 2007), and migrants also tend to form networks with people from the same hometown, the impact of rainfall on migration intention could be eventually translated into an impact on network size of migrants. Specifically, it means that rainfall is negatively correlated with migrants' network size, given the negative correlation between rainfall and migration intention.

In order to keep the strength of the first-stage estimation, I use the average daily rainfall in home counties between April and August two years before the survey as the instrumental variable in the following analysis.[22] Figure 2.2 shows that the instrumental variable is negatively correlated with network size, which verifies the aforementioned conjecture.

[22] The network information is derived from contacts made during Chinese New Year, which is usually in January and February. It takes time to migrate and form networks. Hence, there is not enough time for the spring and summer rainfall in the previous year to affect networks, so it is not a strong IV.

However, in relation to the validity condition there are a number of potential reasons for rainfall to be invalid. First, rainfall may have a direct effect on mental health. Too many cloudy or rainy days may make people depressed. For this reason, I include the average daily rainfall in the previous 10 years and its squared term as the control variables in the extended model of the cross-sectional IV estimations. I assume that the direct effect of rainfall on mental health is mainly shaped by long-term rainfall (i.e., average daily rainfall in the previous 10 years in the cross-sectional model or the individual fixed effect in the fixed effect model) and, after controlling for it, the transitory rainfall in migrants' home counties two years before (i.e., the instrumental variable) does not have much direct effect on the current mental health of migrants in urban areas.[23]

Second, rainfall may affect migrants' mental health because it affects agricultural income in their hometown. Since income may affect migrants' mental health, the omitted variable of previous agricultural income may bias the IV estimates. However, two arguments could address this concern. First, the literature provides evidence that individuals can adapt to external income shocks (Tella et al., 2010; Frijters et al., 2011). Thus, I assume that income shocks that occurred two years before do not affect current mental health much. The second argument is that even if individuals cannot fully adapt to income shocks, I can reasonably predict the direction of this bias. As shown in Figure 2.2 and Table 2.2, spring and summer rainfall negatively correlates with social networks but positively correlates with agricultural income. If an increase in income is associated with fewer mental health problems, then these relationships imply that the IV estimate would contain positive bias and is the lower bound of the beneficial effect of social networks.[24]

Third, for new migrants, the rainfall in their home counties may be correlated with their unobserved preference for city life. In home regions where rainfall promotes agricultural output, people who are the most adaptive to city life choose to move; but in home regions with drought many people would migrate to the city, even if some of them do not adapt well to city life. Similar to the second concern, this potential violation of the validity condition makes the effect of social networks to be underestimated. To address this issue, I follow Munshi's (2003) strategy of using the fixed effect instrumental variable model (FEIV) to control for this

[23] Note that in the main analysis, I control for the long-term rainfall or individual fixed effect, so I essentially use the residual rainfall which cannot be explained by the long-term rainfall to identify the effect. In the robustness check, I also explicitly define the residual rainfall as the difference between the rainfall two years before and the average rainfall between 1980 and 2012. I take this variable as the instrumental variable. The results remain similar.

[24] One caveat here is that if relative concern about others' income shock dominates the effect of their own income shock, then the bias is negative.

unobserved preference, which assumes that this unobserved preference for city life is largely time-invariant. Given this, the FEIV estimates are my preferred estimates.

Another strategy to mitigate these three concerns is to restrict the sample to those migrants who first migrated to the city three years or more before the survey, and never returned to their hometowns to stay for more than three continuous months. This group of respondents may be less likely to be directly affected by rainfall in their hometown and less dependent on agricultural income. Also, since they migrated before the rainfall actually occurred, rainfall could be less correlated with their unobserved preference. Therefore, I expect that the IV estimates contain smaller biases for this sample. I will conduct the robustness check to assess whether the results for this group of migrants are similar to the results from the full sample.

I conduct a falsification test in Table 2.3, which examines the impact of the rainfall instrument on the number of contacts with city Hukou. Since rainfall two years before in the hometown should have nothing to do with whether a migrant makes friends with urban local people, rainfall is supposed to have no effect on the number of contacted people with city Hukou during the Chinese New Year. If there is a significant impact, then we are concerned with the validity of the rainfall IV, because it may capture factors which can directly affect the city life of migrants. I show the results of this falsification test in Panel 1. I also show the estimated impact of the rainfall IV on the number of contacts living in the city, which includes both the people with and without city Hukou, in Panel 2 as a comparison. The results suggest that the rainfall IV does not have significant impact on the number of contacts with city Hukou, and the estimates are also small in magnitude.

Distance between the home village and its closest traffic station

The second instrumental variable is the distance between a migrant's home village and its closest traffic station. It is sourced directly from the question asking respondents "distance between your home village and the nearest traffic station (coach, train or dock)" in the RUMIC migrant household survey. This instrumental variable can be correlated with the size of a migrant's networks, through two channels. First, as a factor determining the cost of migration, distance may affect villagers' intention to migrate. Second, traffic stations are usually built in populated areas. Villages closer to traffic stations often have larger populations than those further away. These two channels can affect how many people migrate from a source village and thereby influence the potential network size of the migrants who stayed in the city. Specifically, these two channels predict that the larger the distance between a migrant's home village and its closest traffic station, the smaller his/her networks. To account for the non-linear relationship between distance and size of networks, in the following estimation I use the inverse of one plus this distance as the instrumental variable.[25]

The validity of this instrumental variable relies on the assumption that the distance between the home village and its closest traffic station is not correlated with the error term in Equation (2.1) or (2.2). This assumption may not hold if there are omitted variables correlating with both the error term in Equation (2.1) or (2.2) and the instrumental variable. As the distance between the home village and the closest traffic station is usually correlated with the level of regional development and geographic factors, and these variables may affect the mental health of the villagers, I include the characteristics of home village, such as the daily wages of unskilled labour in the village, the inverse of one kilometre plus the distance between the village and its closest county seat, whether there is a medical centre in the village and whether the village is located in a mountainous area in the extended model to avoid this omitted variable problem. I assume that, conditional on these variables, this instrumental variable does not directly affect mental health problems of migrants.

One caveat for the second instrumental variable is that the distance information was not recorded for the longitudinal sample in the 2011 and 2012 waves of the RUMIC migrant household survey. For this sample, I use the distance information from the previous waves as proxies for the current waves. Although this generated instrumental variable would introduce measurement error which makes the 2SLS estimation less efficient, it would not affect the asymptotic consistency and inference (Wooldridge, 2010). Given the large sample used in this study, this issue is not a major concern in the cross-sectional IV estimations. However, the limitation of such extrapolation is real for the FEIV estimation, since there is little within-individual variation on this variable. Given this concern, I do not use this instrumental variable to conduct the FEIV estimation.

In the following analysis, I both separately and jointly use these two instrumental variables. I also conduct FEIV estimation using the rainfall instrumental variable. Since the FEIV estimates can mitigate omitted variable problem and remove individual heterogeneity which makes estimation more efficient, they are my preferred estimates.

[25] The unit of distance is kilometre. Around 5 - 6% of respondents reported the distance as zero kilometres. To include these observations, I add one kilometre to the distance to make the inverse feasible.

2.5 Main Results

In this section, I use OLS, fixed effect model and instrumental variable approach to estimate the effect of social networks on mental health problems. In the baseline model I do not include hometown characteristics, occupation and industry dummies as control variables, and I include these variables as covariates in the extended model. The standard errors in the cross-sectional estimations are clustered at the home-county level, and the standard errors in the fixed effect estimations are clustered at the individual level.

2.5.1 OLS and fixed effect results

Table 2.4 presents the OLS estimates for the repeated cross-sectional and pooled cross-sectional samples. The first two columns show the results using the repeated cross-sectional sample; and the other two columns show the results obtained from the pooled cross-sectional sample.

Table 2.4 suggests that migrants' networks are significantly and negatively correlated with their mental health problems in both the repeated cross-sectional sample and the pooled cross-sectional sample. The estimates of the baseline model in the repeated cross-sectional sample suggest that one additional person greeted in the Chinese Lunar New Year while currently living in urban areas is associated with a reduction of the Likert score by 0.016, which is equivalent to 0.4% of the standard deviation of the Likert score.

Comparing the estimates across columns, I find that the estimates are similar between the baseline models and extended models. The estimates slightly drop by 0.001 Likert points from the baseline models to the extended models. This suggests that the mechanism for social networks affecting mental health does not strongly depend on hometown characteristics, occupation and industry. I also find that the OLS estimates are robust in the pooled cross-sectional sample estimates. They show similar magnitudes and significance level to the estimates in the repeated cross-sectional sample.

The associations of other control variables are also interesting. Given that the results are largely consistent, my discussion below focuses on the repeated cross-sectional sample in Columns (1) and (2). Men have less mental health problems than women. The years since the first migration show an inverted U-shape relationship with mental health problems. As the years increase, the Likert score first increases and then decreases. This non-linear assimilation process of mental health is opposite to the findings on migrants' wage assimilation (Zhang et al., 2010). More educated migrants tend to be mentally healthier. Relative to those with an education below junior high school, the Likert score for those with junior high school education decreases by around 0.4, and for those who gained senior high school degrees or above, the Likert score decreases by 0.7 or more. Married people have better mental health than single people, whereas divorced people have worse mental health. This is perhaps because spouses usually provide emotional support to each other (Smith and Christakis, 2008), or because mentally healthier people are selected for marriage. The self-rated unhealthy level is strongly correlated with the Likert score, which is similar to Akay et al.'s (2012) finding. Economic factors are also a critical predictor of mental health. Income is negatively and significantly associated with mental health problems. These OLS associations are the common findings in the literature (e.g., Gardner and Oswald, 2006; Akay et al., 2012).

Table 2.5 shows the FE estimates. The estimates suggest that social networks are still significantly and negatively correlated with mental health problems, even though the individual fixed effects have been controlled for. Compared to the OLS estimates in Table 2.4, I find that the magnitudes of the FE estimates become smaller. The estimates are around -0.008 to -0.01, which is equivalent to 0.2% of the standard deviation of the Likert score.

Given that the average network size is around 13 people, the OLS and FE estimates above do not indicate a large effect of social networks. However, these estimates are likely to be biased downwardly, due to measurement error and the possibility that mentally unhealthy people may choose to migrate to a city with large existing networks. In the following, I show that using the instrumental variable approach which can correct the endogeneity issue and measurement error problem, the magnitudes of the estimates enlarge substantially.

2.5.2 IV results

Tables 2.6 to 2.7 show the cross-sectional instrumental variable estimations. In these estimations, I both separately and jointly use the instrumental variables mentioned in Section 2.4 and repeat the exercises on both the repeated cross-sectional sample and pooled cross-sectional sample. For each estimation, the relevant test statistics - Kleibergen-Paap rk Wald F statistic for the weak IV test (Kleibergen and Paap, 2006) and Hansen's J test for the over-identification test (Hansen, 1982) - are shown at the bottom of each set of the IV results to assess the reliability of the IV estimation.

I begin by discussing the first stage results of 2SLS estimations in Table 2.6. Panels 1 and 2 show the first-stage results using the instrumental variables individually. They suggest that both the two instrumental variables are strongly correlated with the endogenous variable. In particular, Panel 1 shows that a 1-mm increase in the average daily rainfall between April and August two years before reduces the number of people greeted during the recent Chinese New Year by 0.3-0.6 person. This means that, on average, rainfall contributes 1.3-2.8 persons to a network, accounting for around 10%-21% of migrants' average networks.[26] In Panel 2 the distance between home village and the closest traffic station shows a non-linear and negative relationship with the size of migrants' urban networks. The closer the traffic station, the larger the network. The weak IV test statistics of these two individual instrumental variables both exceed 10 except the baseline model using the repeated cross-sectional sample in Panel 1. However, even in this case, after controlling for the characteristics of job and home village, the statistic of the weak IV test also exceeds 10 in the extended model. This suggests that these two instrumental variables are strong enough to give reliable estimates. Panel 3 shows the first-stage results using these two instrumental variables jointly. Compared with the first two panels, both the coefficients and the statistical significance of these two instrumental variables in Panel 3 remain similar, and the weak IV test statistics are also above 10. This suggests that these two instrumental variables are not correlated with each other, and each of them provides different identification information to the second stage regression. Hence, jointly using these two instrumental variables makes the 2SLS estimation efficient.

Table 2.7 shows the results of the second-stage regressions. I show the OLS estimates in Panel 1 as a comparison, and the IV estimates are listed in the rest of the three panels. In Panels 2 to 4, all the results suggest that social networks help relieve mental health problems. In particular, the estimates from the rainfall instrumental variable (in Panel 2) range from -0.029 to -0.094, and the estimates from the distance instrumental variable (in Panel 3) range from -0.124 to -0.243. Since different instrumental variables give different local average treatment effects, it is natural that the magnitudes of the estimates differ between Panels 2 and 3. The most important message from these two panels is that all the estimates have negative signs, indicating that social networks help reduce mental health problems.[27] In terms of the magnitude of the effect, these estimates indicate that an extra network member reduces the Likert score by 0.6% to 5% of its standard deviation. In Panel 4, the estimation jointly uses the two instrumental variables. The results are all significant at the 5% level or 1% level. The magnitudes of the estimates indicate

[26] These calculations are based on the fact that the mean value of the rainfall instrumental variable is 4.7 mm and the average size of migrants' network is 13, as shown in Table 2.1.

[27] It is a puzzle to me that the coefficients increase in Panel 2 but decrease in Panel 3 from the repeated cross-sectional sample to the pooled cross-sectional sample. However, given the large standard error, these changes may be induced by noise in the data.

that an additional network member reduces the Likert score by -0.107 to -0.158 point, which is equivalent to 2% to 4% of the standard deviation of the Likert score. Since jointly using the two instrumental variables makes the estimation more efficient and covers a larger complier group in the setting of local average treatment effect, the results in Panel 4 are preferred to the other panels.[28]

Table 2.8 provides the FEIV estimates. Since the information on the distance instrumental variable is unavailable for the longitudinal sample in the 2011 and 2012 waves of the migrant survey, the FEIV estimation employs only rainfall as the instrumental variable. FEIV estimation has two advantages over the cross-sectional estimation above. First, as the individual heterogeneity of mental health is potentially large, controlling for the individual fixed effects could enhance the efficiency of the estimation. Second, controlling for the individual fixed effects could also reduce the bias caused by the unobserved preferences for city life as argued in Section 2.4. Thus, the FEIV estimates are more internally valid and are my preferred results. However, we should also note that the respondents in the longitudinal sample may not be representative. The migrants in the longitudinal sample tend to be more socio-economically advantaged, better established in the city and less mobile than those in the repeated cross-sectional sample (Xue, 2015). We need to keep this caveat in mind when we interpret the results.

Panel 2 of Table 2.8 indicates that rainfall is a strong instrumental variable in the FEIV estimations. The coefficients suggest that a 1-mm increase in the rainfall IV reduces the number of contacted people living in the city by 0.6 to 0.7 persons, which is slightly larger than the cross-sectional IV results. The statistics for the weak IV test are greater than 10. Panel 3 suggests again that social networks could significantly reduce mental health problems for the longitudinal sample. The FEIV estimates are around -0.174 to -0.168, which is equivalent around 4% of the standard deviation of Likert score. The FEIV estimates are larger than the IV estimates in Panel 2 of Table 2.7. This is perhaps partly because FEIV estimation removes the positive bias caused by unobserved preferences for city life, and partly because the sample used is a selected group of migrants.

In summary, Tables 2.7 and 2.8 suggest that larger social networks reduce migrants' mental health problems. The effect is statistically significant except when only the rainfall instrumental variable is used in the estimation. In terms of the magnitude of the effect, the cross-sectional estimations suggest that the effect of average network size is equivalent to 8% to 69% of the

[28] Because the over-identification test cannot test exogeneity when the effect of social networks is heterogeneous, I do not comment on them. I only show them in the table for reference.

standard deviation of Likert score, and the FEIV estimations suggest that the effect of average network size is equivalent to half of the standard deviation of Likert score. These results indicate that the effect of social networks is also economically sizeable.[29]

Relative to the OLS and fixed effect results, the magnitude of IV results is larger. This observation is consistent with Munshi's (2003) study, which finds that the IV estimates of network effect on employment and obtaining higher paid jobs are larger than the OLS and fixed effect associations. The larger IV estimates may be caused by two reasons. First, less mentally healthy people probably endogenously move to cities where they have larger networks. Migration is a stressful process, and potential migrants probably realise this, so it is possible that migrants with more mental health problems choose to migrate to a city with larger potential networks in case that they need help. Second, measurement error in social networks is large which creates substantial attenuation bias in OLS and fixed effect estimates. As mentioned in Section 2.3, in the data 52% of respondents rounded their answers on social network to a multiple of five. Given these two possibilities, the OLS and FE estimates can be seriously biased downwards.

2.5.3 Potential heterogeneous effect

In Table 2.9 I explore the potential heterogeneity to better understand the effect of social networks. In particular, Panels 2 to 4 of Table 2.9 show the results from three sub-samples, and the results from the full sample are shown in Panel 1 for ease of comparison. All the cross-sectional results in Table 2.9 are estimated from the repeated cross-sectional sample, and all the results are estimated from the extended model.

First, I explore the heterogeneous effect across different network sizes. As shown in Figure 2.1, the downward relationship between social networks and mental health problems mainly concentrates in migrants whose network size is not more than 50. Theoretically, it is possible that the marginal return to social networks is diminishing, so social networks may have a stronger effect for migrants who have smaller networks. I directly test whether this conjecture is true in Panel 2. In particular, I restrict the sample to migrants whose network size is not larger than 50 and then estimate the effect of social networks. The results confirm the conjecture. The OLS and FE correlations increase from -0.015 and -0.008 in the full sample to -0.024 and -0.012 in the restricted sample, respectively. The IV and FEIV estimates also become larger. In

[29] Unfortunately, I did not find any existing work that uses similar measures for mental health problems and social networks, so I am unable to compare these estimates with the literature.

the full sample, the IV and FEIV estimates are -0.128 and -0.168 respectively, but the corresponding figures are -0.141 and -0.297 in the restricted sample. Although the IV and FEIV estimates are not statistically significant, probably due to the reduced sample size and weak IV problem, the Anderson-Rubin Wald test (a test robust to weak IV problem) suggests that the estimates are significant at the 5% level. These results suggest that the beneficial effect of social networks is stronger for migrants with smaller networks.

Second, I investigate whether the effect differs according to access to social welfare in the city. The "guest worker" system in China strictly controls the social welfare to which migrants are entitled (Meng, 2012). A large proportion of migrants have no access to social welfare in the city. In the repeated cross-sectional sample, 70% of migrants have no access to unemployment insurance, employment injury insurance, pensions and house accumulation funds – none of them.[30] Lack of social welfare can make migrants vulnerable to life shocks and thereby mental health problems. In the repeated cross-sectional sample, I find that migrants who have no access to social welfare have significantly more mental health problems than migrants who do. In particular, the Likert scores of the migrants with and without the social welfare are 7.90 and 7.24, respectively. The difference in the Likert score between these two groups of migrants is significant at the 1% level.[31] Given this, it is important to understand whether, and to what extent, social networks can protect mental health of the migrants who do not have access to social welfare. I test this in Panel 3 by restricting the sample to migrants who have no access to unemployment insurance, employment injury insurance, pensions and house accumulation funds. Although the OLS estimates are similar, the FE, IV and FEIV estimates in Panel 3 are all larger than the estimates in Panel 1. The FE, IV and FEIV increase from -0.008, -0.128 and -0.168 in the full sample to -0.013, -0.164 and -0.265 in the restricted sample. This suggests that social networks are more protective among migrants with no access to social welfare.[32]

Last, I search the gender difference in the effect of social networks. Kawachi and Berkman (2001) stress that social networks may have differential effect on mental health between men and women. On the one hand, relative to men, women are more vulnerable to contagion of mental health problems, because they tend to be more sympathetic to others' stressful events and tend to provide more help to others which may incur large psychological costs on

[30] The pooled cross-sectional sample and longitudinal sample have similar proportions of migrants who do not have access to social welfare.

[31] Similar differences in mental health problems can be found in the pooled cross-sectional sample and longitudinal sample.

[32] Among the migrants without access to social welfare, 28% of them are self-employed. Thus, I include self-employment as a control variable in regression. Note that this control variable also appears in all the other regressions.

themselves. On the other hand, women tend to seek more help from their networks than men, when they experience difficulty. I examine the potential gender difference in Panel 4 by restricting the sample to male migrants. The results suggest that the IV and FEIV estimates are smaller from this restricted sample than those from the full sample. Specifically, the IV and FEIV estimates are -0.079 and -0.098 in Panel 4 , but they are -0.128 and -0.168 in Panel 1. This indicates that social networks play a less protective role in male migrants' mental health, and female migrants benefit more from their social networks.

2.5.4 Robustness check

I check the robustness of the results in Table 2.10. In this section the cross-sectional estimates are obtained from the repeated cross-sectional sample, and all the exercises are conducted in the extended models.[33]

Panels 1 and 2 assess the sensitivity of the results which are generated by the rainfall instrumental variable. Specifically, Panel 1 considers the potential non-linear effect of rainfall on social networks. In reality, the impact of rainfall on migrants' social networks may not be linear, and the excessive rainfall and rainfall deficit may have different effects on social networks. Missing this non-linear effect may lead to a model misspecification problem and make the estimates unreliable. Panel 1 tests whether the results change if we consider the non-linear effect of rainfall. In particular, I explicitly define excessive rainfall and rainfall deficit, and take them and their square and cubic terms as instrumental variables.[34] The results in Panel 1 are generally similar to the main results in Panel 2 of Table 2.7 and Panel 3 of Table 2.8, and the cross-sectional IV estimate become even larger. This suggests that the results are robust to considering the non-linear effect of rainfall on social networks.

Panel 2 restricts the sample to migrants who first migrated to the city three years before or earlier, and never returned to their hometown to stay for more than three continuous months. Note that the instrumental variable used is the rainfall which happened two years ago before the

[33] Panels 1 and 2 focus on the validity of the rainfall instrumental variables, so the reported results in these panels are from the estimations which only use the rainfall instrumental variable. I also checked the results when the two instrumental variables are jointly used, and the estimates are significant at the 5% level.

[34] I define these two variables as $excessive\ rainfall = (IV - LR\ rainfall) \times 1(IV - LR\ rainfall > 0)$ and $rainfall\ deficit = (IV - LR\ rainfall) \times 1(IV - LR\ rainfall < 0) \times (-1)$. IV is the rainfall instrumental variable used, and $LR\ rainfall$ is the long-term rainfall, which is calculated as the average daily rainfall between April and August from 1980 to 2012.

survey, and this group of migrants migrated three years or earlier before the survey. Hence, this group of migrants is less likely to be directly affected by the instrumental variable. To examine whether the results above are driven by the direct effect of the instrumental variable, I use this sample to estimate the effect of social networks. If the results are driven by the direct effect of rainfall, we expect to see that the estimates from this sample are smaller than those from the full sample. Because of a large reduction in sample size, the IV and FEIV estimates in Panel 2 are not statistically significant. However, the magnitude of the FEIV estimate is similar to that in Panel 3 of Table 2.8, and the IV estimate becomes even larger than the estimate in Panel 2 of Table 2.7. This indicates that the main results may not be driven by the direct effect of the instrumental variable.

Panel 3 includes the weekly hours worked, remittance ratio, network size in rural areas and home-county fixed effects as the control variables.[35] These variables may be endogenous to mental health problems, and including these variables may reduce the estimation efficiency, so I do not include these variables in the main analysis. However, these variables may be correlated with migrants' social networks as well. Thus, including these variables as control variables may be helpful in reducing the bias caused by the omitted variable problem. I examine whether including variables changes the results in Panel 3. The results are similar to the main results in Panel 5 of Table 2.7 and Panel 3 of Table 2.8. This indicates that omitting these variables does not cause bias.

Panel 4 takes the GHQ score of GHQ 12 as the dependent variable. As discussed in Section 2.3, GHQ score is also a frequently used measure in the literature (e.g., Clark and Oswald, 1994). I test whether the results are sensitive to the choice of dependent variable. Same as the main results in Table 2.7 and Table 2.8, the results in Panel 4 suggest that social networks help reduce mental health problems. This indicates that the choice of dependent variable does not alter the results substantively.

Panel 5 includes all household members in the analysis. I include only respondents who answered the questions about social networks and home villages in the main analysis to avoid the potential measurement error problem. In Panel 5, I test whether the results can be extended to all household members. The results suggest that social networks significantly improve mental health for all household members.

[35] The remittance ratio is defined as the ratio of the remittance over the household income, and the network size in the rural areas is defined as the difference between the total number of contacted people in the recent Chinese New Year and the number of contacted people currently living in cities.

2.6 Conclusion

This chapter investigates the relationship between social networks and mental health problems in the context of Chinese rural-to-urban migration. The OLS and fixed effect estimates indicate that larger social networks in urban areas are significantly correlated with fewer mental health problems for migrant workers.

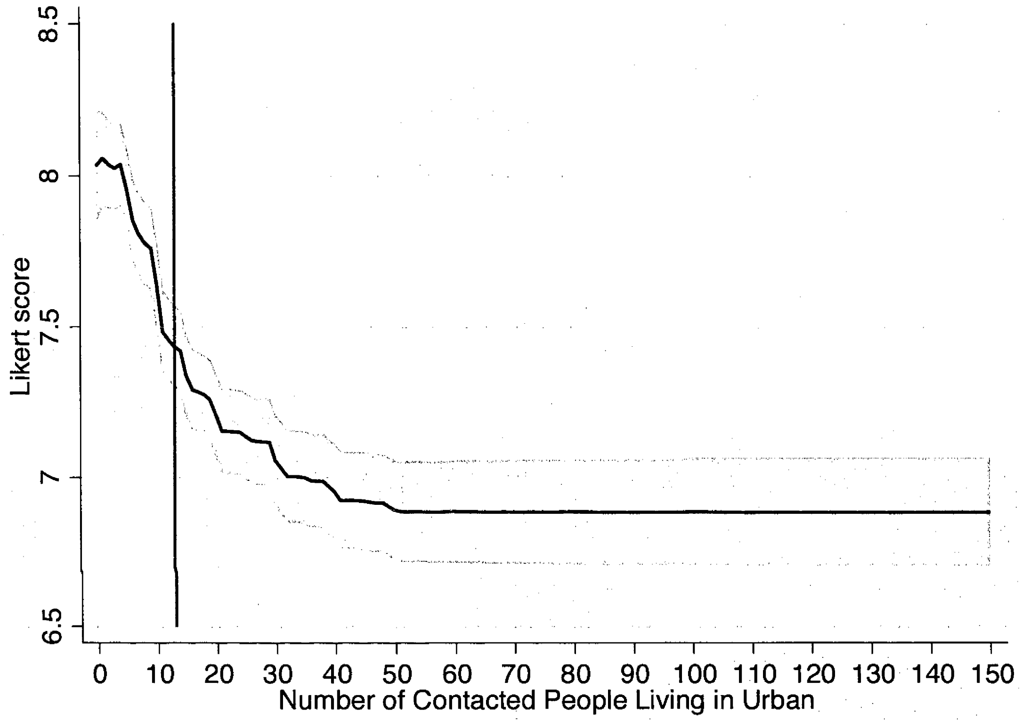
To mitigate the endogenous bias, I adopt the instrumental variable approach. Specifically, I use the previous spring and summer rainfall in the home county and the distance between home village and its closest traffic station as the instrumental variables. Both the two instrumental variables give qualitatively similar results and suggest social networks benefit mental health in the cross-sectional estimations. When these two instrumental variables are jointly used in the estimation, the coefficients become significant at the 5% level. The fixed effect IV estimations give similar results to the cross-sectional IV estimations, again indicating the beneficial effect of social networks. The magnitudes of the IV results are non-trivial. My preferred results indicate that the effect of average network size is equivalent to 30% and 50% of the standard deviation of Likert score, depending on the sample and model used.

In the heterogeneity analysis, I find that social networks have a stronger beneficial effect on migrants with smaller networks or with no access to social welfare. Moreover, females benefit more from their social networks than males.

Because of the constraints of the data, this chapter leaves one gap for future research. In this chapter, I do not explicitly consider the quality of networks. In future research, this gap deserves our attention.

Figures and Tables

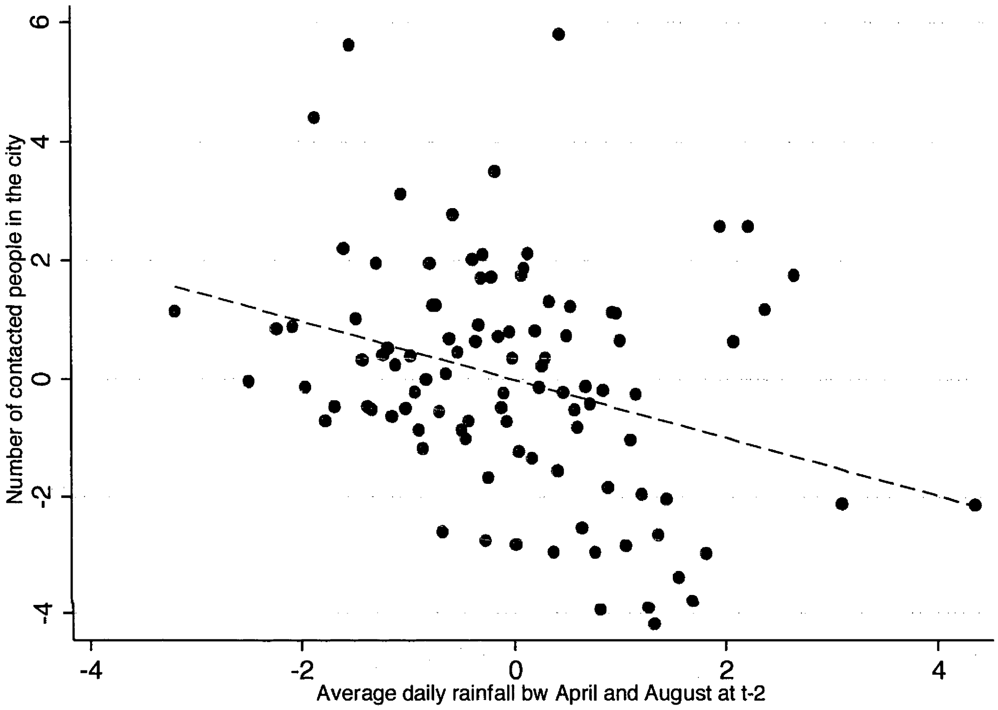
Figure 2.1 Unconditional relationship between social networks and mental health problems



Note: The vertical line denotes the average network size.

Source: Repeated cross-sectional sample from the 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Figure 2.2 Conditional relationship between rainfall IV variable and size of network



Note: I partition the sample into 100 cells according to the magnitude of rainfall variable. Each dot in the graph represents average rainfall and network size within one cell. Both network size and rainfall are adjusted by the characteristics in the extended model.

Source: Repeated cross-sectional sample from the 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.1 Summary statistics

	Panel A Repeated cross-sectional sample		Panel B Longitudinal Sample		Panel C Pooled cross-sectional Sample	
VARIABLE	Mean	S.D.	Mean	S.D.	Mean	S.D.
Likert score	7.64	4.50	7.68	4.55	7.70	4.54
# of contacted people living in urban areas	13.00	17.47	13.66	17.50	13.39	17.51
Age	30.40	10.35	33.25	10.12	31.66	10.40
Years since the first migration	7.83	6.66	10.06	6.86	8.90	6.90
# of family members over 16 years old	1.45	0.71	1.69	0.83	1.57	0.79
Monthly income (yuan)	1901	1136	2227	1286	2093	1245
Daily wage of unskilled labour at home village (yuan)	50.21	22.82	59.78	25.45	56.33	25.19
# of children	0.76	0.90	0.95	0.89	0.84	0.90
Height (cm)	166.57	7.16	166.44	7.22	166.45	7.20
Unhealthy level (self-rated health)	1.77	0.72	1.85	0.74	1.81	0.73
Average daily rainfall from t-10 to t-1 (1mm)	3.01	0.89	2.93	0.83	2.98	0.87
Distance btw home village and the closest county (km)	25.78	36.07	25.06	33.76	25.01	35.43
Growth of GDP in destination cities (%)	11.67	2.59	11.90	2.76	11.67	2.71
Growth of real minimum wage in destination cities (%)	4.98	8.58	6.34	9.82	5.93	9.36
Distance btw home village and the closest traffic centre (km)	16.40	32.78	15.71	30.78	16.15	31.92
Average daily rainfall btw Apr and Aug at t-2 (0.1mm)	4.71	1.89	4.73	1.71	4.74	1.80
Male	0.64		0.65		0.64	
Self-employment	0.17		0.27		0.22	
Below junior high school education	0.20		0.21		0.21	
Junior high school education	0.44		0.43		0.44	
Below senior high school or equivalent education	0.06		0.05		0.05	
Senior high school or equivalent education	0.24		0.24		0.24	

Above senior high school education	0.06	0.07	0.06
Married	0.52	0.65	0.57
Divorced	0.02	0.02	0.02
Death of family member	0.02	0.03	0.03
Home village has medical station	0.89	0.89	0.89
Home village is in a mountainous area	0.23	0.20	0.21
Observations	10827	8362	17254

Note: Pooled cross-sectional sample consists of all the observations across waves. Repeated cross-sectional sample consists of 2008 wave and random refreshments in each follow-up wave after 2008. Longitudinal sample consists of individuals who appeared in two or more waves.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.2 OLS estimates of impact of rainfall on migration choice

	(1)	(2)	(3)	(4)
	Log household net agricultural income per capita in previous year	Intention to migrate in 3 months	Intention to migrate in 12 months	Migration for business over 12 months in the previous year
Daily rainfall between Apr and Aug in year t-1		-0.005**	-0.005**	
		(0.002)	(0.003)	
Daily rainfall between Apr and Aug in year t-2	0.016***	-0.009***	-0.009***	-0.007***
	(0.005)	(0.002)	(0.003)	(0.002)
Daily rainfall between Apr and Aug in year t-3	0.015***	-0.004*	-0.004	-0.004***
	(0.006)	(0.002)	(0.003)	(0.001)
Daily rainfall between Apr and Aug in year t-4	0.003			-0.000
	(0.004)			(0.001)
Observations	12991	31621	31621	34092
Adjusted R-squared	0.278	0.134	0.142	0.065

Note: Standard errors are clustered at household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. Number of family members over 16 years, male dummy, age, squared age, education dummies, dummies of marriage status, number of children, height, dummies of self-rated health and dummy of death of family member in the last 12 months, year dummies, county dummies and constant are included. The reference group is unmarried females with education below junior high school and excellent health. In the first column each household is an observation, and in the last three columns each household member is an observation. The daily rainfall variable is in 1 mm.

Source: 2008 and 2009 waves of the RUMIC rural household survey.

