University of Toronto
Department of Economics

Working Paper 693

Migration Costs, Sorting, and the Agricultural Productivity Gap

By Qingen Gai, Naijia Guo, Bingjing Li, Qinghua Shi and Xiaodong Zhu

April 17, 2021

Migration Costs, Sorting, and the Agricultural Productivity Gap

Qingen Gai Naijia Guo Bingjing Li Qinghua Shi
Xiaodong Zhu*

First version: January 2020† This version: April, 2021

Abstract

We use a unique panel dataset and a policy experiment as an instrument to estimate the impact of policy-induced migration cost reductions on rural-to-urban migration and the associated increase in labor earnings for migrant workers in China. Our estimation shows that there exist both large migration costs and a large underlying productivity difference between rural agricultural and urban non-agricultural sectors in China. More than half of the observed labor earnings gap between the two sectors can be attributed to the underlying productivity difference, and less than half of the gap can be attributed to sorting of workers. We also structurally estimate a general equilibrium Roy model and use it to quantify the effects of reducing migration costs on the observed sectoral productivity difference, migration, and aggregate productivity. If we implement a *hukou* policy reform by setting the *hukou* liberalization index in all regions of China to the level of the most liberal region, the observed agricultural productivity gap would decrease by more than 30%, the migrant share would increase by about 9%, and the aggregate productivity would increase by 1.1%. In contrast, in a partial equilibrium in which the underlying productivity difference does not change with migration cost, the *hukou* policy reform would reduce the observed agricultural productivity gap by only 9%, the migrant share would increase by more than 50%, and the aggregate productivity would increase by 6.8%.

JEL Classification: E24, J24, J61, O11, O15

Keywords: Migration cost; sorting; agricultural productivity gap; panel data; China; general equilibrium Roy model

[†] The first version of the paper was entitled “Selective Migration and Agricultural Productivity Gap: Evidence from China”.

[*] Gai: Shanghai University of Finance and Economics, email: gai.qingen@shufe.edu.cn; Guo: Chinese University of Hong Kong, email: guonaijia@cuhk.edu.hk; Li: National University of Singapore, email: ecslib@nus.edu.sg; Shi: Shanghai Jiaotong University, email: shq@sjtu.edu.cn; and Zhu: University of Toronto, email: xiaodong.zhu@utoronto.ca. We would like to thank Kory Kroft, David Lagakos, Rachel Ngai, Albert Park, Chris Pissarides, Diego Restuccia, Todd Schoellman, Aloysius Siow, Daniel Trefler, Yuanyuan Wan, Lin Zhong, and the seminar participants of the London School of Economics, Jinan University, Peking University, Tsinghua University, University College London, and University of Toronto, and conference participants of the 2019 Asian Meeting of the Econometric Society, China and World Economy 2019 Annual Meeting, 2019 International Symposium on Contemporary Labor Economics, 2020 China Economics Summer Institute, 2020 European Economics Association Congress, and 2021 STEG Annual Conference for their constructive comments. Qingen Gai and Qinghua Shi acknowledge the financial support of the National Natural Science Foundation of China (71773076, 71973094, and 72073087). Xiaodong Zhu acknowledges financial support from the Social Science and Humanity Research Council of Canada.

1 Introduction

There are large gaps in value-added per worker between the agricultural and non-agricultural sectors in developing countries, a phenomenon known in the literature as the agricultural productivity gap (APG). Sectoral labor productivity gaps remain sizeable, even after controlling for observable sectoral differences in worker characteristics, such as human capital and working hours (Gollin et al., 2014). Because a large portion of the labor force in poor countries work in agriculture, the APG is also the main reason for the large disparity in aggregate labor productivity between rich and poor countries (Gollin et al., 2002; Caselli, 2005; Restuccia et al., 2008). Therefore, understanding the sources of the APG is important for understanding why developing countries lag behind in aggregate productivity, and for designing policies that may help reduce income disparities between developing and developed countries.

There are two competing explanations for the large APG in developing countries. One explanation refers to differences in unobserved worker characteristics and sorting.[1] Another explanation focuses on barriers to worker mobility between the two sectors, which prevent farmers from migrating to the more productive non-agricultural sector.[2] In the former case, efficient sorting implies that there is little room for policy makers to improve welfare by reallocating workers out of agriculture. In contrast, in the latter case, the APG reflects a combination of the underlying sectoral productivity gap and barriers to switching sectors, and policies that reduce the barriers could help improve aggregate productivity in the developing countries.

Of course, these two explanations are not mutually exclusive. As pointed out by Lagakos (2020) and Donovan and Schoellman (2020), it is likely that both sorting and mobility barriers are important in accounting for the observed APG, and the research challenge is to identify these two sources empirically and to quantitatively estimate their contributions to the APG. We tackle the challenge in this paper. First, we use a unique large panel dataset and a policy experiment in China to empirically estimate the average migration cost of marginal workers affected by the policy and the underlying average labor productivity difference between the two sectors without imposing strong functional form assumptions. These estimates can tell us not only if there exist significant migration barriers, but also how much of the observed APG can be attributed to sorting and the underlying sectoral productivity difference, respectively. These reduced-form estimates, however, cannot tell us why there exists an underlying sectoral productivity difference and how barriers to migration affect the productivity difference and sorting. To address these

[1] See, e.g., Beegle et al. (2011), Lagakos and Waugh (2013), Young (2013), Herrendorf and Schoellman (2018), Alvarez (2020), and Hamory et al. (2021).

[2] See, e.g., Restuccia et al. (2008), Bryan et al. (2014), Munshi and Rosenzweig (2016), Lagakos et al. (2018), Ngai et al. (2019), Tombe and Zhu (2019), Hao et al. (2020), Lagakos et al. (2020), Imbert and Papp (2020).

questions, we then use the same panel dataset to structurally estimate a general equilibrium Roy model, which we use to determine the relative contributions of migration costs and sorting to the observed APG, and to quantify the effects of reducing migration costs on the underlying sectoral productivity difference, migration, and aggregate productivity.

China is an excellent case study for three reasons. First, both the sectoral income gap and the explicit policies restricting rural-to-urban migration are well documented (Ngai et al., 2019; Tombe and Zhu, 2019; and Hao et al., 2020). Second, there is a unique large panel dataset, the annual National Fixed Point Survey (NFP) of agriculture, that tracks around 80,000 rural agricultural workers and rural-to-urban migrant workers from 2003 to 2012. Finally, there has been a policy change that serves as a policy experiment to help identify the effect of changes in migration costs empirically.

Specifically, the policy experiment is the gradual county-by-county roll-out of the New Rural Pension Scheme (NRPS) between 2009 and 2012. Existing studies show that the new pension scheme increases elderly consumption of healthcare services and reduces their reliance on the eldercare provided by their children (Zhang and Chen, 2014; Eggleston et al., 2016; Chen et al., 2018). The studies also show that the new pension scheme reduces elderly labor supply in farm work and increases their time spent with their grandchildren (Jiao, 2016; Huang and Zhang, 2020). Through these two channels, the new pension scheme helps reduce the migration costs of the elderlies' adult children, but has no direct impact on their labor earnings in the two sectors. Therefore, the policy experiment can serve as an instrument for estimating the migration returns of workers who switched sectors due to the policy – the local average treatment effect (LATE). Our estimation yields a LATE estimate of 79 log point difference in annual earnings between the non-agricultural and agricultural sectors. We show theoretically that this LATE estimate also provides an estimate of the average migration cost (as a percentage of non-agricultural earnings) for those migrant workers who were affected by the policy on the margin. So, the estimation result also implies that, prior to the implementation of the new rural pension scheme, these migrant workers faced migration costs that were around 55% of their potential non-agricultural earnings.

Having the policy experiment as the instrument, we can also use the control function approach suggested by Card (2001) and Cornelissen et al. (2016) to estimate the average treatment effect (ATE) of migration. This estimate corresponds theoretically to an increase in the labor productivity for an average rural worker if she moves from the agricultural sector to the non-agricultural sector in an urban area. We call this increase in productivity the *underlying APG*. Our control function estimation yields an underlying productivity difference that ranges from 38 to 46 log points. The results suggest that there is a substantial underlying labor productivity gap between the agricultural and non-agricultural sectors in China that is not due to worker sorting. In comparison, the OLS estimation of the APG that controls for observed worker characteristics but not selection based on unobserved characteristics yields an estimate of 68 log points. So, our estimation results also imply that sorting of workers based on unobserved characteristics accounts for less than half of the observed APG in China, with the rest accounted for by the underlying productivity difference.

In summary, our reduced-form estimation using the panel data and a policy experiment as the instrument shows that, in China, rural residents face significant barriers to migration from the rural agricultural sector to the urban non-agricultural sector, and there is a large underlying labor productivity difference between the two sectors. It also shows that the underlying productivity difference accounts for more than half of the observed APG in China.

Why is there a large underlying sectoral productivity difference? What are the sources of the migration barriers? How would reductions in migration barriers affect the underlying productivity difference, sorting, and aggregate productivity? To address these questions, we then develop and structurally estimate a general equilibrium Roy model. Since our dataset is an origin-based survey of rural residents, we model these individuals' sectoral choices carefully, but assume that urban residents always work in the non-agricultural sector. Like the standard Roy model, we assume rural residents have heterogeneous comparative advantage with respect to working in the two sectors. In the model, a rural resident who decides to migrate to the urban non-agricultural sector faces a migration cost. We allow for heterogeneous migration costs across individuals and assume that migration costs are time-invariant functions of location, policies, and individual characteristics such as gender, age, and education level. We consider two measures of policies. One is a dummy variable that indicates if the NRPS had been implemented in the individual's county of residence, and the other is an index that measures how easy it is to get (*hukou*) residency status in the potential destinations. We conjecture that individuals living in a rural county that had already implemented the new rural pension scheme and have elderly in the family, or living close to cities with a less stringent *hukou* policy are likely to face lower migration costs. Finally, we allow for idiosyncratic shocks to migration costs and human capital in the two sectors that are i.i.d. across individuals and time to capture rich income and migration dynamics observed in the panel data.

We use the maximum likelihood method to estimate the structural model. The structural estimation yields similar results to those from our reduced form estimation. The estimated underlying productivity gap is 59 log points and the average proportional migration cost faced by all workers with rural *hukou* is 39% of their potential non-agricultural earnings. More important, our structural estimation reveals significant heterogeneity in migration costs across locations and individuals with different characteristics. It shows that the migration costs are lower for men, highly educated workers, younger workers, and workers with an elderly family member above age 60 in the household. Our estimation also shows that *hukou* policy and the NRPS both have a significant negative effect on migration costs. A rural individual living in a county with a higher *hukou* liberalization index or with an elderly in the household and the NRPS implemented in the village faces much lower migration costs than the average rural resident. We also find that abilities are more dispersed in agriculture than in non-agriculture.

We next extend our model into a general equilibrium framework to allow for changes in the underlying productivity gap in the counterfactual analysis. If we implement a *hukou* reform by setting the *hukou* liberalization index in all regions of China to the level of the most liberal region, the observed APG would decrease by more than 30%, the migrant share would increase by about 9%, and the aggregate productivity would increase by 1.1%. In contrast, in a partial equilibrium in which the underlying productivity difference does not change with migration cost, the *hukou* policy reform would reduce the observed agricultural productivity gap by only 9%, increase the migrant share by more than 50%, and increase the aggregate productivity by 6.8%. Our results suggest that taking into account the general equilibrium effect of reductions in the rural-to-urban migration cost on the relative price of agriculture is important for evaluating their impact on the observed APG, migration, and aggregate productivity. We also quantify the impact of hypothetical reductions in the average migration cost faced by all rural individuals and find similar differences in their effects between partial and general equilibrium. Finally, we quantify the impact of sectoral productivity changes and find that quantitatively, the change in agricultural productivity is important for migration, but the change in non-agricultural productivity is important for aggregate productivity.