Table 2.3 A falsification test on rainfall instrumental variable

	Cross-sectional model		FE model
	repeated cross-sectional sample	pooled cross-sectional sample	longitudinal sample
Panel 1 # of contacted people with city Hukou			
Average daily rainfall bw Apr. and Aug. at t-2 (1mm)	0.063	0.014	-0.035
	(0.064)	(0.057)	(0.095)
Observations	10821	17244	8356
Adjusted R-squared	0.055	0.057	0.018
Panel 2 # of contacted people living in the city			
Average daily rainfall bw Apr. and Aug. at t-2 (1mm)	-0.482***	-0.599***	-0.701***
	(0.133)	(0.126)	(0.189)
Observations	10827	17254	8362
Adjusted R-squared	0.098	0.088	0.032

Note: Standard errors are clustered at the home county level in the cross-sectional model and clustered at the individual level in the FE model. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. In the cross-sectional model, age, squared age, male dummy, years since the first migration and its square, education attainment dummies, dummies of marriage status, number of children, height, self-rated health, dummy of death of family member in the last 12 months, log(1+monthly wage), self-employment, number of family members above 16 years, growth of GDP and minimum wage at destination cities, daily wage of unskilled labour at home village, whether home village is at mountainous area, inverse of 1 + the distance between home village and the closest county, having medical station at home village, average daily rainfall from t-10 to t-1 and its square, industry and occupation dummies, year dummies and destination city dummies are included as control variables. In the FE model, dummies of marriage status, number of children, height, self-rated health, dummy of death of family member in the last 12 months, log(1+ monthly wage), self-employment, number of family members above 16 years, growth of GDP and minimum wage at destination cities, daily wage of unskilled labour at home village, industry and occupation dummies as control variables. These control variables are the same as those in the extended model in Table 2.4. Repeated cross-sectional sample consists of 2008 waves and random refreshments for each wave after 2008. Pooled cross-sectional sample consists of all the observations across waves. Longitudinal sample consists of the individuals appearing in two or more waves.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.4 OLS estimates of network effect on mental health problems

	Panel A		Panel B	
	Repeated cross-sectional sample		Pooled cross-sectional sample	
	Baseline	Extended	Baseline	Extended
	(1)	(2)	(3)	(4)
Number of contacted people living in cities	-0.016***	-0.015***	-0.016***	-0.015***
	(0.002)	(0.002)	(0.002)	(0.002)
Age	0.063**	0.061**	0.059*	0.054
	(0.030)	(0.030)	(0.034)	(0.034)
Squared age *10^(-2)	-0.069*	-0.064	-0.070	-0.063
	(0.040)	(0.040)	(0.046)	(0.046)
Male	-0.527***	-0.557***	-0.654***	-0.678***
	(0.101)	(0.102)	(0.124)	(0.125)
Years since the first migration	0.023	0.026	0.025	0.028
	(0.019)	(0.019)	(0.023)	(0.023)
Squared year since the first migration *10^(-2)	-0.105*	-0.109*	-0.107	-0.112
	(0.063)	(0.063)	(0.081)	(0.081)
Junior high school education	-0.433***	-0.422***	-0.405***	-0.376***
	(0.103)	(0.102)	(0.123)	(0.123)
Below senior high school education or equivalent	-0.556***	-0.511***	-0.612***	-0.534***
	(0.190)	(0.189)	(0.207)	(0.207)
Senior high school education or equivalent	-0.801***	-0.744***	-0.812***	-0.743***
	(0.119)	(0.119)	(0.133)	(0.134)
Above senior high school education	-1.152***	-0.974***	-1.317***	-1.145***
	(0.167)	(0.168)	(0.203)	(0.205)
Married	-0.486***	-0.477***	-0.523***	-0.503***
	(0.134)	(0.133)	(0.162)	(0.162)
Divorced	0.576**	0.555**	0.176	0.213
	(0.275)	(0.274)	(0.352)	(0.353)
Number of children	-0.079	-0.073	-0.071	-0.049
	(0.067)	(0.067)	(0.082)	(0.081)
Height (cm)	-0.010	-0.011	-0.004	-0.005
	(0.007)	(0.007)	(0.008)	(0.008)
Unhealthy level	1.477***	1.473***	1.519***	1.511***
	(0.057)	(0.057)	(0.066)	(0.066)
Log (1+monthly wage)	-0.282***	-0.245***	-0.261***	-0.219***
	(0.049)	(0.049)	(0.062)	(0.064)
Self-employment	-0.024	0.019	-0.038	0.071
	(0.105)	(0.138)	(0.128)	(0.175)
Number of household members over 16 years	-0.068	-0.069	-0.087	-0.093
	(0.060)	(0.061)	(0.072)	(0.073)
Death of family member	0.118	0.127	0.341	0.332
	(0.200)	(0.200)	(0.262)	(0.263)
Growth of GDP in destination cities (%)	0.143***	0.139***	0.137***	0.130***
	(0.024)	(0.024)	(0.030)	(0.030)
Growth of real minimum wage in destination cities (%)	-0.006	-0.005	-0.004	-0.004

	(0.004)	(0.004)	(0.006)	(0.006)
Daily wage of unskilled labour in home village (yuan)		-0.004**		0.000
		(0.002)		(0.003)
Home village is in a mountainous area		0.109		0.051
		(0.093)		(0.110)
Inverse of 1 + the distance btw home village and the closest county		-0.240		-0.234
		(0.195)		(0.247)
Home village has medical centre		-0.271**		-0.560***
		(0.113)		(0.143)
Average daily rainfall from t-10 to t-1		-0.332		-0.339
		(0.224)		(0.265)
Squared average daily rainfall from t-10 to t-1		0.045		0.057
		(0.034)		(0.040)
Industry and occupation dummies		Yes		Yes
City and year dummies	Yes	Yes	Yes	Yes
Observations	17254	17254	10827	10827
Adjusted R-squared	0.108	0.111	0.122	0.126

Note: Standard errors are clustered at home county level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. All the regressions include year dummies, destination city dummies and constant. Repeated cross-sectional sample consists of 2008 wave and random refreshments in each follow-up wave after 2008. Pooled cross-sectional sample consists of all the observations across waves. Longitudinal sample consists of individuals who appeared in two or more waves. The daily rainfall variable is in 1 mm. The reference group is unmarried females with education below junior high school. Unhealthy level is from the self-rated health question "what is your current health status compared with the same age group", and the answers range from "very good health" (1), "good health" (2), "just so so" (3), "poor health" (4) and "very poor health" (5).

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.5 FE estimates of network effect on mental health problems

	Longitudinal sample	
	Baseline	Extended
	(1)	(2)
Number of contacted people living in cities	-0.010***	-0.008**
	(0.004)	(0.004)
Married	0.129	0.144
	(0.313)	(0.314)
Divorced	0.671	0.675
	(0.656)	(0.668)
Number of children	0.017	0.022
	(0.180)	(0.178)
Unhealthy level	1.148***	1.136***
	(0.093)	(0.093)
Height (cm)	-0.034	-0.034
	(0.030)	(0.030)
Log (1+monthly wage)	-0.134	-0.135
	(0.107)	(0.106)
Self-employment	-0.389	-0.543*
	(0.297)	(0.315)
Number of household members over 16 years	-0.151	-0.148
	(0.136)	(0.134)
Death of family member	0.333	0.338
	(0.348)	(0.347)
Growth of GDP in destination cities (%)	0.105***	0.105***
	(0.039)	(0.039)
Growth of real minimum wage in destination cities (%)	-0.007	-0.007
	(0.005)	(0.005)
Daily wage of unskilled labour in home village (yuan)		-0.010***
		(0.003)
Industry and occupation dummies		Yes
Year dummies	Yes	Yes
Observations	8362	8362
Adjusted R-squared	0.048	0.056

Note: Standard errors are clustered at individual level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. Both the regressions include year dummies and constant. Longitudinal sample consists of individuals who appeared in two or more waves. The reference group is unmarried females with education below junior high school. Unhealthy level is from the self-rated health question "what is your current health status compared with the same age group", and the answers range from "very good health" (1), "good health" (2), "just so so" (3), "poor health" (4) and "very poor health" (5).

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.6 The first-stage results of the cross-sectional IV estimation

	Repeated cross-sectional sample		Pooled cross-sectional Sample	
	Baseline	Extended	Baseline	Extended
	(1)	(2)	(3)	(4)
Panel 1 Using average daily rainfall btw Apr and Aug at t-2 as IV				
Rainfall	-0.277***	-0.482***	-0.305***	-0.599***
	(0.102)	(0.133)	(0.091)	(0.126)
weak IV test statistics	7.434	13.144	11.282	22.703
Panel 2 Using distance to the closest transportation centre as IV				
Inverse of 1 + the distance	3.143***	2.423***	3.217***	2.939***
	(0.693)	(0.761)	(0.766)	(0.846)
weak IV test statistics	20.564	10.137	17.635	12.079
Panel 3 Using the two IVs				
Inverse of 1 + the distance	3.107***	2.372***	3.177***	2.867***
	(0.695)	(0.761)	(0.764)	(0.841)
Rainfall	-0.267***	-0.473***	-0.294***	-0.586***
	(0.102)	(0.133)	(0.090)	(0.125)
Weak IV test statistics	13.725	10.478	12.393	14.270
Observations	10827	10827	17254	17254

Note: Standard errors are clustered at home county level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The regression specifications are the same as Table 2.4. Repeated cross-sectional sample consists of 2008 wave and random refreshments in each follow-up wave after 2008. Pooled cross-sectional sample consists of all the observations across waves. The daily rainfall variable is in 0.1 mm. The weak IV test statistics are Kleibergen-Paap rk Wald F statistic.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.7 The second-stage results of the cross-sectional IV estimation

	Repeated cross-sectional sample		Pooled cross-sectional Sample	
	Baseline	Extended	Baseline	Extended
	(1)	(2)	(3)	(4)
Panel 1 OLS				
Number of contacted people living in cities	-0.016***	-0.015***	-0.016***	-0.015***
	(0.002)	(0.002)	(0.002)	(0.002)
Panel 2 Using average daily rainfall btw Apr. and Aug. at t-2 as IV				
Number of contacted people living in cities	-0.074	-0.029	-0.060	-0.094*
	(0.101)	(0.078)	(0.077)	(0.054)
Panel 3 Using distance to the closest transportation centre as IV				
Number of contacted people living in cities	-0.185***	-0.243**	-0.124**	-0.128**
	(0.064)	(0.104)	(0.049)	(0.060)
Panel 4 Using the two IVs				
Number of contacted people living in cities	-0.158***	-0.128**	-0.107***	-0.110***
	(0.054)	(0.059)	(0.041)	(0.041)
P-value of over-identification test	0.377	0.067	0.489	0.650
Observations	10827	10827	17254	17254

Note: Standard errors are clustered at home county level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The regression specifications are the same as Table 4. Repeated cross-sectional sample consists of 2008 wave and random refreshments in each follow-up wave after 2008. Pooled cross-sectional sample consists of all the observations across waves. The daily rainfall variable is in 0.1 mm. The over-identification test is the Hansen's J test.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.8 FEIV estimates of network effect on mental health problems

	Baseline (1)	Extended (2)
Panel 1 Fixed effect		
Number of contacted people living in cities	-0.010*** (0.004)	-0.008** (0.004)
Panel 2 1st-stage results of FEIV		
Average daily rainfall btw Apr and Aug at t-2	-0.654*** (0.188)	-0.701*** (0.189)
Weak IV test statistics	12.128	13.722
Panel 3 2nd-stage results of FEIV		
Number of contacted people living in cities	-0.174** (0.085)	-0.168** (0.078)
Observations	8362	8362

Note: Standard errors are clustered at individual level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The regression specifications are the same as Table 2.5. Only the longitudinal sample is used in this table, which consists of individuals who appeared in two or more waves. The daily rainfall variable is in 1 mm. The weak IV test statistics are Kleibergen-Paap rk Wald F statistic.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.9 Heterogeneous effect of network on mental health problems

	OLS	FE	IV	FEIV
Panel 1 all sample				
Number of contacted people living in cities	-0.015*** (0.002)	-0.008** (0.004)	-0.128** (0.059)	-0.168** (0.078)
Weak IV test statistics			10.478	13.722
P value of over-id test			0.067	
Observations	10827	8362	10827	8362
Panel 2 migrants with smaller network[a]				
Number of contacted people living in cities	-0.024*** (0.004)	-0.012* (0.006)	-0.141 (0.115)	-0.297 (0.181)
Weak IV test statistics			7.387	7.427
P value of over-id test			0.011	
Observations	10451	7942	10451	7942
Panel 3 migrants with no access to welfare				
Number of contacted people living in cities	-0.014*** (0.003)	-0.013** (0.005)	-0.164*** (0.059)	-0.265** (0.110)
Weak IV test statistics			12.551	10.992
P value of over-id test			0.010	
Observations	7632	4908	7632	4908
Panel 4 male migrants				
Number of contacted people living in cities	-0.015*** (0.003)	-0.007 (0.004)	-0.079 (0.071)	-0.098 (0.077)
Weak IV test statistics			5.577	11.049
P value of over-id test			0.286	
Observations	6928	5439	6928	5439

Note: Standard errors are clustered at home county level in OLS and IV estimates and at individual level at FE and FEIV estimates. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The OLS and IV regression specifications are the same as the extended model in Table 2.4, and the FE and FEIV regression specifications are the same as the extended model in Table 2.5. All the regressions of OLS and IV estimations use the repeated cross-sectional sample which consists of 2008 wave and random refreshments for each follow-up wave after 2008, and the regressions of FE and FEIV estimation use the longitudinal sample which consists of individuals who appeared two or more waves. The weak IV test statistics is the Kleibergen-Paap rk Wald F statistic. The over-identification test is the Hansen's J test. All the results are from the extended model.

a smaller networks is defined as no more 50 contacted people in cities.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.10 Robustness check

	OLS	FE	IV	FEIV
Panel 1 Non-linearity of rainfall IV				
Number of contacted people living in cities			-0.075	-0.141**
			(0.046)	(0.059)
Weak IV test statistics			4.140	3.681
P value of over-identification test			0.383	0.425
Observations			10827	8362
Panel 2 Long-term migrants				
Number of contacted people living in cities	-0.016***	-0.006	-0.088	-0.161
	(0.003)	(0.005)	(0.131)	(0.106)
Weak IV test statistics			4.388	6.821
Observations	6510	5584	6510	5584
Panel 3 Adding additional control				
Number of contacted people living in cities	-0.016***	-0.008**	-0.146*	-0.172**
	(0.003)	(0.004)	(0.080)	(0.083)
Weak IV test statistics			5.730	13.418
P value of over-identification test			0.030	
Observations	10706	8185	10706	8185
Panel 4 Mental health problem measure -- GHQ score				
Number of contacted people living in cities	-0.005***	-0.001	-0.036**	-0.072**
	(0.001)	(0.001)	(0.017)	(0.030)
Weak IV test statistics			14.391	13.722
P value of over-identification test			0.727	
Observations	10827	8362	10827	8362
Panel 5 All the household members				
Number of contacted people living in cities	-0.015***	-0.009***	-0.145**	-0.235**
	(0.003)	(0.003)	(0.060)	(0.094)
Weak IV test statistics			11.557	12.520
P value of over-identification test			0.020	
Observations	12705	9901	12660	9901

Note: Standard errors are clustered at home county level in OLS and IV estimates and at individual level in FE and FEIV estimates. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The OLS and IV regression specifications are the same as the extended model in Table 2.4, and the FE and FEIV regression specifications are the same as the extended model in Table 5. All the regressions of OLS and IV estimations use the repeated cross-sectional sample which consists of the 2008 wave and random refreshments for each follow-up wave after 2008, and the regressions of the FE and FEIV estimation use the longitudinal sample, which consists individuals who appear in two or more waves. The weak IV test statistics is the Kleibergen-Paap rk Wald F statistic. The over-identification test is the Hansen's J test.

Panel 1 includes the linear, square and cubic terms of excessive rainfall and rainfall deficit as instrumental variables, where the excessive rainfall= (rainfall two year ago-long-run mean of rainfall)*1(rainfall two year ago-long-run mean of rainfall>=0) and the rainfall deficit= (rainfall two year ago-long-run mean of rainfall)*1(rainfall two year ago-long-run mean of rainfall<0)*(-1). Only these rainfall instruments are used in this panel.

Panel 2 restricts the sample to observations who migrated to cities three or more years before, and never went back to their hometowns to stay for more than three months. Only the rainfall instrument is used in this panel.

Panel 3 includes weekly working hours, the ratio of remittance over income, network size in rural areas and home county fixed effects as control variables in addition to the variables in the extended models. Both the two instruments are jointly used in this panel.

Panel 4 uses the Likert score of GHQ 12 as the dependent variable. Both the two instruments are jointly used in this panel.

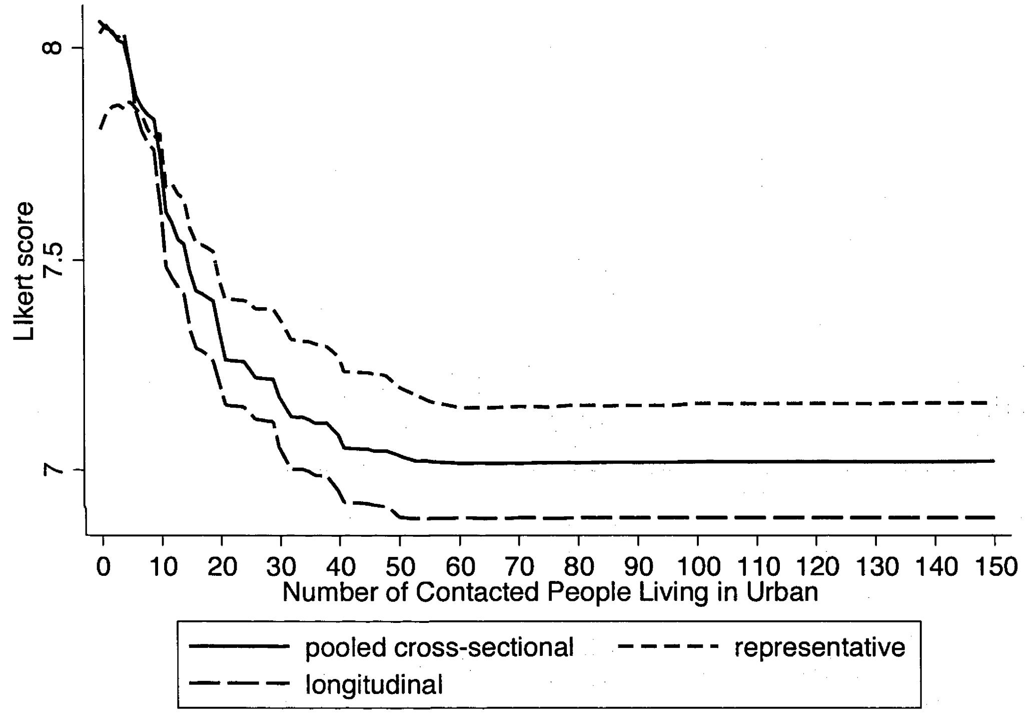
Panel 5 includes all the household member in the sample. Both the two instruments are jointly used in this panel.

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Appendix

Appendix A: Figures and tables

Figure 2.A.1 Unconditional relationships between Likert score and networks for three samples



Source: The 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.A.1 Selected results of cross-sectional IV estimation using the repeated cross-sectional sample

	Using distance IV		Using rainfall IV		Jointly use 2 IVs	
	1st-stage	2nd-stage	1st-stage	2nd-stage	1st-stage	2nd-stage
Inverse of 1 + the distance btw home village and the closest traffic hub	2.423***				2.372***	
	(0.761)				(0.761)	
average daily rainfall btw Apr and Aug at t-2			-0.482***		-0.473***	
			(0.133)		(0.133)	
Number of contacted people living in cities		-0.243**		-0.029		-0.128**
		(0.104)		(0.078)		(0.059)
Age	-0.154	0.017	-0.157	0.052	-0.148	0.036
	(0.125)	(0.047)	(0.125)	(0.036)	(0.125)	(0.038)
Squared age *10^(-2)	0.033	-0.052	0.039	-0.062	0.026	-0.057
	(0.155)	(0.059)	(0.155)	(0.046)	(0.155)	(0.050)
Male	0.722	-0.513***	0.718	-0.668***	0.712	-0.597***
	(0.491)	(0.177)	(0.492)	(0.137)	(0.491)	(0.139)
Years since the first migration	0.279***	0.093**	0.284***	0.032	0.276***	0.060**
	(0.072)	(0.041)	(0.073)	(0.032)	(0.073)	(0.030)
Squared year since the first migration *10^(-2)	-0.780***	-0.292**	-0.795***	-0.123	-0.781***	-0.201**
	(0.239)	(0.128)	(0.239)	(0.104)	(0.239)	(0.098)
Junior high school education	0.490	-0.258*	0.535	-0.368***	0.508	-0.317**
	(0.380)	(0.154)	(0.381)	(0.130)	(0.380)	(0.131)
Below senior high school education or equivalent	2.037**	-0.066	2.095**	-0.505*	2.072**	-0.302
	(0.842)	(0.348)	(0.845)	(0.269)	(0.844)	(0.258)
Senior high school education or equivalent	2.952***	-0.057	3.023***	-0.701**	2.958***	-0.403*
	(0.485)	(0.358)	(0.485)	(0.275)	(0.485)	(0.229)
Above senior high school education	5.984***	0.228	6.049***	-1.061**	5.993***	-0.466
	(1.030)	(0.707)	(1.026)	(0.513)	(1.028)	(0.428)

Married	-1.015*	-0.735***	-1.039*	-0.517***	-1.034*	-0.618***
	(0.613)	(0.239)	(0.612)	(0.181)	(0.613)	(0.187)
Divorced	0.413	0.304	0.359	0.219	0.374	0.258
	(1.223)	(0.449)	(1.225)	(0.353)	(1.225)	(0.378)
Number of children	-0.419	-0.143	-0.409	-0.055	-0.416	-0.096
	(0.304)	(0.118)	(0.303)	(0.086)	(0.304)	(0.093)
Height (cm)	0.052	0.007	0.051	-0.004	0.052	0.001
	(0.033)	(0.013)	(0.033)	(0.009)	(0.033)	(0.010)
Unhealthy level	0.580**	1.644***	0.595**	1.519***	0.589**	1.577***
	(0.249)	(0.102)	(0.248)	(0.079)	(0.248)	(0.076)
Log (1+monthly wage)	1.139***	0.041	1.149***	-0.203*	1.141***	-0.090
	(0.194)	(0.145)	(0.193)	(0.112)	(0.193)	(0.096)
Self-employment	-0.008	0.072	-0.029	0.071	-0.041	0.072
	(0.774)	(0.246)	(0.774)	(0.174)	(0.774)	(0.194)
Number of household members over 16 years	1.205***	0.180	1.205***	-0.076	1.209***	0.042
	(0.303)	(0.159)	(0.302)	(0.122)	(0.303)	(0.111)
Death of family member	3.572**	1.157**	3.570**	0.382	3.512**	0.740**
	(1.579)	(0.576)	(1.581)	(0.383)	(1.582)	(0.371)
Growth of GDP in destination cities (%)	-0.492***	0.017	-0.484***	0.124**	-0.477***	0.074*
	(0.125)	(0.066)	(0.124)	(0.049)	(0.124)	(0.045)
Growth of real minimum wage in destination cities (%)	0.105***	0.019	0.099***	-0.003	0.102***	0.007
	(0.033)	(0.016)	(0.033)	(0.010)	(0.033)	(0.010)
Daily wage of unskilled labour in home village (yuan)	0.058***	0.014*	0.059***	0.001	0.058***	0.007
	(0.012)	(0.007)	(0.012)	(0.005)	(0.012)	(0.005)
Home village is in a mountainous area	0.096	0.062	-0.018	0.052	0.031	0.057
	(0.413)	(0.145)	(0.415)	(0.110)	(0.414)	(0.120)
Inverse of 1 + the distance btw home village and the closest county	1.465	0.418	2.901***	-0.194	1.522	0.089
	(1.120)	(0.471)	(1.044)	(0.326)	(1.120)	(0.328)

Home village has medical centre	1.071**	-0.308	1.118**	-0.544***	1.082**	-0.435***
	(0.530)	(0.206)	(0.528)	(0.167)	(0.529)	(0.159)
Average daily rainfall from t-10 to t-1	-0.071	-0.372	0.341	-0.341	0.403	-0.355
	(1.095)	(0.373)	(1.116)	(0.265)	(1.117)	(0.297)
Squared average daily rainfall from t-10 to t-1	0.000	0.059	0.044	0.057	0.035	0.058
	(0.158)	(0.056)	(0.160)	(0.040)	(0.160)	(0.045)
Observations	10827	10827	10827	10827	10827	10827

Note: Standard errors are clustered at home county level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. All the regressions include industry dummies, occupation dummies, year dummies, destination city dummies and constant. Repeated cross-sectional sample consists of 2008 wave and random refreshments in each follow-up wave after 2008. The daily rainfall variable is in 1 mm. The reference group is unmarried females with education below junior high school. Unhealthy level is from the self-rated health question "what is your current health status compared with the same age group", and the answers range from "very good health" (1), "good health" (2), "just so so" (3), "poor health" (4) and "very poor health" (5).

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Table 2.A.2 Selected results of FEIV estimation

	1st-stage	2nd-stage
average daily rainfall bw Apr. and Aug. at t-2 (1mm)	-0.701***	
	(0.189)	
# of contacted people living in urban		-0.168**
		(0.078)
Married	0.476	0.232
	(1.468)	(0.402)
Divorced	-1.578	0.401
	(2.735)	(0.796)
Number of children	0.325	0.084
	(0.669)	(0.211)
Unhealthy level	0.773**	1.264***
	(0.311)	(0.122)
Height (cm)	0.238**	0.005
	(0.103)	(0.039)
Log (1+monthly wage)	0.939***	0.026
	(0.330)	(0.139)
Self-employement	2.102*	-0.203
	(1.277)	(0.416)
Number of household members over 16 years	0.471	-0.068
	(0.658)	(0.167)
Death of family member	0.381	0.421
	(1.293)	(0.411)
Growth of GDP in destination cities (%)	0.793***	0.226***
	(0.167)	(0.077)
Growth of real minimum wage in destination cities (%)	0.022	-0.003
	(0.023)	(0.006)
Daily wage of unskilled labour in home village (yuan)	0.072***	0.002
	(0.014)	(0.007)
Observations	8362	8362

Note: Standard errors are clustered at individual level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. Both the regressions include industry dummies, occupation dummies, year dummies and constant. Longitudinal sample consists of individuals who appeared in two or more waves. The reference group is unmarried females with education below junior high school. Unhealthy level is from the self-rated health question "what is your current health status compared with the same age group", and the answers range from "very good health" (1), "good health" (2), "just so so" (3), "poor health" (4) and "very poor health" (5).

Source: 2008, 2009, 2011 and 2012 waves of the RUMIC migrant household survey.

Appendix B: Data appendix

General Health Questionnaire 12

The questions in General Health Questionnaire 12 are as follows: "In the last few weeks have you ever had the following feelings?

1. Have you been able to concentrate on whatever you're doing?
   (1) been able to concentrate; (2) attention occasionally diverted; (3) attention sometimes diverted; (4) attention frequently diverted, not been able to concentrate;
2. Have you lost much sleep over worry?
   (1) never; (2) occasionally; (3) fairly often; (4) very often;
3. Have you felt that you were playing a useful part in things?
   (1) true so; (2) to some extent; (3) rarely; (4) not at all;
4. Have you felt capable of making decisions about things?
   (1) very capable; (2) quite capable; (3) not quite capable; (4) not capable at all;
5. Have you felt constantly under strain?
   (1) never; (2) slightly; (3) considerably; (4) seriously;
6. Have you felt you couldn't overcome your difficulties?
   (1) never; (2) slightly; (3) considerably; (4) seriously;
7. Have you felt your normal day-to-day activities are interesting?
   (1) very interesting; (2) fairly interesting; (3) not very interesting; (4) not interesting at all;
8. Have you been able to face up to problems?
   (1) always; (2) most of the time; (3) sometimes; (4) rarely;
9. Have you been feeling unhappy or depressed?
   (1) never; (2) slightly; (3) considerably; (4) seriously;
10. Have you been losing confidence in yourself?
    (1) never; (2) slightly; (3) considerably; (4) seriously;
11. Have you been thinking of yourself as a worthless person?
    (1) never; (2) slightly; (3) considerably; (4) seriously;
12. Have you been feeling reasonably happy, all things considered?
    (1) very happy; (2) fairly happy; (3) not so happy; (4) not happy at all"

Data appendix in Table 2.2

The sample used in Table 2.2 is extracted from the 2008 and 2009 waves of the RUMIC rural household survey. This survey covers 82 counties in 9 provinces in China. I restrict the sample to households whose agricultural income per household member in the previous year is not more than 50000 yuan to reduce the potential measurement error. I also exclude respondents who are younger than 16 or older than 65, because these respondents are unlikely to migrate. The rainfall data in Table 2.2 is constructed in the way described in Section 2.3.

Chapter 3 Can contact help to improve attitudes towards migrants? Evidence from urban China

3.1. Introduction

Economics theory suggests that migration enhances economic efficiency and improves social welfare; however, in reality, local people are often hostile to migrants. A large body of economics literature attributes this hostility to competition for economic resources and to the cultural differences that exist between migrants and local people (e.g., Mayda, 2006; Dustmann and Preston, 2007; Facchini and Mayda, 2009). The extant research helps us understand why such conflict exists, but does not directly answer the question of how to reduce it. Migration brings many benefits; thus, policy-makers are keenly interested in reducing conflict between locals and migrants. Governments which have a better understanding of the issue can establish more effective migration policies that both improve economic conditions and maintain social harmony. Motivated by this concern, in this chapter I examine whether inter-personal contact between locals and migrants can improve locals' attitudes towards migrants. I investigate this question in the context of the willingness of urban locals to interact with migrants in China.

Scholars in sociology and psychology argue that inter-personal contact helps reduce prejudice and improve attitudes by providing more information, eliminating biased views and generating affective ties (Allport, 1954; Pettigrew, 1998). An abundance of empirical studies test this hypothesis and find that more inter-personal contact is associated with better attitudes. Nonetheless, these studies mainly identify the correlation, rather than the causality (Pettigrew, 1998). In addition, the sociology and psychology literatures tend to focus on discriminations against homosexuals, racial and religious minorities (e.g., Herek and Glunt, 1993; Pettigrew, 1998; Tropp, 2007; Bevelander and Otterbeck, 2010), where economic conflict between majority and minority groups is not the main reason for prejudice. Different from these types of discriminations, economic concerns, such as taking away jobs and driving down wages, are at the centre of the hostility from locals towards migrants. Whether the inter-personal contact could reduce local people's prejudice towards migrants given the pressure of economic competition still remains unclear.

China has witnessed tremendous labour mobility in the past two decades. Since the early 1990s, a huge workforce has moved rapidly from rural to urban areas and from underdeveloped to developed regions to pursue a better life. The number of migrants who stay in urban areas for more than half a year increased sharply from 62 million in 1993 to 166 million in 2013. Given the potential economic and cultural clashes between migrants and locals, such unprecedented large-scale migration may arouse the fears of local people and worsen their attitudes towards migrants.

In addition to the large inflow of migrants, economic reform has also brought other dramatic changes to the lives of urban local people. Previously urban residents in China enjoyed a cradle-to-grave social welfare system. They had a lifetime job, highly subsided housing, medical care and education. There was little uncertainty in their lives and little competition in the labour market. However, China's economic reforms dismantled the social welfare system and established a new system which emphasises individual responsibility. The urban locals were no longer guaranteed a lifetime job and other benefits (e.g., Cai et al., 2006; Meng, 2007). This change is likely to aggravate the concerns of urban residents about the economic competition caused by migrants, making them even more hostile to migrants.[36] All these circumstances make China a unique case study for investigating the effect of inter-personal contact on improving urban local's attitudes towards migrants in a society with significant social transformation.

In this chapter, I examine the impact of experience in inter-personal contact with migrants on urban locals' attitudes towards interacting with migrants using the 2005 China General Social Survey. One advantage of this survey is that it provides sophisticated measures of attitudes, ranging from willingness to have non-intimate interactions with migrants (i.e., being their colleagues or living in the same community as them) to willingness to have intimate relationships with migrants (i.e., as next-door neighbour of them, inviting them home as guests or having relatives or children marrying or being in a relationship with them). This finer-grained categorization of attitudes allows us to see whether previous experience of interacting with migrants has a similar effect on all dimensions of attitudes or if the effect varies according to the level of intimacy.

Using OLS regression I first show that previous contact experience with migrants is significantly correlated with stronger willingness to interact with them, for all types of interactions. Then, I use Lewbel's (2012) heteroskedasticity identification approach to alleviate the endogeneity bias between previous contact experience and current attitudes, and the results suggest that previous contact with migrants significantly improves urban residents' willingness

[36] Knight and Yueh (2009) find that the urban workers who were more likely to be laid-off in the reform were less willing to work with migrants and tended to view migrants as competitors.

to work with and live in the same community as migrants, but has no significant effect on their willingness to have intimate interactions with migrants. The contrast between these results and the OLS estimates suggests that the correlations uncovered in the OLS estimations on intimate interactions may arise from the omitted variables.

This chapter makes three contributions. First, it adds a less studied determinant, inter-personal contact with migrants, to the literature on the formation of attitudes towards migrants (e.g., Mayda, 2006; Dustmann and Preston, 2007; Facchini and Mayda, 2009). Second, in relation to the literature on contact theory, the differences between the OLS estimates and the estimates from Lewbel's (2012) heteroskedasticity identification approach suggest that the endogeneity problem may cause large bias in this issue, so mitigating endogeneity bias could be important for this topic. Last, the findings in this chapter shed light on policy-making. On the one hand, the results suggest that governments could take measures to promote communication between locals and migrants to reduce segregation in working and living places. On the other hand, the insignificant contact effect on intimate interactions suggests that governments cannot simply rely on contact alone if they want to reduce segregation in intimate interactions.

The rest of this chapter is organised as follows. Section 3.2 provides a literature review, and Section 3.3 presents the empirical methodology. Section 3.4 describes the data and the general pattern in summary statistics. The results are shown in Section 3.5, and Section 3.6 concludes.

3.2 Literature review

3.2.1 Why are natives hostile to migrants?

Migration brings cultural diversity and enhances economic efficiency; however, native people tend to hold negative attitudes towards migrants.[37] For example, the 1995 National Identity Module of the International Social Survey Programme (ISSP) found that only 7.4% of citizens from 22 countries agreed with increasing immigrant numbers. The 1995-1997 World Value Survey (WVS) found that 46.7% of native people from 44 mostly developing countries said that the government should "place strict limits on the number of foreigners" or "prohibit people coming here from other countries" (Mayda, 2006). The 2002-2003 European Social Survey reported that 35.5% of citizens from 21 higher-income countries "would like a few immigrants or none" (Facchini and Mayda, 2009).

[37] In this chapter, "locals" and "natives" are interchangeable.

Individual preferences about immigration can be grouped into economic and non-economic considerations. Economic considerations are those to do with labour market competition and the burden on the local welfare system. Natives often believe that migrants will take away jobs and depress wages and over-consume public resources via the welfare system; thus, migrants are perceived as a threat to locals' economic opportunities. Non-economic considerations reflect natives' preferences for a familiar culture, ideology, social norms and so on. An inflow of migrants inevitably brings new cultures and ideologies to the destination region. Thus, conservative natives may dislike migrants (see the discussions in Citrin et al., 1997; Scheve and Slaughter, 2001; O'Rourke and Sinnott, 2006; Facchini and Mayda, 2009).

A large body of empirical studies has tested these determinants using various datasets and methodologies in the context of developed countries. For example, Scheve and Slaughter (2001) found that low-skilled workers in the US prefer less immigration, suggesting that local labour market competition is a concern for natives, based on the 1992, 1994 and 1996 National Election Studies Surveys. Mayda (2006) and O'Rourke and Sinnott (2006) found that both economic and non-economic factors play a role in the formation of attitudes towards immigrants, based on cross-country studies from ISSP (1995) and WVS (1995-1997). Dustmann and Preston (2007) used factor analysis to explicitly model the channels through which labour market competition, welfare concerns and racial and cultural considerations affected British natives' attitudes towards immigrants, based on the British Social Attitudes Survey. They found that although all three channels are indispensable in explaining attitude, welfare concerns is more important than concerns about labour market competition, and the effect of racial and cultural prejudice is very large on the ethnically different immigrants (i.e., West Indian and Asian).[38]

Most of the extant literature in this field is based on western society. By contrast, little attention has been paid to the Chinese context, and there is no consensus in the literature. Knight and Yueh (2009) analysed data from an urban household survey covering 13 cities in China, and found that locals who were less-educated and unemployed had significantly worse attitudes to migrants, than others in the study. In contrast, Nielsen et al.(2006) did not find that education was positively significantly correlated with better local attitudes to migrants. Neither of these studies tested the effect of non-economic factors.

[38] For further discussion of different determinants of natives' attitudes towards immigrants, please refer to Hainmueller and Hiscox (2007), Facchini and Mayda (2008), Facchini et al. (2011) and Facchini and Mayda (2012).

3.2.2 Contact hypothesis and reducing prejudice

Studies which test the effect of inter-personal contact on reducing discrimination mainly appear in the sociology and psychology literature. The effect of inter-personal contact is usually explained by the contact hypothesis. This states that if contacts (1) "lead to a sense of equality in social status", (2) "occur in ordinary purposeful pursuits", (3) "avoid artificiality" and (4) "enjoy the sanction of the community in which they occur", they can effectively encourage friendliness (Allport, 1954) and thereby improve the attitudes of the ingroup towards the outgroup.[39]

Over the past sixty years, the contact hypothesis has been used to analyse changing attitudes towards various groups, such as homosexuals, Muslims, and different ethnic groups. Herek and Glunt (1993) found that interpersonal contact is strongly associated with positive attitudes towards gay men. Bevelander and Otterbeck (2010) explored attitudes towards Muslims and found that knowing a Muslim is significantly correlated with a positive attitude towards them. Tropp (2007) reported that interracial contact is associated with greater interracial closeness between black and white, and that this association is stronger for the white than black. Contact is also associated with better attitudes towards the disabled and the mentally ill (Pettigrew, 1998).

Recent studies suggest that contact hypothesis can work in fairly general situations. Pettigrew et al.'s (2011) meta-analysis showed that even though Allport's four conditions (see above) are essential, they are not indispensable for yielding a positive effect on attitudes. Contact can still be helpful, even without these four conditions. Pettigrew et al. (2011) also noted that the contact effect produced in one situation could be generalized to other situations and that some forms of indirect contact can yield positive views, such as learning from a friend who has friends in the outgroup and interacting via media (also see Dovidio et al., 2011; Pettigrew et al., 2011).

The sociology and psychology literatures show that contact can improve intergroup attitudes; however, in some cases contact can also generate negative attitudes towards outgroup members. The positive effect generated by contact arises from gaining more positive information about the outgroup, generating affective ties and then adjusting behaviour (Pettigrew, 1998). But this mechanism will fail if, in the contact, the outgroup member is a true threat to the ingroup. In this case, the contact actually has a harmful effect (e.g., Allport, 1954; Pettigrew et al., 2011).

[39] Ingroup is the social group to which an individual belongs, and outgroup is the social group to which an individual does not belong.

Generally speaking, contact will only improve attitudes if the behaviour of outgroup members are actually better than the original stereotype of the outgroup.

To my knowledge very few studies apply the contact hypothesis to China. There are only two studies in English focusing on the rural-to-urban migration in China. Nielsen et al. (2006) collected a sample of 835 urban residents from six cities in Jiangsu Province, and found that contact does not significantly improve attitudes of the urban residents. Nielsen and Smyth (2011) found that having previous contact experience with an urban local friend is significantly correlated with better attitudes towards urban locals, based on a sample of 548 rural migrants from Fuzhou City. However, these two papers suffer from two problems. First, the surveys used in these two studies are confined to very small regions, which may limit the external validity of the results. Second, the two studies used very simple attitude measures. In particular, they only ask urban residents whether they hold negative attitudes towards migrants (Nielsen et al., 2006), or ask migrants whether they get along with urban locals (Nielsen and Smyth, 2011). These simple questions cannot examine contact effects on different dimensions of attitudes. Given these problems, I use a large representative survey, 2005 China General Social Survey, which provides sophisticated attitude variables, to complement the literature.

3.3 Empirical methodology

In this chapter, the relationship between attitudes and contact is modelled as follows:

$$\text{Attitude}_i = \beta_1 \text{Contact}_i + \beta_2 X_i + \beta_3 \mu_i + \epsilon_{1i} \qquad (3.1),$$

$$\text{Contact}_i = \gamma_1 X_i + \gamma_2 \mu_i + \epsilon_{2i} \qquad (3.2),$$

where the Attitude_i is the current attitude variable measuring willingness to interact with migrants, Contact_i is the variable showing whether the respondent had previous contact experience with migrants, X_i is a vector of observed demographic, economic and non-economic variables including regional fixed effects (i.e., city or county/district fixed effects) which plausibly affect both the attitudes and contact experience, μ_i measures the effect of the unobserved common factors, which could be correlated with both Attitude_i and Contact_i, such as open-mindedness, sense of fairness and initial attitude toward migrants, and ϵ_{1i} and ϵ_{2i} are the idiosyncratic random disturbances. I assume X_i are exogenous to $\beta_3\mu_i + \epsilon_{1i}$ and $\gamma_2\mu_i + \epsilon_{2i}$ (i.e., $E[X_i(\beta_3\mu_i + \epsilon_{1i})] = E[X_i(\gamma_2\mu_i + \epsilon_{2i})] = 0$), and μ_i, ϵ_{1i} and ϵ_{2i} are pairwise uncorrelated.

The goal is to estimate β_1. If $E[\text{Contact}_i(\beta_3\mu_i + \epsilon_{1i})] = 0$, then β_1 is identifiable by the OLS estimation. However, due to the omitted variables, $E[\text{Contact}_i(\beta_3\mu_i + \epsilon_{1i})] = \beta_3\gamma_2\mu_i^2 \neq 0$, the

OLS estimator is biased and β_1 is unidentifiable.[40] To mitigate the endogeneity bias, I proceed in a stepwise fashion.

Initially ignoring the endogeneity, I estimate the OLS regressions, controlling for a set of basic demographic and economic variables which are commonly used in the literature. These are age, gender, employment status, education, party membership, socio-economic status and health. Then I expand the set of control variables to add other important covariates in the regressions, including tolerances of others' negative behaviour and familiarity with neighbours, which are typically not available in other surveys but helpful for capturing individual preferences in attitudes. The comparison of the results between these two regressions may shed light on the extent of omitted variable bias. If the results are sensitive to the inclusion of the additional variables, then the correlation between attitude and contact is likely to be driven by these additional variables.

Even if the results are robust to the kitchen-sink specification, the possible existence of other omitted variables could still bias the OLS estimate. In the following steps, I use both the conventional instrumental variable (IV) estimation and Lewbel's (2012) heteroskedasticity identification approach to deal with endogeneity. For the conventional IV estimation, a candidate of the instrumental variable is the county/district-level migrant population share for people aged from 18 to 45.[41] A valid IV must satisfy two assumptions: 1) the IV is strongly correlated with the endogenous variable, and 2) it does not have a direct effect on the dependent variable. Because migrants work and live in urban areas, urban locals inevitably come into contact with migrants in their daily life. In a region where there are more migrants, locals are more likely to have had previous interaction with migrants. Hence, this IV could satisfy the first assumption. However, the validity of the second assumption is controversial. If the respondents are xenophobic, or the migrants endogenously choose a destination region where the locals are

[40] Please note that Contact_i is a measure of past experience which cannot be directly affected by current attitudes. This means that there is no contemporaneous reverse causality from Attitude_i to Contact_i. Hence, in the following main analysis, I show the results from the triangle model assuming no contemporaneous reverse causality (as shown in Equations 3.1 and 3.2). In Table 3.A.1, I also check the robustness of the results by estimating the simultaneous model which allows the contemporaneous reverse causality. The results are similar, and the effect of contemporaneous reverse causality from attitude to contact experience is statistically insignificant.

[41] Because young people tend to contact each other more, I set the age interval between 18 and 45 to keep the IV strong.

welcoming, then the second assumption is violated, and the IV estimator is inconsistent.[42] Given this, the IV estimates should be interpreted cautiously in the following analysis.

To prevent the direct effect of the instrumental variable from causing the bias, I use the heteroskedasticity identification technique proposed by Lewbel (2012) to estimate the contact effect. This method has appeared in a number of publications as the strategy to deal with endogeneity problem (e.g. Emran and Hou, 2013; Huang et al, 2009; Kelly et al, 2011; Sabia, 2007; Smyth and Mishra, 2014; Zhao, 2015). One advantage of the heteroskedasticity identification technique is that it can control for the direct effect of the instrumental variable. Instead of assuming the exogeneity of the instrumental variables, this method imposes an assumption on higher order moments, and the identification comes from the heteroskedasticity in the equation of $\mathrm{Contact_i}$. Specifically, in this method I estimate the following model:

$$\mathrm{Attitude_i} = \beta_1 \mathrm{Contact_i} + \beta_2 \mathrm{X_i} + \beta_3 \mathrm{Z_{Hi}} + \beta_4' \mu_i' + \epsilon_{1i}' \qquad (3.3),$$

$$\mathrm{Contact_i} = \gamma_1 \mathrm{X_i} + \gamma_2 \mathrm{Z}_{Hi} + \gamma_3' \mu_i' + \epsilon_{2i}' \qquad (3.4),$$

where Z_{Hi} is the county/district-level migrant population share, which is used as the instrumental variable in the conventional IV estimation. Comparing with Equations (3.1) and (3.2), we could see $\beta_3 \mu_i + \epsilon_{1i} = \beta_3 \mathrm{Z}_{Hi} + \beta_4' \mu_i' + \epsilon_{1i}'$ and $\gamma_2 \mu_i + \epsilon_{2i} = \gamma_2 \mathrm{Z}_{Hi} + \gamma_3' \mu_i' + \epsilon_{2i}'$. Equations (3.3) and (3.4) explicitly model the impact of the instrumental variable, and μ_i' is the common factor which is free of the effect of the local migrant share.

In this model, the identifying assumptions are $cov[Z_{Hi}, (\beta_4' \mu_i' + \epsilon_{1i}')(\gamma_3' \mu_i' + \epsilon_{2i}')] = 0$ and $cov[Z_{Hi}, (\gamma_3' \mu_i' + \epsilon_{2i}')^2] \neq 0$. If these two assumptions are satisfied, then β_1 could be consistently estimated through the following steps:

1. Run OLS regression on Equation (3.4) and calculate residuals $\hat{\gamma}_{2i}$[43]

2. Generate$(\mathrm{Z}_{Hi} - \overline{Z_H})\ \hat{\gamma}_{2i}$ from the sample, where $\overline{Z_H}$ is the sample mean of Z_{Hi}.

3. Take $(\mathrm{Z}_{Hi} - \overline{Z_H})\ \hat{\gamma}_{2i}$ as the instrumental variable to run 2SLS regression on Equation (3.3) and then obtain $\widehat{\beta_1}$.

[42] There are some evidences in the literature confirming this possibility, although the results are still mixed. Stein et al. (2000) reported that "whites residing in areas with high concentrations of minority populations have significantly more negative attitudes towards minorities". Lennox (2012) founds that a non-white concentration in the local area helps reduce the racial prejudice of white. Even though no extant study focuses on the Chinese context, these two studies can warrant the concern of the validity of the instrumental variable in this case.

[43] In Equation (3.4) the observed variables are $\mathrm{X_i}$ and Z_{Hi}, so $\hat{\gamma}_{2i}$is the predicted values of $\gamma_3' \mu_i' + \epsilon_{2i}'$.

In this chapter, I argue that the county/district-level migrant population share Z_{Hi} could be a good candidate satisfying the identifying assumptions. First, the assumption $cov[Z_{Hi}, (\beta_4'\mu_i' + \epsilon_{1i}')(\gamma_3'\mu_i' + \epsilon_{2i}')] = 0$ could be guaranteed by a set of mild and plausible sub-assumptions: 1) $cov[Z_{Hi}, \mu_i'^2] = 0$, and 2) μ_i', ϵ_{1i}' and ϵ_{2i}' are uncorrelated to each other conditional on Z_{Hi}. Regarding the first sub-assumption, since μ_i' represents the individual's unobserved preference, such as open-mindedness and sense of fairness, which is free of the direct effect of Z_{Hi} as shown in Equations (3.3) and (3.4), and Z_{Hi} is the local aggregate information, it is reasonable to believe that the collective outcome Z_{Hi} is irrelevant to the individual preference μ_i'. However, one concern we should note is that the heterogeneous impact of migrants on urban locals' attitudes may cause the correlation between Z_{Hi} and $\mu_i'^2$. For example, a larger migrant share may improve the attitudes of well educated urban locals, due to the complementary effect on the labour market, but worsen the attitudes of urban locals with less education, due to the substitution effect. Missing this differential effect would causes $cov[Z_{Hi}, \mu_i'^2] \neq 0$. One way to remove this potential heterogeneous impact of the local migrant population share on μ_i' is to include the interactions between the individual characteristics and Z_{Hi}. To reduce the scope of the potential heterogeneous impact as much as possible, I include the interactions between Z_{Hi} and all the individual characteristics except the county/city fixed effects in the control variables. I show below that the results are robust to the inclusion of interactions, which suggests that the effect of the heterogeneous impact of migrants may be limited. Given this, it is reasonable to assume $cov[Z_{Hi}, \mu_i'^2] = 0$.[44] Regarding the second sub-assumption, since μ_i' is the common factor, and ϵ_{1i}' and ϵ_{2i}' are idiosyncratic factors, there is no reason to expect that they are correlated with each other conditional on the local aggregate variable Z_{Hi}.

Given the above assumption, whether $cov\left[Z_{Hi}, (\gamma_3'\mu_i' + \epsilon_{2i}')^2\right] = cov\left[Z_{Hi}, \epsilon_{2i}'^2\right] \neq 0$ could be empirically tested. In his paper, Lewbel (2012) suggests that this condition could be tested by Breusch and Pagan (1979)'s method. If $cov\left[Z_{Hi}, (\gamma_3'\mu_i' + \epsilon_{2i}')^2\right] \neq 0$ is also satisfied, then the causal effect of contact β_1 could be estimated.[45][46]

[44] As a placebo test, I test heteroskedasticity on observed attitudes variables which may be both correlated with attitudes and contact experience (i.e., tolerances to negative behaviours) with respect to the migrant population share Z_{Hi}. No heteroskedasticity can be found at 10% level in *tolerance to negative non-social behaviours*. In *tolerance of negative social behaviours*, only in the samples of *general attitude* and *living in the same community as migrants* heteroskedasticity can be found at 10% but cannot be found at 5%. *The tolerances of the negative behaviours*, *general attitude* and *living in the same community as migrants* are defined in Section 3.4.

[45] Lewbel's (2012) heteroskedasticity identification approach could be extended to the linear probability model where both the dependent variable and endogenous variable are binary variables. Proof is in the Appendix. Also several

The validity of Lewbel's (2012) heteroskedasticity identification approach can also be examined by comparing its estimates (of Equations 3.3 and 3.4) with the IV estimates (of Equations 3.1 and 3.2). As discussed above, Lewbel's (2012) heteroskedasticity identification approach can be used to estimate the β_1 and the direct effect of the instrumental variable. If Lewbel's (2012) heteroskedasticity identification approach is valid, we can expect to see that the estimated direct effect of the instrumental variable from Lewbel's heteroskedasticity identification approach corroborates the difference between the estimated β_1s from the IV estimation and Lewbel's heteroskedasticity identification approach. In particular, if the correlation between the instrumental variable and endogenous variable is positive and the direct effect of the instrumental variable is positive (negative), then the IV estimates of β_1 should be larger (smaller) than the one from Lewbel's heteroskedasticity identification approach. Similarly, if the correlation between the instrumental variable and endogenous variable is negative and the direct effect of the instrumental variable is positive (negative), then the IV estimates of β_1 should be smaller (larger) than the one from Lewbel's heteroskedasticity identification approach. If these patterns cannot be observed in the following results, then Lewbel's heteroskedasticity identification approach is probably invalid.[47]

As pointed out by Lewbel (2012), a potential problem with this method is that the identification is based on second order moments, so the estimates may be sensitive and less reliable. To cope with this concern I construct a slightly different IV (i.e., the count-level migrant population share without age limit) in the robustness check to see whether the results change significantly.

3.4 Data and descriptive statistics

3.4.1 Data

The data used are drawn from the China General Social Survey (CGSS). The CGSS is the Chinese version of the General Social Survey, and it aims to track social development in China.

published articles have applied Lewbel's (2012) heteroskedasticity identification approach in this case (e.g., Sabia, 2007; Kelly and Markowitz, 2009; Kelly et al., 2011)

[46] The assumption $cov\left[Z_{Hi}, (\gamma_3'\mu_i' + \epsilon_{2i}')^2\right] \neq 0$ may arise partly because the binary nature of Contact_i, so in the following results all the standard errors are the heteroskedasticity-robust standard errors, and I use Kleibergen-Paap statistics, which are computed from the heteroskedasticity-robust standard errors, to assess the strength of the first-stage estimations.

[47] Note that this is not a formal test, because even if the heteroskedasticity identification approach is invalid, it is still possible to observe these patterns.

It is jointly conducted by the Renmin University of China and the Hong Kong University of Science and Technology. Currently, there are four waves of CGSS published, and each wave is a cross-sectional dataset. In this chapter I use only the 2005 wave because the variables on contact experience with migrants are only available in this wave.

The CGSS 2005 is a large representative survey. It covers 125 counties and districts across 28 provinces, municipalities and autonomous regions in mainland China.[48] For sampling efficiency, the survey adopts a stratified sampling strategy. Nine strata are designed according to the difference in social and economic development. The samples in each stratum could well represent the corresponding subpopulation, and if the total samples from all the stratums are used, we can obtain reliable and accurate estimates about the general population in China.[49] The CGSS also uses a multi-stage sampling method, which starts sampling the counties in the sampling frame, and progresses to sampling households. One member of each household is randomly selected to be interviewed. The selected respondents must be aged above 18.[50] [51]

The 2005 CGSS records detailed information on the attitude variables and contact experiences of urban local residents. The attitude variables come from a set of questions regarding willingness to interact with migrant. The questions are

"In recent years, the population of migrants has grown. What are your attitudes towards migrants?[52]

- Are you willing to work with migrant(s)?
- Are you willing to live in the same communities as migrant(s)?
- Are you willing to have migrant(s) as your next-door neighbour(s)?
- Are you willing to invite migrant(s) to visit your home?
- Are you willing to have somebody among your children or relatives marrying or being in a relationship with migrant(s)?"

[48] Only Tibet autonomous region, Ningxia Hui autonomous region and Qinghai province are not included in the survey. These provinces are not primary destinations for migrants.

[49] We must note that sub-districts with a non-agricultural population share of less than 11.34% and townships with a non-agricultural population share of more than 43.37% are excluded from the sampling framework. This means that the samples are biased to advanced urban areas and under-developed rural areas. Table 5 in the CGSS Manual shows that the bias among rural samples is relatively large, but the bias for the urban sample is tiny.

[50] Note that the weights in the 2005 CGSS are derived from the 2005 1% Population Survey, which makes estimates sensible to the population in 2005. The sampling framework is designed according to the 2000 Census.

[51] Note that the results are similar with and without using weights.

[52] Note that in urban Chinese's view point, "migrants" is often considered to be rural-to-urban migrants. In addition, according to 2005 1% population survey 70% of migrants are rural migrants.

The answer to each question has three options: "yes", "no" and "not answered", which is recoded as 1 ("yes"), 0 ("no") and missing value ("not answered"). These are my dependent variables. Besides these five specific questions, I create a variable indicating general attitude to integrating with migrants. It defines as 1 if the respondent answered "yes" to any one of these five questions, otherwise 0.[53]

The contact measure, the explanatory variable of interest, is derived from a set of questions indicating real life experiences, as follows:

"In real life, have you had these experiences?

- Have you ever worked with migrant(s)?
- Have you ever lived in the same communities as migrant(s)?
- Have you ever had migrant(s) as your next-door neighbour(s)?
- Have you ever invited migrant(s) to visit your home?
- Is there anybody among your children or relatives who has ever married or been in a relationship with migrant(s)?"

The answers to these questions also consisted of three choices: "yes", "no" and "not answered" and are coded in the same way as the dependent variables.

Using OLS, Table 3.1 explores the partial correlation between each contact variable and each attitude variable. The results suggest that different contact experiences have different correlations with the attitude variables. Take the willingness to work with migrants as an example. In Column (2) the willingness is only significantly correlated with the contact experiences that the respondents had worked with migrants and had invited migrant(s) to visit their home, and the correlations with the other three contact experiences are insignificant and small. These different correlations suggest that each specific contact experiences may have a different impact on the attitudes. Therefore, in the ideal case we should put all the contact experiences in the regression to explore their effects. However, there are two practical reasons why I do not take this approach. First, putting all these five endogenous variables in the same regression causes the under-identification problem in the conventional IV estimation, and results in the weak IV problem in the Lewbel's (2012) heteroskedasticity identification approach, due to multicollinearity. Second, due to different missing values in the different questions, controlling for all these five endogenous variables would decrease the sample size by

[53] One weakness of these dependent variables is that each variable contains different missing values, so the samples are inconsistent across the dependent variables. In the robustness check, I test whether the results change if I use the consistent samples. For the main results, I use the inconsistent samples.

10%, which not only reduces estimation efficiency, but can also cause bias if the missing values do not appear randomly. To avoid these two problems, I unify these five experience questions into one single measure. Similar to the sociology literature (e.g., Herek and Glunt, 1993; Tropp, 2007; Bevelander and Otterbeck, 2010), I use a binary measure in the main analysis. Specifically, if the respondent had previous contact with migrants in any of these five forms, this single binary contact measure is 1, otherwise it is 0. Although this measure cannot distinguish the effects of different contact forms, it can roughly identify whether having contact with a migrant influences respondents' attitudes towards migrants, which is still helpful for understanding the effect of contact. In the robustness check, I construct another measure which may differentiate the intimacy level of contact experience, and the results are similar. For details please refer to Section 3.5.3.[54]

The other explanatory variables are also extracted from the 2005 CGSS. They include age, employment status, Hukou type, subjective health measure, Communist Party membership, gender and public sector employment.[55] The respondents' general socio-economic condition is captured by self-identified socioeconomic status.

Another advantage of the 2005 CGSS is its rich measures of non-economic variables. The survey asked respondents about their attitudes towards various social issues and also recorded their social behaviour. In this chapter, I create two indices measuring urban locals' tolerance of the individual behaviours which are potentially responsible for the prevailing negative stereotypes of migrants and thereby may affect attitudes towards migrants. The first measure is derived from 10 questions on the respondents' view of behaviours which are harmful to others, such as talking loudly in public occasion and spitting. Each question asks whether the respondent is antipathetic towards one particular behaviour. I equally weight these 10 questions and unify them in a single index, which ranges from 0 to 40. The second measure is sourced from 6 questions regarding behaviours which may be not socially acceptable but are not harmful to others, such as patronising a prostitute and watching adult video. Each question asks whether the respondent agrees that the particular behaviour is an individual choice and others should not criticise. Using the same way as the first measure, I combine these 6 questions and create a tolerance measure of these behaviours, which ranges from 0 to 24. For both these two measure,

[54] An alternative way to create a unified measure is adding these five contact experiences together. However, this measure cannot reveal the differential effect of each contact experience either. In addition, this measure results in the weak IV problem in Lewbel's (2012) heteroskedasticity identification approach. Thus, I do not choose this measure in the following analysis.

[55] "Hukou" refers to the household registration system in China. Usually there are two types of Hukou: rural (agricultural) Hukou and urban (non-agricultural) Hukou.

larger values mean more tolerance. I also use familiarity with neighbours data to measure respondents' open-mindedness and social skills, ranging from 0 ("very unfamiliar with neighbours") to 4 ("very familiar with neighbours"). Larger values mean that respondents are more familiar with their neighbourhood. Please refer to Appendix for details about these variable constructions. These two measures not only help to capture individual heterogeneity in attitude formation, making the estimation more efficient, but also may reduce the bias induced by omitted variable problem.

The instrumental variable used in the conventional IV estimation and Lewbel's (2012) heteroskedasticity identification approach is constructed from the 2005 1% Population Survey. It is calculated as the ratio of migrant population divided by the local population aged 18 to 45 at county/district-level in the main results. In the robustness check, I recalculate the ratio without applying the age limit.

This survey has 6098 respondents interviewed in urban areas. Since I focus on urban local residents, the sample is restricted to respondents who live in urban areas and whose residential addresses match their Hukou addresses.[56] This leaves 5382 respondents. I further trim the samples to respondents aged between 18 and 65, and exclude students and those engaged in agricultural jobs, to isolate the effect to people who are in the urban labour force. This results in 4058 respondents remained. Of these 4058 respondents 3577 provide necessary data in the main variables and covariates, which constitute the final sample for my analysis.

Note that this sample includes a small number of respondents (4% of the whole sample) who have rural Hukou, but permanently live in local urban areas.[57] These people are natives and they would probably consider themselves to be urban locals. To give a complete picture of contact effect I include them in the sample of the main analysis. I control for Hukou type in the regression to account for possible differences between people with rural Hukou and those with urban Hukou. I also exclude observations with rural Hukou in the robustness check, and the results are similar (see section 3.5.3).

[56] I do not include respondents with Lanyin Hukou and Zililiang Hukou, because these Hukous are specially designed for migrants who live in destination cities for a limited time. These people are unlikely to be treated as locals, even though they have access to similar welfare benefits.

[57] The emergence of this group is a result of China's rapid urbanisation. Cities such as Beijing, Shanghai, Guangzhou and Shenzhen have expanded enormously, urbanising previously rural areas. Some of the native people in these areas converted their Hukou to urban Hukou. Some were reluctant to do so, because they can enjoy the same city amenities as others and also own the land.

3.4.2 Descriptive statistics

Table 3.2 presents the descriptive statistics. Panels A and B show the summary statistics of attitude variables and contact experience variables, respectively. Two facts can be seen from Panels A and B. First, respondents who had contact experience with migrants are much more willing to interact with migrants than those who did not have. There is a 0.22 increase in willingness to interact with migrants across all forms of the five interactions. All the specific attitudes increase as well. Of these, the largest increase is in willingness to work with migrants (0.24); the smallest increase is in willingness to have relatives and/or children in relationships with migrants (0.2). Second, although 84% of respondents are willing to interact with migrants in any of the five specific contact forms, a large proportion are only willing to interact in non-intimate ways.[58] As intimacy increases, willingness drops. For example, 84% of respondents who have had previous contact with migrants are willing to work with them again and 60% of respondents who have had no previous contact are willing to work with migrants. In contrast, only 54% of respondents who have had previous contact with migrants are happy for their children or relatives to marry or be in a relationship with migrants and only 35% of respondents who have had no contact are happy for such intimate relationships to develop. A similar pattern can also be observed in the contact experience variables presented in Panel B. 76% of respondents with contact experience had previously worked with migrants, but only 22% had had children or relatives marrying or being in relationships with migrants.

Panel C presents the summary statistics of the other variables, revealing several differences between the two groups. Respondents who have contact experience with migrants tend to be younger, healthier, better educated and more likely to be fully employed. They are also less likely to work in the stated-owned or collective sector, more tolerant of the negative behaviours which may be harmful to others, but less tolerant of the behaviours which are negative but not harmful to others. Finally, they are more familiar with neighbours.

3.5 Main results

3.5.1 OLS estimation

[58] The percentage "84%" comes from my own calculation using the data.

This section presents the empirical results of the analysis of 2005 CGSS data, using the procedure described in Section 3.3. Table 3.3 shows the OLS estimates for different attitude variables. There are three regressions for each attitude variable. The first regression controls for the set of explanatory variables typically controlled for in other studies (e.g., Mayda, 2006; Dustmann and Preston, 2007; Facchini and Mayda, 2009), but it does not control for the contact experience. The second regression adds the contact experience and keeps the other control variables the same as the first regression. The last regression incorporates non-economic factors and keeps the other control variables the same as the second regression. These non-economic factors were not used in the previous literature, but are likely to be important in the sense of correlating with both the attitude variables and contact experience. Comparing these three specifications allows me to check whether contact experience is important in explaining attitudes and sensitive to non-economic determinants.

Table 3.3 shows that previous contact experience with migrants is significantly correlated with better attitudes towards them. The results suggest that previous contact is associated with a 20% greater likelihood of being willing to interact with migrants in any one of the five specific forms (see Columns 2 and 3 in Table 3.3.1). The coefficients of the regressions on specific attitudes show that contact is significantly correlated with each dimension of attitude. The associations range from 0.15 ("having children/relatives marrying or being in a relationship with migrant(s)") to 0.22 ("inviting migrant(s) to visit home"). These associations are insensitive to the additional variables in the last specification, demonstrating that these variables do not drive the correlations.

The changes in the Adjusted-R^2 across specifications tell a similar story. Comparing the first two specifications, we can see that including contact experience in the regression leads the Adjusted-R^2 for general attitude to increase by 28%. Among the specific items, the increases are also large in the first four specific items, ranging from 10% ("having migrant(s) as next-door neighbour(s)") to 23% ("working with migrant(s)"). The increase for "having children/relatives marrying or being in a relationship with migrant(s)" is relatively small: 6%. In contrast, the additional non-economic variables (i.e., tolerance variables and familiarity with neighbours) make a much smaller contribution in the last specification. The extra variables generate only a 3% increase in the Adjusted-R^2 for general attitudes. The increases range from 0% for "inviting migrant(s) to visit home" to 3% for "living in the same community as migrant(s)". These findings suggest that contact experience is potentially an important determinant of attitudes, at least for non-intimate interactions.[59]

The other control variables also reveal interesting patterns. Age shows a U-shaped relationship with the willingness to interact with migrants. As age increases, the willingness first decreases and then increases. Less healthy respondents tend to be less willing to interact with migrants, although the correlation is insignificant regarding the willingness to have children or relatives marrying or being in a relationship with migrants. Respondents in the category of urban Hukou tend to be significantly less willing to live in the same community as migrants and have children or relatives marrying or being in a relationship with migrants. Higher educated respondents are significantly more likely to be willing to work with migrants. This is perhaps because migrants complement the higher educated urban locals more in the labour market, or the higher educated people are more open and thus welcome migrants more. Finally, non-economic factors are critical in explaining some dimensions of attitudes. The results suggest that those who are more tolerant of negative behaviours which may be harmful to others are more willing to work with migrants and live in the same community as migrants.

3.5.2 Conventional IV estimation and Lewbel's heteroskedasticity identification approach

Tables 3.4 and 3.5 present the conventional IV estimates and Lewbel's (2012) heteroskedasticity identification estimates. All the regressions in these two tables include the same individual-level characteristics as those in the third specification in Table 3.3, but there are differences in the fixed effects included as explained in the table.[60]

Table 3.4 gives the results of the first-stage estimation. Recall that the IV used here is the county/district-level population share of migrants. Consistent with the argument in Section 3.3, the local population share of migrants is positively correlated with having contact experience with migrants in the conventional IV estimation. One percentage point increase in the local population share of migrants is associated with an additional chance of 0.6 percentage point that urban locals will have had contact experience with migrants. Panels B and C suggest that an increase in local population share of migrants narrows the variation of contact experience with

[59] In the unreported results, the same pattern of Adjusted-R^2 could be seen when the second specification does not have contact measure but has non-economic factors.

[60] Note that due to multicollinearity, we cannot control for the local migrant population share when the county (or district) fixed effects are controlled for.

migrants. Intuitively, if local people live in a place with a large number of migrants, they more or less have contact experience with migrants. Hence, compared to the people living in a place with not many migrants, it is possible that the variation of contact experience among the locals living with many migrants is smaller.

One advantage of the heteroskedasticity identification approach over the conventional IV estimation is that it produces stronger IV. In the conventional IV estimation, the Kleibergen-Paap weak IV test statistics (the F-statistics in testing the strength of IV when the error is non-i.i.d) are all below 10, while the Kleibergen-Paap statistics for the Lewbel IV are all above 20. Staiger and Stock (1997) point out that if the F-statistic is lower than 10, the instrumental variable is weak, making the second-stage estimation and inference imprecise. Given this concern, the Lewbel's heteroskedasticity identification approach is more credible here.

Table 3.5 shows the results of the second-stage estimation. For ease of comparison, Panel A shows the OLS estimates. Panel B shows the conventional IV estimates. Panels C to E show the estimates obtained from Lewbel's (2012) heteroskedasticity identification approach. In particular, Panel C includes the local population share of migrants and city fixed effect as control variables; and Panel D includes county fixed effects but does not control for the local population share of migrants. In addition to the control variables in Panel D, Panel E controls for the interactions between the local population share of migrants and each individual characteristic except fixed effects.

The conventional IV estimates are all insignificant except for the dependent variable "having children/relatives marrying or being in a relationship with migrant(s)". The standard errors for all estimates are large, which may be a result of the low strength of the instrumental variables. The magnitudes of the IV estimates are much larger than those for the OLS estimates. For example, the estimates for "having migrant(s) as next-door neighbour(s)" and "having children/relatives marrying or being in a relationship with migrant(s)" are 1.14 and -1.62 respectively, which are almost six and ten times as large as the corresponding OLS estimates. The large magnitudes of the IV estimates may be caused by the direct effect of the instrumental variable on the dependent variables, because migrants might endogenously choose to live where they are more welcomed, and locals might become more xenophobic when living among more migrants. This possibility is supported by the estimates from the heteroskedasticity identification approach. Panel C suggests that the instrumental variable indeed directly affects attitude outcomes. The coefficients of the local population share of migrants are non-ignorably large, and the signs of the coefficients could also explain the differences between the IV and heteroskedasticity identification estimates. In particular, when the coefficients of the instrumental variable are positive, the estimated contact effects of the IV estimation are larger than those of Lewbel's (2012) heteroskedasticity identification approach (see Columns 1 to 4). When the coefficients of the instrumental variable are negative, the IV estimates of contact effect are smaller than the estimates of the Lewbel's (2012) heteroskedasticity identification approach (see Columns 5 and 6). These findings may partly support the validity of the Lewbel's (2012) heteroskedasticity identification approach as argued in Section 3.3.

Given the concerns on the weak IV problem and the validity of the IV, we should rely more on the estimates from the Lewbel's (2012) heteroskedasticity identification in Panels C to E. In contrast to conventional IV estimates, the standard errors of the estimates from Lewbel's (2012) method are much smaller, which means that the estimation is more accurate. Comparing Panels D and E reveals that including interactions between the individual-level variables and local population share of migrants does not change the results much, which indicates that the heterogeneous impact of local population share of migrants may not undermine the validity of the heteroskedasticity identification approach.