Our study contributes to the literature that examines the roles of labor mobility barriers and sorting in accounting for the observed agricultural productivity gap. See Lagakos (2020) for a recent survey of this literature. In particular, Lagakos and Waugh (2013), Tombe and Zhu (2019), and Hao et al. (2020) use general equilibrium Roy models to quantify the role of selection and migration barriers in accounting for the observed APG. To do so, they impose strong and restrictive assumptions about the distributions of unobserved individual abilities or preferences. Thus the quantitative results could be sensitive to functional form assumptions. To get around this, Herrendorf and Schoellman (2018), Alvarez (2020), Lagakos et al. (2020), and Hamory et al. (2021) try to control for the selection effect by using individual fixed effect regressions to estimate the migration returns of those who did migrate. However, Pulido and Święcki (2018) points out that controlling for individual fixed effects does not solve the selection problem if individuals' unobserved abilities are different in the two sectors and they sort into the two sectors according to their comparative advantage. They propose a Roy model of comparative advantage and sectoral choice and structurally estimate the model using panel data. Their identification, however, still depends heavily on their functional form assumptions. One of our paper's main contributions is that it exploits a quasi-natural policy experiment as an instrument to solve the identification problem and estimate the average treatment effect (ATE) of migration and the average migration cost of the treated individuals (LATE) without imposing strong functional form assumptions.[3] The empirical methods we use are well known in the labor literature (see, e.g., Heckman and Honore (1990), Card (2001) and Cornelissen et al. (2016)), but have so far not been applied in the APG literature. Our paper helps to bridge the gap.

Another main contribution of our study is estimating a general equilibrium Roy model that incorporates migration costs that vary across locations, individual characteristics, and policy environment. Both Lagakos et al. (2020) and Schoellman (2020) argue that heterogeneous migration costs are important to reconcile different pieces of evidence on the returns to migration in the literature. We show, using Chinese data, that migration costs are indeed heterogeneous and vary systematically with policy environment and individuals' gender, age, education level, and family structure. By linking migration cost to policy environment, we can also quantify the effects on aggregate real income and productivity of counterfactual policies that reduce rural-to-urban migration costs in China. By using detailed micro-data to discipline the general equilibrium model of migration, our paper is also related to Lagakos et al. (2018), which uses results from a micro field experiment to calibrate its general equilibrium model of migration in Bangladesh.

Finally, our study is also related to the literature on misallocation and aggregate productivity in China. See, e.g., Hsieh and Klenow (2009), Song et al. (2011), Brandt et al. (2013), Adamopoulos et al. (2017), Ngai et al. (2019), and Tombe and Zhu (2019). In particular, Adamopoulos et al. (2017) also uses the NFP panel data and a general equilibrium Roy model to examine misallocation in China. Their focus, however, is on how the frictions within agriculture affect the occupational choices of workers, while our focus is on the effects of rural-to-urban migration costs. Another difference is that they use the household-level data prior to 2003, while we use the data on individual migrant workers for the 10-year period starting from 2003.

The remainder of the paper proceeds as follows. Section 2 discusses the institutional background and. Section 3 presents a generalized Roy model and discusses our empirical strategy

[3] There are a small number of recent papers that employ field and natural experiments to identify the return to migration, such as Bryan et al. (2014) and Nakamura et al. (2016). Our study complements these papers, but also highlights how to make use of quasi-experimental variation to identify the underlying agricultural productivity gap.

for dealing with selection bias. Section 4 and 5 move onto the reduced-form empirical analysis and structural estimation, respectively. Finally, Section 6 embeds our structural model into a general equilibrium Roy model and conducts quantitative analysis. Section 7 concludes.

2 Institutional Background and Data

2.1 The Hukou System and Origin-based Hukou Index

Under China's household registration system, each Chinese citizen is assigned a *hukou* (registration status), classified as "agricultural (rural)" or "non-agricultural (urban)" in a specific administrative unit that is at or lower than the county or city level. The system is like an internal passport system, where individuals' access to public services is tied to having local *hukou* status. Individuals need approval from local governments to change their *hukou*'s category (agricultural or non-agricultural) or location, and it is extremely difficult to obtain such approval. Due to these institutional barriers, most rural-to-urban migrant workers are without urban *hukou* and therefore have limited access to local public services, such as health care, schooling and social security. Consequently, many migrant workers leave their children and elders behind in the rural areas. In recent years, there have been some policy reforms that relaxed the restrictions imposed by the *hukou* system, but the degree and timing of the liberalization varies across cities.[4]

For our empirical analysis, we construct an origin-based annual Hukou Index for all prefectures in China for the period of 2003-2012. Fan (2019) constructed a destination-based prefecture-level Hukou Reform Index for the period of 1997-2010, with a higher value of the index reflecting better prospects of long-term settlement for migrant workers at a particular destination city in a particular year. We follow his methodology and extend his index to 2012. We then construct our origin-based Hukou Index as follows: For each origination prefecture, we use the pre-determined out-migration flows to weight the Hukou Reform Index across all destinations. The information of pre-determined bilateral migration flows among prefectures are obtained from the 2000 Population Census. Our Hukou Index measures how easy it is for migrant workers from a particular prefecture to settle in cities, and it is negatively related to the migration barriers faced by migrant workers from the prefecture.

Table 1 shows that, from 2003 to 2010, both the average and maximum Hukou Indexes are increasing over time, suggesting a general trend of *hukou* policy liberalization. After 2010,

[4] Chan (2019) provides a detailed and up-to-date discussion of the system and its reforms, and Hao et al. (2020) presents an up-to-date summary of the internal migration patterns in China based on China's population census data.

Table 1: Hukou Index: Summary Statistics

Year	Mean	Std	Min	Max
2003	0.098	0.056	0.013	0.342
2004	0.120	0.075	0.013	0.475
2005	0.120	0.075	0.013	0.475
2006	0.123	0.076	0.013	0.475
2007	0.137	0.090	0.017	0.603
2008	0.136	0.086	0.017	0.579
2009	0.142	0.086	0.017	0.580
2010	0.153	0.099	0.024	0.678
2011	0.137	0.074	0.025	0.424
2012	0.144	0.075	0.025	0.406

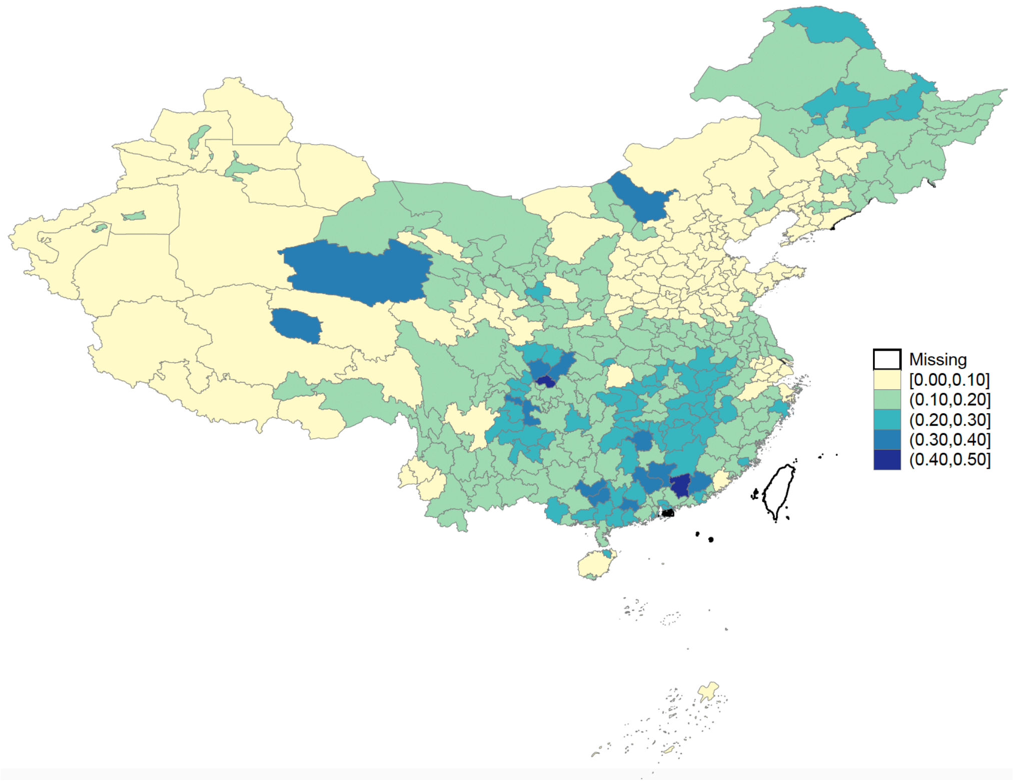
however, both the average and maximum Hukou Indexes fall back from their peak 2010 values. In these later years, many first-tier cities tightened their *hukou* policy restrictions in an attempt to control their city's booming population. There are also large variations in *hukou* policy across prefectures in China. Figure 1 plots the geographic distribution of the Hukou Index in 2012, which ranges from 0.025 for Ngari prefecture in Tibet to 0.406 for Heyuan prefecture in the coastal province of Guangdong. Note that the values of the Hukou Index in areas near Beijing and Shanghai are generally low due to the stringent population control polices in these two first-tier cities.

2.2 The New Rural Pension Scheme

No pension system was in place for rural China until September 2009, when the Chinese government began to gradually roll out the New Rural Pension Scheme (NRPS) across the country. By the end of 2012, the NRPS was introduced to all rural counties in mainland China. Huang and Zhang (2020) compiled the data on the timing of NRPS coverage across counties in China. Based on their data, Figure 2 plots the NRPS's county coverage rate over time across villages in our sample.

Upon the introduction of the NRPS to a county, all people aged 16 years or older with rural *hukou* in the county can participate in the scheme on a voluntary basis. All of the enrollees aged 60 years or older at the start of the NRPS are eligible to receive the basic pension benefit of 660 RMB (about 108 USD) per year, *regardless of previous earnings or income.* Enrollees aged 45 and above need to pay the premiums continuously until they reach age 60 and enrollees under age 45 need to pay the premiums continuously for at least 15 years, before they can claim any pension benefits. Participants can choose from 100, 200, 300, 400 or 500 RMB as the

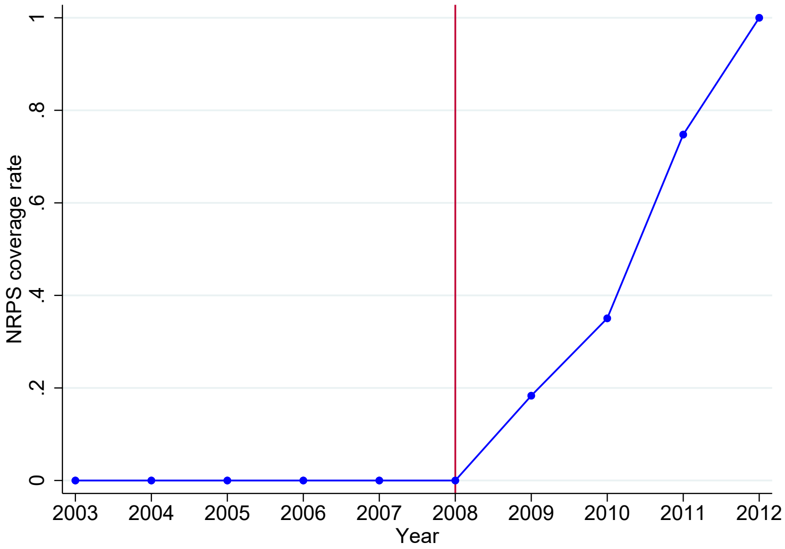
Figure 1: Geographic Distribution of Hukou Index in 2012



level of their annual contribution. Pensioners can claim the pension benefits after age 60 and the pension benefits consist of two parts: one is from the accumulated fund in the individual's account and the other is the basic pension benefit.

Since many migrant workers leave their children and elders behind in their rural homes, the introduction of the NRPS lowers the intangible migration cost faced by working-age rural workers through the eldercare and childcare channels. The existing literature shows that, with the new pension plan, the elders increase healthcare service consumption and rely less on the eldercare provided by their children (Zhang and Chen, 2014; Eggleston et al., 2016), and reallocate time from farm work to non-farm home production and to taking care of their grandchildren (Jiao, 2016; Huang and Zhang, 2020). These channels in effect reduce the migration costs associated with dependent care and non-farm home production. We will use the data on the timing of the introduction of the NRPS as an indicator of policy shocks for our empirical analysis.

Figure 2: NRPS Coverage Rate



2.3 Origin-based Panel Data on Migration and Income

2.3.1 Description of the NFP Data

The main data we use in this paper is the annual National Fixed Point (NFP) Survey conducted by the Research Center of Rural Economy (RCRE) of the Chinese Ministry of Agriculture and Rural Affairs. The survey covers rural households in more than 300 villages from all 31 mainland provinces. The villages were selected for their representativeness based on region, income, cropping pattern, population, and so on.[5] It is designed to be a longitudinal survey, following the same households over time, and has been conducted annually since 1986, with the exceptions of 1992 and 1994 due to funding difficulties. The data have recently been used by several researchers studying China's agriculture. See, e.g., Adamopoulos et al. (2017), Kinnan et al. (2018), Chari et al. (2020), and Tian et al. (2020). Benjamin et al. (2005) provide a detailed description of the data and suggest that the data are of good quality.