In Panels C to E, an interesting pattern of the estimates of contact effect is that while the standard errors across different attitude outcomes remain constant, the coefficients become smaller and less significant as the intimacy level of the interaction increases. The coefficients on willingness to "work with migrant(s)" and "live in the same community as migrant(s)" are around 0.31 to 0.33 and 0.27 to 0.28, and both of them are significant. However, the coefficients on willingness to "have migrant(s) as next-door neighbour(s)" are reduced to around 0.13 to 0.15 and become insignificant. As for willingness to "invite migrant(s) to visit home" and "have children/relatives marrying or being in a relationship with migrant(s)", the coefficients further drop to around 0 and 0.04 to 0.07 and are insignificant. This suggests that contact can only improve willingness to engage in non-intimate interactions, but has no significant effect on willingness to engage in intimate interactions. Comparing these results with the OLS results, we can see that the OLS associations on intimate interaction willingness mostly come from self-selection, and that the OLS associations on non-intimate interaction willingness are close to the causal effects. This suggests that the omitted variables, such as open-mindedness and sense of fairness, play probably more important roles in the willingness to engage in intimate relationships than non-intimate relationships.[61]

[61] Please note that the heteroskedasticity identification estimates are slightly larger than the OLS estimates for willingness to engage in non-intimate relationships. These differences may be caused by measurement error and the finite-sample bias in the estimates from the heteroskedasticity identification approach. However, this is not a major

In Panel C of Table 3.5, it is interesting to see that local population share of migrants is positively correlated with the willingness to engage in the non-intimate relationships, but negatively correlated with the willingness to engage in the intimate relationships (see Columns 5 and 6). One *conjecture* for this pattern to arise is that migrants tend to move to economically advanced regions with more job opportunities.[62] Since migrants and urban locals tend to be segregated into different occupations in the urban labour market (e.g., Meng and Zhang, 2001), migrants may not place large adverse effect on urban locals, and instead, they may have complementary effect on urban locals in the labour market.[63] This complementary effect may be larger in more economically advanced areas.[64] Thus, urban locals in more economically advanced areas tend to be more willing to associate with migrants in non-intimate ways (e.g., working together). This may explain why the coefficients on the first three specific attitudes are positive (from "working with migrants" to "having migrant(s) as next-door neighbour(s)"). However, the segregation in the urban labour market also causes wage differential between migrants and urban locals. Migrants are usually segregated into lower-paid jobs (Meng and Zhang, 2001; Lee, 2012), so migrants and urban locals may have unequal economic status. This unequal economic status could also be larger in more economically advanced cities, because in these cities urban locals are wealthier. When it comes to the formation of intimate relationships, especially for marriage, economic factor is an important consideration. Given the more unequal economic status, local people in economically advanced areas may be reluctant to integrate with migrants in intimate ways because of the potential burden induced by association with them (e.g., borrowing money). Hence, the coefficients of the local population share of migrants turn negative in Columns (5) and (6).

The above analysis suggests that for the average urban locals previous contact experience significantly increases their willingness to interact with migrants in non-intimate relationships, and has no harmful effect on the willingness to engage in intimate relationships. However,

concern because the difference between the OLS estimates and the heteroskedasticity identification estimates is around or less than the standard error of the heteroskedasticity identification estimates.

[62] According to 2005 1% Population Survey, 54% of migrants clustered at the regions of Pearl River Delta and Yangtze River Delta where the economies are more dynamic than other areas in China.

[63] Meng and Zhang (2010) find that rural migrants have insignificant positive effect on employment and wages of urban locals, which indicates the possible existence of complementary effect.

[64] Although there is no systematic study on the complementarity between migrants and urban locals in China, some studies suggest that there is less segregation in less economically advanced areas. For example, the urban labour market is found to be more integrated between migrants and urban locals in Sichuan, a less economically advanced province, than that in Guangdong, an economically advanced province (CCER, 1998a,b). Since in a more segregated market migrants may complement urban locals better, it is possible that the complementary effect of migrants is stronger in more economically advanced areas.

whether this is true for the respondents who may compete with migrants in the labour market is still unclear. Table 3.6 explores the contact effect on urban locals who may be competitors of migrants in the labour market. As migrants tend to work in the private sector and compete with less-educated locals, I mainly look at the respondents with high school education or below (in Panel A) and the respondents working in the private sector (in Panel B).[65] The results for these two groups of respondents are qualitatively similar to those in Table 3.5. Although the magnitudes of the coefficients drop slightly probably due to the economic pressure migrants place, contact still significantly improves willingness to engage in non-intimate interactions and has no significant effect on intimate interactions with migrants for the competitors of migrants.[66]

A question naturally arises from the findings shown in Tables 3.5 and 3.6: Why does contact not improve willingness to engage in intimate interactions? A possible explanation is that there exist some differences between migrants and urban locals, and these differences may hinder the contact effect. For example, migrants and urban locals tend to have different economic conditions. As mentioned before, several studies suggest that migrant workers tend to earn less than urban residents (e.g., Meng and Zhang, 2001; Lee, 2012). The unequal economic status may constrain social activities of migrants, which makes urban local not willing to interact with migrants in intimate relationships. Migrants may also have different customs from urban locals. For instance, rural migrants may prefer sons over daughters (Lei and Pals, 2011) and have a more conservative view of gender roles than urban locals. The 2006 CGSS shows that male migrants are more likely to believe that the responsibility of a husband is to make money and the responsibility of a wife is to take care of the family than urban male residents. They also think that women should be first to be dismissed in an economic downturn (see Table 3.A.2). These differences usually do not directly affect non-intimate interactions, but they could be important concerns for intimate interactions, especially forming friendships and marriages. During the contact, urban locals can obtain this information, which may offset the beneficial effect of contact. Thus, in this case contact may have a nil effect on willingness to engage in intimate interactions.[67]

In order to test this explanation, I estimate the contact effect for respondents who permanently live in urban areas but hold rural Hukou, in Table 3.7. Chinese cities have expanded rapidly

[65] In 2005, only 5% of migrants had college education or above, and only 10% of migrants worked in the state-owned or collective sector, according to the 1% Population Survey. Therefore, migrants mainly compete with urban locals who have high school education or below and who work in private sector.

[66] The results in Table 3.6 are robust to including interactions between individual variables and migrant share.

[67] Note that the differences listed above are just some examples. It is possible that many other differences cause the null effect on willingness to engage in intimate interactions.

over the past two decades.[68] Many rural areas urbanised not a long time before the survey period, including, most likely, the homes of respondents in the Table 3.7 sample. These respondents have probably lived in a rural environment for a long time and likely share the culture of (mostly rural) migrants. Thus, these respondents should be less different from migrants than the other urban locals. If the differences between migrants and urban locals are the main cause of the null effect of contact for intimate interactions, then we can expect to see that the estimates of contact effect on intimate interactions will be larger in Table 3.7 than that in Table 3.5. Indeed, in Table 3.7 we can see that contact not only significantly improves willingness to engage in non-intimate interactions, but also has a large beneficial effect on intimate interactions among natives with rural Hukou. This suggests that the differences between migrants and urban locals may be responsible for the nil effect of contact on intimate interactions.[69]

3.5.3 Robustness check

In this sub-section, I test robustness of results in five different aspects. First, as mentioned in Section 3.3, Lewbel's (2012) heteroskedasticity identification approach is based on the second order moments, so the estimates may be sensitive. To test whether the results are sensitive to the choice of the instrumental variable used, I slightly change the construction of the instrumental variable. I calculate the county/district-level migrant population share without any age limit, and re-conduct Lewbel's (2012) heteroskedasticity identification approach to see whether the results alter significantly. The results are reported in Panel A of Table 3.8 and similar to the results of the main analysis in Table 3.5, except that the standard errors are slightly larger.

Second, I exclude 4% of the observations who have rural Hukou to check whether the results are sensitive to this exclusion. The results shown in Panel B are similar to the main results in Panel D of Table 3.5. This indicates that the contact effect works among respondents who hold urban Hukou.

Third, people's attitudes may differ across different income groups (e.g., Facchini and Mayda, 2009). In the main analysis, I do not include income in the estimation because it may be endogenous and has missing values which could result in a sample selection problem. To test

[68] The area of build districts in China increased from 7438 km^2 in 1980 to 12252.9 km^2 in 1990 and to 40533.8 km^2 in 2010 (Wang et al., 2012; Cai et al., 2013). Area of build districts refers to the urban areas with municipal utilities and public facilities. It is a usual measure to assess the level of urbanisation.

[69] However, this analysis cannot rule out the possibility that the larger contact effect on intimate interactions in Table 3.7 is caused by other reasons. I also acknowledge that the problem of small size in Table 3.7.

whether the exclusion of income drives my results, Panel C presents results with log average income per family member as an additional control variable in the regression. The results are robust to this inclusion.

Fourth, the original data analysis revealed that contact effect is different across different attitude variables. However, this conclusion is drawn from different samples, due to the fact that for each attitude variable the observations with missing value are not the same. To rule out the possibility that inconsistent samples lead to my results, I re-estimate the effect using a sample which is consistent across different dependent variables. The results in Panel D suggest that this difference is robust to the sample construction.

Last, as previously mentioned, the binary contact measure cannot differentiate the intimacy level of contact. Given this, I construct a new contact measure to shed light on how the intimacy level of contact affects willingness in Panel E. The new measure contains four categories: has had no previous contact with migrants (0), has lived in the same community as migrants (1), has worked with migrants or had migrant neighbours next-door (2), and has invited migrants to visit home, or has had children/relatives marrying or being in relationships with migrants (3). The measure picks up the largest number for respondents who have had multiple contact experiences. The results in Panel E are generally similar to the main results.[70]

3.6 Conclusion

In this chapter, I examine the relationship between interpersonal contact and the attitudes of urban locals to migrants, using a large representative survey. I find that contact significantly improves willingness to engage in non-intimate interactions with migrants, but its effect is insignificant on willingness to engage in intimate interactions, though not harmful. The results are robust among those respondents who tend to be migrant's competitors in the labour market.

One limitation in the chapter is the contact measure. Due to the multicollinearity problem, this chapter cannot disentangle the effects of each detailed contact experience and cannot provide evidence on which type of contact experience is the most helpful for reducing discrimination. This problem should be addressed in the future study. This chapter also leaves a gap. Social segregation and discrimination may also be caused by the hostility of migrants towards locals.

[70] However, it should be noted that I am not able to construct a perfect measure for contact intimacy due to data constraint. This measure is only an illustrating example to show the results are robust.

Hence, using contact to improve the attitudes of locals is not enough for eliminating all the segregation. Future research should consider whether migrants discriminate against locals and how this can be rectified.

Figures and tables

Table 3.1 OLS estimates of the effects of disaggregated contact measures on attitudes

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Worked with migrants	0.116***	0.174***	0.107***	0.096***	0.105***	0.033
	(0.016)	(0.020)	(0.023)	(0.023)	(0.023)	(0.024)
Lived in the same community as migrants	0.038**	0.023	0.093***	0.016	-0.018	-0.014
	(0.018)	(0.021)	(0.025)	(0.025)	(0.025)	(0.026)
Had migrant neighbours next-door	0.024*	0.022	0.069***	0.146***	0.046**	0.059**
	(0.013)	(0.017)	(0.020)	(0.021)	(0.021)	(0.023)
Had migrants visited their home	0.089***	0.132***	0.128***	0.203***	0.398***	0.188***
	(0.011)	(0.016)	(0.020)	(0.021)	(0.021)	(0.023)
Had kids/relatives marry or be in a relationship with migrants	0.022*	-0.031	0.029	0.067***	0.040*	0.282***
	(0.013)	(0.020)	(0.023)	(0.024)	(0.024)	(0.026)
Observations	3464	3467	3481	3469	3466	3309
Adjusted R^2	0.207	0.216	0.205	0.246	0.295	0.271

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The control variables include age, squared age, male, health dummies, Hukou status, employment dummies, education attainment, party membership, ownership of workplace and county fixed effect. These control variables are the same as the second specification of Table 3.3.

Source: 2005 China General Social Survey.

Table 3.2 Summary statistics

	Had previous contact with migrant(s)			No previous contact with migrant(s)			Difference
	N	Mean	S.D.	N	Mean	S.D.	(2)-(5)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A Dependent variables							
General attitude	2953	0.903	0.297	511	0.679	0.468	0.224***
Working with migrants	2951	0.842	0.365	516	0.600	0.49	0.243***
Living in the same community as migrants	2957	0.758	0.429	524	0.517	0.5	0.241***
Having migrant neighbours next-door	2946	0.660	0.474	523	0.440	0.497	0.220***
Having migrants visited their home	2941	0.615	0.487	525	0.373	0.484	0.242***
Have children/relatives marrying or being in a relationship with migrants	2808	0.543	0.498	501	0.345	0.476	0.198***
Panel B Detailed contact variables							
Worked with migrant(s)	3031	0.758	0.428	546	-	-	-
Lived in the same community as migrant(s)	3031	0.854	0.354	546	-	-	-
Had migrant neighbour(s) next-door	3031	0.448	0.497	546	-	-	-
Had migrant(s) visit their home	3031	0.400	0.490	546	-	-	-
Had children/relatives marrying or being in a relationship with migrant(s)	3031	0.216	0.412	546	-	-	-
Panel C Other variables							
Age	3031	41.421	11.504	546	43.64	10.976	-2.220***
Male	3031	0.483	0.500	546	0.471	0.500	0.012
Good health	3031	0.661	0.474	546	0.627	0.484	0.034

Fair health	3031	0.245	0.43	546	0.232	0.422	0.013
Bad health	3031	0.094	0.292	546	0.142	0.349	-0.047**
Urban Hukou	3031	0.948	0.221	546	0.953	0.211	-0.005
Fully employed	3031	0.528	0.499	546	0.457	0.499	0.070**
Non-fully employed	3031	0.285	0.452	546	0.329	0.470	-0.043*
Retired	3031	0.187	0.390	546	0.214	0.411	-0.027
Education attainment	3031	2.608	0.910	546	2.415	0.948	0.193***
Self-identified member of upper class	3031	0.062	0.241	546	0.057	0.232	0.005
Communist party member	3031	0.144	0.351	546	0.141	0.348	0.003
Work in state-owned or collective sector	3031	0.651	0.477	546	0.735	0.442	-0.085***
Tolerance of negative social behaviour	3031	6.838	4.929	546	7.482	5.718	-0.644**
Tolerance of negative non-social behaviour	3031	7.502	4.561	546	6.979	4.449	0.523**
Familiarity with neighbours	3031	2.501	0.945	546	2.657	0.923	-0.156***
Local population share of migrants	3031	0.266	0.189	546	0.194	0.177	0.072***

Note: *Non-fully employed* is defined as those who are unemployed and employed part-time. *Tolerance of negative social behaviour* is defined as tolerance of behaviours which are harmful to others. *Tolerance of negative non-social behaviour* is defined as tolerance of behaviours which are not harmful to others but may not be socially acceptable.

Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.3.1 OLS estimates in attitudes towards migrants

	General attitudes			Working with migrant(s)		
	(1)	(2)	(3)	(4)	(5)	(6)
Previous contact with migrants		0.204***	0.206***		0.211***	0.212***
		(0.024)	(0.024)		(0.026)	(0.026)
Age	-0.009**	-0.008**	-0.008**	-0.011**	-0.010**	-0.010**
	(0.004)	(0.004)	(0.004)	(0.005)	(0.005)	(0.005)
Squared age	0.000**	0.000**	0.000**	0.000**	0.000*	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.029**	0.027**	0.021*	0.026*	0.024	0.018
	(0.012)	(0.012)	(0.012)	(0.016)	(0.015)	(0.015)
Fair health	-0.027*	-0.034**	-0.033**	-0.020	-0.027	-0.027
	(0.016)	(0.016)	(0.015)	(0.018)	(0.018)	(0.018)
Bad health	-0.060**	-0.058**	-0.054**	-0.065**	-0.063**	-0.062**
	(0.026)	(0.026)	(0.026)	(0.029)	(0.029)	(0.029)
Urban Hukou	-0.027	-0.027	-0.024	-0.016	-0.018	-0.018
	(0.027)	(0.025)	(0.024)	(0.037)	(0.036)	(0.035)
Not fully employed	0.006	0.006	0.005	-0.014	-0.013	-0.014
	(0.016)	(0.015)	(0.015)	(0.019)	(0.019)	(0.019)
Retired	-0.019	-0.021	-0.021	-0.030	-0.031	-0.032
	(0.025)	(0.025)	(0.025)	(0.029)	(0.029)	(0.029)
Education attainment	0.018**	0.015*	0.018**	0.022**	0.019**	0.021**
	(0.008)	(0.008)	(0.008)	(0.010)	(0.009)	(0.010)
Self-identified as member of upper class	-0.030	-0.031	-0.030	-0.043	-0.043	-0.042
	(0.021)	(0.020)	(0.021)	(0.035)	(0.035)	(0.035)
Communist party member	0.022	0.019	0.018	-0.003	-0.004	-0.004
	(0.017)	(0.017)	(0.017)	(0.023)	(0.023)	(0.023)
Work in state-owned or collective sector	-0.014	-0.009	-0.007	-0.007	-0.001	-0.000
	(0.015)	(0.014)	(0.014)	(0.018)	(0.018)	(0.017)
Tolerance of negative social behaviour			0.005***			0.004**
			(0.001)			(0.002)
Tolerance of negative non-social behaviour			0.000			0.002
			(0.002)			(0.002)
Familiarity with neighbours			0.004			0.009
			(0.007)			(0.009)
Observations	3464	3464	3464	3467	3467	3467
Adjusted R-squared	0.145	0.185	0.190	0.137	0.169	0.171

Note: Robust standard errors in parentheses: * p < 0.10, ** p < 0.05, *** p < 0.01. *Non-fully employed* is defined as those who are unemployed and employed part-time. *Tolerance of negative social behaviour* is defined as tolerance of the behaviours which are harmful to others. *Tolerance of negative non-social behaviour* is defined as tolerance of the behaviours which are not harmful to others but may not be socially acceptable. County fixed effects are included in the regressions.

Source: 2005 China General Social Survey.

Table 3.3.2 OLS estimates in attitudes towards migrants

	Living in the same community as migrant(s)			Having migrant(s) as next-door neighbour(s)		
	(1)	(2)	(3)	(4)	(5)	(6)
Previous contact with migrants		0.210***	0.213***		0.175***	0.177***
		(0.028)	(0.028)		(0.028)	(0.028)
Age	-0.018***	-0.017***	-0.019***	-0.016***	-0.015**	-0.016***
	(0.005)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
Squared age	0.000***	0.000***	0.000***	0.000**	0.000**	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.030*	0.027	0.022	0.013	0.011	0.009
	(0.018)	(0.018)	(0.018)	(0.019)	(0.019)	(0.019)
Fair health	-0.042*	-0.050**	-0.048**	-0.062***	-0.068***	-0.066***
	(0.022)	(0.022)	(0.022)	(0.023)	(0.023)	(0.023)
Bad health	-0.088***	-0.085***	-0.081**	-0.110***	-0.107***	-0.105***
	(0.033)	(0.033)	(0.032)	(0.033)	(0.033)	(0.033)
Urban Hukou	-0.075**	-0.077**	-0.078**	-0.052	-0.052	-0.055
	(0.036)	(0.035)	(0.034)	(0.041)	(0.040)	(0.040)
Not fully employed	0.001	0.001	-0.002	-0.001	-0.001	-0.003
	(0.022)	(0.022)	(0.022)	(0.024)	(0.023)	(0.023)
Retired	0.008	0.004	0.001	-0.011	-0.013	-0.015
	(0.033)	(0.032)	(0.032)	(0.034)	(0.034)	(0.034)
Education attainment	0.019*	0.016	0.021*	0.009	0.007	0.009
	(0.012)	(0.011)	(0.012)	(0.012)	(0.012)	(0.012)
Self-identified as member of upper class	0.008	0.011	0.011	0.026	0.030	0.030
	(0.033)	(0.033)	(0.032)	(0.038)	(0.037)	(0.037)
Communist party member	-0.002	-0.004	-0.008	0.034	0.032	0.029
	(0.027)	(0.027)	(0.027)	(0.029)	(0.029)	(0.029)
Work in state-owned or collective sector	-0.013	-0.007	-0.008	-0.027	-0.023	-0.024
	(0.021)	(0.021)	(0.021)	(0.023)	(0.023)	(0.023)
Tolerance of negative social behaviour			0.006***			0.003
			(0.002)			(0.002)
Tolerance of negative non-social behaviour			0.000			-0.000
			(0.002)			(0.002)
Familiarity with neighbours			0.025**			0.020*
			(0.010)			(0.010)
Observations	3481	3481	3481	3469	3469	3469
Adjusted R-squared	0.134	0.159	0.164	0.144	0.159	0.160

Note: Robust standard errors in parentheses: * p < 0.10, ** p < 0.05, *** p < 0.01. *Non-fully employed* is defined as those who are unemployed and employed part-time. *Tolerance of negative social behaviour* is defined as tolerance of the behaviours which are harmful to others. *Tolerance of negative non-social behaviour* is defined as tolerance of the behaviours which are not harmful to others but may not be socially acceptable. County fixed effects are included in the regressions.

Source: 2005 China General Social.

Table 3.3.3 OLS estimates in attitudes towards migrants

	Inviting migrant(s) to visit home			Having kids/relatives to marrying or being in a relationship with migrant(s)		
	(1)	(2)	(3)	(4)	(5)	(6)
Previous contact with migrants		0.217***	0.218***		0.154***	0.153***
		(0.029)	(0.028)		(0.028)	(0.028)
Age	-0.014**	-0.013**	-0.013**	-0.025***	-0.025***	-0.024***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
Squared age	0.000**	0.000**	0.000**	0.000***	0.000***	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.022	0.020	0.016	0.035*	0.033*	0.025
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
Fair health	-0.089***	-0.096***	-0.095***	-0.018	-0.023	-0.023
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Bad health	-0.136***	-0.131***	-0.130***	-0.026	-0.023	-0.024
	(0.034)	(0.034)	(0.034)	(0.034)	(0.033)	(0.033)
Urban Hukou	-0.051	-0.052	-0.052	-0.133***	-0.134***	-0.132***
	(0.046)	(0.045)	(0.045)	(0.047)	(0.047)	(0.047)
Not fully employed	-0.004	-0.004	-0.005	0.009	0.010	0.010
	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)	(0.025)
Retired	-0.019	-0.021	-0.022	0.011	0.010	0.010
	(0.036)	(0.035)	(0.035)	(0.036)	(0.036)	(0.036)
Education attainment	-0.014	-0.018	-0.016	0.013	0.010	0.011
	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)	(0.013)
Self-identified as member of upper class	0.040	0.043	0.045	0.030	0.032	0.034
	(0.040)	(0.039)	(0.040)	(0.041)	(0.041)	(0.041)
Communist party member	0.028	0.025	0.024	0.044	0.042	0.044
	(0.030)	(0.030)	(0.030)	(0.031)	(0.030)	(0.030)
Work in state-owned or collective sector	-0.018	-0.013	-0.012	-0.012	-0.009	-0.007
	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)	(0.024)
Tolerance of negative social behaviour			0.003			0.003
			(0.002)			(0.002)
Tolerance of negative non-social behaviour			0.002			0.004*
			(0.002)			(0.002)
Familiarity with neighbours			0.009			-0.001
			(0.011)			(0.011)
Observations	3466	3466	3466	3309	3309	3309
Adjusted R-squared	0.124	0.146	0.146	0.167	0.177	0.179

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. *Non-fully employed* is defined as those who are unemployed and employed part-time. *Tolerance of negative social behaviour* is defined as tolerance of the behaviours which are harmful to others. *Tolerance of negative non-social behaviour* is defined as tolerance of the behaviours which are not harmful to others but may not be socially acceptable. County fixed effects are included in the regressions.

Source: 2005 China General Social Survey.

Table 3.4 The first-stage results of conventional IV estimates and Lewbel heteroskedasticity identification estimates

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A Conventional IV						
Local population share of migrants	0.576***	0.574***	0.591***	0.573***	0.547**	0.576***
	(0.211)	(0.214)	(0.209)	(0.213)	(0.213)	(0.215)
Weak IV test statistics	7.454	7.191	7.968	7.255	6.605	7.159
Panel B Heteroskedasticity Identification - City fixed effects included						
Lewbel IV	-1.263***	-1.292***	-1.512***	-1.300***	-1.317***	-1.368***
	(0.265)	(0.265)	(0.271)	(0.261)	(0.261)	(0.275)
Weak IV test statistics	22.759	23.819	31.159	24.700	25.409	24.759
Panel C Heteroskedasticity Identification - County fixed effects included						
Lewbel IV	-1.263***	-1.290***	-1.511***	-1.298***	-1.316***	-1.366***
	(0.266)	(0.266)	(0.272)	(0.263)	(0.262)	(0.276)
Weak IV test statistics	22.545	23.537	30.863	24.443	25.16	24.439
Panel D Heteroskedasticity Identification - County fixed effects and interactions included						
Lewbel IV	-1.319***	-1.346***	-1.559***	-1.360***	-1.373***	-1.439***
	(0.255)	(0.254)	(0.260)	(0.251)	(0.252)	(0.263)
Weak IV test statistics	26.797	28.042	35.889	29.168	29.671	30.021
Observations	3464	3467	3481	3469	3466	3309

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. City fixed effects are included in the regressions of Panels A and B respectively, and county fixed effects are included in the regressions of Panel C. In addition to the control variables in Panel C, Panel D also controls for interactions between the local migrant population share and each control variables except fixed effect. The other control variables are the same to Specification (3) in Table 3.3. The Breusch-Pagan heteroskedasticity tests for all the regressions are rejected at 1% level. Weak IV test is Kleibergen-Paap test.

Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.5 The second-stage results of conventional IV estimates and Lewbel heteroskedasticity identification estimates

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A OLS						
Previous contact with migrants	0.206***	0.212***	0.213***	0.177***	0.218***	0.153***
	(0.024)	(0.026)	(0.028)	(0.028)	(0.028)	(0.028)
Panel B Conventional IV						
Previous contact with migrants	0.555	0.904	0.613	1.144	-0.115	-1.618**
	(0.408)	(0.576)	(0.531)	(0.716)	(0.568)	(0.763)
Panel C Heteroskedasticity Identification - City fixed effects included						
Previous contact with migrants	0.325***	0.314***	0.271***	0.128	0.009	0.053
	(0.098)	(0.102)	(0.097)	(0.116)	(0.123)	(0.110)
Local population share of migrants	0.132	0.339	0.202	0.581*	-0.068	-0.961***
	(0.227)	(0.284)	(0.301)	(0.319)	(0.320)	(0.303)
Panel D Heteroskedasticity Identification - County fixed effects included						
Previous contact with migrants	0.328***	0.320***	0.272***	0.126	-0.001	0.041
	(0.098)	(0.102)	(0.096)	(0.116)	(0.124)	(0.110)
Panel E Heteroskedasticity Identification - County fixed effects and interactions included						
Previous contact with migrants	0.337***	0.333***	0.284***	0.148	0.012	0.072
	(0.094)	(0.099)	(0.093)	(0.109)	(0.117)	(0.100)
Observations	3464	3467	3481	3469	3466	3309

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. City fixed effects are included in the regressions of Panels B and C respectively, and county fixed effects are included in the regressions of Panels A, D and E. Panel E also includes interactions between the local migrants population share and each control variables except fixed effect in addition to the covariates in Panel D. The other control variables are the same to Specification (3) in Table 3.3. The Breusch-Pagan heteroskedasticity tests for all the regressions are rejected at 1% level.
Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.6 Lewbel's heteroskedasticity identification estimates on respondents who may compete with migrants in the labour market

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Panel A Respondents with high school education or below						
Previous contact with migrants	0.319***	0.312***	0.232**	0.041	-0.040	0.054
	(0.108)	(0.114)	(0.105)	(0.129)	(0.137)	(0.118)
Observations	2822	2826	2846	2842	2832	2704
Weak IV test statistics	19.253	19.739	29.393	21.396	21.652	21.554
Panel B Respondents in private sector						
Previous contact with migrants	0.311*	0.286*	0.231*	-0.109	-0.261	-0.160
	(0.161)	(0.160)	(0.122)	(0.250)	(0.285)	(0.220)
Observations	1083	1086	1081	1073	1081	1028
Weak IV test statistics	5.242	5.988	18.659	5.483	5.788	5.849

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The control variables are the same as those included in Panel D of Table 3.5. The Breusch-Pagan heteroscedasticity tests for all the regressions are rejected at 1% level. Weak IV test is Kleibergen-Paap test.

Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.7 Lewbel's heteroskedasticity identification estimates on urban natives with rural Hukou

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Previous contact with migrants	0.146*	0.369**	0.277*	0.288**	0.311**	0.347
	(0.081)	(0.149)	(0.164)	(0.137)	(0.146)	(0.259)
Observations	144	141	139	141	140	130
Weak IV test statistics	132.534	91.681	95.346	105.429	138.508	101.276

Note: Robust standard errors in parentheses: * p < 0.10, ** p < 0.05, *** p < 0.01. The control variables are the same as those included in Panel D of Table 3.5. The Breusch-Pagan heteroscedasticity tests for all the regressions are rejected at 1% level. Weak IV test is Kleibergen-Paap test.
Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.8 Robustness check

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Panel A Lewbel IV without age limit						
Previous contact with	0.344***	0.319***	0.279***	0.108	-0.046	0.019
migrants	(0.107)	(0.111)	(0.105)	(0.128)	(0.141)	(0.124)
Observations	3464	3467	3481	3469	3466	3309
Weak IV test statistics	17.598	18.382	23.671	18.802	19.528	19.063
Panel B Exclude sample with rural Hukou						
Previous contact with	0.307***	0.308***	0.268***	0.137	0.004	0.024
migrants	(0.102)	(0.108)	(0.101)	(0.121)	(0.127)	(0.117)
Observations	3320	3326	3342	3328	3326	3179
Weak IV test statistics	21.927	22.58	29.129	23.428	24.689	23.792
Panel C Control for log average income per household member						
Previous contact with	0.294***	0.316***	0.275***	0.121	-0.024	0.016
migrants	(0.102)	(0.108)	(0.105)	(0.122)	(0.133)	(0.121)
Observations	3312	3316	3330	3320	3315	3166
Weak IV test statistics	20.165	21.164	26.832	21.897	22.347	21.706
Panel D Consistent sample across dependent variables						
Previous contact with	0.325***	0.346***	0.322***	0.115	0.083	0.036
migrants	(0.094)	(0.100)	(0.106)	(0.109)	(0.112)	(0.104)
Observations	3131	3131	3131	3131	3131	3131

Weak IV test statistics	28.038	28.038	28.038	28.038	28.038	28.038
Panel E Contact measure differentiating intimacy level						
Previous contact with	0.126***	0.096*	0.098*	0.039	-0.106	0.006
migrants	(0.046)	(0.052)	(0.052)	(0.064)	(0.085)	(0.064)
Observations	3464	3467	3481	3469	3466	3309
Weak IV test statistics	15.802	16.075	21.035	16.352	16.74	17.252

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The control variables are the same as those included in Panel D of Table 3.5. The Breusch-Pagan heteroscedasticity tests for all the regressions are rejected at 1% level. Weak IV test is Kleibergen-Paap test.
Source: 2005 China General Social Survey and 2005 1% Population Survey.

Appendix

Appendix A: Figures and tables

Table 3.A.1 Lewbel's heteroskedasticity identification estimates from the simultaneous model

	General attitude	working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Panel A Migrant population share and city fixed effect included						
Contact effect on attitudes	0.230***	0.233***	0.221***	0.123	0.088	0.083
	(0.077)	(0.083)	(0.085)	(0.087)	(0.092)	(0.125)
Attitudes effect on contact	0.141	0.112	0.052	-0.173	0.314	0.113
	(0.110)	(0.097)	(0.266)	(0.258)	(0.426)	(0.682)
P-value of heteroscedasticity test on contact	0.000	0.000	0.000	0.000	0.000	0.000
P-value of heteroscedasticity test on attitude	0.000	0.000	0.003	0.000	0.045	0.041
Panel B County fixed effect included						
Contact effect on attitudes	0.228***	0.234***	0.218***	0.123	0.069	0.108
	(0.078)	(0.083)	(0.083)	(0.085)	(0.098)	(0.178)
Attitudes effect on contact	0.147	0.118	0.060	-0.183	0.449	0.459
	(0.111)	(0.098)	(0.274)	(0.266)	(0.544)	(1.337)
P-value of heteroscedasticity test on contact	0.000	0.000	0.000	0.000	0.000	0.000

P-value of heteroscedasticity test on attitude	0.000	0.000	0.005	0.000	0.141	0.123
Observations	3464	3467	3481	3469	3466	3309

Note: Robust standard errors in parentheses: * p < 0.10, ** p < 0.05, *** p < 0.01. The exclusion restrictions are $cov[Z_{Hi}, (\beta_4'\mu_i' + \epsilon_{1i}')(\gamma_3'\mu_i' + \epsilon_{2i}')] = 0$, $cov[Z_{Hi}^2, (\beta_4'\mu_i' + \epsilon_{1i}')(\gamma_3'\mu_i' + \epsilon_{2i}')] = 0$ and $cov[Z_{Hi}^3, (\beta_4'\mu_i' + \epsilon_{1i}')(\gamma_3'\mu_i' + \epsilon_{2i}')] = 0$, where Z is the county-level migrant population share aged from 18 to 45. City or county fixed effects are included in the regressions. The other control variables are the same as Specification 3 in Table 3.3. P-value of the heteroskedasticity test is derived from Breusch-Pagan Test.

Source: 2005 China General Social Survey and 2005 1% Population Survey.

Table 3.A.2 Differences in gender view between male migrants and male urban locals

	Migrant	Urban locals with urban Hukou	Urban locals with rural Hukou	Difference: (2)-(1)	Difference: (3)-(1)
	(1)	(2)	(3)	(4)	(5)
The responsibility of wife is to take care of family	3.41	3.67	3.39	0.26**	0.02
Women should be dismissed first in the economic downturn	4.35	4.60	4.50	0.25**	0.15

Note: P-value is calculated using robust standard error. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. The gender role in the first row is from the question that “Do you agree that the responsibility of husband is to make money, and the responsibility of wife is to take care of family.” The gender role in the first second is from the question that “Do you agree that women should be dismissed first in the economic downturn.” The answers to these questions are “strongly agree” (1), “fairly agree” (2), “sort of agree” (3), “neither” (4), “sort of disagree” (5), “fairly disagree” (6), “strongly disagree” (7).

Source: 2006 China General Social Survey.