The survey contains village-level, household-level, and, since 2003, individual-level questionnaires. At the village level, it collects information that includes population, collective assets, village leader, etc, and at the household level, it surveys households' agricultural production, consumption, asset accumulation, employment, and income. Most existing studies use the data for the years prior to 2003, which do not include detailed information about individual household members. Due to the restrictions imposed by the *hukou* system, rural-urban migra-

[5] In Table A.1, we show that the workers in the 2005 wave of the NFP share similar characteristics with the workers with rural *hukou* in the 2005 China 1% Population Sampling Survey (mini census).

tion in China is mostly temporary in nature and few households migrate to cities as a whole. It is therefore critical to have information about individual household members for studying rural-urban migration in China.

Unique to this study, we have access to annual waves of the data between 2003 to 2012 that include an individual-level questionnaire in the survey. It asks for information on individuals' age, gender, schooling attainment, industry of work, working days, etc. Most important, it asks whether an individual migrated outside the township of her/his *hukou* residence for work during each year of the survey. For those who answered yes, the survey also asks about their earnings from working as a migrant worker. In each year of our sample period, the survey covers approximately 20,000 households and 80,000 individuals from 350 villages in mainland China.

For studying rural-urban migration, the NFP data have several advantages over other data that are commonly employed in the studies on internal migration in China. Relative to repeated cross-sectional data, such as the population census, the panel structure of the NFP better serves identification purposes. Another advantage of the NFP over the population censuses is that the NFP provides detailed information on individual income, whereas only the 2005 population census includes income information. Different from other longitudinal surveys, such as the Longitudinal Survey on Rural Urban Migration in China (RUMiC) and the China Family Panel Study (CFPS), the NFP has a much more comprehensive sample coverage in both geographic and time dimensions. It tracks both rural residents and migrants annually over 10 years. In particular, given that it is an origin-based survey, its attrition rate is relatively low. In the raw NFP data, 30% can be tracked for one year, 14% for two years, 10% for three years, 8% for four years, and 38% for five or more years.[6] In contrast, destination-based surveys of migrants such as the RUMiC have very high attrition rates.

One drawback of the NFP data is that they include limited information on migration destinations. We can only know whether a migrant is within home county, within home province, or outside home province. For the surveys after 2009, we know the destination provinces but not the destination cities. Hence, our analysis focuses on migration from the rural agricultural sector to the urban non-agricultural sector, instead of spatial movements among provinces and cities. For analyzing spatial allocation of labor, population census data are more suitable.

2.3.2 Construction of Key Variables

Now, we formally introduce some key variables constructed from the NFP data that are important to our analysis. More details are provided in Appendix A.

[6] See Table A.2 for details.

Sector of Employment and Migration. We define an individual as working in the non-agricultural (na) sector in a particular year if she/he worked more than 180 days out of town during that year, and working in the agricultural (a) sector otherwise. This classification aligns with the definition of migrant workers by the National Bureau of Statistics (NBS) of China. For workers who worked in town , but reported working in the non-agricultural sector, the NFP unfortunately does not have information about their non-agricultural earnings. We thus treat them as agricultural workers with the implicit assumption that a rural worker earns the same wage in agriculture and *local* non-agriculture. Given our definition, we shall use "migration" and "working in the non-agricultural sector" interchangeably throughout the paper.

Nominal Agricultural Earnings. The NFP survey provides detailed information on household agricultural production, including all inputs and output at the crop level. We compute the gross output for each type of crop as the production multiplied by the corresponding market price in that year. Intermediate inputs such as fertilizers and pesticides are also valued by their market prices. We subtract expenditures on intermediate inputs from the gross output to obtain the value-added for each type of crop. We aggregate the value-added of all crops to the household level, which is then allocated to each household member based on the formula below:

$$\text{Individual earnings in } a = \frac{\text{Individual's working days in } a}{\text{HH's working days in } a} \times \text{Household's value-added from } a$$

Specifically, we construct individual earnings from agricultural production by apportioning household agricultural earnings to each household member according to the number of working days they each allocated to agricultural production. The annual income of rural workers is the product of individual agricultural daily earnings and total within-town working days.

Nominal Non-agricultural Earnings. The NFP survey also asks each household member the number of days they worked out of town and the corresponding earnings. Non-agricultural annual earnings is defined as the earnings made when individuals work outside of their home town.

Real Earnings. We deflate all nominal earnings into 2003 Beijing prices using province-level spatial price deflators constructed by Brandt and Holz (2006), so that the measures reflect the real incomes from different sectors. For workers in agriculture, we deflate their annual earnings by the rural price index of the province in which their village is located. For workers in the out-of-town non-agricultural sector, their migration destination is unobserved during the period of 2003-2008. To deflate their incomes, we proceed as follows. First, we use the 2000 Population Census to calculate the shares of out-migrants to different provinces for each prefecture. Second, we map the villages to prefectures, and based on the predetermined migration shares, construct the weighted average of urban price indices across different destination provinces for each village. The annual earnings of out-migrants is deflated by this weighted urban price index. For the remainder of the paper, all earnings refer to real annual earnings unless stated otherwise. The total annual income is the sum of the earnings from agriculture and non-agriculture.

2.3.3 Basic Facts

Our analysis focuses on the sample of individuals aged between 20 and 54 with no more than 12 years of schooling, and who appear at least two times in our sample period of 2003 to 2012. We make the age restriction because we want to focus on those of the working-age population who have finished schooling but are not close to the eligible age (60) for receiving the rural pension income. We also exclude individuals with more than 12 years of schooling because there are very few of them in the data. We additionally restrict the sample to those who can be observed for at least two years, as our individual fixed-effect model requires repeated observations. After the restriction, we obtain 51,688 individuals with 234,031 individual-year observations. Among them, 25% are tracked for two years, 18% for three years, 14% for four years, and 43% for five or more years. We trim the sample at the top 1% and bottom 1% of the annual income distribution in the agricultural and non-agricultural sector, respectively.

Table 2 reports summary statistics of the data. About 30% of workers in our sample migrated out of town to work in non-agriculture at some time during the sample period. The means of log annual earning in agriculture and non-agriculture are 8.63 and 9.26, respectively, which implies that the raw average income gap between the agricultural workers and migrant workers in the non-agricultural sector is 63 log points. The variance of log annual earnings is smaller for the migrant workers than that for the agricultural workers. Note that we are comparing agricultural workers to migrant workers who were born in rural areas, not to the whole population of non-agricultural sector workers, which would also include urban residents. Most of these migrant workers work in low-skill manufacturing and service jobs,[7] which may explain the lower dispersion of their earnings.

Table 2 also shows that, in general, migrant workers are younger and healthier, have higher educational attainment, and and are more likely to be male and have an elderly household member aged 60 or above. The differences between agricultural and migrant workers suggest that there is sorting of workers along these observable individual and household characteristics. It is likely that there is also sorting along other unobserved or hard-to-measure characteristics. We next present an empirical framework for dealing with the issue of worker sorting or selection

[7] See Figure A.2 in Appendix A.

in estimating the underlying APG.

Table 2: Summary Statistics

Sample:	All	Non-agri	Agri
ln Daily wage	3.499	3.571	3.468
	(0.906)	(0.628)	(1.002)
ln Annual income	8.827	9.265	8.635
	(1.001)	(0.628)	(1.078)
Total working days	237.387	303.441	208.493
	(101.408)	(44.075)	(105.777)
Share of working days in:			
Within-town agri production	0.554	0.036	0.780
	(0.435)	(0.078)	(0.318)
Within-town non-agri production	0.122	0.006	0.173
	(0.258)	(0.032)	(0.294)
Out-of-town	0.324	0.958	0.047
	(0.443)	(0.086)	(0.163)
Age	37.981	31.855	40.660
	(10.091)	(8.865)	(9.403)
Years of Schooling	7.277	8.149	6.895
	(2.431)	(2.045)	(2.488)
Female	0.469	0.330	0.530
	(0.499)	(0.470)	(0.499)
Poor health status	0.012	0.003	0.015
	(0.107)	(0.057)	(0.122)
Agricultural Hukou	0.976	0.962	0.983
	(0.151)	(0.192)	(0.129)
Arable land per capita	2.184	1.394	2.530
	(2.855)	(1.674)	(3.178)
Household with an elderly aged ≥60	0.279	0.343	0.251
	(0.448)	(0.475)	(0.433)
Number of observations	234031	71218	162813
Share of workers	1	0.304	0.696

Notes: Standard deviation in parentheses.

3 A Framework for Empirical Analysis

This section presents a generalized Roy model of rural agriculture to urban non-agriculture migration that will serve as a framework for our empirical analysis of migration costs, sorting, and the APG. Some of the propositions that we state in this section are not new and well-known in the literature on generalized Roy models. However, we think it is useful to present them in our context to clarify what objects of interests are estimated by different empirical methods, respectively, in the APG literature. Furthermore, we clarify how the selection bias is affected by migration costs.

3.1 Technologies and Labor Earnings in the Two Sectors

There are two sectors, agricultural and non-agricultural, which are denoted by a and na, respectively. The production technologies of the two sectors are:

$$Y_a = A_a H_a, \quad Y_{na} = A_{na} H_{na}. \tag{1}$$

Here H_j represents the total efficiency units of labor in sector $j = a, na$. The real wage per efficiency unit of labor in the two sectors are:

$$w_a = p_a A_a, \quad w_{na} = p_{na} A_{na}, \tag{2}$$

where p_j is the price of the sector-j good relative to the price of consumption, $j = a, na$.

Each worker is endowed with a vector of observed characteristics $\boldsymbol{X}$, and a vector of unobserved "individual productivity" denoted by $\boldsymbol{U} = (U_a, U_{na})$. The latter represents the innate abilities of being a worker in the agricultural and non-agricultural sectors, respectively. Without loss of generality, we normalize the mean of $\boldsymbol{U}$ to zero. We assume that an individual worker's efficiency units of labor in the two sectors are given by the following human capital functions:

$$h_a(\boldsymbol{X}, \boldsymbol{U}) = \exp(\boldsymbol{X}\beta + U_a), \quad h_{na}(\boldsymbol{X}, \boldsymbol{U}) = \exp(\boldsymbol{X}\beta + U_{na}). \tag{3}$$

So, the worker's real potential earnings in the two sectors are:

$$y_a(\boldsymbol{X}, \boldsymbol{U}) = w_a \exp(\boldsymbol{X}\beta + U_a), \quad y_{na}(\boldsymbol{X}, \boldsymbol{U}) = w_{na} \exp(\boldsymbol{X}\beta + U_{na}). \tag{4}$$

3.2 Productivity Differences, Migration Costs and Sorting

The agricultural and non-agricultural sectors are located in the rural and urban areas, respectively.[8] A worker can always choose to work in the agricultural sector. If she chooses to work in the non-agricultural sector, however, she has to pay a migration cost that is proportional to her wage in the non-agricultural sector. So, her net income is $(1-\theta)y_{na}(\boldsymbol{X}, \boldsymbol{U})$, where θ is the proportional migration cost. Let $M_c = -\ln(1-\theta)$ be a monotonic transformation of θ. We

[8] We ignore non-agricultural production in rural areas because the NFP data do not have good information about worker earnings from rural non-agricultural jobs. We introduce rural non-agricultural production in our general equilibrium analysis in Section 6.

assume that it takes the following form:

$$M_c = m(\boldsymbol{X}, \boldsymbol{Z}),$$

where $\boldsymbol{Z}$ is an observable vector that represents the policy environment the worker faces. Henceforth, we shall refer to $m(\boldsymbol{X}, \boldsymbol{Z})$ simply as the migration cost.