Table 3.A.3 Selected first-stage results of Lewbel's heteroskedasticity identification estimation

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
Lewbel IV	-1.263***	-1.290***	-1.511***	-1.298***	-1.316***	-1.366***
	(0.266)	(0.266)	(0.272)	(0.263)	(0.262)	(0.276)
Age	-0.003	-0.005	-0.003	-0.004	-0.004	-0.003
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
Squared age	0.000	0.000	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Male	0.002	0.003	0.004	0.003	0.001	0.001
	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)	(0.014)
Fair health	0.032*	0.033**	0.039**	0.031*	0.029*	0.036**
	(0.017)	(0.017)	(0.016)	(0.017)	(0.017)	(0.017)
Bad health	-0.016	-0.016	-0.015	-0.017	-0.023	-0.015
	(0.026)	(0.026)	(0.025)	(0.026)	(0.025)	(0.025)
Urban Hukou	-0.020	-0.015	-0.014	-0.020	-0.019	-0.013
	(0.034)	(0.034)	(0.033)	(0.034)	(0.034)	(0.034)
Not fully employed	0.000	-0.002	0.001	0.001	0.002	-0.004
	(0.016)	(0.016)	(0.015)	(0.015)	(0.015)	(0.016)
Retired	0.003	-0.002	0.008	0.003	-0.002	-0.005
	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)	(0.027)
Education attainment	0.016*	0.014	0.016*	0.016*	0.014	0.018*
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.010)
Self-identified as member of upper class	-0.002	-0.005	-0.023	-0.023	-0.020	-0.015
	(0.027)	(0.027)	(0.027)	(0.028)	(0.028)	(0.028)
Communist party member	0.017	0.011	0.011	0.014	0.012	0.011

	(0.022)	(0.022)	(0.022)	(0.022)	(0.022)	(0.023)
Work in state-owned or collective sector	-0.024	-0.022	-0.023	-0.019	-0.019	-0.018
	(0.015)	(0.015)	(0.015)	(0.016)	(0.016)	(0.016)
Tolerance of negative social behaviour	-0.002	-0.002	-0.002	-0.002	-0.002	-0.001
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Tolerance of negative non-social behaviour	0.002	0.002	0.002	0.002	0.002	0.002
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Familiarity with neighbours	-0.013*	-0.015**	-0.011	-0.011	-0.014*	-0.012
	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)	(0.008)
Observations	3464	3467	3481	3469	3466	3309

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Non-fully employed is defined as those who are unemployed and employed part-time. Tolerance of negative social behaviour is defined as tolerance of the behaviours which are harmful to others. Tolerance of negative non-social behaviour is defined as tolerance of the behaviours which are not harmful to others but may not be socially acceptable. County fixed effects are included in the regressions.

Source: 2005 China General Social Survey.

Table 3.A.4 Selected second-stage results of Lewbel's heteroskedasticity identification estimation

	General attitude	Working with migrant(s)	Living in the same community as migrant(s)	Having migrant(s) as next-door neighbour(s)	Inviting migrant(s) to visit home	Having kids/relatives to marrying or being in a relationship with migrant(s)
previous contact with migrants	0.328***	0.320***	0.272***	0.126	-0.001	0.041
	(0.098)	(0.102)	(0.096)	(0.116)	(0.124)	(0.110)
age	-0.008**	-0.010**	-0.018***	-0.016***	-0.014**	-0.025***
	(0.004)	(0.005)	(0.005)	(0.006)	(0.006)	(0.006)
squared age	0.000**	0.000*	0.000***	0.000**	0.000**	0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
male	0.020*	0.017	0.021	0.009	0.017	0.026
	(0.012)	(0.015)	(0.018)	(0.019)	(0.020)	(0.020)
fair health	-0.037**	-0.030*	-0.050**	-0.065***	-0.089***	-0.019
	(0.016)	(0.018)	(0.022)	(0.023)	(0.024)	(0.024)
bad health	-0.052**	-0.060**	-0.080**	-0.106***	-0.135***	-0.026
	(0.026)	(0.029)	(0.032)	(0.032)	(0.034)	(0.033)
urban hukou	-0.024	-0.018	-0.078**	-0.054	-0.052	-0.132***
	(0.024)	(0.035)	(0.033)	(0.039)	(0.045)	(0.046)
not fully employed	0.005	-0.014	-0.002	-0.003	-0.004	0.009
	(0.015)	(0.019)	(0.021)	(0.023)	(0.024)	(0.025)
retired	-0.023	-0.033	-0.000	-0.015	-0.020	0.011
	(0.025)	(0.028)	(0.031)	(0.033)	(0.035)	(0.035)
education attainment	0.016**	0.020**	0.020*	0.010	-0.013	0.013
	(0.008)	(0.010)	(0.011)	(0.012)	(0.013)	(0.013)
self-identified as upper class group member	-0.030	-0.042	0.012	0.029	0.041	0.033
	(0.021)	(0.035)	(0.032)	(0.037)	(0.039)	(0.040)
community party member	0.016	-0.006	-0.009	0.030	0.027	0.045
	(0.017)	(0.022)	(0.026)	(0.029)	(0.030)	(0.030)

work in state owned or collective sector	-0.004	0.003	-0.007	-0.025	-0.017	-0.009
	(0.014)	(0.017)	(0.020)	(0.022)	(0.024)	(0.024)
tolerance to negative social behaviour	0.006***	0.004**	0.006***	0.003	0.003	0.003
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
tolerance to negative non-social behaviour	0.000	0.002	0.000	-0.000	0.002	0.004*
	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
familarity with neighbors	0.005	0.011	0.026***	0.019*	0.007	-0.002
	(0.007)	(0.009)	(0.010)	(0.010)	(0.011)	(0.011)
Observations	3464	3467	3481	3469	3466	3309

Note: Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Non-fully employed is defined as those who are unemployed and employed part-time. Tolerance of negative social behaviour is defined as tolerance of the behaviours which are harmful to others. Tolerance of negative non-social behaviour is defined as tolerance of the behaviours which are not harmful to others but may not be socially acceptable. County fixed effects are included in the regressions.

Source: 2005 China General Social Survey.

Appendix B: Lewbel's (2012) heteroskedasticity approach

This appendix discusses the Lewbel's heteroskedasticity identification approach (2012) in the case of triangle linear probability model (LPM) where the dependent and endogenous variables are binary.

Consider a triangle model with binary variables y_1 and y_2 as follows

$$y_1 = \beta_0 y_2 + \beta_1 X + \epsilon_1 \qquad (3.B.1),$$

and

$$y_2 = \gamma_1 X + \epsilon_2 \qquad (3.B.2),$$

where $\epsilon_1 = \beta_2\mu + v_1 + e_1, \epsilon_2 = \gamma_2\mu + v_2 + e_2, y_1$ and y_2 could only be 0 or 1.

Assumption 1 : $\Pr(y_1 = 1) = E(y_1) = \beta_0 y_2 + \beta_1 X + \beta_2\mu + v_1$ and $\Pr(y_2 = 1) = E(y_2) = \gamma_1 X + \gamma_2\mu + v_2$.

This assumption follows the LPM framework. In this model, μ is the common factor between y_1 and y_2 which causes the endogeneity. v_j and e_j are the idiosyncratic disturbances which are specific to $y_j, j = 1, 2$. The difference between v_j and e_j is that v_j affects $E(y_j)$, but e_j does not. e_j is the binary residual from the LPM. Under this assumption, the expectation of y_1 is *only* determined by y_2, X, μ and v_1; and the expectation of y_2 is *only* determined by X, μ and v_2.

The results below immediately follow Assumption 1.

$$e_1 = \begin{cases} 1 - \beta_0 y_2 - \beta_1 X - \beta_2\mu - v_1, & y_1 = 1 \\ -\beta_0 y_2 - \beta_1 X - \beta_2\mu - v_1, & y_1 = 0 \end{cases}, \qquad (3.B.3),$$

and

$$e_2 = \begin{cases} 1 - \gamma_1 X - \gamma_2\mu - v_2, & y_2 = 1 \\ -\gamma_1 X - \gamma_2\mu - v_2, & y_2 = 0 \end{cases} \qquad (3.B.4).$$

Assumption 2: $y = (y_1, y_2)'$ and X are random vectors. $E(Xy'), E(Xy_1y'), E(Xy_2y')$,and $E(XX')$ are finite and identified from data. $E(XX')$ is nonsingular.

Assumption 3: $Cov(Z, \mu^2) = 0, Cov(Z, v_1v_2) = 0$ and $Cov(Z, \mu v_j) = 0$ where $j = 1, 2$.

Assumption 2 is consistent with Assumption A1 in Lewbel (2012). Assumption 3 is similar to the exclusion restriction in the conventional IV estimation, which is not testable but has been argued to plausibly hold in Section 3.3.

Lewbel (2012) shows that if $Cov(Z, \epsilon_1\epsilon_2) = 0$ and $Cov(Z, \epsilon_2^2) \neq 0$ then β_0 is identifiable. In the following I argue that under Assumptions 1 to 3 $Cov(Z, \epsilon_1\epsilon_2) = 0$ and $Cov(Z, \epsilon_2^2)$ is not necessarily zero when y_1 and y_2 are binary variables. Therefore, whether the heteroskedasticity identification is feasible in this case depends on $Cov(Z, \epsilon_2^2) = 0$ or not, which we can empirically test. For the sake of simplicity, I use the following notation in the analysis.

Notation: Let $I_1 = \{Z, X, y_2, \mu, v_1, v_2\}$, $I_2 = \{Z, X, \mu, v_1, v_2\}$, $a \equiv \beta_0 + \beta_1 X + \beta_2\mu + v_1$, $b \equiv \beta_1 X + \beta_2\mu + v_1$ and $p_2 \equiv \Pr(y_2 = 1|I_2) = \gamma_1 X + \gamma_2\mu + v_2$.

Proposition 1: Under Assumptions 1 to 3, $Cov(Z, \epsilon_1\epsilon_2) = 0$ holds for the model of Equations (3.B.1) and (3.B.2)

Proof: Substituting $\epsilon_1 = \beta_2\mu + v_1 + e_1$ and $\epsilon_2 = \gamma_2\mu + v_2 + e_2$ in $Cov(Z, \epsilon_1\epsilon_2) = 0$ shows that

$$\begin{aligned} Cov(Z, \epsilon_1\epsilon_2) &= Cov(Z, \beta_2\gamma_2\mu^2 + \gamma_2\mu v_1 + \beta_2\mu v_2 + \gamma_2\mu e_1 + \beta_2\mu e_2 + v_1 e_2 + v_2 e_1 + v_1 v_2 \\ &\quad + e_1 e_2) \\ &= Cov(Z, \gamma_2\mu e_1 + \beta_2\mu e_2 + v_1 e_2 + v_2 e_1 + e_1 e_2), \end{aligned}$$

where the second equality follows Assumption 3.

Since

$$\begin{aligned} E(e_1|I_1) &= (1 - \beta_0 y_2 - \beta_1 X - \beta_2\mu - v_1)(\beta_0 y_2 + \beta_1 X + \beta_2\mu + v_1) \\ &\quad +(-\beta_0 y_2 - \beta_1 X - \beta_2\mu - v_1)(1 - \beta_0 y_2 - \beta_1 X - \beta_2\mu - v_1) \\ &= 0, \end{aligned}$$

then

$$E(\mu e_1) = E_{I_1}[E(\mu e_1|I_1)] = E_{I_1}[\mu E(e_1|I_1)] = E_{I_1}[\mu \times 0] = 0,$$

and

$$E(Z\mu e_1) = E_{I_1}[E(Z\mu e_1|I_1)] = E_{I_1}[Z\mu E(e_1|I_1)] = E_{I_1}[Z\mu \times 0] = 0.$$

It follows that $Cov(Z, \mu e_1) = E(Z\mu e_1) - EZE(\mu e_1) = 0$. Similarly, $Cov(Z, \mu e_2) = Cov(Z, v_1 e_2) = Cov(Z, v_2 e_1) = 0$. Hence, $Cov(Z, \epsilon_1\epsilon_2)$ further simplifies to $Cov(Z, e_1 e_2)$.

Since

$$\Pr(y_1 = 1, y_2 = 1|I_2) = \Pr(y_1 = 1|y_2 = 1; I_2)\Pr(y_2 = 1|I_2) = ap_2,$$
$$\Pr(y_1 = 1, y_2 = 0|I_2) = \Pr(y_1 = 1|y_2 = 0; I_2)\Pr(y_2 = 0|I_2) = b(1 - p_2),$$
$$\Pr(y_1 = 0, y_2 = 1|I_2) = \Pr(y_1 = 0|y_2 = 1; I_2)\Pr(y_2 = 1|I_2) = (1 - a)p_2,$$

and

$$\Pr(y_1 = 0, y_2 = 0|I_2) = \Pr(y_1 = 0|y_2 = 0; I_2)\Pr(y_2 = 0|I_2) = (1 - b)(1 - p_2),$$
$$\begin{aligned} E(e_1e_2|I_2) &= E(e_1e_2|y_1 = 1, y_2 = 1; I_2)\Pr(y_1 = 1, y_2 = 1|I_2) \\ &+ E(e_1e_2|y_1 = 1, y_2 = 0; I_2)\Pr(y_1 = 1, y_2 = 0|I_2) \\ &+ E(e_1e_2|y_1 = 0, y_2 = 1; I_2)\Pr(y_1 = 0, y_2 = 1|I_2) \\ &+ E(e_1e_2|y_1 = 0, y_2 = 0; I_2)\Pr(y_1 = 0, y_2 = 0|I_2) \\ &= (1 - a)(1 - p_2)ap_2 + (1 - b)(-p_2)b(1 - p_2) \\ &+ (-a)(1 - p_2)(1 - a)p_2 + (-b)(-p_2)(1 - b)(1 - p_2) \\ &= 0. \end{aligned}$$

By law of iterated expectation, $E(e_1e_2) = E_{I_2}[E(e_1e_2|I_2)] = 0$. Similarly, $E(Ze_1e_2) = E_{I_2}[E(Ze_1e_2|I_2)] = E_{I_2}[ZE(e_1e_2|I_2)] = 0$.

Therefore, $Cov(Z, \epsilon_1\epsilon_2) = Cov(Z, e_1e_2) = E(Ze_1e_2) - EZE(e_1e_2)$=0 holds for Equations (3.B.1) and (3.B.2).

Proposition 2: Under Assumptions 1 to 3, $Cov(Z, \epsilon_2^2) = Cov(Z, v_2^2) + E[Zp_2(1 - p_2)] - EZE[p_2(1 - p_2)]$ holds for the model of Equations (3.B.1) and (3.B.2).

Proof: Since $E(Ze_2^2) = E_{I_2}[E(Ze_2^2|I_2)] = E[Zp_2(1 - p_2)]$ and $E(e_2^2) = E_{I_2}[E(e_2^2|I_2)] = E[p_2(1 - p_2)]$,

$$\begin{aligned} Cov(Z, \epsilon_2^2) &= Cov(Z, \mu^2 + v_2^2 + e_2^2 + 2\mu v_2 + 2\mu e_2 + 2v_2e_2) \\ &= Cov(Z, v_2^2 + e_2^2) \\ &= Cov(Z, v_2^2) + E(Ze_2^2) - EZE(e_2^2) \\ &= Cov(Z, v_2^2) + E[Zp_2(1 - p_2)] - EZE[p_2(1 - p_2)]\,. \end{aligned}$$

Proposition 2 shows that the heteroskedasticity comes from both the specific factor v_2^2 and the binary nature of y_2. Based on Propositions 1 and 2, if $Cov(Z, \epsilon_2^2) \neq 0$, Theorem 1 in Lewbel (2012) can be applied to estimate this triangle model.

Appendix C: Data Appendix

This section provides additional details about variable construction (see also Section 4). Tolerance of negative social behaviour comes from the questions: “If others have the following behaviours, what are your opinions?” The behaviours are “talking loudly in public occasion”, “smoking in front of or near non-smoker”, “spitting”, “throwing rubbish”, “swearing”, “jumping the queue”, “crossing the road without following traffic light and pedestrian lines”, “not being punctual”, “breaking one’s word”, “not caring about the senior, sick, disabled, pregnant and young people”. The respondents were asked to evaluate their attitudes to each behaviour from “not antipathy” (1) to “very antipathy” (5). The tolerance is measured as the difference between 50 and the summation of all these items.

Tolerance of negative non-social behaviour is constructed from the question “How do you agree with the following sentences I am going to read?”. The sentences are:

- “Cohabitation before marriage is an individual choice, and others should not criticise”;
- “Homosexual love is an individual choice, and others should not criticise”;
- “Reading adult books/watching adult video is an individual choice, and others should not criticise”;
- “Patronising a prostitute is an individual choice, and others should not criticise”;
- “Joining superstitious activities is an individual choice, and others should not criticise”;
- “Suicide is an individual choice, and others should not criticise”.

There are five options for each sentence, ranging from “completely disagree” (1) to “completely agree” (5). The index is calculated as the summation of the options across all the items and then minus six for normalisation.

Neighbourhood familiarity is extracted from the question “How familiar are you with your neighbours and people living in the same district/village as you?”. The answers are ranged from “very unfamiliar” (0) to “very familiar” (4).

Chapter 4 Who are the movers and who are the stayers? Attrition in the Migrant Household Survey of the Rural-to-Urban Migration in China Project: 2008-2013

4.1 Introduction

Rural-to-urban migration is one of the most important drivers of economic growth in developing countries. During migration, surplus labour in rural areas moves to the city to work in high return sectors. This process helps to efficiently allocate labour resources and expand the secondary and tertiary sectors, which, consequently, results in economic growth.

Across the world, more than 800 million rural dwellers have moved to urban areas since 1950 (FAO, 2004). One fifth of these rural-to-urban migrants live in China. Since the Chinese government eased restrictions on movement, the number of rural migrants has increased from 26 million in 1986 to 166 million in 2013 (see Figure 1.1). Rural migrants now account for more than one third of the urban labour force (Frijters et al., 2011b) and 18% of the rural population.[71] This massive rural-to-urban migration significantly stimulates the growth of Chinese economy. It is estimated that labour reallocation was responsible for 20% of economic growth, from 1978 to 1997 (Cai and Wang, 1999).

Rural-to-urban migration in China is widely recognised as important, but empirical studies are constrained by the available datasets. Most of the extant studies are based on cross-sectional datasets that contain migrant information in destination areas, such as China General Social Survey (e.g., Hu et al., 2011), China Household Income Project (e.g., D'emurger et al., 2009), Survey of Occupational Mobility and Migration (e.g., Zhang, 2010), Survey of Floating Population (e.g., Roberts, 2001) and self-collected data (e.g., Chen and Feng, 2013). One common drawback of these datasets is that their cross-sectional nature means that researchers cannot control for the individual fixed effects nor identify the effect of past behaviour on current behaviour. This limits the scope of research which can be conducted using these datasets, and makes thorough investigation impossible. Some studies utilise the longitudinal rural household surveys which are conducted in out-migration areas to cope with this problem (e.g., Giles and

[71] In China, 935 million people hold a rural household registration (NBS, 2012), according to the 2010 population census. Thus, migrants account for 18% of rural population.

Mu, 2007; Giulietti et al., 2014). However, these rural household surveys may suffer from the selection problem because they may omit information on households which have migrated entirely (Bilsborrow et al., 1984). According to NBS (2014), around a fifth of rural migrants moved with their entire households between 2008 and 2013. The absence of these migrants means that the rural household surveys are likely to suffer from a large selectivity bias.

Given the aforementioned concerns, a longitudinal migrant survey conducted in destination areas is in great need to facilitate the studies on the internal migration in China. In 2008, the Rural-to-Urban Migration in China (RUMIC) Project initiated a longitudinal migrant household survey covering 15 major migration destination cities. The survey collects detailed information on the city life of the rural migrants. To mitigate the problem of selection bias, the survey also collects substantial information on pre-migration life, the first migration experience, left-behind families and hometown characteristics. Currently the RUMIC migrant household survey is the largest and longest longitudinal survey of rural-to-urban migrants in China. It has become valuable infrastructure to the migration research. To date, more than two hundred researchers have requested the data and used them in studies of well-being, migration and labour market dynamics (see a summary in Akguc et al., 2014).[72]

However, migrants tend to be highly mobile, so the RUMIC migrant household survey has a high attrition rate. Between the first two waves (2008-2009), 64% of households left the survey. The overall attrition rate then gradually decreased to 52% between the second and the third waves (2009-2010), 43% between the third and fourth waves (2010-2011), 33% between the fourth and fifth waves (2011-2012) and 38% between the fifth and sixth waves (2012-2013). To properly interpret the results based on this survey and provide experience to the future survey conduction, it is necessary to understand the formation and impact of the attrition.

This chapter studies attrition in the first six waves of the RUMIC migrant household survey. In particular, I focus on three questions. What types of respondents are more likely to leave the survey sample? Does attrition bias the statistical estimates? Are the respondents who remain in both the initial and follow-up surveys still representative of the migrant population at the time of the follow-up survey? The answers to the last two questions depend on the cases and models studied, so I illustrate them in terms of the earnings equation, which is a focus in labour economics.

[72] I thank Corrado Giulietti for providing the number of data requests.

First, I find that respondents who remain in the RUMIC migrant household survey tend to be socio-economically better off and more established than those who leave. They economically gain more from migration and are more willing to stay in cities. They are less mobile within cities and more likely to be self-employed. These findings depict a general picture of those who stay in the survey, so that empirical researchers can better understand the characteristics of the non-attritors and the external validity of the results when the non-attriting sample is used (i.e., to what types of migrants the results are more applicable). The findings also helps survey designers recognise the types of migrants they need to put more effort into tracking, if they want to reduce the survey attrition rate.

As to the second and third questions, I find evidence that attrition does bias the estimates and that the sample of respondents who remain is not representative of the general migrant population at the time of follow-up surveys. However, in the examples of earnings equation, attrition and sample representativeness do not always generate large biases in the regression coefficients of the individual-level characteristics that are most relevant to research and policy interests; indeed, the existence and magnitude of bias are case-dependent. This suggests that practitioners should neither ignore nor overly worry about the possible existence of bias caused by attrition and (un)representativeness. Instead, practitioners should evaluate the bias according to their own cases.

The next section describes the RUMIC migrant household survey. Section 4.3 reviews the general pattern of attrition in the survey. Sections 4.4 to 4.6 present the results, answering the three questions mentioned above. Section 4.7 concludes with discussion.

4.2 The RUMIC Migrant Household Survey

The RUMIC Project was established to study rural-to-urban migration in China. The project includes three different longitudinal surveys: the urban household survey, the rural household survey and the migrant household survey. The rural household survey is conducted on out-migration areas and includes both the migrant and non-migrant members of the households. The migrant household survey is conducted on in-migration areas and only includes migrant households. This chapter focuses on the migrant household survey, as the other two surveys have low attrition rates.[73]

[73] For example, the attrition rates between the first two waves were 1% in the rural household survey and 5.7% in the urban household survey. See the data description at http://idsc.iza.org/?page=27&id=58.

Currently, the RUMIC migrant household survey is the largest migrant household survey in China. Each wave it consists of around 5000 households from 15 cities located in the largest migrant-sending or migrant-receiving provinces in China.[74] These 15 cities are destinations to many migrants. According to the 2005 1% Population Survey, the rural migrants in these 15 cities account for 38% of total migrants in China.[75] The RUMIC migrant household survey is also the longest longitudinal migrant household survey in China. The baseline wave was conducted in 2008, and follow-up waves have been conducted annually since 2009. At the time of writing, the seventh wave of the survey is underway. Such a long panel survey not only allows us to control for individual time-invariant characteristics, but also to closely scrutinise the rapidly-changing city life of migrants.

The RUMIC survey has two advantages over previous migrant surveys in China. First, it provides a more representative sample of migrants. A common problem with previous surveys is that their sampling frames are largely based on residential address; thus, overlooking migrants who do not have a formal address (e.g., living in workplaces or workplace dormitories). By contrast, the RUMIC migrant household survey uses a sampling frame that is established on the census of migrants' workplaces. This sampling frame includes migrants living in workplaces, which, hence, makes it possible to collect a representative sample of migrants.[76] [77]

Second, the RUMIC migrant household survey contains many pieces of unique information which other surveys do not typically observe. Specifically, it provides rich data on the psychological characteristics of migrants (e.g., mental health problems, trust and risk aversion), the left-behind families of migrants (e.g., left-behind children, spouse, and parents), and hometown characteristics of migrants. This information helps mitigate the problem of omitted variables in the substantive research. Regarding the attrition literature, it also offers us a unique opportunity to make a more in-depth examination of the predictors of attrition.

[74] These cities are Guangzhou, Dongguan, Shenzhen, Zhengzhou, Luoyang, Hefei, Bengbu, Chongqing, Shanghai, Nanjing, Wuxi, Hangzhou, Ningbo, Wuhan and Chengdu.

[75] In 2005, there are 283 cities in China.

[76] Another advantage of the sampling frame of the RUMIC migrant household survey is that it includes also street workers and taxi drivers who are often omitted in other surveys. See Gong et al. (2008) and Kong (2010) for more details of sampling strategy.

[77] The RUMIC project conducted three censuses. The first census was conducted in 2008 and used for the sampling in 2008 and 2009 waves; the second census was conducted in 2009 and used for the sampling in 2010 to 2012 waves, and the third census was conducted in 2012 and used for the sampling in 2013 to 2014 waves.

Internal migrants are highly mobile, so the RUMIC project has devised several ways to keep track of respondents. In the interview, the detailed contact information of respondents is recorded, including their working and residential addresses, phone numbers and contact details of their three associates. Using the recorded information the survey company is required to contact the respondents twice or three times by either visiting their working or residential address or making a phone call to them in the period after the previous wave ends but before the new follow-up wave starts. During the same period, the RUMIC project also runs three lotteries to encourage the respondents to stay in the survey and keep in touch with the survey company.[78] After the follow-up survey formally begins, the fieldworkers are required to track all the respondents who were in the previous wave by visiting their working or residential addresses first. If the respondents are not founded in the visit, they are contacted by phone. If the respondents are founded in the survey cities, the fieldworkers would arrange interviews with them. To reduce the attrition rate, the tracking process continues throughout the whole survey period.[79] If the respondents cannot be contacted or are out of scope during the whole survey period, then they attrited.

It would be ideal if the respondents who moved out of the survey cities could also be tracked and interviewed. However, this is not financially viable, given that respondents may move to regions far away from the survey cities.[80] Therefore, the follow-up waves only re-interview the respondents who remain in the survey cities. This strategy inevitably leads to sample attrition. Since 2009, the survey has adopted a split panel design to alleviate the impact of this attrition. The follow-up survey consists of two parts. One part includes the households tracked from the previous wave (called "old sample") and the other includes random refreshments from each city (called "new sample"). The size of the new sample in each city depends on the number of households tracked and is designed such that the total sample size of the old and new samples is about the same as the original sample size in the baseline wave. This unique survey design offers three opportunities to us. First, we can use the old sample to construct a longitudinal

[78] The prize ranged from 50 Yuan to 2000 Yuan in 2008 wave and increases in the later waves according to how long the respondents stay in the survey. The lottery is designed so that around 4.6% of households win a prize each wave.

[79] The surveys usually start around the middle of March and end before the middle of September. Migrants tend to return to their hometown during the Spring Festival (January-February), so the survey period is deliberately chosen to guarantee that most migrants have come back to the cities. However, the 2013 follow-up survey was delayed, for financial reasons of the survey company. In 2013, all the city surveys ended in December, except for Guangzhou and Shanghai which ended in 23rd January, 2014.

[80] Note that the respondents in 2008-2012 waves are from 1780 source counties. On average, these home counties are 392km away from destination cities. The home counties of 12% of respondents (3017) are over 1000km from their destination cities. Clearly, it would be extremely costly to track migrants who return home. Although there is not much information on what are the other cities the migrants moved to, the cost to track these migrants is also likely to be prohibitively expensive.

sample through which we can conduct fixed effect estimation or explore dynamic changes. Second, the baseline wave and new sample in the follow-up waves can constitute a repeated cross-sectional dataset which is free of the attrition problem. Third, the new sample can serve as a benchmark to roughly check whether the old sample is representative of the migrant population at the time of follow-up, because the new sample consists of random refreshments within each city, every wave.

4.3 General picture of attrition

Many respondents attrited in the RUMIC migrant household survey, despite the substantial tracking efforts described in the previous section. Table 4.1 shows the general pattern of attrition. The first column shows the attrition rate of the original 2008 sample, and the other columns show the attrition rates of the new samples in subsequent waves. The attrition rate is defined as the ratio of number of lost households or individuals over the total number which appeared in the previous wave. I show the attrition rate at both the household level (see Panel 1) and individual level (see Panels 2 and 3).[81]

The three panels of Table 4.1 all suggest that the survey has a very high attrition rate. In the first follow-up waves, the attrition rates vary from 43% to 64% across years. The attrition rate drops continuously in subsequent follow-up waves. By the fifth follow-up wave, the attrition rates have fallen to 18% to 19%. This drop is probably because the most mobile respondents have already left and those who remain have settled into city life. Another possibility is that the survey gains credibility over time, so that respondents are more willing to participate. This pattern is consistent with other surveys (e.g., PSID, see Fitzgerald et al. (1998) and Zabel (1998)).

Overall, the attrition rate in the RUMIC migrant household survey is much higher than those in other household surveys which do not focus solely on migrants. Table 4.2 lists the attrition rates for several normal household surveys, in both developed and developing countries. The usual attrition rate between the baseline wave and the first follow-up wave is around 5% to 21%, much lower than the attrition rate in the RUMIC migrant household survey. However, it is important to note that migrant-specific surveys often have higher attrition rates than normal household surveys. Indeed, Bilsborrow et al. (1984) gives two examples. They mention that

[81] Table 4.1 excludes 2195 observations (3.8%) or 677 individuals (2.2%) who exited the survey and then came back ("non-absorbing attrition"). I also exclude new entrants in old households from the individual-level analysis.

only one third of households in a migrant survey that was conducted in Bangkok could be tracked to the same dwelling six to seven months after the initial contact. Similarly, half of the respondents in an Iranian migrant survey had left the sample within four months.[82] Therefore, although high, the attrition rate in the RUMIC migrant household survey is not abnormal.

Theoretically, there are three reasons why a respondent might leave the survey: return to hometown, move to other cities, and stay in the survey city but refuse to be interviewed. Table 4.3 presents the distribution of attrition for the 2012 RUMIC migrant survey.[83] It shows that 8.7% of the attrited respondents returned to their hometown, 5% went to other cities, 17.6% refused to participate in the survey and 68.7% were lost because the fieldworkers could not contact them. Of the households which could not be contacted, 62% had an invalid phone number, 3.5% always kept their mobile phone off during the contact period, and 34.5% did not answer phone call or did not have a phone.[84] [85] It is likely that these households went to other cities or returned to their hometown. In China, if mobile phone users use a mobile phone number registered in another city, they are charged a substantial roaming fee.[86] Thus, when respondents move to other cities, they are likely to stop or avoid using their old mobile phone number and change to a local number. If a mobile phone number has been out of use or out of credit for certain amount of time, it is deactivated and put back on the market for new users. In this case, the fieldworkers would find the phone number is invalid.[87]

Given these facts, Table 4.3 reveals two useful messages. First, mobility is the main reason of attrition. The majority of attrition is likely to be caused by respondents moving out of the survey

[82] Unfortunately, there is no information on survey methodology about these two surveys, so I am unable to make detailed comparison with the RUMIC migrant household survey.

[83] The survey company was asked to record why their attempt to track a respondent failed, but from 2008 to 2013, they only recoded this data for the 2012 wave. As well, information from Zhengzhou in the 2012 wave was lost.

[84] "Invalid phone number" includes cases where the phone number was no longer used, did not exist, or had changed hands. If the phone number exists but the respondent did not answer phone call, then this case belongs to "did not answer phone call or did not have a phone' rather than 'invalid phone number" in Table 4.3.

[85] Note that all the households in the category of "lose contact" cannot be found in their working and residential addresses.

[86] For instance, consider a mobile phone number registered with Shanghai China Mobile. If the phone number is used in Shanghai, a phone call costs 0.2 yuan per minute. If the phone number is used in another city, the call incurs an additional roaming fee of 0.4 or 0.6 yuan per minute, depending whether the call is received or made.

[87] An invalid phone number could also mean that the respondent deliberately provided a false phone number. However, the fieldworkers reported that only 2.9% of households did not cooperate with them well in the interview or their answers are not trustworthy. Hence, this possibility is unlikely to explain why such a large proportion of respondents failed to be contacted.

scope. By contrast, refusal to participate only accounts for a small part of attrition.[88] Second, recording a mobile phone number does not guarantee that the respondent can be tracked, because of high roaming fees. This is an important consideration for designing future surveys that track migrant movements across cities. It would be helpful to record contact details which are portable, but not affected by a change in phone number, such as Weichat (the Chinese version of *What's up*).

Table 4.1 shows that the RUMIC migrant household survey has high attrition rates and Table 4.3 indicates that such large attrition rates are likely to be caused by migrant mobility. A question naturally arises: what makes migrants so mobile that this survey has such a high attrition rate? There are several potential answers.

The age structure of migrants may contribute to the high attrition rate. In general, young people are more mobile (Olsen, 2005). For example, the attrition rate for individuals aged 20 to 24 years is 23.4%, which is almost double the overall attrition rate (13.2%) in the HILDA survey (Watson and Wooden, 2004). In the CFPS survey, the attrition rate for individuals aged 16 and 25 years is 28.8%. It is also higher than the overall attrition rate (21.4%). Respondents in the RUMIC migrant survey tend to be younger than the general Chinese population. The average age of a RUMIC respondent is 32 years, and 37% of respondents are aged 16 to 25 years. The corresponding figures for the CFPS survey are 46 years and 14.2%. Thus, the age profile of the RUMIC survey might be explaining part of its high attrition rate.

China's unique context may also contribute to migrant mobility and consequently the RUMIC survey's high attrition rate. First, the majority of migrants are not eligible for welfare in the city. They usually return to their hometown when they are unemployed, sick, or old. They may also tend to relocate to other cities to find jobs that provide better benefit, such as social insurances. Both these movements push them out of the survey scope. Second, the children of migrants are often restricted to access the local city schools, so one parent sometimes stay home to look after them. This institutional barrier means that many migrants are separated from their left-behind families, which reduces their length of stay in the city (Meng, 2012) and increases their mobility. Third, migrants change jobs more often than urban residents (Knight and Yueh, 2004). Changing jobs can mean changing residential address, so high job mobility may also contribute to high attrition rates.

[88] Note that the refusal rate is only 6.3% if the households that were successfully tracked are taken into account.

4.4 Attrition analysis

Given the survey's high attrition rate, it is important to understand which respondents are more likely to leave. The answer to this question will help us understand the external validity of the results obtained from the old sample. Further, knowing which types of migrants tend to exit will guide future survey design. To answer this question, I first make mean comparison between the attritors and non-attritors on several characteristics, to give a general picture of the non-attritors. Then I extend the analysis to the multivariate framework to examine the predictors of attrition. I use the characteristics in the initial wave when respondents first entered this survey to conduct these two exercises.

As 2013 wave is the latest follow-up survey available to me, I study the attrition pattern of respondents who first entered the survey between 2008 and 2012. There are 24574 respondents who participated the survey during this period. In the following analysis, I restrict the sample to respondents with absorbing attrition who are aged between 16 and 65 years, since this sample represents the primary labour force and is usually the focus of empirical studies (e.g., Frijters et al., 2011a; Meng and Xue, 2014).[89] This sample restriction leaves 21990 individuals in the analysis. I further restrict the sample to respondents who did not provide missing data in a set of variables used in the mean comparison analysis, to make the sample consistent across variables.[90] [91] Finally, I have a sample of 18530 respondents for my analysis. As we mainly use the longitudinal sample to implement Fixed Effect estimation or examine the behaviour dynamics of migrants across waves, I focus on two types of attrition: respondents who left the survey in the first follow-up wave after they entered the survey (called immediate attrition), and respondents who did not stay in the survey for all the first six waves (called long-term attrition).[92] Please refer to Appendix for details of the variable construction.

[89] Absorbing attrition means that once the respondent exits the survey, he/she will never come back to the survey. There are 2.2% of respondents with non-absorbing attrition.