A worker will choose to migrate to the non-agricultural sector if the following inequality holds:

$$(1-\theta)y_{na}(\boldsymbol{X}, \boldsymbol{U}) = y_{na}(\boldsymbol{X}, \boldsymbol{U})\exp(-m(\boldsymbol{X}, \boldsymbol{Z})) > y_a(\boldsymbol{X}, \boldsymbol{U}),$$

and stays in the agricultural sector otherwise. Let $R = \ln(w_{na}/w_a)$ be the underlying real wage difference between the agricultural and non-agricultural sectors, which we will simply refer to as the *underlying APG*. Then, from equation (4), the inequality above is equivalent to

$$U_{na} - U_a > m(\boldsymbol{X}, \boldsymbol{Z}) - R. \tag{5}$$

Migration condition (5) says that a rural worker will migrate to the non-agricultural sector if and only if her comparative advantage in the non-agricultural sector is higher than the net migration cost $m(\boldsymbol{X}, \boldsymbol{Z}) - R$.

We assume that $\boldsymbol{U}$ is i.i.d. across individual workers and independent of $(\boldsymbol{X}, \boldsymbol{Z})$. We also assume that $V = U_{na} - U_a$ has a continuous and strictly increasing distribution function. Let $F(\cdot)$ be the CDF of V. Conditional on $(\boldsymbol{X}, \boldsymbol{Z})$, the proportion of workers who migrate to the non-agricultural sector is

$$\pi_{na}(\boldsymbol{X}, \boldsymbol{Z}) = 1 - F\left(m(\boldsymbol{X}, \boldsymbol{Z}) - R\right), \tag{6}$$

and the aggregate proportion of workers who migrate to the non-agricultural sector is

$$\bar{\pi}_{na} = 1 - E\left[F\left(m(\boldsymbol{X}, \boldsymbol{Z}) - R\right)\right]. \tag{7}$$

3.3 Selection Bias of Observed APG

The observed log earnings are given by

$$\ln y(\boldsymbol{X}, \boldsymbol{U}) = \ln(w_a) + \mathbf{1}(j = na)R + \boldsymbol{X}\beta + U_a + \mathbf{1}(j = na)(U_{na} - U_a). \tag{8}$$

Let R_{OLS} be the observed difference in average log earnings of agricultural and non-agricultural workers, or *observed APG*. We have

$$R_{\text{OLS}} = E\left[\ln\left(y_{na}(\boldsymbol{X},\boldsymbol{U})\right)|V > m(\boldsymbol{X},\boldsymbol{Z}) - R\right] - E\left[\ln\left(y_{a}(\boldsymbol{X},\boldsymbol{U})\right)|V \leq m(\boldsymbol{X},\boldsymbol{Z}) - R\right].$$

Again, from (4), we have,

$$R_{\text{OLS}} = R + \underbrace{E\left[U_{na}|V > m(\boldsymbol{X},\boldsymbol{Z}) - R\right] - E\left[U_a|V \leq m(\boldsymbol{X},\boldsymbol{Z}) - R\right]}_{\text{selection bias}}. \tag{9}$$

Due to heterogeneous innate abilities and sorting, the observed APG is generally different from the underlying APG. The last two terms in equation (9) show the selection bias or the effect of sorting on the deviation of the observed APG from the underlying APG. The following proposition shows how the migration cost affects the selection bias.

Proposition 1: *If $E[U_{na}|U_{na} - U_a > x]$ and $E[U_a|U_a - U_{na} > x]$ are both increasing functions of x for $x \in (-\infty, \infty)$, then the selection bias $R_{OLS} - R$ is increasing with $m(\boldsymbol{X},\boldsymbol{Z}) - R$.*
Proof: All proofs of propositions in this paper are in Appendix B.

Intuitively, the assumption in Proposition 1 requires that an individual's comparative advantage and absolute advantage are positively correlated for both sectors. This assumption holds, for example, if $(\exp(U_a), \exp(U_{na}))$ has a bi-variate Fréchet distribution or (U_a, U_{na}) has a bi-variate normal distribution and $Corr(U_a, U_{na}) < \min\left\{\frac{\sigma_{na}}{\sigma_a}, \frac{\sigma_a}{\sigma_{na}}\right\}$. Under this assumption, the selection bias term in (9) is increasing with the net migration cost. Therefore, the larger the net migration cost, the more likely it is that the observed APG is higher than the underlying APG.

If U_a and U_{na} have a symmetric joint distribution, as is often assumed in the quantitative migration literature (e.g. Bryan and Morten, 2019; Tombe and Zhu, 2019; and Hao et al., 2020), we can further characterize the selection bias as follows.

Proposition 2: *If the joint distribution of U_a and U_{na} are symmetric with respect to U_a and U_{na}, then the selection bias is zero if the net migration cost $m(\boldsymbol{X},\boldsymbol{Z}) - R$ is zero. If, in addition, the assumption in Proposition 1 holds, then, the selection bias $R_{OLS} - R$ is positive, zero, or negative if and only if net migration cost $m(\boldsymbol{X},\boldsymbol{Z}) - R$ is positive, zero, or negative, respectively.*

Note that the symmetry assumption in Proposition 2 has a very strong implication that the observed APG has a selection bias if and only if there is non-zero net migration cost. Therefore, a positive observed APG implies that either the underlying APG is positive or the net migration cost is positive.

Finally, if (U_a, U_{na}) follows a bi-variate normal distribution, we have the following well-known expression for the selection bias (see, e.g., Heckman and Honore, 1990).

$$R_{\text{OLS}} - R = \sigma_{na}\rho_{na,v}\frac{\phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}{\Phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})} + \sigma_a\rho_{a,v}\frac{\phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}{1-\Phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}, \tag{10}$$

where σ_a, σ_{na}, and σ_v are the standard deviations of U_a, U_{na}, and $V = U_{na} - U_a$, respectively, and $\rho_{a,v}$ and $\rho_{na,v}$ are the correlations of V with U_a and U_{na}, respectively. In the special case where $m(\boldsymbol{X}, \mathbf{Z}) = R$, equation (10) becomes:

$$R_{\text{OLS}} - R = \sqrt{\frac{2}{\pi}}\left(\sigma_{na}\rho_{na,v} + \sigma_a\rho_{a,v}\right) = \sqrt{\frac{2}{\pi}}\frac{\sigma_{na}^2 - \sigma_a^2}{\sigma_v}$$

So, in the case of zero net migration cost, the observed APG has a positive bias if and only if the dispersion of innate abilities is larger in the non-agricultural sector than in the agricultural sector. If the dispersion is actually larger in the agricultural sector, the observed APG underestimates the underlying APG. If $Corr(U_a, U_{na}) < \min\left\{\frac{\sigma_{na}}{\sigma_a}, \frac{\sigma_a}{\sigma_{na}}\right\}$, the assumption of Proposition 1 holds. In this case, if the net migration cost is sufficiently large, the observed APG will overestimate the underlying APG.

In summary, the selection bias depends critically on both the distribution of abilities and the net migration costs faced by individuals. Next, we turn to the empirical methods for dealing with the selection bias problem.

3.4 Empirical Methods

In the literature on APG, there are two commonly used methods in dealing with the selection bias problem. The first method assumes that the distribution of (U_a, U_{na}) or $(\exp(U_a), \exp(U_{na}))$ takes a particular functional form, e.g., a multivariate Fréchet or multivariate normal distribution, and uses the moment matching method to estimate the distribution parameters, underlying APG, and migration costs. See, e.g., Lagakos and Waugh (2013), Adamopoulos et al. (2017), Pulido and Święcki (2018), Tombe and Zhu (2019), and Hao et al. (2020). As pointed out by Heckman and Honore (1990), however, the identification of Roy models is not robust to alternative distribution assumptions, and the estimation results are also not robust, depending critically on the functional form assumptions.

More recently, several authors have adopted a second method, using the observed labor returns of new migrant workers or sector switchers in panel data as estimates of the APGs. See, e.g. Herrendorf and Schoellman (2018), Alvarez (2020), and Hamory et al. (2021). While this method does not rely on strong functional form assumptions, it is not clear what the observed labor returns of sector switchers really measure. Both Pulido and Świecki (2018) and Lagakos et al. (2020) provide examples showing that these returns may over- or under- estimate the underlying APGs if the shocks that caused workers to switch sectors are correlated with individual comparative advantages. Also, Schoellman (2020) argues heuristically that, if the income or migration cost shocks are independent of individual comparative advantages, the estimated return to migration for switchers is not the underlying APG but a measure of the average migration cost faced by the switchers before the shocks hit.

So, neither of the two commonly used methods in the APG literature is ideal for dealing with the selection bias problem. We consider a different method in this paper. The model we presented belongs to a class of models that are called generalized Roy models. There is an extensive literature in labor economics and applied econometrics on the identification and estimation of generalized Roy models. See, e.g., Card (2001), Eisenhauer et al. (2015), and Cornelissen et al. (2016). We apply the insights from this literature for identification and estimation of our model. Using the terminology of this literature, the underlying APG is the average treatment effect (ATE) of migration:

$$R = E\left[\ln\left(y_{na}(\boldsymbol{X},\boldsymbol{U})\right) - \ln\left(y_{a}(\boldsymbol{X},\boldsymbol{U})\right)\right].$$

To control for selection bias, the literature suggests using either field or natural experiments. For the case of China, we will use the gradual implementation of the NRPS as a policy experiment and a control function approach to estimate the ATE or the underlying APG. Specifically, we estimate equation (8) controlling for proxies for the selection terms $E[U_a|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]$ and $E[(U_{na}-U_a)|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]$. The basic idea is to make some assumptions about the nature of the covariances between unobserved components U_{na} and U_a and the observable variables $\mathbf{1}(j=na)$, $\boldsymbol{X}$, and $\boldsymbol{Z}$.[9] Following Card (2001) and Cornelissen et al. (2016), the proxies are the transformations of the residuals obtained from the selection equation with the policy variable $\mathbf{Z}$ as the excluded cost shifters. If we assume a joint normal distribution for $\mathbf{U}$, the approach

[9] Equation (8) implies that

$$\begin{aligned} E[\ln y(\boldsymbol{X},\boldsymbol{U})|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}] = &\ln(w_a) + \mathbf{1}(j=na)R + \boldsymbol{X}\beta \\ &+ E[U_a|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}] + \mathbf{1}(j=na)E[(U_{na}-U_a)|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]. \end{aligned}$$

With the controls for $E[U_a|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]$ and $E[(U_{na}-U_a)|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]$, one may identify R by OLS estimation and obtain the ATE.

can also be modified by explicitly accounting for the binary nature of the endogenous variable and replacing the selection terms by a generalized residual based on the inverse Mills ratio from a first stage probit regression (Wooldridge, 2015).[10] We will present the results of the control function approach with and without the normality assumption.

Using the policy experiment, we also estimate the local average treatment effect (LATE) that reveals the average labor return of workers whose migration decisions are marginally affected by the policy. We can show that, under the exclusion assumption of the policy instrument, the LATE estimate of return to migration is also an estimate of the average migration cost faced by these marginal workers. To see this, consider a change in policy variable $\boldsymbol{Z}$ that reduces the migration cost and satisfies the exclusion restriction; i.e., $\Delta m = m(\boldsymbol{X},\boldsymbol{Z}) - m(\boldsymbol{X},\boldsymbol{Z'}) > 0$ is independent of $\boldsymbol{U}$. Then, the LATE estimate of the return to migration can be formally written as follows:

$$R_{\text{LATE}} = E\left[\ln\left(y_{na}(\boldsymbol{X},\boldsymbol{U})\right) - \ln\left(y_a(\boldsymbol{X},\boldsymbol{U})\right)|m(\boldsymbol{X},\boldsymbol{Z}) - R - \Delta m < V < m(\boldsymbol{X},\boldsymbol{Z}) - R\right]. \tag{11}$$

We prove the following proposition in Appendix B.

Proposition 3: *If the migration cost change Δm is independent of individual comparative advantage in the non-agricultural sector, $V = U_{na} - U_a$, then,*

$$\lim_{\Delta m \to 0} R_{\text{LATE}} = \frac{E\left[m(\boldsymbol{X},\boldsymbol{Z})f\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right]}{E\left[f\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right]}, \tag{12}$$

where $f(.)$ is the PDF of V.

Intuitively, a small migration cost change only induces workers who are ex-ante indifferent between the a and na sectors to migrate. When they switch sectors, the change in income reveals their baseline migration cost.

[10] Specifically, with the joint normal assumption of $\boldsymbol{U}$,

$$\begin{aligned}&E[U_a|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}] + \mathbf{1}(j=na)E[(U_{na}-U_a)|\mathbf{1}(j=na),\boldsymbol{X},\boldsymbol{Z}]\\ &= \sigma_a\rho_{a,v}(1-\mathbf{1}(j=na))\frac{\phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}{1-\Phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})} + \sigma_{na}\rho_{na,v}\mathbf{1}(j=na)\frac{\phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}{\Phi(\frac{R-m(\boldsymbol{X},\boldsymbol{Z})}{\sigma_v})}.\end{aligned}$$

4 Reduced-Form Analysis

Having laid out the empirical framework, we now turn to the empirical analysis of rural-to-urban migration in China. We start with a simple cross-sectional comparison of the labor productivity in the two sectors.