[90] These variables are age, gender, years of schooling, education attainment, marriage status, weekly hours worked, monthly earnings, hypothetical monthly earnings in rural origin, access to social insurances, employment status, years since the first migration, willingness to stay in cities, whether respondents are likely to move residential address in the next 12 months. I also repeated the following analysis without applying this sample restriction. The results are similar.

[91] Note that due to the changes of questionnaire there are still some changes in the sample size across variables in the mean comparison analysis. All the changes are explained in Table 4.4

[92] In other words, the respondents with the immediate attrition only stay in the survey for one wave, and the respondents with long-term attrition remain in the survey for less than six waves.

4.4.1 Mean comparison

Table 4.4 examines mean differences between the attritors and non-attritors across different variables. Panel A shows the results for respondents who stayed in the survey for at least two waves and those who only stayed in the survey for one wave (immediate attrition), based on the characteristics in the initial waves when respondents first entered the survey. The sample comprises all individuals who entered the survey from 2008 to 2012. Panel B shows the differences in the characteristics of the 2008 wave between respondents who were always in and those who ever exited the survey during the first six waves (long-term attrition). The sample in Panel B only includes the individuals in the 2008 wave.

Table 4.4 includes a set of variables which are typically considered in other attrition analysis, such as age, gender, education and other demographic and socio-economic characteristics (e.g., Fitzgerald et al., 1998; Falaris, 2003; Velsquez et al., 2011; Thomas et al., 2012). Apart from these variables, I also include several variables which are unobserved in other studies but which are important to migrant mobility. In particular, these variables include the information on migration experience, potential income opportunity in hometown, willingness to live in cities, and psychological and behavioural preferences. These variables either reflect a preference for migration or proxy the (opportunity) cost of migration; thus, they may affect the mobility of respondents, and hence be relevant to attrition. For simplicity, I group these variables into five categories: (i) demographic and household structure variables, (ii) health, psychological and behavioural preferences, (iii) economic performance and welfare, (iv) work related variables, and (v) other variables.[93]

The discussion begins with Panel A.

<u>Demographic and Household Structure Variables</u> *(Panel 1 of Table 4.4)*

The demographic and household structure variables suggest that non-attritors are older and more likely to be married and live with their children and spouses, than attritors. On average, non-attritors are two years older and 12 percentage points more likely to be married than attritors. For those who are married, non-attritors are 10 percentage points more likely to live with their spouses, and for those who have children younger than 16 years, non-attritors are 14 percentage points more likely live with their children, than attritors. In China, due to institutional barrier, many migrants are unable to migrate with their family. The fact that a higher proportion of non-attritors living with their spouses and children suggests that they may be socio-economically

[93] Note that the wage and earnings related variables are adjusted for CPI.

better off so that they have higher chance to overcome the institution barrier. Abraham et al. (2006) point out that living with spouse and children is also an indication of better integration into the local community.

Figure 4.1 illustrates attrition rates at different ages. It suggests a U-shaped relationship between age and attrition. Attrition rates are highest for the younger and older age groups and lowest for the mid-30s to mid-40s age group. Since around 90% of respondents are 46 years old or younger, the mean comparison is dominated by the downward slope between age and attrition. Hence, the attritors are younger than the non-attritors, on average.

Health and Psychological Preferences (Panel 2 of Table 4.4)

The health and psychological preferences data reveal mixed differences between attritors and non-attritors. On the one hand, non-attritors tend to have worse physical health than attritors (see height and self-rated health). On the other hand, the GHQ score, a measure of mental health problems (see the details in Appendix), suggests that non-attritors enjoy better mental health than attritors. In addition, non-attritors tend to be less risk-loving, but more trusting. Although statistically significant, it should be noted that the magnitudes of these differences are not large. For example, the difference in GHQ score only accounts for 7% to 8% of its standard deviation.

Economic Performance and Welfare (Panel 3 of Table 4.4)

Looking at economic performance and welfare, non-attritors tend to have better labour market outcomes in cities and worse income opportunities in their rural origins, than attritors. Thus, migration offers more benefits to non-attritors than attritors. On average, non-attritors in cities work more weekly hours and earn significantly more income than attritors. These gaps widen even further when I exclude unemployed migrants and unpaid family helpers.[94] The data on hypothetical monthly earnings in rural origin suggests that non-attritors would have earned significantly less than attritors had they stayed in their hometowns, indicating that there is less opportunity cost for them to migrate. If we consider earnings in cities and hometowns separately, the magnitudes of the differences between non-attritors and attritors are not large. But if we consider the income gains created by migration, then the difference is large. Non-attritors gain monthly 130 yuan more than attritors, which accounts for 13% and 15% of income gain for non-attritors and attritors, respectively. Figure 4.2 explores the differences in income gains at different ages. Panel 1 shows that the earnings gain of non-attritors is greater than that of

[94] Note that non-attritors had insignificantly lower income than attritors in 2011 and 2012, as shown in Table 4.A.1

attritors at almost all ages, except the very young (16 years) or the old (above 57 years). Finally, non-attritors are more likely to have social insurances.[95]

<u>Work Related Variables</u> *(Panel 4 of Table 4.4)*

There is a significant difference in employment status between non-attritors and attritors. Non-attritors are 14 percentage points more likely to be self-employed than attritors. There are two possible reasons for this difference. First, self-employed migrants are less mobile and more likely to stay in the same city than salaried workers, because they may invest more time and effort into establishing city-based business and social networks. Second, it might be easier for self-employed migrants to participate in the survey because they have more flexible work schedules than salaried workers. For these reasons, self-employed migrants are less likely to attrite than salaried workers.

Non-attritors are 9 percentage points more likely to work in a small work unit. Since self-employed migrants and non-paid family helpers are less likely to attrite and usually work in small workplace, after excluding them, the difference is much smaller (2.3 percentage points), but still statistically significant. This moderate difference may be because it is difficult to track migrants in large workplaces, for three reasons. First, the increases in city prices of labour and land may force some factories to relocate out of the survey area, to the edge of the survey cities or to smaller cities and towns where production costs are lower (see wage increase of migrants and relevant discussion in Meng, 2014). Second, some large work units are very mobile. For example, construction companies move from one project site to another, and possibly move out of the survey area. These two possibilities make it hard to track migrants employed in large workplaces. Third, according to the enumerators, large workplaces are less willing to participate the survey, making it difficult to interview their migrant employees.

<u>Other Migration-Related Variables</u> *(Panel 5 of Table 4.4)*

The differences on the other migration-related variables suggest that non-attritors tend to migrate to cities earlier and migrate within their home province. These results suggest that non-attritors may be better assimilated into cities, have accumulated more skills for city jobs and be

[95] In Table 4.4, for the salary workers ‘social insurances’ is defined as having medical health insurance, pension, work injury insurance and unemployment insurance at the same time, and for self-employed and non-paid family helpers it is defined as having medical health insurance and pension.

more familiar with local communities.[96] Additionally, non-attritors are more willing to stay in cities permanently and less likely to move to other places within the next twelve months.

The results of the long-term attrition in Panel B are similar to the results of the immediate attrition, but with larger magnitudes of difference for many variables, such as the demographic and migration-related characteristics.[97] [98]

4.4.2 Multivariate analysis of attrition

Table 4.5 extends the analysis to the multivariate setting. I use the linear probability model to examine the determinants of attrition. Panels A and B investigate immediate and long-term attritions, respectively. In the linear probability model, the dependent variable is the dummy variable of whether the respondent attrited and the independent variables are the characteristics in the initial wave in which the respondent first joined the survey. The sample used is the same as Table 4.4 except excluding the unemployed, as no work related information is available for these respondents.[99]

I include the variables discussed in Table 4.4 as the explanatory variables. To account for the regional and time differences, I add the destination city and survey wave fixed effects in the explanatory variables.

Many of the findings in Table 4.4 re-appear in Table 4.5. The coefficients of age again show a U-shaped relationship with attrition. The presence of spouse and children is negatively correlated with the probability of attriting. The estimated income at rural origin is positively associated with propensity to attrite, and access to social insurances is associated with less

[96] The 2012 survey wave collected information on how many neighbour households the respondent knows in the city. This variable is significantly and positively correlated with years since first migration after adjusting for city fixed effects. It verifies the argument that years since the first migration may reflect familiarity with the local community.

[97] The main differences between these two panels lie in access to social insurances. In contrast to Panel A, the attritors in Panel B are more likely to be covered by social insurances than the non-attritors. However, this difference disappears in the multivariate analysis in Table 4.5, which suggests that this difference is driven by correlation with other factors.

[98] The 2008 survey wave did not collect data about risk and trust, so Panel B does not compare these two variables.

[99] The 2010 survey wave did not collect health data (self-rated health and height and mental health problems) and the 2008 survey wave did not collect risk and trust data, so the sample sizes vary across the Panel A regression specifications.

likelihood of attriting. Self-employed migrants and migrants employed in small workplaces are less likely to attrite. Finally, the coefficients in years since the first migration also a show U-shaped relationship; attrition is negatively correlated with willingness to stay in cities and positively correlated with intention to move elsewhere.

The main differences between Tables 4.4 and 4.5 are that the variables of educational attainment become significant, and the variables of marital status, health, psychological and behavioural preferences and urban labour market performances become insignificant. This change is caused by multicollinearity. Due to the negative correlations between age and education and between age and attrition, the unconditional correlation of education is small and insignificant (see the analysis of mean comparison), but variables of educational attainment become significant when age is controlled for.[100] Since the variables of marital status, health, psychological preferences and urban labour market performances are positively correlated with education, these variables become insignificant after controlling for education. This suggests that education explains the differences on these variables in the analysis of mean comparison, and that educational attainment is an important predictor of attrition, when other variables are kept constant.

The results presented in Panels A and B of Table 4.5 are qualitatively similar; however, the coefficients of some variables for long-term attrition are smaller and less significant (e.g., education dummy variables, being self-employed and access to social insurances). This indicates that these observed variables have less predictive power for long-term attrition, and whether the respondent always stays in the survey depends more on unobserved factors.

In summary, the results in Tables 4.4 and 4.5 suggest that several differences exist between non-attritors and attritors. First, non-attritors seem to have better socio-economic conditions and be more established than attritors. They are more likely to be highly educated, covered by social insurances and have stayed longer in the city. These differences also result in better labour market outcomes and better mental health for non-attritors. Second, the non-attritors are inclined to stay in the city more. In terms of the individual preference, the non-attritors tend to prefer staying in cities permanently. They also tend to bring their families to the city. From the point of view of economic motivation, the non-attritors gain more from their migration, as the potential income they would have earned at hometown is lower. Third, non-attritors are less likely to change residential address within the next twelve months. This is probably because they live in larger households and have higher moving costs. Last, non-attritors are more likely to be self-employed. These differences give a general picture of the non-attritors. However, the low R-

[100] The unconditional differences on educational attainments are also not significant in the unreported results.

square tells us that these observed variables only explain a limited part of the attrition; unobserved variables account for more of the difference between non-attritors and attritors.

4.5 Does attrition bias the estimates?

The above analysis suggests that attrition in the RUMIC migrant household survey is not random. Does this non-random attrition necessarily bias the regression estimates? In this section, I take earnings regression as an example to answer this question.

I use the test proposed by Becketti et al. (1988) (usually called BGLW test) to test for the existence of attrition bias. This approach is often used in the literature (e.g., Fitzgerald et al., 1998; Alderman et al., 2001; Falaris, 2003). In particular, I run regressions of log monthly earnings against the characteristics in the waves in which the respondents first entered the survey, and examine whether the regression coefficients differ significantly between attritors and non-attritors. If the coefficients are found to be significantly different and the unobserved factors in the earnings equation are correlated across waves, then the attrition would cause bias.

Table 4.6 shows the results of the earnings regression where a set of conventional human capital variables are taken as the explanatory variables, and Panels A and B examine immediate and long-term attrition, respectively. I separate male and female cases, to avoid model misspecification and to be consistent with the extant literature (e.g., Fitzgerald et al., 1998; Falaris, 2003). The sample used is slightly different from Table 4.5 due to missing values in different covariates, and in order to be consistent with the extant literature (e.g., Fitzgerald et al., 1998) the respondents who reported zero wage are also excluded in Table 4.6.

The results of immediate attrition in Table 4.6 suggest that attrition bias possibly exists, but does not greatly affect the estimates on individual characteristics. F-tests show that the coefficients are unequal between attritors and non-attritors, indicating that attrition bias probably exists; however, the magnitude of the bias on the individual characteristics (i.e., the variables except city dummies, survey wave dummies and constant) which are usually the focus of substantive research is mostly insignificant and small. The only difference on the individual characteristics which is significant at the 5% level is “married” in the female sample. The coefficients change from -0.114 among non-attritors to -0.049 among attritors. This change does not cause either different coefficient signs or a substantive change on significance level between non-attritors and attritors.

Relative to immediate attrition, long-term attrition generates larger differences in some coefficients between the non-attritors and attritors, though no difference on the coefficients of individual characteristics are significant at the 5% level. In particular, the differences on the coefficients of "age" and "divorce" are large, which changes the significance level or sign of the coefficients. This suggests that if the respondents who always stay in the survey are used to study the age-earning profile or impact of marriage dissolution, the results may suffer from attrition bias and cannot be generalised to the original sample of the baseline wave.

This example carries three important messages. First, the results of immediate attrition suggest that attrition in the RUMIC migrant household survey does not necessarily bias the regression estimates of the individual-level characteristics which are the subject of most research and policy interest. Second, the difference in the results between immediate and long-term attrition indicates that whether attrition generates large bias is case-dependent. Practitioners should test attrition bias according to their own case. Third, for longitudinal income related-studies, the balanced sample may be at more risk of attrition bias than the unbalanced sample.

4.6 Is the sample of non-attritors representative of the general migrant population at the time of follow-up surveys?

Fitzgerald et al.'s (1998) study shows that attrition in the PSID survey is non-random, but the sample of non-attritors is still roughly representative of the general population in 1989, twenty years after the baseline wave. This finding suggests that the sample of non-attritors in the longitudinal survey is possible to maintain the representativeness of the population at the time of the follow-up survey, even if attrition is not random. This section examines whether this possibility holds true for the RUMIC migrant household survey. This piece of analysis tells us whether attrition causes incomparability between the sample of non-attritors and the general migrant population at the time of the follow-up surveys.

I use the new sample (i.e., the random refreshment within cities) in the follow-up waves of the RUMIC migrant household survey as the benchmark to check the representativeness of the non-attritors.[101] I first compare the mean of some key characteristics and then take earnings

[101] It is very difficult to find an external dataset which contains representative information on migrants for the same time period as the RUMIC migrant household survey. Some other surveys include migrants, for instance the 2010 wave of the China Family Panel Survey (CFPS) and the 2011 pilot of the China Labour Force Dynamic Survey

regression as an example to compare whether the coefficients are significantly different between the new sample and non-attritors. If the non-attritors are representative in all the cities, then there should be no significant differences in the mean comparison results and regression coefficients; otherwise, the non-attritors are unrepresentative. As the distribution of sample size of the new sample across cities and years is different from that of the non-attritors, in the following analysis I re-weight the new sample so that its distribution of sample size the same as that of the non-attritors. This avoids the possibility that different sample size distributions generate differences in the mean comparisons and regression coefficients.[102]

4.6.1 Mean comparison

Table 4.7 shows the differences between the non-attritors and the new samples for the 2009, 2011 and 2013 waves. The comparison for the other waves is shown in Table 4.A.2. Several variables show significant differences between the new sample and non-attritors and these differences are consistent across waves. The non-attritors are older and more likely to be married and live with their spouses and children. They tend to rate themselves as less healthy, probably because they are older. Non-attritors tend to work more than the new sample (when unemployed migrants are excluded) and earn more (when both unemployed migrants and non-paid family helpers are excluded). In terms of employment status the non-attritors are more likely to be self-employed. Interestingly, they are also more likely to be unemployed, even though the magnitudes of difference are not very large. This may be because in the households of the non-attritors there tend to be other well-established members supporting them during periods of unemployment in the city. Finally, non-attritors tend to migrate earlier than the new sample; they are more willing to stay in cities permanently and are less likely to change residential address in the next twelve months. In addition to the differences which are consistent across waves, there are some variables sporadically showing differences in some waves, such as male, mental health problems and psychological preferences. Overall, these differences suggest that the old sample (i.e., the sample of non-attritors) is not representative of the migrant population in the years when the follow-up surveys were conducted.[103]

(CLDS). But these surveys have very small migrant samples in the same destination provinces as the RUMIC survey (616 migrants in CFPS and 180 migrants in CLDS) so they are not desirable benchmarks.

[102] The results are generally similar if the non-attritors are re-weighted according to the distribution of the new sample or the analysis is unweighted.

[103] Tables 4.4 and 4.7 reveal an interesting pattern: the new sample could partially replace the attritors from the previous waves. Comparing these two tables suggests that the difference between the new and old samples is in the same direction as the difference between the attritors and non-attritors. For example, in Table 4.7 the new sample is younger and less likely to be married and self-employed; Table 4.4 shows that the attritors also have these properties.

4.6.2 Regression comparison

Tables 4.8 and 4.9 use the same earnings regression as Table 4.6 to examine whether the regression coefficients are different between the new sample and the non-attritors for males and females, respectively. This examination allows us to evaluate the differences between these two samples on the unobserved part. Tables 4.8 and 4.9 report the results for the 2009, 2011 and 2013 waves. The results for the other waves are shown in Tables 4.A.3 and 4.A.4.

In Tables 4.8 and 4.9, F-tests significantly reject equality of all the coefficients, suggesting that the sample of non-attritors is not representative; however, whether this pervasively biases the coefficients of individual-level attributes depends on the case. For example, the coefficient differences in the Table 4.9 regressions for females are largely insignificant and small, but there are several large and/or significant coefficient differences for males in Table 4.8 (e.g., age-related coefficients in the 2013 wave and coefficients on divorce in all waves). This example shows that in some cases it is possible for the coefficients of individual-level characteristics to be equal between the old and new samples. If so, practitioners can combine these two samples and control for the interactions between city fixed effects and the indicator variable of the new sample to improve estimation efficiency. However, practitioners should test whether the estimated coefficients are significantly different or not, before pooling the old and new samples.

4.7 Concluding remarks

This chapter investigates attrition in the migrant household survey of the RUMIC project. To my knowledge, it is the first study to extend the attrition literature to a survey which focuses solely on migrants. The RUMIC migrant household survey has a much higher attrition rate than normal household surveys, likely because of the mobility of migrants.

This chapter allows us to draw several important conclusions about attrition in the RUMIC migrant household survey. First, non-attritors and attritors are different. Non-attritors tend to be better off than attritors and enjoy higher economic gains from migration. They are more willing to stay in cities, are less mobile and are more likely to be self-employed. This finding provides the predictors of attrition and sheds light on which types of migrants are more likely to leave the

However, the magnitudes of difference between these two tables are not same because the population of migrants may be changing and the random refreshments in the follow-up waves includes "non-attritors".

survey. If the estimation needs a non-attritor sample (e.g., fixed effect model), then the characteristics of the non-attritors tell practitioners external validity of results (i.e., which sorts of migrants the estimates may apply to). This finding also tells us that in a regression where the dependent variable is likely to be affected by these predictors, if we do not control for these predictors, then the estimation is likely to suffer from attrition bias.

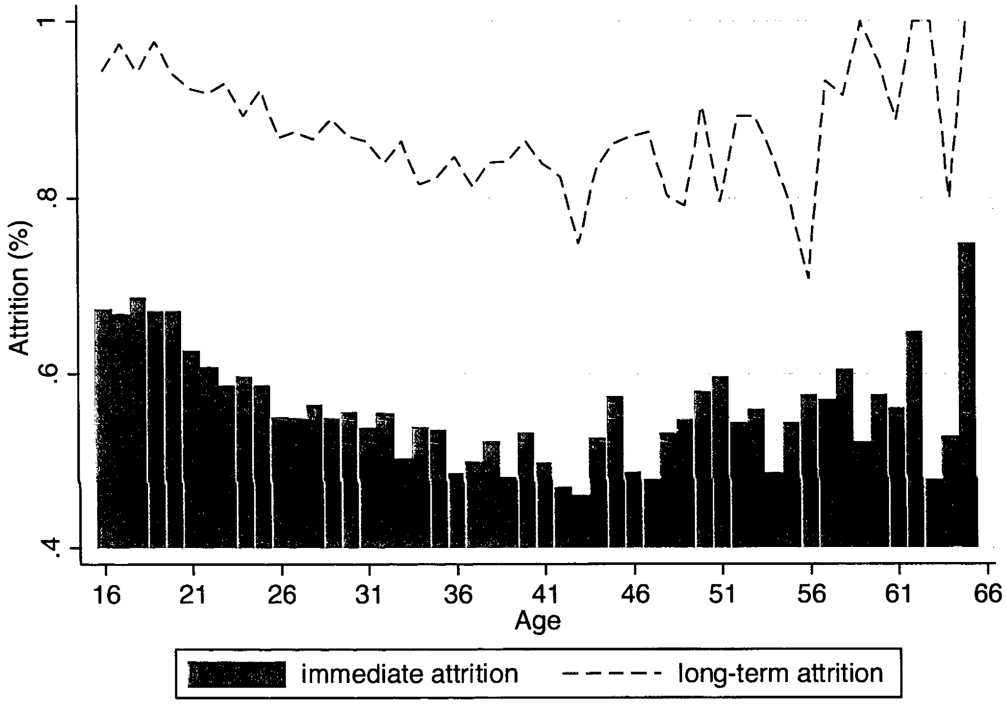
Second, whether attrition significantly biases the estimates is case-dependent. In the example of earnings regression, the BGLW test indicates that the sample of immediate attrition seems to suffer from small attrition bias, but the sample of long-run attrition does not. This suggests that practitioners should evaluate attrition bias according to their own cases. For income-related studies, the example of earnings regression suggests that the balanced panel sample may suffer from larger attrition bias than the unbalanced panel sample.

Third, comparing the sample of non-attritors and the new sample reveals that the non-attritor sample may not be representative of the migrant population at the time of follow-up waves. However, the regression analysis suggests that the impact of sample representativeness on the regression coefficients is case-dependent. In some cases, only a few coefficients of interest are affected by the (un)representativeness of the non-attritor sample. This indicates that in these cases, we can combine the new sample and the non-attritor sample, and control for interactions between the indicator variable of the new sample and variables which are differentially associated with outcome variables (e.g., city fixed effect) to increase the estimation efficiency.

Since the new sample consists of random refreshments in each city and each wave, if the attrition bias is found to be large in the sample of non-attritors, one suggestion is that practitioners could simply use the new sample to obtain the estimates free of attrition bias.

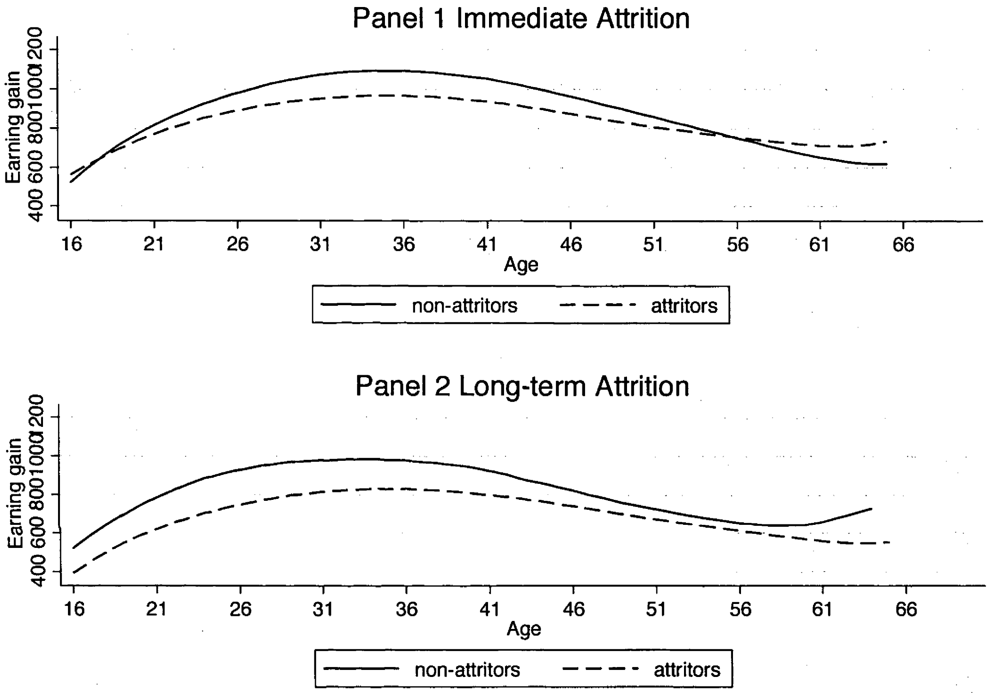
Figures and tables

Figure 4.1 Attrition rate by age



Source: 2008-2013 waves of the RUMIC migrant household survey.

Figure 4.2 Earnings gain by attritors and non-attritors



Note: The graphs are generated by LOWESS command in Stata 13.

Source: 2008-2013 waves of the RUMIC migrant household survey.

Table 4.1 General pattern of the attrition

	2008 Sample	2009 Sample	2010 Sample	2011 Sample	2012 Sample
Panel 1 households					
1 year after the survey	64%	58%	59%	47%	56%
2 years after the survey	40%	33%	38%	38%	
3 years after the survey	21%	22%	33%		
4 years after the survey	18%	24%			
5 years after the survey	19%				
Panel 2 individuals					
1 year after the survey	62%	57%	56%	43%	53%
2 years after the survey	38%	31%	35%	36%	
3 years after the survey	20%	21%	33%		
4 years after the survey	17%	24%			
5 years after the survey	19%				
Panel 3 individuals aged between 16 and 65					
1 year after the survey	62%	59%	57%	44%	54%
2 years after the survey	41%	30%	35%	36%	
3 years after the survey	18%	21%	34%		
4 years after the survey	17%	25%			
5 years after the survey	18%				

Note: Panel 1 excludes 93 (0.5%) households which have non-absorbing attrition or whose initial waves in the survey cannot be found. Panel 2 excludes 2195 (3.8%) observations or 677 (2.2%) individuals which have non-absorbing attrition or whose initial waves in the survey cannot be found, and also exclude 4325 (7.5%) or 2401 (7.8%) individuals who were new entrants in old households. Panel 3 restricts the individuals who are between 16 and 65 based on Panel 2.

Source: 2008- 2013 waves of the RUMIC migrant household survey.

Table 4.2 Attrition rates in normal household surveys

Dataset	Country	Attrition Rate	Period Gap	Reference
Michigan Panel Study of Income Dynamics (PSID)	U.S.	12%	1 year	Fitzgerald et al. (1998)
British Household Panel Survey (BHPS)	U.K.	13%	1 year	Uhrig (2008)
Household, Income and Labour Dynamics in Australia survey (HILDA)	Australia	13%	1 year	Watson and Wooden (2004)
Mexican Family Life Survey (MxFLS)	Mexico	11%	3 years	Velasquez et al. (2011)
Indonesia Family Life Survey (IFLS)	Indonesia	5%	4 years	Thomas et al. (1998)
China Family Panel Survey (CFPS)	China	21%	2 years	Author's own calculation
China Health and Nutrition Survey (CHNS)	China	13%	2 years	Liang (2011)

Note: “Period Gap” is the time between the baseline survey and the first follow-up survey.

Table 4.3 Distribution of attrition reasons

Panel A Overall Distribution			
Return migrants	Go to other cities	Refusal	Lose contact
8.70%	5.00%	17.60%	68.70%
Panel B Distribution within Losing Contact			
Invalid phone number	Phone off	Did not answer phone call or did not have a phone	
62.00%	3.50%	34.50%	

Note: All the households in the category of “lose contact” cannot be found in their working and residential addresses.

Source: 2012 RUMIC migrant household survey, excluding Zhengzhou data due to missing.

Table 4.4 Mean comparison between attritors and non-attritors

	Panel A: Immediate attrition				Panel B: Long-term attrition			
	Non-attritor	Attritor	Diff		Non-attritor	Attritor	Diff	
Panel 1 Demographic and household structure variables								
Age	32.70	30.60	2.00	***	34.20	30.80	3.30	***
	(7998)	(10532)			(795)	(5690)		
Male (%)	56.30	58.00	-1.70	**	56.70	57.30	-0.50	
	(7998)	(10532)			(795)	(5690)		
Years of schooling	9.40	9.40	0.00		9.10	9.30	-0.20	**
	(7998)	(10532)			(795)	(5690)		
Married (%)	67.00	54.90	12.00	***	80.80	60.10	20.60	***
	(7998)	(10532)			(795)	(5690)		
Divorced (%)	1.20	1.20	0.00		0.90	1.00	-0.10	
	(7998)	(10532)			(795)	(5690)		
Presence of spouse in the household (%)[a]	80.30	70.80	9.50	***	88.00	76.30	11.70	***
	(5356)	(5785)			(642)	(3420)		
Presence of child under 16 in the household (%)[b]	55.00	40.80	14.10	***	65.00	45.30	19.70	***
	(3328)	(3437)			(440)	(2250)		
Panel 2 Health and psychological preferences								
Height (cm)[c]	165.80	166.20	-0.30	***	165.40	166.00	-0.60	**
	(6882)	(9032)			(795)	(5686)		
Good health or better (%)[c]	84.40	85.30	-0.90		81.30	85.20	-4.00	***
	(6625)	(8721)			(795)	(5690)		
Mental health problems - GHQ score[d]	0.90	1.00	-0.10	***	0.90	1.00	-0.10	
	(5443)	(7334)			(692)	(4986)		

Risk-loving[e]	4.10	4.30	-0.10	***				
	(4398)	(5264)						
Trust (%)[e]	51.70	49.20	2.50	**				
	(4399)	(5264)						
Panel 3 Economic performance and welfare								
Weekly hours worked	63.40	60.10	3.30	***	68.10	60.90	7.20	***
	(7998)	(10532)			(795)	(5690)		
Weekly hours worked[f]	64.60	61.20	3.40	***	69.80	62.50	7.20	***
	(7847)	(10336)			(776)	(5544)		
Monthly earnings	1815.00	1733.00	82.00	***	1446.70	1422.20	24.50	
	(7998)	(10532)			(795)	(5690)		
Monthly earnings[g]	1900.60	1807.70	92.90	***	1622.10	1531.10	91.00	**
	(7638)	(10097)			(709)	(5285)		
Hypothetical monthly earnings in rural origin	798.20	846.60	-48.40	***	508.90	667.00	-158.10	***
	(7998)	(10532)			(795)	(5690)		
Earnings gain from migration	1016.80	886.40	130.40	***	937.70	755.20	182.60	***
	(7998)	(10532)			(795)	(5690)		
Access to social insurances[h]	13.10	11.20	1.90	***	6.30	9.10	-2.80	***
	(7998)	(10532)			(795)	(5690)		
Panel 4 Work-related variables								
Self-employment (%)	29.60	15.20	14.40	***	42.40	17.90	24.50	***
	(7998)	(10532)			(795)	(5690)		
Non-paid family helper (%)	2.60	2.30	0.30		8.40	4.60	3.90	***
	(7998)	(10532)			(795)	(5690)		
Unemployed (%)	1.90	1.90	0.00		2.40	2.60	-0.20	
	(7998)	(10532)			(795)	(5690)		
Less than 50 workers in the workplace[f] (%)	65.40	55.80	9.60	***	76.90	60.50	16.50	***

	(7835)	(10319)			(776)	(5541)		
Less than 50 workers in the workplace[i] (%)	49.00	46.70	2.30	***	52.70	49.00	3.70	
	(5260)	(8481)			(372)	(4263)		
Panel 5 Other migration-related variables								
Years since first migration	8.60	7.20	1.40	***	9.40	7.50	1.90	***
	(7998)	(10532)			(795)	(5690)		
Migrated from the same prov as destination (%)	55.90	47.80	8.20	***	75.70	54.40	21.40	***
	(7998)	(10532)			(795)	(5690)		
Willingness to stay in cities permanently (%)	64.90	54.10	10.80	***	71.20	58.50	12.70	***
	(7998)	(10532)			(795)	(5690)		
Likely to move (%)	9.20	16.50	-7.30	***	7.40	17.80	-10.40	***
	(7998)	(10532)			(795)	(5690)		

Note: * significant at 10% level; ** significant at 5% level; *** significant at 1% level. Panel A includes the individuals who first entered the survey in 2008 to 2012, and Panel B includes the individuals in the 2008 wave. The sample is restricted to migrants with absorbing attrition aged between 16 years and 65 years. The characteristics compared are those in the initial waves when respondents first entered the survey. The sample size is in parenthesis.

a. excluding respondents who are not married.

b. excluding respondents who do not have children under 16 years.

c. no information in the 2010 sample.

d.no information in the 2010 sample and respondents who were not present in the interviews in the other waves.

e.no information in the 2008 sample and respondents who were not present in the interviews in the other waves.

f. excluding unemployed.

g. excluding non-paid family helpers and unemployed.

h. having access to all of unemployment insurance, a pension, injury insurance and medical insurance for salary workers, and having access to both of a pension and medical insurance for self-employed and non-paid family helpers.

i. excluding unemployed, self-employed and non-paid family helpers.

Source: 2008-2013 waves of the RUMIC migrant household survey.