4.1 Cross-Sectional Estimation of Returns to Migration

We estimate the following regression equation:

$$\ln y_{ihjt} = \gamma_1 NonAgri_{ihjt} + X_{ihjt}\gamma_2 + \varphi_j + \varphi_{pt} + \nu_{ihjt}, \tag{13}$$

where y_{ihjt} denotes the year-t annual earnings of individual i who belongs to household h in village j; $NonAgri_{ihjt}$ is a binary indicator for employment in sector na. X_{ihjt} is a vector of individual and household characteristics, including four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, type of *hukou*, a dummy indicating whether there is an elderly aged 60 or above residing in the household, and the share of months in year t that the NRPS has been in effect; φ_j denotes the village fixed effects, which absorbs all time-invariant village-specific determinants of income; we also include province×year fixed effects φ_{pt}, which flexibly control for unobserved income shocks at the province level. Standard errors are clustered at the village×year-level to account for unobserved shocks that are correlated across individuals residing in the same village in the same year.

Table 3 reports the OLS regression results. We find in column (1) that, unconditional on inidividual characteristics, annual earnings in sector na are on average 64 log points higher than those in sector a. As is shown in Column (2), the estimate changes slightly to 68 when the individual controls are included. Columns (3) includes three indicator variables which are defined based on the migration status in period $t-1$ and t: a-to-na switchers, na-to-a switchers, and sector-na stayers. The stayers in sector a constitute the omitted group. Therefore, the estimates reflect the income gaps relative to the stayers in agriculture. The income gap is 60 log points for a-to-na switchers, and 68 log points for sector-na stayers. This finding suggests that a large portion of the income gains is realized upon migration. Interestingly, relative to sector-a stayers, na-to-a switchers have a lower annual income, suggesting that there are factors other than income, such as idiosyncratic shocks to migration costs or preferences, that also affect workers' migration decisions.

Table 3: Sector of Employment and Annual Earnings: OLS

Dep. Var.: ln Annual Earnings	(1)	(2)	(3)
NonAgri	0.6449***	0.6814***	
	(0.0122)	(0.0119)	
a-to-*na* switchers			0.5979***
			(0.0145)
na-to-*a* switchers			-0.0405**
			(0.0183)
Sector-*na* stayers			0.6757***
			(0.0140)
Individual and household controls	N	Y	Y
Province× Year FE	Y	Y	Y
Village FE	Y	Y	Y
Observations	234,025	234,025	157,985
R-squared	0.3422	0.3904	0.3868

Notes: Individual controls include four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, type of Hukou, a dummy indicating whether there is an elderly aged 60 or above residing in the household, and the share of months in year t that the NRPS has been in effect. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

4.2 Individual Fixed Effect Estimation of Returns to Migration

Table 4 repeats the regression analysis in Table 3, but further controls for individual fixed effects. This approach has recently been adopted in the APG literature (Herrendorf and Schoellman, 2018; Alvarez, 2020; and Hamory et al., 2021) to address the potential selection bias problem under the assumption that selection on sector of employment is only determined by time-invariant individual characteristics which have the same effect on potential earnings across sectors. If for some reason high-ability workers are more likely to work in the non-agricultural sector, then the observed APG would be due to the difference in average ability of workers in the two sectors, and thus an individual fixed-effect regression could control for this selection bias. They therefore argue that the estimated labor return to migration after controlling for individual fixed-effects is a better measure of the APG or migration barriers.

Columns (1) and (2) of Table 4 show that our fixed effect (FE) estimate of the income gap between sector *na* and sector *a* for China is 69 log points, which is statistically indistinguishable from the OLS estimate of 68 log points. This finding is in contrast with the findings in the studies we cited above that use data from some other countries. Hamory et al. (2021) shows that, after controlling for individual fixed effects the estimated APG drops from 36 log points

Table 4: Sector of Employment and Annual Earnings:
Individual Fixed Effects

Dep. Var.: ln Annual Earnings	(1)	(2)
NonAgri	0.6916***	
	(0.0137)	
a-to-*na* switchers		0.6221***
		(0.0184)
na-to-*a* switchers		0.0084
		(0.0189)
Sector-*na* stayers		0.6998***
		(0.0185)
Individual and household controls	Y	Y
Province× Year FE	Y	Y
Individual FE	Y	Y
Observations	234,025	144,049
R-squared	0.6810	0.6920

Notes: Individual controls include four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, type of Hukou, a dummy indicating whether there is an elderly aged 60 or above residing in the household, and the share of months in year t that the NRPS has been in effect. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

to 24 log points for Indonesia, and from 48 log points to 22 log points for Kenya. Alvarez (2020) shows that controlling for individual fixed effects also leads to a large reduction in the estimated income gap between the manufacturing sector and the agricultural sector in Brazil, from 48 log points to 9 log points, as well as a large reduction in the estimated income gap between the service sector and the agricultural sector in Brazil, from 48 log points to 4 log points. Using the data from the US, Herrendorf and Schoellman (2018) finds that the wage gains based on switchers is only 6%, much lower than the cross-sectional wage gap of 76%.[11] These results suggest that the labor returns to migration are small in many countries. Our estimates for China, in contrast, suggest a large return to migration in China. As we have shown in Proposition 3 of Section 3, if all the sector switches are driven by exogenous shocks to migration costs, the FE estimate reveals the average migration cost faced by switchers before

[11] Our result is also different from the findings in Lagakos et al. (2020), which uses the CFPS data to estimate the return from switching sectors in China and finds that the cross-sectional OLS estimate is significantly higher than the FE estimate. However, they use per capita consumption rather than real income as the dependent variable, which is probably a lower bound for income gains, because income elasticity of consumption is generally less than 1. In fact, when we use the real earning data from the CFPS, we obtain an OLS estimate of 1.09 and FE estimate of 1.29. The details of the regression analysis using the CFPS data are discussed in Appendix D.1.

the shocks hit. Hence one interpretation of the difference in the results between China and the other countries studied in the literature is that migration costs are larger in China than in these other countries due to China's rigid *hukou* system that explicitly restricts rural-to-urban migration.

A caveat of the above interpretation is that sectoral switch could be endogenous, and hence the FE estimate would be biased. More specifically, if workers have heterogeneous comparative advantage and if (for example) the reductions in migration costs are larger for individuals with a higher comparative advantage in the non-agricultural sector, then, sector switchers are more likely to have a higher return to migration. In such a case, the FE estimate captures neither the underlying APG nor the baseline average migration cost. To address this problem, we need to find exogenous shocks to migration costs that are uncorrelated with migrant workers' potential earnings. The gradual county-by-county implementations of the NRPS in China constitute such shocks.

4.3 The NRPS and IV Estimation of Migration Costs

As is discussed in Section 2, the migration costs of the younger household members may be altered due to the NRPS through the channels of eldercare or childcare. Moreover, the effect of the NRPS on sector choice may vary by household depending on the presence of elderly aged 60 or above who are entitled to the NRPS pension benefits. Therefore, our IV strategy employs $Elder60_{hjt} \times NRPS_{jt}$ to generate exogenous variation in $NonAgri_{ihjt}$,

The first-stage regression is:

$$NonAgri_{ihjt} = \beta_1 Elder60_{hjt} \times NRPS_{jt} + X_{ihjt}\beta_2 + \varphi_j + \varphi_{pt} + \nu_{ihjt}, \tag{14}$$

where $NRPS_{jt}$ captures the share of months in year t that the NRPS covers the elderly in village j. $Elder60_{hjt}$ is an indicator variable that equals one if there is an elderly aged 60 or above residing in the household. Note that X_{ihjt} contains $NRPS_{jt}$ and $Elder60_{hjt}$ to account for their independent effects on sectoral choice. The second-stage of the IV estimation is:

$$\ln y_{ihjt} = \gamma_1 \widehat{NonAgri}_{ihjt} + X_{ihjt}\gamma_2 + \varphi_j + \varphi_{pt} + u_{ihjt}, \tag{15}$$

where $\widehat{NonAgri}_{ihjt}$ is predicted value from the first-stage regression in the IV framework.

Conceptually, instrumenting for the sector of employment with the interaction term $Elder60_{hjt} \times NRPS_{jt}$ is similar to a triple-difference estimation strategy. A simple difference-in-difference estimation would capture the change in the likelihood of non-agricultural employment induced by the implementation of the NRPS, with the identification stemming from the differential timing of the onset of the NRPS across regions. The triple-differences makes an additional comparison across households with and without an elderly aged 60 or above, which adds the advantage of differencing out the village-specific shocks to migration costs or to incomes that coincides in timing with the introduction of the NRPS. The triple-difference approach addresses the concern that the new pension plan may have been rolled out across the country in an endogenous way such that the villages which received the NRPS earlier may have had different trends in income and migration.

The exclusion restriction for the instrument is

$$Cov(Elder60_{hjt} \times NRPS_{jt}, u_{ihjt} | X_{ihjt}, \varphi_j, \varphi_{pt}) = 0.$$

This requires that, conditional on all the observables, (i) the NRPS does not directly affect income differently for individuals in households with an elderly aged 60 or above relative to those without, other than its differential effect on the sector choice across households, and (ii) the NRPS is uncorrelated with any other village-specific unobserved shocks that affect income differently for individuals in households with an elderly aged 60 or above relative to those without. The exclusion restriction is plausibly valid in our context – there is little reason to think that cash transfers received by the elderly would change younger household members' innate abilities for working in different sectors. Despite this consideration, we provide further evidence to substantiate the identification assumptions in the following discussion.

The IV estimate captures the local average treatment effect (LATE), i.e., the difference in potential earnings between the two sectors for a-to-na switchers because of an exogenous reduction in migration costs induced by the NRPS (i.e., compilers). As we have shown in Section 3, the LATE estimate is an estimate of the average (proportional) migration cost of the marginal workers whose sectoral choice was affected by the NRPS policy.

Column (1) of Table 5 reports the first-stage regression result. We find that, in response to the implementation of the NRPS, younger members from households with an elderly aged 60 or above are 4 percentage points more likely to work in the non-agricultural sector relative to those from households without an elderly. Column (2) estimates the reduced form relationship between log earnings and the NRPS. We find that the introduction of the NRPS raises annual earnings for workers from households with an elderly by 3 log points more than those without an elderly dependent. Column (3) shows the second-stage regression result. The IV estimate implies that working in the non-agricultural sector increases annual earnings by 79 log points, which is even larger than the OLS and individual FE estimates.[12] The result indicates that the

[12] We conduct Monte Carlo simulations in Appendix C to present the possible scenarios where the IV estimate is larger than the OLS estimate.

baseline average migration costs faced by the switchers is around 55% ($= 1 - \exp(-0.79)$) of the earnings in the non-agricultural sector. The Kleibergen-Paap F statistic is 21.52, which is above the Stock-Yogo 10 percent threshold for weak instruments. In column (4), we conduct a mediation analysis by including $NonAgri_{ihjt}$ and $Elder60_{hjt} \times NRPS_{jt}$ simultaneously in the earning equation. We show that, conditional on the sector of employment, $Elder60_{hjt} \times NRPS_{jt}$ no longer has an independent effect on income; the estimated coefficient is insignificant in both economic and statistical terms. The finding provides strong supportive evidence for the exclusion restriction, indicating that the NRPS only affects earnings through the channel of switching employment sector.