Table 4.5 Linear probability model of attrition

	Panel A			Panel B
	Immediate attrition			Long-term attrition
Demographic and household structure variables				
Age	-0.006**	-0.010***	-0.015***	-0.009**
	(0.003)	(0.004)	(0.005)	(0.003)
Squared age/100	0.007	0.012**	0.018***	0.010**
	(0.004)	(0.005)	(0.006)	(0.005)
Male	0.002	-0.008	-0.001	-0.004
	(0.006)	(0.012)	(0.016)	(0.011)
Junior high school	-0.010	0.000	0.001	-0.019
	(0.010)	(0.012)	(0.016)	(0.012)
Below senior high school or equivalent	-0.080***	-0.081***	-0.070**	-0.007
	(0.019)	(0.021)	(0.030)	(0.019)
Senior high school or equivalent	-0.040***	-0.037***	-0.017	-0.028*
	(0.012)	(0.014)	(0.018)	(0.014)
Above senior high school	-0.086***	-0.074***	-0.054*	-0.062**
	(0.019)	(0.022)	(0.028)	(0.027)
Married	0.018	0.017	0.017	0.026
	(0.015)	(0.017)	(0.022)	(0.016)
Divorced	0.011	0.001	0.032	0.014
	(0.034)	(0.038)	(0.045)	(0.043)
Presence of spouse in the household	-0.055***	-0.063***	-0.059***	-0.043***
	(0.013)	(0.015)	(0.020)	(0.014)
Presence of child under 16 years in the household	-0.042***	-0.042**	-0.042*	-0.029
	(0.015)	(0.017)	(0.022)	(0.019)
Health and psychological preferences				
Height (cm)		-0.000	-0.001	0.000
		(0.001)	(0.001)	(0.001)
Good health		0.015	-0.009	0.016
		(0.010)	(0.014)	(0.010)
Average health		-0.006	0.003	0.007
		(0.015)	(0.020)	(0.016)
Poor health or worse		-0.069*	-0.028	-0.052
		(0.037)	(0.047)	(0.050)
Mental health problems-GHQ score		0.006**	0.006	0.002
		(0.003)	(0.004)	(0.003)
Risk-loving			0.003	
			(0.002)	
Trust			-0.004	
			(0.012)	
Economic performance and welfare				
Weekly hours worked	-0.001*	-0.000	0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)
Log (1+monthly earnings)	-0.002	-0.005	0.002	-0.009
	(0.007)	(0.009)	(0.011)	(0.011)
Log (1+hypothetical monthly earnings) at origin	0.009***	0.007**	0.007**	0.008**

	(0.002)	(0.003)	(0.003)	(0.003)
access to social insurances[a]	-0.091***	-0.107***	-0.105***	-0.027*
	(0.013)	(0.015)	(0.018)	(0.016)
Work-related variables				
Self-employed	-0.105***	-0.094***	-0.071***	-0.093***
	(0.013)	(0.016)	(0.021)	(0.018)
Non-paid family helper	-0.054	-0.099	-0.006	-0.125
	(0.057)	(0.074)	(0.121)	(0.082)
Less than 50 people in the workplace	-0.016*	-0.024**	-0.026*	-0.005
	(0.009)	(0.010)	(0.014)	(0.010)
Other variables				
Years since first migration	-0.014***	-0.013***	-0.013***	-0.005**
	(0.002)	(0.002)	(0.003)	(0.002)
Squared years since first migration/100	0.048***	0.048***	0.048***	0.022**
	(0.008)	(0.009)	(0.012)	(0.010)
Migrated from the same province as destination	-0.005	-0.013	-0.043***	-0.032***
	(0.011)	(0.012)	(0.016)	(0.012)
Willingness to stay in cities permanently	-0.039***	-0.033***	-0.038***	-0.018*
	(0.009)	(0.010)	(0.013)	(0.010)
Likely to move	0.086***	0.094***	0.109***	0.043***
	(0.012)	(0.013)	(0.018)	(0.011)
City dummies and survey wave dummies	Yes	Yes	Yes	Yes
Observations	18154	12605	7027	5577
Adjusted R-squared	0.118	0.117	0.114	0.157

Note: * significant at 10% level; ** significant at 5% level; *** significant at 1% level. Panel A includes the individuals who first entered the survey in 2008 to 2012. Panel B includes the individuals in the 2008 wave. The sample is restricted to migrants with absorbing attrition aged between 16 years and 65 years.

a. having access to all of unemployment insurance, a pension, injury insurance and medical insurance for salary workers, and having access to both of a pension and medical insurance for self-employed and non-paid family helpers.

Source: 2008-2013 waves of the RUMIC migrant household survey.

Table 4.6 BGLW test on the earnings equation

Panel A Immediate Attrition

	Male			Female		
	Non-attritor	Attritor	Diff	Non-attritor	Attritor	Diff
Age	0.029***	0.041***	-0.011*	0.031***	0.029***	0.002
	(0.005)	(0.004)	(0.007)	(0.006)	(0.005)	(0.008)
Square of age	-0.000***	-0.001***	0.000	-0.000***	-0.000***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.025***	0.021***	0.004	0.021***	0.024***	-0.003
	(0.003)	(0.003)	(0.004)	(0.003)	(0.002)	(0.004)
Married	0.016	0.041**	-0.025	-0.114***	-0.049**	-0.065**
	(0.021)	(0.017)	(0.028)	(0.025)	(0.020)	(0.032)
Divorced	0.010	-0.020	0.030	-0.010	0.013	-0.022
	(0.050)	(0.046)	(0.068)	(0.061)	(0.055)	(0.082)
Years since first migration	0.028***	0.024***	0.005	0.022***	0.021***	0.001
	(0.004)	(0.003)	(0.005)	(0.004)	(0.004)	(0.006)
Square of years since first migration	-0.001***	-0.001***	-0.000	-0.001***	-0.001***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self-employment	0.157***	0.162***	-0.005	0.259***	0.318***	-0.060*
	(0.020)	(0.024)	(0.031)	(0.021)	(0.024)	(0.032)
City dummies and survey wave dummies	Yes	Yes		Yes	Yes	
Observations	4843	6606		3496	4457	
F Test1			5.09***			2.12**
F Test2			3.76***			3.02***

Panel B Long-term Attrition

	Male			Female		
	Non-attritor	Attritor	Diff	Non-attritor	Attritor	Diff
Age	-0.002	0.041***	-0.043*	0.005	0.038***	-0.033
	(0.022)	(0.006)	(0.023)	(0.028)	(0.008)	(0.029)
Square of age	-0.000	-0.001***	0.001*	-0.000	-0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.029***	0.023***	0.005	0.031***	0.035***	-0.003
	(0.011)	(0.003)	(0.011)	(0.012)	(0.004)	(0.012)
Married	0.050	0.042*	0.007	-0.153	-0.044	-0.109
	(0.076)	(0.024)	(0.080)	(0.127)	(0.027)	(0.130)
Divorced	0.202	0.005	0.197	-0.436**	-0.097	-0.338*
	(0.171)	(0.065)	(0.183)	(0.174)	(0.077)	(0.190)
Years since first migration	0.031**	0.029***	0.002	0.034**	0.024***	0.010
	(0.014)	(0.005)	(0.015)	(0.016)	(0.006)	(0.017)
Square of years since first migration	-0.001*	-0.001***	-0.000	-0.001	-0.001***	-0.000
	(0.001)	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)
Self-employment	0.226***	0.261***	-0.035	0.302***	0.343***	-0.041
	(0.054)	(0.025)	(0.059)	(0.071)	(0.032)	(0.077)
City dummies	Yes	Yes		Yes	Yes	
Observations	464	3440		291	2232	
F Test1			1.25			1.23
F Test2			1.70**			1.94***

Note: Standard errors are clustered at the household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. City fixed effects are included in both panels, and Panel A also includes survey wave fixed effects. Test 1 tests the equality of coefficients of individual-level characteristics listed in the table, and Test 2 tests the equality of all the coefficients including constant term, destination city fixed effects and survey wave fixed effect. Panel A includes the individuals who first entered the survey in 2008 to 2012. Panel B includes the individuals in the 2008 wave. Source: 2008-2013 waves of the RUMIC migrant household survey.

Table 4.7 Mean comparison between the new sample and old sample (non-attritor)

	Panel A: 2009			Panel B: 2011			Panel C: 2013		
	New	Old	Diff	New	Old	Diff	New	Old	Diff
Panel 1 Demographic and household structure variables									
Age	31.77	33.51	-1.75 ***	31.86	34.24	-2.37 ***	32.39	36.55	-4.16 ***
	(4153)	(2379)		(2695)	(3720)		(1618)	(3461)	
Male (%)	57.08	58.26	-1.18	53.83	56.26	-2.43 *	52.65	55.45	-2.79
	(4153)	(2379)		(2695)	(3720)		(1618)	(3461)	
Years of schooling	9.45	9.26	0.19 ***	9.39	9.52	-0.13	10.03	9.36	0.67 ***
	(4153)	(2379)		(2695)	(3720)		(1618)	(3461)	
Married (%)	61.15	74.27	-13.12 ***	57.48	74.44	-16.96 ***	56.66	75.99	-19.33 ***
	(4153)	(2379)		(2695)	(3720)		(1618)	(3461)	
Divorced (%)	1.24	1.26	-0.02	2.06	1.56	0.5	2.06	1.91	0.15
	(4153)	(2379)		(2695)	(3720)		(1618)	(3461)	
Presence of spouse in the household (%)[a]	73.77	83.36	-9.59 ***	77.21	84.11	-6.9 ***	70.29	84.98	-14.7 ***
	(2492)	(1767)		(1546)	(2769)		(1025)	(2630)	
Presence of child under 16 years in the household (%)[b]	51.68	64.28	-12.6 ***	50.89	62.62	-11.73 ***	38.54	66.23	-27.7 ***
	(1517)	(1103)		(859)	(1589)		(546)	(1389)	
Panel 2 Health and psychological preferences									
Height (cm)[c]	165.88	165.87	0.01	165.58	165.73	-0.15	166.16	165.7	0.47
	(4143)	(2378)		(2690)	(3706)		(1609)	(3451)	
Good health or better (%)[c]	84.05	76.08	7.96 ***	88.01	81.4	6.61 ***	89.92	86.16	3.76 ***
	(4153)	(2379)		(2695)	(3720)		(1616)	(3455)	
Mental health problems - GHQ score[d]	0.92	1.18	-0.25 ***	1.1	1.03	0.07	0.72	0.94	-0.22 ***
	(3342)	(1857)		(2136)	(2781)		(1245)	(2422)	
Risk-loving[e]	4.33	4.1	0.23 ***	4.33	4.27	0.06	3.77	3.67	0.1
	(3342)	(1857)		(2141)	(2781)		(1247)	(2429)	
Trust (%)[e]	58.36	53.74	4.61 ***	53.44	54.8	-1.36	45.6	42.94	2.66

	(3342)	(1857)			(2141)	(2781)			(1243)	(2429)		
Panel 3 Economic performance and welfare												
Weekly hours worked	61.65	62.46	-0.81		63.32	61.35	1.97	***	59.49	61.85	-2.36	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Weekly hours worked[f]	63.37	65.6	-2.24	***	63.66	64.34	-0.68		60.37	65.6	-5.23	***
	(4034)	(2265)			(2678)	(3547)			(1590)	(3263)		
Monthly earnings	1601.9	1696.96	-95.05	***	2121.1	2135.99	-14.89		2333.88	2405.07	-71.2	
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Monthly earnings[g]	1670.86	1816.86	-146	***	2150.04	2261.85	-111.81	***	2370.93	2573.09	-202.16	***
	(3982)	(2222)			(2661)	(3513)			(1586)	(3235)		
Hypothetical monthly earnings in rural origin	663.72	702.15	-38.43	***	932.43	1025.33	-92.9	***	1361.78	1328.89	32.9	
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Earnings gain from migration	938.19	994.81	-56.62	**	1188.67	1110.66	78.01	**	972.09	1076.19	-104.09	*
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Access to social insurances[h]	8.1	11.06	-2.96	***	12.55	17.04	-4.49	***	24.07	25.8	-1.73	
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Panel 4 Work-related variables												
Self-employment (%)	24.27	36.4	-12.13	***	25.38	35.4	-10.02	***	22.19	39.32	-17.13	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Non-paid family helper (%)	1.42	1.81	-0.39		0.81	0.91	-0.1		0.11	0.81	-0.7	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Unemployed (%)	2.71	4.79	-2.09	***	0.53	4.65	-4.12	***	1.45	5.72	-4.27	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Less than 50 workers in the workplace[f] (%)	61.9	68.04	-6.14	***	65.43	68.03	-2.61	**	67.93	68.74	-0.81	
	(4028)	(2256)			(2667)	(3535)			(1589)	(3247)		
Less than 50 workers in the workplace[i] (%)	48.74	49.56	-0.82		54.11	49.09	5.03	***	59.21	46.06	13.16	***
	(3086)	(1350)			(2043)	(2190)			(1156)	(1865)		
Panel 5 Other migration-related variables												
Years since first migration	8	9.68	-1.69	***	8.09	10.51	-2.42	***	8.58	12.13	-3.55	***

	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Migrated from the same prov as destination (%)	58.53	59.94	-1.41		56.02	60.81	-4.79	***	60.66	64.81	-4.15	*
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Willingness to stay in cities permanently (%)	66.29	69.4	-3.11	**	62.16	67.37	-5.21	***	61.1	70.04	-8.93	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		
Likely to move (%)	12.23	11.1	1.14		8.74	6.48	2.27	***	10.84	5.55	5.29	***
	(4153)	(2379)			(2695)	(3720)			(1618)	(3461)		

Note: * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The sample is restricted to migrants aged between 16 years and 65 years and did not provide missing data on age, gender, years of schooling, education attainment, marriage status, weekly hours worked, monthly earnings, hypothetical monthly earnings in rural origin, access to social insurances, employment status, years since the first migration, willingness to stay in cities, whether respondents are likely to move residential address in the next 12 months.. The sample size is in parenthesis.

a. excluding respondents who are not married.

b. excluding respondents who do not have children under 16 years.

c. no information in the 2010 sample.

d.no information in the 2010 sample and respondents who were not present in the interviews in the other waves.

e.no information in the 2008 sample and respondents who were not present in the interviews in the other waves.

f. excluding unemployed.

g. excluding non-paid family helpers and unemployed.

h. having access to all of unemployment insurance, a pension, injury insurance and medical insurance for salary workers, and having access to both of a pension and medical insurance for self-employed and non-paid family helpers.

i. excluding unemployed, self-employed and non-paid family helpers.

Source: 2009, 2011 and 2013 waves of the RUMIC migrant household survey.

Table 4.8 Test of coefficient equality between the new sample and non-attritors on the earnings regression -- Males

	Panel A 2009 wave			Panel B 2011 wave			Panel C 2013 wave		
	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff
Age	0.031***	0.024**	0.007	0.039***	0.017**	0.022*	0.052***	0.011	0.041***
	(0.009)	(0.011)	(0.014)	(0.010)	(0.008)	(0.013)	(0.011)	(0.008)	(0.014)
Square of age	-0.000***	-0.000***	-0.000	-0.001***	-0.000***	-0.000	-0.001***	-0.000***	-0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.029***	0.034***	-0.005	0.025***	0.028***	-0.002	0.018**	0.029***	-0.011
	(0.006)	(0.006)	(0.009)	(0.006)	(0.004)	(0.007)	(0.007)	(0.004)	(0.008)
Married	0.017	0.135***	-0.117*	-0.028	0.053*	-0.081	0.107**	0.108***	-0.001
	(0.039)	(0.049)	(0.062)	(0.041)	(0.028)	(0.050)	(0.054)	(0.030)	(0.062)
Divorced	-0.047	0.180	-0.227	-0.088	0.079	-0.167	0.031	0.248***	-0.217*
	(0.075)	(0.128)	(0.148)	(0.087)	(0.092)	(0.126)	(0.097)	(0.083)	(0.127)
Years since first migration	0.038***	0.014	0.024**	0.017**	0.027***	-0.010	0.021***	0.016***	0.005
	(0.008)	(0.009)	(0.012)	(0.007)	(0.006)	(0.009)	(0.008)	(0.006)	(0.010)
Square of years since first migration	-0.001***	-0.001	-0.000	-0.000	-0.001***	0.000	-0.001**	-0.000**	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self-employment	0.048	0.064*	-0.015	0.078*	0.034	0.044	-0.072	0.043*	-0.115
	(0.042)	(0.038)	(0.057)	(0.042)	(0.025)	(0.049)	(0.065)	(0.025)	(0.070)
City dummies	Yes	Yes		Yes	Yes		Yes	Yes	
Observations	2537	1439		1617	2301		1100	2411	
F Test1			3.29***			1.74*			3.25***
F Test2			2.30***			3.14***			2.60***

Note: Standard errors are clustered at the household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The destination city fixed effects and constant are included in the regressions. Test 1 tests the equality of coefficients of individual-level characteristics listed in the table, and Test 2 tests the equality of all the coefficients including constant terms and destination city fixed effects. The new sample is re-weighted to make its distribution of sample size across city-year combinations the same as the non-attritors.

Source: 2009, 2011 and 2013 waves of the RUMIC migrant household survey.

Table 4.9 Test of coefficient equality between the new sample and non-attritors on the earnings regression -- Females

	Panel A 2009 wave			Panel B 2011 wave			Panel C 2013 wave		
	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff
Age	0.040***	0.016	0.025	0.017	0.028***	-0.011	-0.002	0.037***	-0.039
	(0.010)	(0.014)	(0.017)	(0.011)	(0.010)	(0.015)	(0.032)	(0.010)	(0.034)
Square of age	-0.001***	-0.000*	-0.000	-0.000**	-0.000***	0.000	-0.000	-0.001***	0.001
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.022***	0.025***	-0.003	0.015***	0.024***	-0.009	0.019**	0.027***	-0.007
	(0.005)	(0.006)	(0.008)	(0.005)	(0.004)	(0.007)	(0.008)	(0.004)	(0.009)
Married	-0.092**	-0.077	-0.015	-0.086*	-0.117***	0.031	0.194	-0.080**	0.274
	(0.038)	(0.049)	(0.062)	(0.049)	(0.037)	(0.062)	(0.171)	(0.036)	(0.175)
Divorced	-0.007	0.014	-0.021	0.094	0.174*	-0.080	0.427	-0.038	0.465
	(0.075)	(0.116)	(0.138)	(0.064)	(0.104)	(0.123)	(0.376)	(0.081)	(0.384)
Years since first migration	0.021***	0.004	0.017	0.014	0.012**	0.002	0.017	0.011	0.006
	(0.006)	(0.011)	(0.013)	(0.009)	(0.006)	(0.011)	(0.012)	(0.007)	(0.014)
Square of years since first migration	-0.001*	0.000	-0.001	-0.000	-0.000*	0.000	-0.001	-0.000	-0.000
	(0.000)	(0.000)	(0.001)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self-employment	0.284***	0.326***	-0.042	0.219***	0.185***	0.033	0.107*	0.169***	-0.062
	(0.034)	(0.038)	(0.051)	(0.043)	(0.028)	(0.051)	(0.056)	(0.028)	(0.063)
City dummies	Yes	Yes		Yes	Yes		Yes	Yes	
Observations	1758	948		1266	1671		871	1782	
F Test1			2.66***			0.82			1.16
F Test2			2.21***			1.87***			1.97***

Note: Standard errors are clustered at the household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The destination city fixed effects and constant are included in the regressions. Test 1 tests the equality of coefficients of individual-level characteristics listed in the table, and Test 2 tests the equality of all the coefficients including constant terms and destination city fixed effects. The new sample is re-weighted to make its distribution of sample size across city-year combinations the same as the non-attritors.

Source: 2009, 2011 and 2013 waves of the RUMIC migrant household survey.

Appendix

Appendix A: Figures and tables

Table 4.A.1 Mean comparison between attritors and non-attritors across waves

	Panel A: 2008				Panel B: 2009				Panel C: 2010				Panel D: 2011				Panel E: 2012			
	Non-attritor	Attritor	Diff		Non-attritor	Attritor	Diff		Non-attritor	Attritor	Diff		Non-attritor	Attritor	Diff		Non-attritor	Attritor	Diff	
Panel 1 Demographic and household structure variables																				
Age	32.49	30.52	1.97	***	32.47	30.98	1.50	***	32.30	30.56	1.74	***	33.19	29.52	3.66	***	33.29	31.64	1.65	***
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Male (%)	57.61	56.95	0.66		56.55	58.88	-2.34		54.77	60.68	-5.91	***	55.14	56.31	-1.17		56.46	57.42	-0.96	
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Years of schooling	9.26	9.30	-0.04		9.59	9.42	0.17	*	9.58	9.51	0.07		9.29	9.62	-0.33	***	9.49	9.61	-0.12	
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Married (%)	71.13	57.56	13.57	***	65.52	55.90	9.62	***	62.42	50.36	12.06	***	65.36	47.22	18.14	***	68.19	58.97	9.21	***
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Divorced (%)	1.03	0.91	0.12		1.24	1.34	-0.10		1.09	1.49	-0.40		1.66	1.85	-0.19		0.98	1.01	-0.03	
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Presence of spouse in	83.94	73.81	10.13	***	75.37	69.95	5.42	***	80.05	67.65	12.39	***	78.78	65.24	13.54	***	82.01	70.65	11.35	***
the household (%)[a]	(1725)	(2337)			(1161)	(1331)			(857)	(912)			(985)	(561)			(628)	(644)		
Presence of child under	57.73	41.47	16.25	***	51.54	45.57	5.96	**	54.47	36.97	17.50	***	52.99	30.84	22.15	***	56.59	41.71	14.88	***
16 years in the household (%)[b]	(1171)	(1519)			(749)	(768)			(470)	(468)			(551)	(308)			(387)	(374)		
Panel 2 Health and psychological preferences																				
Height (cm)[c]	165.96	165.90	0.06		165.64	166.25	-0.61	***	165.29	167.77	-2.48	***	165.81	166.39	-0.58	**	166.06	166.15	-0.09	
	(2424)	(4057)			(1769)	(2374)			(267)	(323)			(1503)	(1187)			(919)	(1091)		

Good health or	82.47	86.11	-3.63	***	84.14	83.70	0.44						84.94	83.50	1.44		88.82	87.36	1.45	
better (%)^c	(2425)	(4060)			(1772)	(2381)							(1507)	(1188)			(921)	(1092)		
GHQ score	0.89	0.99	-0.10	**	0.92	0.92	0.00						1.05	1.37	-0.32	***	0.80	0.92	-0.12	
	(2117)	(3561)			(1442)	(1900)							(1168)	(968)			(716)	(905)		
Risk-loving^e					4.07	4.22	-0.15	*	3.87	3.98	-0.11		4.17	4.56	-0.39	***	4.45	4.44	0.01	
					(1442)	(1900)			(1070)	(1486)			(1171)	(970)			(715)	(908)		
Trust (%)^e					59.08	57.11	1.98		48.79	43.88	4.91	**	55.76	52.06	3.70	*	34.50	38.55	-4.05	*
					(1442)	(1900)			(1070)	(1486)			(1171)	(970)			(716)	(908)		
Panel 3 Economic performance and welfare																				
Weekly hours worked	64.52	60.15	4.37	***	60.88	60.12	0.75		64.04	58.86	5.18	***	63.38	61.59	1.79	***	64.18	60.27	3.92	***
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Weekly hours worked^f	66.35	61.64	4.71	***	62.90	61.73	1.17	**	64.89	59.78	5.11	***	63.76	62.00	1.76	***	64.18	60.27	3.92	***
	(2358)	(3962)			(1715)	(2319)			(1355)	(1783)			(1498)	(1180)			(921)	(1092)		
Monthly earnings	1467.95	1399.61	68.34	***	1669.77	1662.97	6.81		1908.49	1864.16	44.33		2209.55	2271.19	-61.6		2223.61	2322.56	-99.0	*
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Monthly earnings^g	1614.41	1499.71	114.70	***	1747.69	1729.80	17.88		1959.88	1912.75	47.13		2234.76	2304.16	-69.4		2243.09	2341.86	-98.8	*
	(2205)	(3789)			(1693)	(2289)			(1337)	(1765)			(1490)	(1171)			(913)	(1083)		
Hypothetical monthly	607.69	671.47	-63.78	***	678.34	751.25	-72.9	***	872.71	926.09	-53.4	**	969.80	1114.05	-144	***	1139	1283	-144	***
earnings in rural origin	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Earnings gain	860.26	728.14	132.12	***	991.43	911.72	79.71	**	1035.78	938.07	97.71	**	1239.74	1157.14	82.60	*	1084.63	1039.33	45.30	
from migration	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Access to social	8.78	8.72	0.06		10.50	8.99	1.51		10.63	11.87	-1.24		17.39	13.13	4.25	***	26.17	21.70	4.46	**
insurances^h	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Panel 4 Work-related variables																				
Self-employment (%)	31.09	14.83	16.27	***	25.79	18.31	7.48	***	32.05	13.42	18.63	***	29.20	14.23	14.97	***	30.18	13.83	16.36	***
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Non-paid family	6.31	4.26	2.05	***	1.24	1.26	-0.02		1.31	0.99	0.32		0.53	0.76	-0.23		0.87	0.82	0.04	
helper (%)	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Unemployed (%)	2.76	2.41	0.35		3.22	2.60	0.61		1.31	1.55	-0.24		0.60	0.67	-0.08		0.00	0.00	0.00	

	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Less than 50 workers in the	69.28	58.43	10.85	***	61.14	55.19	5.96	***	67.58	57.00	10.58	***	64.79	50.94	13.85	***	60.94	50.55	10.39	***
workplacef (%)	(2357)	(3960)			(1714)	(2314)			(1354)	(1779)			(1491)	(1176)			(919)	(1090)		
Less than 50 workers	50.41	48.79	1.62		46.28	44.59	1.68		51.45	50.00	1.45		50.91	42.99	7.92	***	44.39	42.37	2.03	
in the workplacei (%)	(1450)	(3185)			(1236)	(1850)			(896)	(1518)			(1045)	(998)			(633)	(930)		
Panel 5 Other migration-related variables																				
Years since	8.54	7.26	1.28	***	8.51	7.49	1.03	***	7.61	6.09	1.52	***	8.96	6.72	2.25	***	9.57	8.37	1.20	***
first migration	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Migrated from the same	60.00	55.17	4.83	***	57.84	48.80	9.04	***	54.33	43.51	10.82	***	51.16	37.37	13.79	***	51.68	36.26	15.42	***
prov as destination (%)	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Willingness to stay in	65.40	56.80	8.60	***	68.57	57.75	10.82	***	65.55	50.97	14.58	***	64.30	43.86	20.44	***	56.68	52.56	4.11	*
cities permanently (%)	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		
Likely to move (%)	11.63	19.48	-7.85	***	8.97	15.41	-6.44	***	11.14	16.18	-5.04	***	7.23	15.24	-8.00	***	3.58	9.62	-6.03	***
	(2425)	(4060)			(1772)	(2381)			(1373)	(1811)			(1507)	(1188)			(921)	(1092)		

Note: * significant at 10% level; ** significant at 5% level; *** significant at 1% level. This table compares respondents who immediately attrited with those who did not for each wave of the survey. From Panels A to E only respondents who first entered in the survey at the particular wave are included in the analysis. The sample is restricted to migrants with absorbing attrition aged between 16 years and 65 years. The characteristics compared are those in the initial waves when respondents first entered the survey. The sample size is in parenthesis.

a. excluding respondents who are not married.

b. excluding respondents who do not have children under 16 years.

c. no information in the 2010 sample.

d.no information in the 2010 sample and respondents who were not present in the interviews in the other waves.

e.no information in the 2008 sample and respondents who were not present in the interviews in the other waves.

f. excluding unemployed.

g. excluding non-paid family helpers and unemployed.

h. having access to all of unemployment insurance, a pension, injury insurance and medical insurance for salary workers, and having access to both of a pension and medical insurance for self-employed and non-paid family helpers.

i. excluding unemployed, self-employed and non-paid family helpers.

Source: 2008-2013 waves of the RUMIC migrant household survey.

Table 4.A.2 Mean Comparison between New Sample and Old Sample (Non-attritor) in 2010 and 2012 Waves

	Panel A: 2010				Panel B: 2012			
	New	Old	Diff		New	Old	Diff	
Panel 1 Demographic and household structure variables								
Age	31.82	33.84	-2.01	***	32.87	35.35	-2.48	***
	(3184)	(3066)			(2013)	(4211)		
Male (%)	57.42	57.99	-0.57		55.9	55.33	0.57	
	(3184)	(3066)			(2013)	(4211)		
Years of schooling	9.56	9.46	0.09		9.28	9.34	-0.06	
	(3184)	(3066)			(2013)	(4211)		
Married (%)	57.51	72.7	-15.19	***	64.05	76.04	-11.99	***
	(3184)	3066			(2013)	(4211)		
Divorced (%)	2.54	1.4	1.14		1.49	1.66	-0.17	
	(3184)	(3066)			(2013)	(4211)		
Presence of spouse in the household (%)[a]	77.04	77.07	-0.03		79.98	83.73	-3.75	**
	(1769)	(2229)			(1272)	(3202)		
Presence of child under 16 years in the household (%)[b]	48.21	59.2	-10.98	***	54.01	66.56	-12.55	***
	(938)	(1343)			(761)	(1839)		
Panel 2 Health and psychological preferences								
Height (cm)[c]	166.03	166.39	-0.36		165.65	165.88	-0.23	
	(590)	(1128)			(2010)	(4202)		
Good health or better (%)[c]					90.4	84.71	5.69	***
					(2013)	(4211)		
Mental health problems - GHQ score[d]					0.67	0.81	-0.14	***
					(1621)	(3091)		
Risk-loving[e]	4.18	3.75	0.42	***	4.3	4.11	0.19	
	(2556)	(2471)			(1623)	(3090)		
Trust (%)[e]	50.82	53.91	-3.08		37.51	53.09	-15.58	***
	(2556)	(2471)			(1624)	(3093)		
Panel 3 Economic performances and welfare								
Weekly hours worked	62.61	61.41	1.19	*	63.09	60.35	2.74	***
	(3184)	(3066)			(2013)	(4211)		
Weekly hours worked[f]	63.47	64.64	-1.17	*	63.09	63.92	-0.82	
	(3138)	(2913)			(2013)	(3976)		
Monthly earnings	1789.16	1821.23	-32.07		2142.89	2184.84	-41.95	
	(3184)	(3066)			(2013)	(4211)		
Monthly earnings[g]	1833.23	1949.68	-116.45	***	2163.26	2358.46	-195.2	***
	(3102)	(2864)			(1996)	(3901)		
Hypothetical monthly earnings in rural origin	794.64	845.45	-50.81	**	1134.77	1128.38	6.39	
	(3184)	(3066)			(2013)	(4211)		
Earnings gain from migration	994.51	975.78	18.73		1008.12	1056.46	-48.34	
	(3184)	(3066)			(2013)	(4211)		
Access to social insurances[h]	8.79	14.71	-5.92	***	17.46	20.45	-2.99	**
	(3184)	(3066)			(2013)	(4211)		
Panel 4 Work-related variables								
Self-employment (%)	25.69	32.91	-7.22	***	26.95	36.67	-9.72	***
	(3184)	(3066)			(2013)	(4211)		
Non-paid family helper (%)	1.04	1.6	-0.56	*	0.94	1.78	-0.84	**

	(3184)	(3066)			(2013)	(4211)		
Unemployed (%)	1.36	4.99	-3.63	***	0	5.58	-5.58	***
	(3184)	(3066)			(2013)	(4211)		
Less than 50 workers in the workplace[f] (%)	67.25	66.91	0.35		64.61	69.06	-4.45	**
	(3133)	(2907)			(2009)	(3969)		
Less than 50 workers in the workplace[i] (%)	55.54	49.27	6.27	***	51.32	48.22	3.1	
	(2414)	(1855)			(1563)	(2356)		
Panel 5 Other migration-related variables								
Years since first migration	7.17	10.17	-2.99	***	9.39	11.58	-2.19	***
	(3184)	(3066)			(2013)	(4211)		
Migrated from the same prov as destination (%)	60.38	62.04	-1.66		55.48	61.32	-5.84	***
	(3184)	(3066)			(2013)	(4211)		
Willingness to stay in cities permanently (%)	63.48	69.11	-5.64	***	57.06	68.77	-11.71	***
	(3184)	(3066)			(2013)	(4211)		
Likely to move (%)	14.09	10.6	3.49	***	5.65	4.99	0.66	
	(3184)	(3066)			(2013)	(4211)		

Note: * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The sample is restricted to migrants aged between 16 years and 65 years and did not provide missing data on age, gender, years of schooling, education attainment, marriage status, weekly hours worked, monthly earnings, hypothetical monthly earnings in rural origin, access to social insurances, employment status, years since the first migration, willingness to stay in cities, whether respondents are likely to move residential address in the next 12 months.. The sample size is in parenthesis.

a. excluding respondents who are not married.

b. excluding respondents who do not have children under 16 years.

c. no information in the 2010 sample.

d.no information in the 2010 sample and respondents who were not present in the interviews in the other waves.

e.no information in the 2008 sample and respondents who were not present in the interviews in the other waves.

f. excluding unemployed.

g. excluding non-paid family helpers and unemployed.

h. having access to all of unemployment insurance, a pension, injury insurance and medical insurance for salary workers, and having access to both of a pension and medical insurance for self-employed and non-paid family helpers.

i. excluding unemployed, self-employed and non-paid family helpers.

Source: 2010 and 2012 waves of the RUMIC migrant household survey.

Table 4.A.3 Additional results on test of coefficient equality between the new sample and old sample (non-attritors) on the earnings regression -- Males

	Panel A 2010 wave			Panel B 2012 wave		
	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff
Age	0.031**	0.016*	0.015	0.052***	0.013	0.039***
	(0.013)	(0.009)	(0.016)	(0.011)	(0.008)	(0.014)
Square of age	-0.000**	-0.000***	-0.000	-0.001***	-0.000***	-0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.020***	0.032***	-0.011	0.021***	0.030***	-0.009
	(0.006)	(0.005)	(0.007)	(0.006)	(0.004)	(0.008)
Married	0.037	0.077**	-0.040	-0.031	0.126***	-0.157***
	(0.043)	(0.031)	(0.053)	(0.047)	(0.032)	(0.057)
Divorced	0.216*	0.042	0.174	-0.143	0.172**	-0.314**
	(0.114)	(0.086)	(0.143)	(0.094)	(0.081)	(0.124)
Years since first migration	0.030***	0.032***	-0.001	0.025***	0.029***	-0.005
	(0.007)	(0.006)	(0.009)	(0.007)	(0.006)	(0.009)
Square of years since first migration	-0.001***	-0.001***	-0.000	-0.000**	-0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self-employment	0.190***	0.136***	0.054	0.108**	0.092***	0.016
	(0.047)	(0.027)	(0.054)	(0.051)	(0.023)	(0.056)
City dummies	Yes	Yes		Yes	Yes	
Observations	2089	1893		1302	2573	
Test1			2.87***			2.08**
Test2			2.25***			2.85***

Note: Standard errors are clustered at the household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The destination city fixed effects and constant are included in the regressions. Test 1 tests the equality of coefficients of individual-level characteristics listed in the table, and Test 2 tests the equality of all the coefficients including constant terms and destination city fixed effects. The new sample is re-weighted to make its distribution of sample size across city-year combinations the same as the non-attritors.

Source: 2010 and 2012 waves of the RUMIC migrant household survey.

Table 4.A.4 Additional results on test of coefficient equality between the new sample and old sample (non-attritors) on the earnings regression --Females

	Panel A 2010 wave			Panel B 2012 wave		
	new sample	Non-attritor	Diff	new sample	Non-attritor	Diff
Age	0.008	0.038***	-0.030*	0.046***	0.024**	0.022
	(0.013)	(0.012)	(0.017)	(0.010)	(0.009)	(0.014)
Square of age	-0.000	-0.001***	0.000**	-0.001***	-0.000***	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Years of schooling	0.013**	0.025***	-0.011*	0.017***	0.029***	-0.012*
	(0.005)	(0.004)	(0.007)	(0.006)	(0.004)	(0.007)
Married	-0.051	-0.118***	0.067	-0.040	-0.031	-0.009
	(0.049)	(0.038)	(0.062)	(0.038)	(0.035)	(0.052)
Divorced	0.217	0.065	0.152	0.326***	0.010	0.316**
	(0.171)	(0.082)	(0.190)	(0.123)	(0.060)	(0.136)
Years since first migration	0.027***	0.008	0.020*	0.009	0.019***	-0.010
	(0.008)	(0.008)	(0.011)	(0.007)	(0.006)	(0.009)
Square of years since first migration	-0.001**	-0.000	-0.001	-0.000	-0.001***	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Self-employment	0.293***	0.358***	-0.065	0.217***	0.203***	0.014
	(0.058)	(0.029)	(0.065)	(0.048)	(0.024)	(0.054)
city dummies	Yes	Yes		Yes	Yes	
Observations	1435	1273		971	1888	
Test1			3.57***			1.16
Test2			2.93***			1.97***

Note: Standard errors are clustered at the household level. * significant at 10% level; ** significant at 5% level; *** significant at 1% level. The destination city fixed effects and constant are included in the regressions. Test 1 tests the equality of coefficients of individual-level characteristics listed in the table, and Test 2 tests the equality of all the coefficients including constant terms and destination city fixed effects. The new sample is re-weighted to make its distribution of sample size across city-year combinations the same as the non-attritors.

Source: 2010 and 2012 waves of the RUMIC migrant household survey.

Appendix B: Details about variable construction

All the individual-level variables are extracted from the migrant household survey of the RUMIC project. This Appendix provides definitions for variables which are not immediately obvious from the survey questions.

Presence of spouse in the household: This is based on relationship to the household head. The survey collects data on the household head and their spouse; so we can directly identify if a spouse is present. However, the survey does not collect data about spouses of other household members. I identify the spouses of other household members from the corresponding categories, according to their relationship with the household head. The corresponding categories for *children* and *grandchildren* are *children-in-law* and *grandchildren-in-law* respectively. The corresponding category for *sibling*, *uncle/aunt* and *nephew/niece* is *other relative*. The corresponding categories for *parent*, *parent-in-law* and *grandparent* are themselves. I identified couples based on gender, marital status, years of marriage and number of children. If a married person in the corresponding category in the same household is of the opposite sex, was married in the same year and has the same number of children (if it is their first marriage) as the individual, then this person is identified as the spouse of the individual. These pieces of information uniquely identified the spouses, except for 34 individuals. For these exceptions, I checked whether these individuals have the same child as other household members. This identified the spouses for 18 individuals. For the remaining 16 individuals, I found only one household member in the corresponding category who was aged within five years, and I took them to be the spouses.

I acknowledge that this identification approach introduces some measurement error. The approach may overestimate the presence of a spouse for household members who are not household heads or spouses, because being of the opposite sex, marrying in the same year and having the same number of children is a necessary rather than sufficient condition of being spouse,. However, *children-in-law*, *grandchildren-in-law*, *grandparent* and *other relative* only account for a small proportion of the total sample (varies from 0.16% to 0.72% of the sample used in mean comparisons and attrition regressions). Thus, very few spouses were identified using the approach described above and the measurement error should be small. To thoroughly address this concern, I use the sample of household head and spouse to replicate all the mean comparison and attrition regression exercises, and the results are very similar.

Presence of child under 16 years in the household: The survey collects information on all children under 16 years belonging to household members. If children live in the migrant household, I use their relationship to the household head and the member ID of their parents to identify whether an individual and his/her spouse has a child under 16 years in the household.

Good health or better: This is from the question "Your current state of health (compared to people your own age): (1) excellent (2) good (3) average (4) poor (5) very poor". If the respondent chooses "excellent" or "good", then this variable is 1; otherwise it is 0.

Poor health or worse: This is from the question "Your current state of health (compared to people your own age): (1) excellent (2) good (3) average (4) poor (5) very poor". If the respondent chooses "poor" or "very poor", then this variable is 1; otherwise it is 0.

Mental health problems - GHQ score: This is constructed from the General Health Questionnaire 12 which has 12 questions. The answer to each question ranges from 1 to 4. The GHQ score counts how many questions were answered 3 or 4. A higher GHQ score indicates that the respondent has more mental health problems.

Risk-loving: This is from the question "Generally, some people prefer to take risk, while others try to avoid any risk. If you were to rank yourself from low to high (as shown by the following chart) with 0 being 'never take risk' and 10 being 'most likely to take risk', which level do you belong to?"

Trust: This is from the question "Generally, do you think that most people are trustworthy? Or do you think you had better be careful when dealing with other people? (1) most people are trustworthy (2) the more careful, the better (3) do not know". If the respondent chooses "most people are trustworthy", then this variable is 1; otherwise it is 0.

Hypothetical monthly earnings in rural origin: The is from the question "If you were still in your home village, how much do you estimate you could earn per month? (Yuan/Month)"

Earnings gain from migration: This is defined as the difference between monthly earnings in cities and hypothetical monthly earnings in rural origin.

Willingness to stay in cities permanently: This is from the question "If policy allowed, how long would you like to stay in the city? (1) 1 year (2) 1-3 years (3) more than 3 years (4) permanently (5) not sure". If the respondent chooses "permanently", then this variable is 1; otherwise it is 0.

Likely to move: This is from the question "How likely is it that you will move out in the next 12 months? (1) very likely (2) likely (3) not sure (4) unlikely (5) very unlikely". If the respondent chooses "very likely" or "likely", then this variable is 1; otherwise it is 0.

Chapter 5 Conclusion

Over the past two decades, China has experienced epic-scale rural-to-urban migration. In 2013, 166 million rural people worked in cities for more than half a year (NBS, 2014). This massive migration is a strong catalyst for stimulating the growth of the Chinese economy. This thesis examines three different but loosely linked aspects of rural-to-urban migration in China. The main findings are summarised as follows.

5.1 Social networks and mental health problems

Mental health is one of the key indicators of individual well-being. Rural migrants are vulnerable to mental health problems (Wong et al., 2008; Qiu et al., 2011; Mou et al., 2011; He and Wong, 2013), so Chapter 2 studies the impact of social networks on the mental health of rural migrants.

I study this issue using data from the migrant household survey in the RUMIC project. I adopt the instrumental variable approach, in order to handle the endogeneity bias between social networks and mental health problems. Specifically, I use past rainfall in home county and distance between home village and its closest traffic hub to instrument the size of the urban networks of rural migrants.

The IV estimates and fixed effect IV estimates suggest that expanding networks helps reduce the mental health problems of rural migrants who live in cities. The results are robust, regardless of whether the instrumental variables are used jointly or individually in the estimations.

Moreover, I find the effect of social networks is heterogeneous across different sub-samples. In particular, social networks have a greater benefit for migrants who have smaller networks or without access to social welfare. Females also benefit more from their social networks than males.

5.2 Contact and willingness to interact with migrants

Social segregation between natives and migrants exists in many countries. One of the key factors inducing social segregation is the negative attitudes of locals towards migrants. Therefore, changing these attitudes could reduce social segregation. Chapter 3 takes the ongoing migration in China as an example to investigate whether interpersonal contact helps improve urban locals' willingness to interact with migrants.

The data used in this chapter are drawn from the 2005 China General Social Survey. This survey provides detailed information on the willingness of locals to interact with migrants, including willingness to have non-intimate interactions (i.e., working with migrants and living in the same community) and willingness to have intimate interactions (i.e., having migrant neighbours, inviting migrant guests home and having relatives or children marry or be in a relationship with migrants). This detailed information allows us to examine the contact effect on different dimensions of attitudes.

I use the heteroskedasticity identification approach proposed by Lewbel (2012) to alleviate the endogeneity bias. The results suggest that contact improves urban locals' willingness to have non-intimate interactions, but has no significant effect on intimate interactions. As to policy-making, this finding indicates that the government should promote contact between migrants and locals to reduce segregation, but will have to consider other measures to reduce segregation in intimate interactions.

5.3 Attrition in the migrant household survey of the RUMIC project

Longitudinal surveys on migrants are rare, because of the itinerant nature of migrants. Not surprising then that the migrant household survey of the RUMIC project has received much attention from the academic community. However, the survey has an attrition problem. Between the first two waves, only 36% of households from the baseline wave remained in the survey. It is important for future research and data collection to understand the nature and consequences of this high attrition rate.

Chapter 4 investigates three attrition-related questions. First, what are the predictors of attrition? Second, does attrition bias estimates? Third, is the sample of the non-attritors representative of the migrant population at the time of the follow-up survey? I take the earnings equation as an example to illustrate the last two questions.

Specifically, I find that the non-attritors tend to be more socio-economically advantaged, have larger income gains from migration, are more driven to stay in cities and are more likely to be self-employed, than attritors. Further, I find that attrition bias possibly exists and that the non-attritor sample is unrepresentative of the general migrant population at the time of follow-up surveys. However, the impact of attrition bias and sample (un)representativeness on regression coefficients are case-dependent. In some cases, attrition bias and sample (un)representativeness have only limited impact on the regression coefficients of the individual-level variables which are most relevant to research and policy interest.

Given this, I recommend that researchers assess the bias according to their own cases. If the evidence of attrition bias is found, the baseline wave and the random refreshment of the follow-up waves could constitute a sample which provides estimates without attrition bias.

5.4 Future research

The current findings provoke several potential directions for future study. First, Chapter 2 describes the protective effect of social networks on mental health, but little is known about the channel of this effect. So identifying the channels through which social networks operate deserves future attention. Another potential direction to gauge the effect of social networks on mental health is to compare the associations between these two variables across three different surveys: the migrant household survey, the urban household survey and the rural household survey in the RUMIC project. This comparison enable us to understand the effect of social networks for the general Chinese population, and also help us to know how different the migrants are from the other two populations.

Second, Chapter 3 only gives one side of the story regarding the contact effect on attitudes. Certainly, urban locals' attitudes towards migrants are important, but so too are migrants' attitudes towards urban locals. Further research on the attitudes of migrants will be important for fully understanding the contact effect and how to reduce social segregation. I would like to use the migration household survey of the RUMIC project to explore this issue in future work.

Bibliography

Abraham, K. G.; Maitland, A.; and Bianchi, S. M., 2006. Nonresponse in the American time use survey who is missing from the data and how much does it matter? Public Opinion Quarterly, 70, 5 (2006), 676–703.

Akay, A.; Bargain, O.; and Zimmermann, K. F., 2012. Relative concerns of rural-to-urban migrants in China. Journal of Economic Behavior & Organization, 81, 2 (2012), 421 – 441.

Akay, A.; Giulietti, C.; Robalino, J.; and Zimmermann, K., 2013. Remittances and well-being among rural-to-urban migrants in China. Review of Economics of the Household, 12, 3 (2013), 1–30.

Akgüç, M.; Giulietti, C.; and Zimmermann, K. F., 2014. The RUMIC longitudinal survey: Fostering research on labor markets in China. IZA Journal of Labor & Development, 3, 1 (2014), 1–14.

Alderman, H.; Behrman, J. R.; Kohler, H.-P.; Maluccio, J. A.; and Watkins, S. C., 2001. Attrition in longitudinal household survey data. Demographic research, 5, 4 (2001), 79–124.

Allport, G. W., 1954. The nature of prejudice. Cambridge, Mass., Addison-Wesley Pub. Co.

Bao, S.; Bodvarsson, r. B.; Hou, J. W.; and Zhao, Y., 2007. Interprovincial migration in China: The effects of investment and migrant networks. IZA Discussion Papers 2924, Institute for the Study of Labor (IZA).

Bartel, A. and Taubman, P., 1979. Health and labor market success: The role of various diseases. The Review of Economics and Statistics, 61, 1 (1979), 1–8.

Bartel, A. and Taubman, P., 1986. Some economic and demographic consequences of mental illness. Journal of Labor Economics, 4, 2 (1986), 243–256.

Becketti, S.; Gould, W.; Lillard, L.; and Welch, F., 1988. The panel study of income dynamics after fourteen years: An evaluation. Journal of Labor Economics, 6, 4 (1988), 472–492.

Bevelander, P. and Otterbeck, J., 2010. Young people's attitudes towards Muslims in Sweden. Ethnic and Racial Studies, 33, 3 (2010), 404–425.

Bhugra, D., 2004. Migration and mental health. Acta Psychiatrica Scandinavica, 109, 4 (2004), 243–258.

Billor, N.; Hadi, A. S.; and Velleman, P. F., 2000. Bacon: blocked adaptive computationally efficient outlier nominators. Computational Statistics & Data Analysis, 34, 3 (2000), 279 – 298.

Bilsborrow, R. E.; Oberai, A. S.; and Standing, G., 1984. Migration surveys in low income countries: guidelines for survey and questionnaire design. International Labour Organisation.

Bjorklund, A., 1985. Unemployment and mental health: Some evidence from panel data. The Journal of Human Resources, 20, 4 (1985), . 469–483.

Brand, E. F.; Lakey, B.; and Berman, S., 1995. A preventive, psychoeducational approach to increase perceived social support. American journal of community psychology, 23, 1 (1995), 117–135.

Breusch, T. S. and Pagan, A. R., 1979. A simple test for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, (1979), 1287–1294.

Cai, F.; Giles, J.; and Meng, X., 2006. How well do children insure parents against low retirement income? An analysis using survey data from urban China. Journal of Public Economics, 90, 12 (2006), 2229–2255.

Cai, F. and Wang, D., 1999. The sustainability of China's economic growth and labour contribution. Jingjiyanjiu (in Chinese), 10, 1999, 62–68.

Cai, F. and Wang, D., 2008. Impacts of internal migration on economic growth and urban development in China. In Migration and Development Within and Across Borders: Research and Policy Perspectives on Internal and International Migration (Eds. J. DeWind and J. Holdaway). IOM International Organization for Migration and The Social Science Research Council, New York.

Cai, J.; Xiong, C.; and Gao, H., 2013. Uncoordinated development and its causes of the population urbanisation and spatial urbanisation in China. Jingjixue dongtai (in Chinese), 6 , 2013, 15–22.

Chan, K.W., 2013. China: Internal migration. The Encyclopaedia of Global Human Migration. Blackwell Publishing Ltd.

Chen, J., 2011. Internal migration and health: Re-examining the healthy migrant phenomenon in China. Social Science & Medicine, 72, 8 (2011), 1294 – 1301.

Chen, Y. and Feng, S., 2013. Access to public schools and the education of migrant children in China. China Economic Review, 26 (2013), 75–88.

China Centre for Economic Research, 1998a. An analysis for the migratory labour supply shortage: The case of Zhongshan at Guangdong province. Gaige (in Chinese), , 5 (1998), 64–73.

China Centre for Economic Research, 1998b. An integrated rural-urban labour market: The case of Mianyang at Sichuan province. Gaige (in Chinese), 5 (1998), 74–83.

Citrin, J.; Green, D. P.; Muste, C.; and Wong, C., 1997. Public opinion toward immigration reform: The role of economic motivations. The Journal of Politics, 59, 3 (1997), . 858–881.

Clark, A. E. and Oswald, A. J., 1994. Unhappiness and unemployment. The Economic Journal, 104, 424 (1994), 648–659.

Cohen, S., 2004. Social relationships and health. American psychologist, 59, 8 (2004), 676.

Cohen, S. and Janicki-Deverts, D., 2009. Can we improve our physical health by altering our social networks? Perspectives on Psychological Science, 4, 4 (2009), 375–378.

Cohen, S. and Wills, T. A., 1985. Stress, social support, and the buffering hypothesis. Psychological Bulletin, 98 (1985), 310–357.

Cornaglia, F.; Crivellaro, E.; andMcNally, S., 2012. Mental health and education decisions. CEE Discussion Papers e0136, Centre for the Economics of Education, LSE.

Démurger, S.; Gurgand, M.; Li, S.; and Yue, X., 2009. Migrants as second-class workers in urban China? A decomposition analysis. Journal of Comparative Economics, 37, 4 (2009), 610–628.

Deng, Q. and Li, S., 2010. Wage structures and inequality among local and migrant workers in urban China. In The Great Migration: Rural-Urban Migration in China and Indonesia (Eds. X. Meng; C. Manning; L. Shi; and T. N. Effendi). Edward Elgar, 74-92.

Doi, Y. and Minowa, M., 2003. Factor structure of the 12-item general health questionnaire in the japanese general adult population. Psychiatry and Clinical Neurosciences, 57, 4 (2003), 379–383.

Dovidio, J. F.; Eller, A.; and Hewstone, M., 2011. Improving intergroup relations through direct, extended and other forms of indirect contact. Group Processes & Intergroup Relations, 14, 2 (2011), 147–160.

Dunbar, R. I., 1993. Coevolution of neocortical size, group size and language in humans. Behavioral and brain sciences, 16, 4 (1993), 681–694.

Dustmann, C. and Preston, I. P., 2007. Racial and economic factors in attitudes to immigration. The BE Journal of Economic Analysis & Policy, 7, 1 (2007).

Emran, M. S. and Hou, Z., 2013. Access to markets and rural poverty: Evidence from household consumption in China. Review of Economics and Statistics, 95, 2 (2013), 682-697.

Ertel, K. A.; Glymour, M. M.; and Berkman, L. F., 2009. Social networks and health: A life course perspective integrating observational and experimental evidence. Journal of Social and Personal Relationships, 26, 1 (2009), 73–92.

Ettner, S. L.; Frank, R. G.; and Kessler, R. C., 1997. The impact of psychiatric disorders on labor market outcomes. Industrial and Labor Relations Review, 51, 1 (1997), 64-81.

Facchini, G. and Mayda, A. M., 2008. From individual attitudes towards migrants to migration policy outcomes: Theory and evidence. Economic Policy, 23, 56 (2008), 651–713.

Facchini, G. and Mayda, A. M., 2009. Does the welfare state affect individual attitudes toward immigrants? Evidence across countries. The Review of Economics and Statistics, 91, 2 (2009), 295–314.

Facchini, G. and Mayda, A. M., 2012. Individual attitudes towards skilled migration: An empirical analysis across countries. The World Economy, 35, 2 (2012), 183–196.

Facchini, G.; Mayda, A. M.; and Mendola, M., 2013. What drives individual attitude towards immigration in South Africa? Review of International Economics, 21, 2 (2013), 326-341.

Falaris, E. M., 2003. The effect of survey attrition in longitudinal surveys: Evidence from Peru, Côte d'ivoire and Vietnam. Journal of Development Economics, 70, 1 (2003), 133–157.

Fitzgerald, J.; Gottschalk, P.; and Moffitt, R., 1998. An analysis of sample attrition in panel data: The Michigan panel study of income dynamics. The Journal of Human Resources, 33, 2 (1998), 251–299.

Frasure-Smith, N.; Lesperance, F.; Prince, R. H.; Verrier, P.; Garber, R. A.; Juneau, M.; Wolfson, C.; and Bourassa, M. G., 1997. Randomised trial of home-based psychosocial nursing intervention for patients recovering from myocardial infarction. The Lancet, 350, 9076 (1997), 473 – 479.

Frijters, P., Lee, L. and Meng, X., 2010. Jobs, working hours and remuneration packages for migrants and urban workers. In The Great Migration: Rural-Urban Migration in China and Indonesia (Eds. X. Meng; C. Manning; L. Shi; and T. N. Effendi). Edward Elgar, 24-73.

Frijters, P.; Johnston, D. W.; and Meng, X., 2009. The mental health cost of long working hours: the case of rural Chinese migrants. Unpublished Manuscript.

Frijters, P.; Johnston, D. W.; and Shields, M. A., 2010. Mental health and labour market participation: Evidence from iv panel data models. IZA Discussion Papers, No. 4883.

Frijters, P.; Kong, T.; and Meng, X., 2011. Migrant entrepreneurs and credit constraints under labour market discrimination. IZA Discussion paper series.

Gardner, J. and Oswald, A. J., 2006. Do divorcing couples become happier by breaking up? Journal of the Royal Statistical Society: Series A (Statistics in Society), 169, 2 (2006), 319–336.

Gardner, J. and Oswald, A. J., 2007. Money and mental wellbeing: A longitudinal study of medium-sized lottery wins. Journal of Health Economics, 26, 1 (2007), 49–60.

Giles, J. andMu, R., 2007. Elderly parent health and the migration decisions of adult children: Evidence from rural China. Demography, 44, 2 (2007), 265–288.

Giles, J. and Yoo, K., 2007. Precautionary behavior, migrant networks, and household consumption decisions: an empirical analysis using household panel data from rural China. The Review of Economics and Statistics, 89, 3 (2007), 534– 551.

Giulietti, C.; Wahba, J.; and Zenou, Y., 2014. Strong versus weak ties in migration. IZA Discussion paper series.

Gong, X.; Kong, S.; Li, S.; and Meng, X., 2008. Rural-urban migrants: A driving force for growth. In China's Dilemma: Economic Growth, the Environment and Climate Change (Eds. L. Song and W.T. Woo), ANU Epress, 110–152.

Goodwin, P. J.; Leszcz, M.; Ennis, M.; Koopmans, J.; Vincent, L.; Guther, H.; Drysdale, E.; Hundleby, M.; Chochinov, H. M.; Navarro, M.; et al., 2001. The effect of group psychosocial suort on survival in metastatic breast cancer. New England Journal of Medicine, 345, 24 (2001), 1719–1726.

Graetz, B., 1991. Multidimensional properties of the general health questionnaire. Social Psychiatry and Psychiatric Epidemiology, 26 (1991), 132–138.

Hainmueller, J. and Hiscox, M. J., 2007. Educated preferences: Explaining attitudes toward immigration in Europe. International Organization, 61, 02 (2007), 399–442.

Hansen, L. P., 1982. Large sample properties of generalized method of moments estimators. Econometrica: Journal of the Econometric Society, (1982), 1029–1054.

Harris, T.; Brown, G. W.; and Robinson, R., 1999. Befriending as an intervention for chronic depression among women in an inner city: Randomised controlled trial. The British Journal of Psychiatry, 174, 3 (1999), 219–224.

He, X. and Wong, D. F. K., 2013. A comparison of female migrant workers' mental health in four cities in China. International Journal of Social Psychiatry, 59, 2 (2013), 114–122.

Heckman, J. J. and Smith, J. A., 1995. Assessing the case for social experiments. The Journal of Economic Perspectives, 9, 2 (1995), 85–110.

Heller, K.; Thompson, M.; Trueba, P.; Hogg, J.; and Vlachos-Weber, I., 1991. Peer support telephone dyads for elderly women: Was this the wrong intervention? American Journal of Community Psychology, 19, 1 (1991), 53–74.

Herek, G. M. and Glunt, E. K., 1993. Interpersonal contact and heterosexuals' attitudes toward gay men: Results from a national survey. The Journal of Sex Research, 30, 3 (1993), . 239–244.

Hill, R. A. and Dunbar, R. I., 2003. Social network size in humans. Human nature, 14, 1 (2003), 53–72.

Hu, F.; Xu, Z.; and Chen, Y., 2011. Circular migration, or permanent stay? Evidence from China's rural–urban migration. China Economic Review, 22, 1 (2011), 64–74.

Huang, H., Lin, Y. and Yeh, C., 2009. Joint determinations of inequality and growth. Economics Letters, 103, 3 (2009), 163-166.

Kawachi, I. and Berkman, L., 2001. Social ties and mental health. Journal of Urban Health, 78 (2001), 458–467.

Kelly, I. R.; Dave, D. M.; Sindelar, J. L.; and Gallo, W. T., 2011. The impact of early occupational choice on health behaviors. Review of Economics of the Household, 12, 4 (2011), 1–34.

Kelly, I. R. and Markowitz, S., 2009. Incentives in obesity and health insurance. Inquiry, 46, 4 (2009), 418–432.

Kessler, R.; Heeringa, S.; Lakoma, M.; Petukhova, M.; Ru, A.; Schoenbaum, M.; Wang, P.; and Zaslavsky, A., 2008. Individual and societal effects of mental disorders on earnings in the united states: results from the national comorbidity survey replication. American Journal of Psychiatry, 165, 6 (2008), 703–711.

Kihc, C.; Rezaki, M.; Rezaki, B.; Kaplan, I.; Ozgen, G.; Sagduyu, A.; and Ozturk, M. O., 1997. General health questionnaire (GHQ12 & GHQ28): psychometric properties and factor structure of the scales in a Turkish primary care sample. Social Psychiatry and Psychiatric Epidemiology, 32, 6 (1997), 327–331.

Kleibergen, F. and Paap, R., 2006. Generalized reduced rank tests using the singular value decomposition. Journal of Econometrics, 133, 1 (2006), 97–126.

Knight, J. and Yueh, L., 2004. Job mobility of residents and migrants in urban China. Journal of Comparative Economics, 32, 4 (2004), 637–660.

Knight, J. and Yueh, L., 2009. Segmentation or competition in China's urban labour market? Cambridge journal of economics, 33, 1 (2009), 79–94.

Kong, T. S., 2010. Rural-urban migration in China: survey design and implementation. In The Great Migration: Rural-Urban Migration in China and Indonesia (Eds. X. Meng; C. Manning; L. Shi; and T. N. Effendi). Edward Elgar, 135-152.

Lee, L., 2012. Decomposing wage differentials between migrant workers and urban workers in urban China's labor markets. China Economic Review, 23, 2 (2012), 461– 470.

Lei, L. and Pals, H., 2011. Son preference in China: Why is it stronger in rural areas? Population Review, 50, 2 (2011), 27-46.

Liang, Y., 2011, A study on tracking rate in longitudinal survey (Zuizhong diaocha zhongde zuizhong chenggonglv yanjiu). Shehuixueyanjiu (in Chinese), 6, 2011,

Lennox, C., 2012. Racial integration, ethnic diversity, and prejudice: empirical evidence from a study of the British national party. Oxford Economic Papers, 64, 3 (2012), 395–416.

Lewbel, A., 2012. Using heteroscedasticity to identify and estimate mismeasured and endogenous regressor models. Journal of Business & Economic Statistics, 30, 1 (2012), 67–80.

Lewis, W. A., 1954. Economic development with unlimited supplies of labour. The Manchester school, 22, 2 (1954), 139–191.

Li, L.; Wang, H..; Ye, X.; Jiang, M.; Lou, Q.; and Hesketh, T., 2007. The mental health status of Chinese rural-urban migrant workers. Social Psychiatry and Psychiatric Epidemiology, 42 (2007), 716–722.

Li, X.; Stanton, B.; Fang, X.; Xiong, Q.; Yu, S.; Lin, D.; Hong, Y.; Zhang, L.; Chen, X.; and Wang, B., 2009. Mental health symptoms among rural-to-urban migrants in China: a comparison with their urban and rural counterparts. World Health and Population, 11, 1 (2009), 24 – 38.

Luo, C. and Yue, X., 2010. Rural-urban migration and poverty in China. In The Great Migration: Rural-Urban Migration in China and Indonesia (Eds. X. Meng; C. Manning; L. Shi; and T. N. Effendi). Edward Elgar, 117-134.

Mayda, A. M., 2006. Who is against immigration? a cross-country investigation of individual attitudes toward immigrants. The Review of Economics and Statistics, 88, 3 (2006), 510–530.

McKenzie, K.; Whitley, R.; and Weich, S., 2002. Social capital and mental health. The British Journal of Psychiatry, 181, 4 (2002), 280–283.

Meng, X., 2007. Wealth accumulation and distribution in urban China. Economic Development and Cultural Change, 55, 4 (2007), 761–791.

Meng, X., 2012. Labor market outcomes and reforms in China. The Journal of Economic Perspectives, 26, 4 (2012), 75–101.

Meng, X., 2014. China's labour market tensions and future urbanisation challenges. In Deepening Reform For China's Long-term Growth and Development (Eds. L. Song, R. Garnaut and F. Cai), ANU Epress, 379–405.

Meng, X. and Zhang, D. 2010. Labour market impact of large scale internal migration on Chinese urban 'native' workers. IZA Discussion Papers 5288, Institute for the Study of Labor (IZA).

Meng, X. and Zhang, D., 2013. The social impact of rural-urban migration on urban 'natives'. Unpublished Manuscript.

Meng, X. and Zhang, J., 2001. The two-tier labor market in urban China: Occupational segregation and wage differentials between urban residents and rural migrants in Shanghai. Journal of Comparative Economics, 29, 3 (2001), 485–504.

Ministry of Health of the People's Republic of China, 2010. Health Statistical Yearbook in China. Peking Union Medical College Press.

Mittelman, M. S.; Ferris, S. H.; Shulman, E.; Steinberg, G.; Ambinder, A.; Mackell, J. A.; and Cohen, J., 1995. A comprehensive suort program: effect on depression in spouse-caregivers of ad patients. The Gerontologist, 35, 6 (1995), 792–802.

Mou, J.; Cheng, J.; Griffiths, S. M.; Wong, S. Y.; Hillier, S.; and Zhang, D., 2011. Internal migration and depressive symptoms among migrant factory workers in Shenzhen, China. Journal of Community Psychology, 39, 2 (2011), 212–230.

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

Munshi, K., 2011. Chapter 23 -- labor and credit networks in developing economies. Vol. 1 of Handbook of Social Economics (Eds. J. Benhabib, A. Bisin and M. Jackson), North-Holland, 1223 -1254.

NBS, 2012. Tabulation on the 2010 population census of the people's republic of China (*Diliuci quanguo renkou pucha huizong shuju*). National Bureau of Statistics of China.

NBS, 2014. 2013 The national monitoring report of migrant workers in 2013 *(2013nian quanguo nongmingong jiance diaocha baogao*. National Bureau of Statistics of China.

Nielsen, I.; Nyland, C.; Smyth, R.; Zhang, M.; and Zhu, C. J., 2006. Effects of intergroup contact on attitudes of Chinese urban residents to migrant workers. Urban Studies, 43, 3 (2006), 475–490.

Nielsen, I. and Smyth, R., 2011. The contact hypothesis in urban China: The perspective of minority-status migrant workers. Journal of Urban Affairs, 33, 4 (2011), 469–481.

Olesen, J.; Gustavsson, A.; Svensson, M.; Wittchen, H.-U.; and Jönsson, B., 2012. The economic cost of brain disorders in Europe. European Journal of Neurology, 19, 1 (2012), 155–162.

Olsen, R. J., 2005. The problem of respondent attrition: Survey methodology is key. Monthly Labor Review, 128 (2005), 63-70.

O'Rourke, K. H. and Sinnott, R., 2006. The determinants of individual attitudes towards immigration. European Journal of Political Economy, 22, 4 (2006), 838–861.

Pettigrew, T. F., 1998. Intergroup contact theory. Annual Review of Psychology, 49, 1 (1998), 65–85.

Pettigrew, T. F.; Tro, L. R.; Wagner, U.; and Christ, O., 2011. Recent advances in intergroup contact theory. International Journal of Intercultural Relations, 35, 3 (2011), 271–280.

Pischke, S., 2007. Lecture notes on measurement error. The London School of Economics and Political Science. econ.lse.ac.uk/staff/spischke/ec524/Merr_new.pdf.

Qiu, P.; Caine, E.; Yang, Y.; Chen, Q.; Li, J.; and Ma, X., 2011. Depression and associated factors in internal migrant workers in China. Journal of Affective Disorders, 134, 1-3 (2011), 198 – 207.

Rivera, B.; Casal, B.; and Currais, L., 2015. Length of stay and mental health of the immigrant population in Spain: evidence of the healthy immigrant effect. Applied Economics, 47, 19 (2015), 1972-1982.

Roberts, K. D., 2001. The determinants of job choice by rural labor migrants in Shanghai. China Economic Review, 12, 1 (2001), 15–39.

Rose, R., 2000. How much does social capital add to individual health? Social Science & Medicine, 51, 9 (2000), 1421 – 1435.

Rozelle, S.; Taylor, J. E.; and DeBrauw, A., 1999. Migration, remittances, and agricultural productivity in China. American Economic Review, (1999), 287–291.

Sabia, J. J., 2007. Early adolescent sex and diminished school attachment: Selection or spillovers? Southern Economic Journal, 74, 1 (2007), 239–268.

Sa, A. L.; Trentham-Dietz, A.; Newcomb, P. A.; Hampton, J. M.; Moinpour, C. M.; and Remington, P. L., 2003. Social networks and quality of life among female long-term colorectal cancer survivors. Cancer, 98, 8 (2003), 1749–1758.

Scheve, K. F. and Slaughter, M. J., 2001. Labor market competition and individual preferences over immigration policy. Review of Economics and Statistics, 83, 1 (2001), 133–145.

Smith, K. P. and Christakis, N. A., 2008. Social networks and health. Annual Review of Sociology, 34 (2008), 405–429.

Smyth, R. and Mishra, V., 2014. Technological Change and Wages in China: Evidence from Matched Employer–Employee Data. Review of Development Economics, 18, 1 (2014), 123-138.

Staiger, D. and Stock, J. H., 1997. Instrumental variables regression with weak instruments. Econometrica, 65, 3 (1997), . 557–586.

Stein, R. M.; Post, S. S.; and Rinden, A. L., 2000. Reconciling context and contact effects on racial attitudes. Political Research Quarterly, 53, 2 (2000), 285–303.

Stillman, S.; McKenzie, D.; and Gibson, J., 2009. Migration and mental health: Evidence from a natural experiment. Journal of Health Economics, 28, 3 (2009), 677–687.

Taylor, J. E.; Rozelle, S.; and De Brauw, A., 2003. Migration and incomes in source communities: A new economics of migration perspective from China. Economic Development and Cultural Change, 52, 1 (2003), 75–101.

Tella, R. D.; New, J. H.-D.; and MacCulloch, R., 2010. Happiness adaptation to income and to status in an individual panel. Journal of Economic Behavior & Organization, 76, 3 (2010), 834 – 852.

Thoits, P. A., 2011. Mechanisms linking social ties and support to physical and mental health. Journal of Health and Social Behavior, 52, 2 (2011), 145–161.

Thomas, D., Frankenberg E., Smith, J.P., Lost but not forgotten: Attrition and follow-up in the Indonesia Family Life Survey. Journal of Human Resources, 36, 3 (2001), 556-592.

Thomas, D.; Witoelar, F.; Frankenberg, E.; Sikoki, B.; Strauss, J.; Sumantri, C.; and Suriastini, W., 2012. Cutting the costs of attrition: Results from the Indonesia Family Life Survey. Journal of Development Economics, 98, 1 (2012), 108–123.

Tropp, L. R., 2007. Perceived discrimination and interracial contact: Predicting interracial closeness among black and white Americans. Social Psychology Quarterly, 70, 1 (2007), 70–81.

Uhrig, S.C. Noah, 2008. The nature and causes of attrition in the British Household Panel Study. ISER Working Paper Series.

Velasquez, A.; Genoni, M.; Rubalcava, L.; Teruel, G.; and Thomas, D., 2011. Attrition in longitudinal surveys: Evidence from the Mexican family life survey. Unpublished Manuscript,.

Wang, L.; Li, C.; Ying, Q.; Cheng, X.; Wang, X.; Li, X.; Hu, L.; Liang, L.; Yu, L.; Huang, H.; et al., 2012. China's urban expansion from 1990 to 2010 determined with satellite remote sensing. Chinese Science Bulletin, 57, 22 (2012), 2802–2812.

Watson, N. and Wooden, M., 2004. Wave 2 survey methodology. HILDA Project Technical Paper Series.

Wong, D.; He, X.; Leung, G.; Lau, Y.; and Chang, Y., 2008. Mental health of migrant workers in China: prevalence and correlates. Social Psychiatry and Psychiatric Epidemiology, 43 (2008), 483–489.

Wooldridge, J. M., 2010. Econometric analysis of cross section and panel data, MIT press.

World Bank, 2009. From Poor Areas to Poor People: China's Evolving Poverty Reduction Agenda.

Wu, Z. and Schimmele, C. M., 2005. The healthy migrant effect on depression: variation over time? Canadian Studies in Population, 32, 2 (2005), 271–295.

Zabel, J. E., 1998. An analysis of attrition in the panel study of income dynamics and the survey of income and program participation with an application to a model of labor market behavior. Journal of Human Resources, 33, 2 (1998), 479–506.

Zhao, G., 2015. Can money 'buy' schooling achievement? Evidence from 19 Chinese cities. China Economic Review, 35, September (2015), 83-104.

Zhang, D.; Meng, X.; and Wang, D., 2010. The dynamic change in wage gap between urban residents and rural migrants in Chinese cities. PMMA Working Paper.

Zhang, K. H. and Song, S., 2003. Rural-urban migration and urbanization in China: Evidence from time-series and cross-section analyses. China Economic Review, 14, 4 (2003), 386–400.