Table 5: Sector of Employment and Annual Earnings: IV Approach

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	NonAgri	ln Annual	ln Annual	ln Annual	ln Annual	ln Annual
		Earnings	Earnings	Earnings	Earnings	Earnings
	First Stage	Reduced Form	2SLS	OLS	2SLS	2SLS
NonAgri			0.7862**	0.6814***		
			(0.3789)	(0.0119)		
Elder60 × NRPS	0.0401***	0.0315**		0.0042		
	(0.0086)	(0.0149)		(0.0150)		
NRPS	0.0019	-0.0506*	-0.0521*	-0.0519*	-0.0497*	-0.0483*
	(0.0097)	(0.0288)	(0.0280)	(0.0279)	(0.0282)	(0.0282)
Elder60	0.0212***	0.0404***	0.0237**	0.0259***	0.0234*	0.0238**
	(0.0026)	(0.0058)	(0.0116)	(0.0057)	(0.0124)	(0.0116)
Hukou Index: below median × NonAgri					0.8941*	
					(0.4837)	
Hukou Index: above median × NonAgri					0.7169*	
					(0.3657)	
Hukou Index: bottom tercile × NonAgri						0.9331**
						(0.4533)
Hukou Index: middle tercile × NonAgri						0.8274**
						(0.3640)
Hukou Index: top tercile × NonAgri						0.5580
						(0.4407)
Individual and household controls	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Village FE	Y	Y	Y	Y	Y	Y
Observations	234,031	234,025	234,025	234,025	228,176	228,176
R-squared	0.3486	0.3272	0.1618	0.3904	0.1599	0.1544
Kleibergen-Paap F-Stat			21.52		8.525	7.114

Notes: Individual controls include four age group dummies (20-29, 30-39, 40-49, and 50-55), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, and type of Hukou. Column (5) includes the instruments for Hukou Index: below median × NonAgri and Hukou Index: above median × NonAgri are Hukou Index: below median × Edler60 × NRPS and Hukou Index: above median × Edler60 × NRPS. The instruments for the specification in column (6) are defined accordingly. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

With the heterogeneity of migration costs across different rural areas in China, the IV estimate captures the weighted average of the LATEs across rural areas, with the weight of an area proportional to the number of workers who are at the margin between migrating and not-migrating. Again, when the NRPS-induced shift in migration cost is sufficiently small, the IV estimate reflects the weighted average of baseline migration costs faced by the NRPS-induced switchers across the rural areas. To further shed light on this interpretation, we group villages into two groups depending on whether the average Hukou Index (which is negatively related to migration barriers) faced by out-migrants in 2009-2012 is above or below the median, and estimate the LATE specific to each group. Column (5) reports the IV regression results. The IV estimates imply that, among the compliers, working in the non-agricultural sector increases annual earnings by 89 log points in regions with high baseline migration cost (i.e., with Hukou Index below median). The corresponding effect is 72 log points for regions with low baseline migration cost (i.e., with Hukou Index above median). In column (6), we further divide the villages into terciles based on the baseline Hukou Index. It is reassuring to find that the IV estimates diminish monotonically with the baseline migration cost.

4.4 Control Function Estimation of Underlying APG

In this subsection, we adopt the approach of Card (2001) and Cornelissen et al. (2016) to estimate the underlying APG using the control function approach.

With the assumption that U_{na} and U_a follow a joint normal distribution, we can estimate equation (8) by the following regression:

$$\begin{aligned}\ln y_{ihjt} =& \gamma_1 NonAgri_{ihjt} + X_{ihjt}\gamma_2 \\ & + \gamma_3 NonAgri_{ihjt} \times \frac{\phi((Z_{ihjt}, W_{ihjt})\zeta)}{\Phi((Z_{ihjt}, W_{ihjt})\zeta)} \\ & + \gamma_4(1 - NonAgri_{ihjt}) \times \frac{\phi((Z_{ihjt}, W_{ihjt})\zeta)}{1-\Phi((Z_{ihjt}, W_{ihjt})\zeta)} + \varphi_j + \varphi_{pt} + \omega_{ihjt},\end{aligned}$$

where Z_{ihjt} corresponds to $Elder60_{hjt} \times NRPS_{jt}$, W_{ihjt} contains all the control variables (including X_{ihjt}, province×year dummies, and village dummies) and ζ is a vector of estimates obtained from the first-stage probit estimation of the selection equation. The control functions $NonAgri \times \frac{\phi((Z,W)\zeta)}{\Phi((Z,W)\zeta)}$ and $(1 - NonAgri) \times \frac{\phi((Z,W)\zeta)}{1-\Phi((Z,W)\zeta)}$ account for the selection bias.[13] Hence, theoretically, $\hat{\gamma}_1^{CF}$ captures the ATE (Wooldridge, 2015; Cornelissen et al., 2016).

Column (1) in Table 6 shows our benchmark estimate of γ_1 using the control function (CF) approach. The CF estimate suggests that annual earnings of the non-agricultural sector is on average 46 log points higher than that of the agricultural sector for workers with average characteristics.

[13] $(Z, W)\zeta$ maps to $R - m(X, Z)$ in the selection terms in the framework of Section 3. In particular, m is a function of Z and W, and R is absorbed by the constant term in W.

Table 6: Sector of Employment and Annual Earnings: Control Function Approach

Dep Var: Ln Annual Earnings	(1) CF	(2) CF	(3) CF	(4) CF
NonAgri	0.4636***	0.4584***	0.3813**	0.4065**
	(0.0334)	(0.1543)	(0.1639)	(0.1604)
NonAgri$\times\frac{\phi((Z,X)\beta)}{\Phi((Z,X)\beta)}$	-0.0132			
	(0.0178)			
(1-NonAgri)$\times\frac{\phi((Z,X)\beta)}{1-\Phi((Z,X)\beta)}$	-0.3040***			
	(0.0245)			
Residual		0.4851***	0.5597***	0.2729*
		(0.1555)	(0.1649)	(0.1639)
Residual × NonAgri		-0.4663***	-0.4664***	-0.2040***
		(0.0351)	(0.0352)	(0.0726)
Residual × Z			0.0032	0.0490
			(0.0605)	(0.1385)
Residual × NonAgri × Z			0.1041	0.1056
			(0.0778)	(0.1963)
Residual^2				-0.5454***
				(0.0591)
Residual^2 × NonAgri				0.5658***
				(0.0858)
Residual^2 × Z				0.1908
				(0.2560)
Residual^2 × NonAgri × Z				-0.2512
				(0.3110)
First-stage specification	Probit	Linear + interactions with Z	Linear + interactions with Z	Linear + interactions with Z
Individual and household controls	Y	Y	Y	Y
NonAgri × Centered individual and household controls	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y
Village FE	Y	Y	Y	Y
Observations	232,961	234,025	234,025	234,025
R-squared	0.3881	0.3923	0.3923	0.3930

Notes: The first-stage specification in column (1) include the IV (NRPS×Elder60), and control variables in the vector X_{ihjt}: four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, type of Hukou, a dummy indicating whether there is an elderly aged 60 or above residing in the household, and the share of months in year t that the NRPS has been in effect. The first stage specification in columns (2)-(4) additionally includes the interaction between the IV and X_{ihjt}. Individual controls include all variables in the vector X_{ihjt}. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

We then extend the control function model in several dimensions so that it depends less on functional form restrictions and demands a less stringent identification assumption. First, we estimate the first-stage selection equation by extending equation (14) with the interactions between the instrument and controls (except for the village and province-year fixed effects), which allows the NRPS to affect migration decisions in a more non-parametric way.[14] Using the residuals obtained from this augmented model ($\hat{\nu}_{ihjt}$), we estimate the following second-stage regression:

$$\ln y_{ihjt} = \gamma_1 NonAgri_{ihjt} + X_{ihjt}\gamma_2 + \eta NonAgri_{ihjt} \times \hat{\nu}_{ihjt} + \psi\hat{\nu}_{ihjt} + \varphi_j + \varphi_{pt} + u_{ihjt}. \tag{16}$$

Under the identification assumption that

$$E[U_{a,ihjt}|\nu_{ihjt}] = \psi\nu_{ihjt} \quad \text{and} \quad E[U_{na,ihjt} - U_{a,ihjt}|\nu_{ihjt}] = \eta\nu_{ihjt}, \tag{17}$$

the estimated coefficient γ_1 reflects the ATE. The regression result is reported in column (2). Second, as is pointed out in Card (2001), in a general setting, changes in the instrumental variable may affect the entire mapping between unobserved abilities and the outcome of interest, which leads to a violation of assumption (17).[15] Following Card (2001), to address the problem, column (3) extends the control function approach by adding an interaction term of the residual with $NonAgri$, and a three-way interaction with $NonAgri \times Z$. Third, in column (4), we further include the quadratic term of the residual, and the corresponding interactions with $NonAgri$, Z, and $NonAgri \times Z$. This specification relaxes the linearity assumption in (17) (Wooldridge, 2015). Across these extended models, the estimates of γ_1 remain stable and range from 0.38 to 0.46.

The control function approach provides an estimate of the underlying agricultural productivity gap in China. Our results show that it is significantly positive. Comparing our benchmark CF estimate of 46 log points with the counterpart OLS estimate of 68 log points (in column (1) of Table 3) implies that the sorting of workers based on unobserved characteristics accounts for 32% (1 - 0.46/0.68) of the observed APG in the NFP data.

[14] We use age group dummies and education group dummies to capture the effects of age and education on the migration decision non-parametrically.

[15] To be clear, in this case,

$$Cov(U_a, \nu|Z=1) \neq Cov(U_a, \nu|Z=0), \quad Cov(U_{na} - U_a, \nu|Z=1) \neq Cov(U_{na} - U_a, \nu|Z=0),$$

which violates assumption (17). Nevertheless, a simple extension of the control function is appropriate with the identification assumption being:

$$E[U_a|\nu] = \eta_0(1-Z)\nu + \eta_1 Z\nu \quad \text{and} \quad E[U_{na} - U_a|\nu] = \psi_0(1-Z)\nu + \psi_1 Z\nu.$$

4.5 Summary

We now take stock of what we have learned from our reduced form estimation results. First, the OLS cross-sectional regression shows that the observed APG in China is 68 log-points, after we control for sectoral differences in observable worker characteristics. Note that this is the difference in average labor productivity between migrant workers and workers in agriculture. If we include the workers with urban *hukou*, the observed APG would be even higher. Second, in contrast to the recent findings for several other countries, the observed APG is virtually the same if we also control for individual fixed effects. We argue that this is likely due to the high barriers to migration, and therefore high returns to migration are needed to induce migration in China. Third, we estimate the local treatment effect of migration induced by the NRPS policy and find that the incomes of NRPS-induced migrants on average increased by 79 log points, which implies that, before policy implementation, the average of the migration costs faced by these migrants were indeed high, around 55% of their annual potential non-agricultural earnings. Fourth, when we divide the sample villages into three groups based on their Hukou Indices and estimate the LATE separately for each group, we find that the LATE estimate is lower for villages with higher Hukou Indices, suggesting that areas facing more liberal *hukou* policies have lower migration costs. Finally, we also use the NRPS policy as an instrument and use the control function approach to estimate the average treatment effect of migration or the underlying APG. Different specifications of the control function all yield significantly positive underlying APG, ranging from 38 log points to 46 log points. Comparing to the OLS estimate of the observed APG reveals that the underlying APG accounts for more than half of the observed APG between agricultural workers and migrant workers in China.

5 Structural Estimation

Although our IV estimate is informative about the causal impact of the NRPS-induced migration on labor income for compliers, it is a local result that may lack external validity. In this section, we estimate a structural Roy model so that we can evaluate the overall impact of the migration policy change. In the structural model, we also allow for migration cost to be a function of the Hukou Index and individual characteristics to capture migration cost heterogeneity across individuals and locations. Our reduced form analysis in the previous section shows that there are sector switchers in both directions, from agriculture to non-agriculture and vice versa, and that those switching from non-agriculture to agriculture generally experience income losses. To better capture these migration dynamics, we also extend our static generalized Roy model from Section 3 by adding the time dimension, introducing idiosyncratic shocks to migration costs and human capital that are i.i.d. across individuals and time, and allowing for differential wage growth in the two sectors.

5.1 Model

Since the model is similar to the model described in Section 3, we just highlight the differences. The real income an individual i receives in sector j at time t is, $y_{j,it} = w_{j,t}h_{j,it}$, for $j = a, na$. Here, $h_{j,it}$ is the efficiency units of labor of the individual and $w_{j,t} = p_{j,t}A_{j,t}$ is the real wage per efficiency unit of labor in sector j at time t, which grows at a constant but sector-specific rate:

$$\ln w_{j,t} = \ln w_j + g_j t, \quad \text{for } j \in \{a, na\}. \tag{18}$$

Thus, the underlying APG is $R_t = \ln(w_{na,t}/w_{a,t}) = \ln(w_{na}/w_a) + (g_{na} - g_a)t$, which may vary over time. The labor efficiency $h_{j,it}$ is assumed to take the following form:

$$h_{j,it} = \exp(\mathbf{X}_{it}\beta + u_{j,i} + \epsilon_{j,it}) \quad \text{for } j \in \{a, na\}. \tag{19}$$

The observable component $\mathbf{X}_{it}$ includes gender, years of schooling, age, and age squared. The unobservable includes a time-invariant component $u_{j,i}$, representing the innate ability of individual i working in sector j. We assume that $(u_{a,i}, u_{na.i})$ is i.i.d. across individuals and follows a bi-variate normal distribution $N(0, \Sigma_u)$, where the standard deviation of $u_{j,i}$ is σ_j and the correlation between $u_{a,i}$ and $u_{na,i}$ is ρ. Different from the static model in Section 3, we introduce productivity shocks, $\epsilon_{a,it}$ and $\epsilon_{na,it}$, in the structural model. We assume that $\epsilon_{a,it}$ and $\epsilon_{na,it}$ are independent of each other, i.i.d across individuals and time, and follow a bi-variate normal distribution $N(0, \Sigma_\epsilon)$.

The migration cost is assumed to take the linear form, $M_{c,it} = (\mathbf{X}_{it}, \mathbf{Z}_{it})\zeta + u_{c,it}$. Here, $\mathbf{X}_{it}$ includes the same set of observed individual characteristics as in the human capital equation. $\mathbf{Z}_{it}$ includes a constant term and two policies: One is the Hukou Index that captures the weighted average of the lenience of *hukou* policies of potential destination cities. The other policy is the NRPS, and we include the share of months in year t since its introduction to the county, whether there is an elderly above age 60 in the household, and the interaction of the two. Finally, $u_{c,it}$ is an idiosyncratic shock to migration cost, which is assumed to be i.i.d. over time and follows a standard normal distribution $N(0, \sigma_c^2)$.

Worker i chooses sector $j \in \{a, na\}$ at each time t to maximize her income net of migration cost. We assume individuals observe their migration cost shocks, but do not observe their productivity shocks when they make their migration decision. We also assume workers are risk neutral, so the migration decision is carried out according to the following rule:

$$D_{it} = \begin{cases} 1, & \text{if } E[y_{na,it}] > E[y_{a,it}]\exp(-M_{c,it}); \\ 0, & \text{otherwise.} \end{cases}$$

Here, the expectations are taken with respect to the productivity shocks, $\epsilon_{a,it}$ and $\epsilon_{na,it}$. Since we have assumed that the shocks are normally distributed with mean zero, we know that $E[y_{j,it}] = \exp(E[\ln y_{j,it}] + \frac{1}{2}\sigma^2_{j,\epsilon})$, $j = a, na$. Thus, the decision rule can be rewritten as follows:

$$D_{it} = \begin{cases} 1, & \text{if } E[\ln y_{na,it}] - E[\ln y_{a,it}] > M_{c,it} + \dfrac{1}{2}(\sigma^2_{a,\epsilon} - \sigma^2_{na,\epsilon}); \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

5.2 Identification and Estimation

We obtain the identification of our Roy model by using the panel data and an instrumental variable. Eisenhauer et al. (2015) proves that the full marginal treatment effect (MTE) curve of a generalized Roy model can be identified using a continuous instrument that provides sufficient variation in the migration costs. Given that our instrument is a discrete variable that does not provide sufficient variation in migration costs for recovering the full MTE curve, we are unable to identify the model non-parametrically with the instrument alone. Therefore, we make use of the panel data and functional form assumptions to further identify the model. Pulido and Świecki (2018) shows that with longitudinal data and functional form assumptions, the generalized Roy model is identified. Panel data on the earnings of stayers in the agricultural (non-agricultural) sector identify the distribution of agricultural (non-agricultural) ability; panel data on the earnings of individuals who switch sectors identify the correlation between agricultural and non-agricultural abilities. The migration cost is identified from the share of workers working in non-agriculture, and the standard deviation of migration cost shocks is identified from the share of workers who switch from non-agriculture to agriculture.[16]

We use Maximum Likelihood to estimate the model. To facilitate the interpretation of the results, we demean all the observables by $\tilde{\mathbf{X}}_{it} = \mathbf{X}_{it} - E(\mathbf{X})$ and $\tilde{\mathbf{Z}}_{it} = \mathbf{Z}_{it} - E(\mathbf{Z})$. Therefore, from equation (18), the log real income of individual i in period t working in sector j becomes $\ln \tilde{w}_j + g_j\tilde{t} + \tilde{\mathbf{X}}_{it}\beta + u_{j,i}$, where $\ln \tilde{w}_j = \ln w_j + E(\mathbf{X})\beta$.

[16] For workers who have been working in non-agriculture, the only reason for them to move back to agriculture would be an increase in the cost working in non-agriculture.

The likelihood is defined as the following

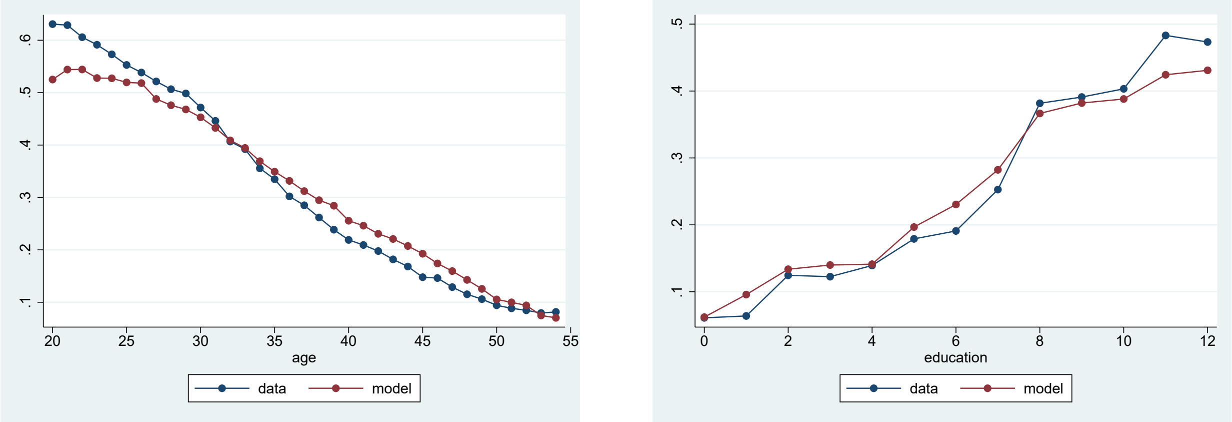
$$
\begin{aligned}
L = \prod_i \prod_t \int & \phi_{\epsilon_j}(\ln y_{j,it} - \ln \tilde{w}_j - g_j \tilde{t} - \tilde{\mathbf{X}}'_{it}\beta - u_{j,i}) \\
& \Phi_{u_c}(R_t + u_{na,i} - u_{a,i} - (\tilde{\mathbf{X}}'_{it}, \tilde{\mathbf{Z}}'_{it})\zeta)^{\mathbf{1}\{D_{it}=1\}} \\
& (1 - \Phi_{u_c}(R_t + u_{na,i} - u_{a,i} - (\tilde{\mathbf{X}}'_{it}, \tilde{\mathbf{Z}}'_{it})\zeta))^{\mathbf{1}\{D_{it}=0\}} d\Phi_{u_{a,i}, u_{na,i}},
\end{aligned}
$$

where ϕ_x and Φ_x are the PDF and CDF of the standardized Normal distribution for variable x, and $y_{j,it}$ is the observed annual real income of a worker i in sector j and year t.

5.3 Model Fit

We show the model fit on sector choice, agricultural income, and non-agricultural income by age, education, gender, and whether the county has the NRPS in Figures 3 to 7. In general, the model fits the data quite well. For example, our model predicts that the share of workers with rural hukou working in the non-agricultural sector (migrant worker share) declines with age, increases with education, and is higher for men than for women. The model also fits well the gender wage gap, education wage premium, and life cycle wage profile for rural workers in the agricultural and non-agricultural sectors. Moreover, the model is able to match the time trends in the migrant worker share and raw APG, as shown in Table 7. Finally, We also use the data simulated from our model to run an OLS regression and an individual fixed-effect regression to estimate the APG, following the same specifications as those in our reduced form analysis. The estimates are reported in Table 8. They are close to the estimates we get using the actual data in Section 4.

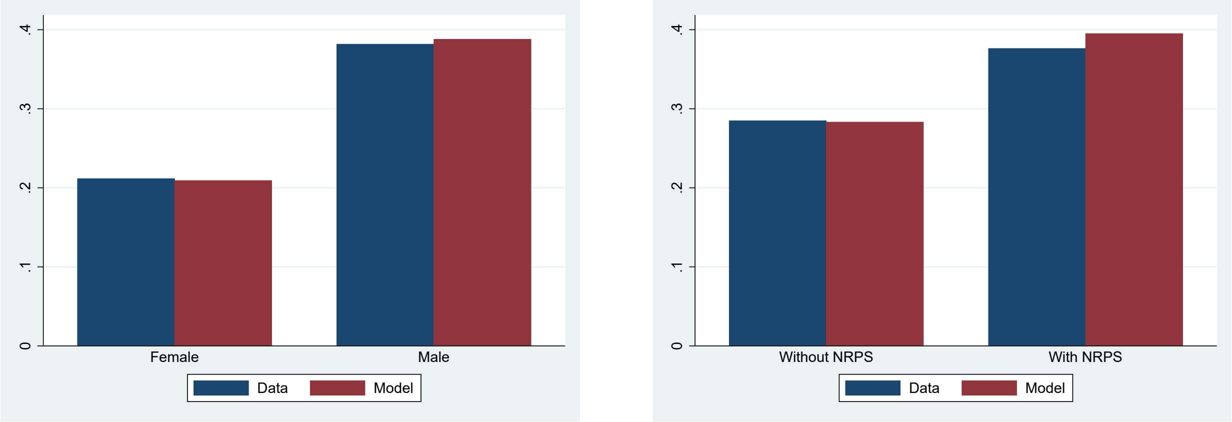
Figure 3: Fraction in the Non-agriculture



(a) Proportion by age

(b) Proportion by education

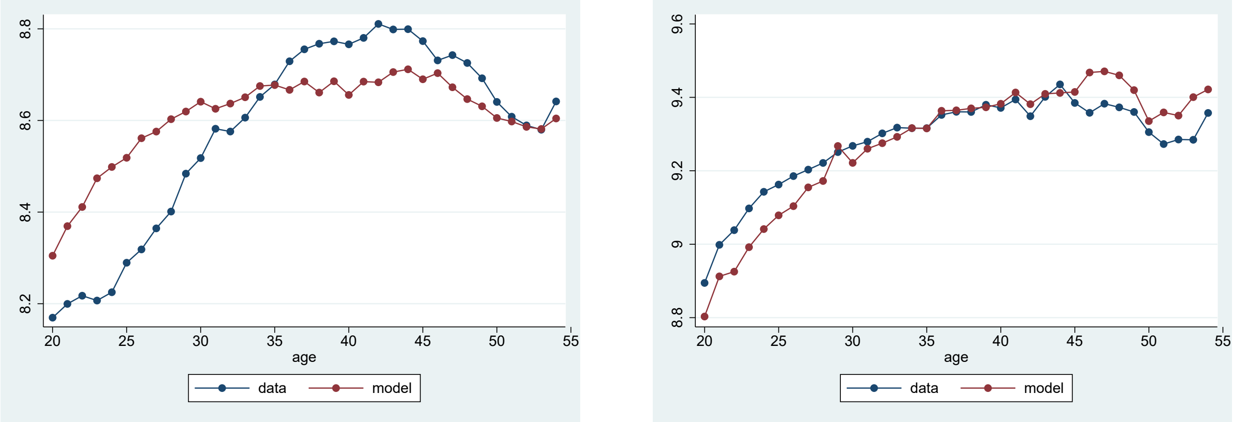
Figure 4: Fraction in Non-agriculture



(a) Proportion by gender

(b) Proportion with/without NRPS

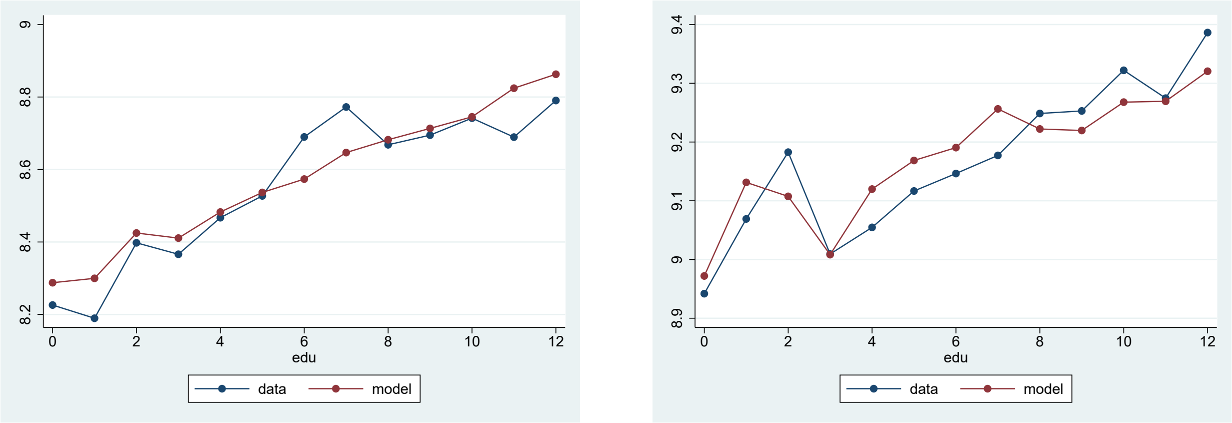
Figure 5: Log Earnings by Age



(a) Agriculture

(b) Non-agriculture

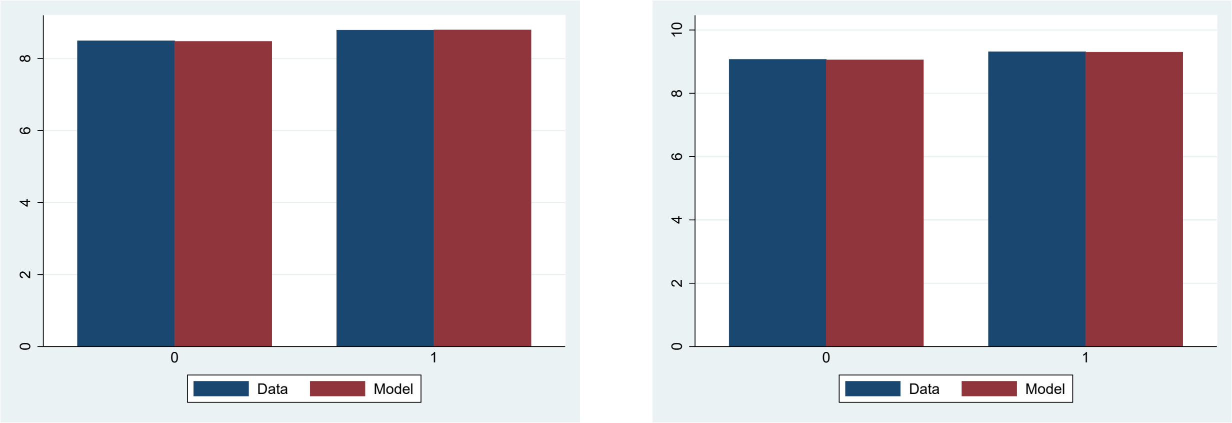
Figure 6: Log Earnings by Education



(a) Agriculture

(b) Non-agriculture

Figure 7: Log Earnings by Gender



(a) Agriculture

(b) Non-agriculture

Table 7: Model Fit: Share of Migrant Workers and Raw APG Over Time

	Share of migrant workers		Raw APG	
	Data	Model	Data	Model
2003	0.216	0.200	0.526	0.452
2004	0.233	0.228	0.343	0.444
2005	0.256	0.253	0.467	0.478
2006	0.283	0.271	0.531	0.484
2007	0.298	0.300	0.475	0.512
2008	0.314	0.319	0.558	0.503
2009	0.368	0.361	0.734	0.529
2010	0.365	0.386	0.497	0.545
2011	0.374	0.395	0.580	0.588
2012	0.375	0.407	0.683	0.614
Total	0.302	0.304	0.600	0.595

Notes: The first two columns report the share of workers with rural hukou working in the urban non-agricultural sector in the data and model, respectively. The next two columns report the raw APG in the data and model, respectively.

Table 8: Model Fit: Sectoral Income Gap

	(1)	(2)	(3)	(4)
	OLS	OLS	Individual FE	Individual FE
	Data	Model	Data	Model
NonAgri	0.681***	0.475***	0.692***	0.705***
	(0.012)	(0.005)	(0.014)	(0.007)
Individual and household controls	Y	Y	Y	Y
Province x Year FE	Y	Y	Y	Y
Individual FE	N	N	Y	Y
Observations	234,025	23,402,500	234,025	23,402,500

Notes: Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

5.4 Estimation Results

Table 9 shows the parameter estimates related to human capital and real wages. The log of the real wage level in the agricultural and non-agricultural sectors are 9.172 and 8.579, respectively, which implies an average underlying APG of 59.3 log points. It is slightly higher than the reduced form estimate using the control function approach (59.3 vs. 46.4). The growth rates of real wages in the agricultural and non-agricultural sectors is 7.4 and 9.9 log points per annum, respectively, which suggests an annual differential growth rate of 2.5 log points.[17] The wage premium for men (compared to women) is 24.0 log points. The return to education is 2.7 log points. The life-cycle human capital has a hump shape, with a peak at age 44.

Table 9: Parameter Estimates (Human Capital)

	Agri	NonAgri
Real wage level	8.579 (0.895)	9.172 (0.850)
Real wage growth rate	0.074 (0.005)	0.099 (0.002)
Human capital – male	0.240 (0.017)	
Human capital – years of education	0.027 (0.001)	
Human capital – age	0.070 (0.003)	
Human capital – age squared	-0.0008 (0.0000)	
SD of wage shock	0.781 (0.134)	0.390 (0.023)
SD of ability	0.576 (0.051)	0.515 (0.049)
Correlation between abilities	0.857 (0.080)	

Notes: Standard errors in parentheses.

Table 9 also shows the estimates on the innate ability distribution. Agricultural ability has a larger standard deviation than non-agricultural ability (57.6 vs. 51.5 log points). We also find a strong positive correlation between agricultural and non-agricultural abilities (0.857). The productivity shock also has a larger standard deviation in agriculture than in non-agriculture (78.1 vs. 39.0 log points). This, together with the larger standard deviation of agricultural ability, explains the larger variance in agricultural income.

[17] In Appendix D.3, we estimate the differential nominal productivity growth across sectors based on the sample of workers who remain the same sector in both $t-1$ and t periods. The estimate is 1.3 log points, which is consistent with the structural estimate.

Table 10 shows the parameter estimates related to migration costs. After demeaning the observables, the constant term of the migration costs reflects the average migration cost. The constant term is 0.566, suggesting that average migration cost accounts for 43% of non-agricultural earnings ($1 - \exp(-0.566)$). The structural estimate of average migration costs is somewhat smaller than the LATE estimate we reported earlier (56.6 vs. 78.6 log points), suggesting that the workers who were induced to migrate by the NRPS may face higher migration costs than that of the average rural worker. Moreover, we find that the migration costs are lower for men, highly educated workers, younger workers, and workers with an elderly over age 60. Besides, Table 10 shows that the Hukou Index has a profound effect on migration costs. A one standard deviation increase in the Hukou Index reduces migration costs by 13% ($0.831 * 0.089/0.566$).[18] Moreover, the NRPS leads to a reduction in the annual migration costs for those with an elderly over age 60 in the household by 2.5% ($(0.036 - 0.022)/0.566$). This is consistent with the literature as the eldercare and childcare effects are mostly observed for the elderly over age 60, who are eligible to receive pension benefits. This result is also consistent with our reduced form IV estimation as we find that most of the effects of the NRPS on migration decision is concentrated on young adults from households with an elderly over age 60.

Table 10: Parameter Estimates (Migration Costs)

	Migration cost
Constant	0.566
	(0.056)
Hukou Index	-0.831
	(0.035)
Male	-0.171
	(0.004)
Years of education	-0.017
	(0.001)
Age	0.000
	(0.000)
Age square	0.000
	(0.000)
NRPS	0.022
	(0.001)
Elderly above 60	-0.008
	(0.000)
NRPS × elderly	-0.036
	(0.000)
SD of migration cost shock	0.157
	(0.007)

Notes: Standard errors in parentheses.

[18] The Hukou Index has a mean of 0.123 and a standard deviation of 0.089.

5.5 Selection

Estimating the model allows us to better understand the selection in terms of unobserved abilities. Table 11 summarizes the average agricultural and non-agricultural abilities by sector choice, and Figure 8 presents the distribution of abilities with the mean abilities normalized to zero. The average agricultural ability of agricultural workers is 6.3 log points, and the average non-agricultural ability of non-agricultural workers is 0.5 log points. Hence, the selection bias is -5.8 log points (0.5 - 6.3), which is relatively small compared to that of the underlying APG (58.5 log points).

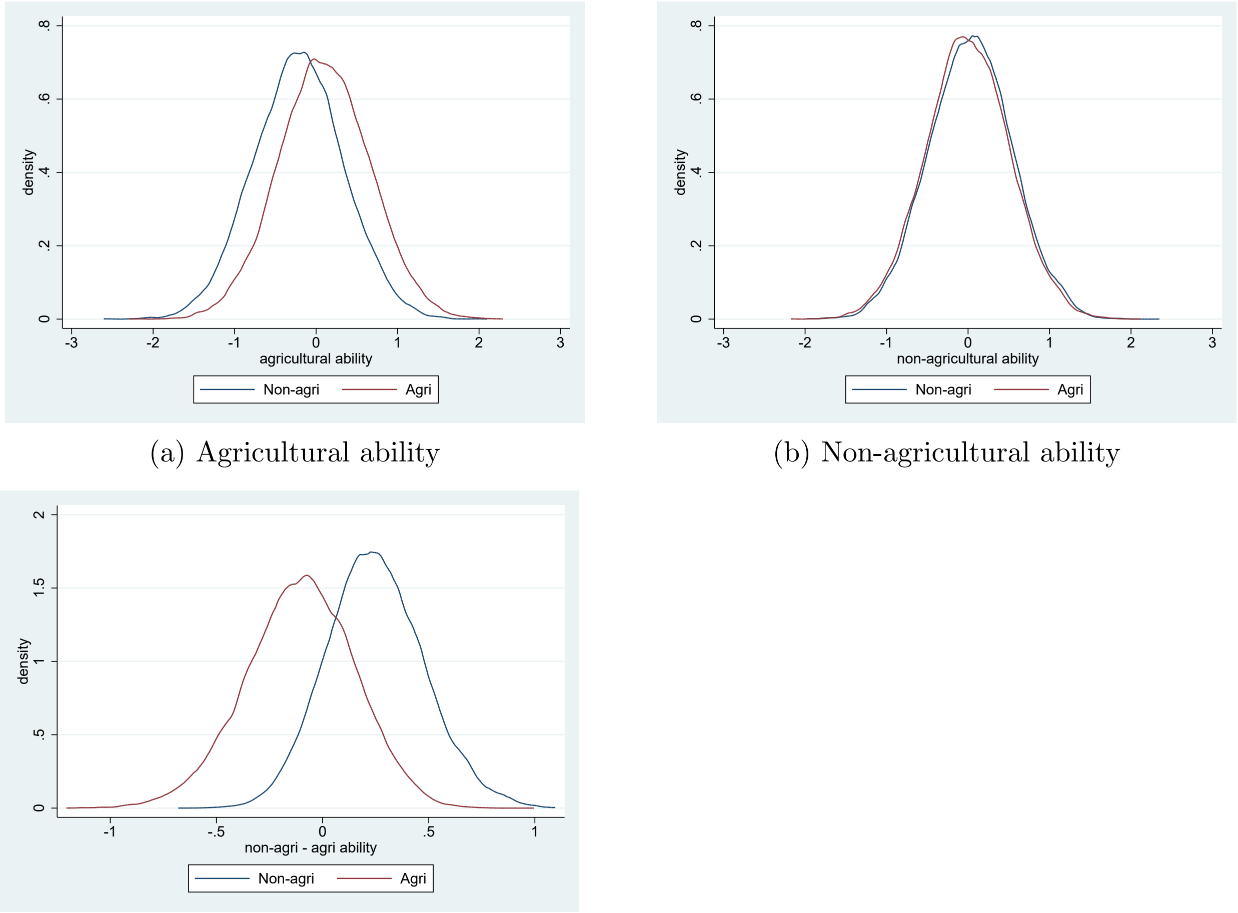
Table 11: Average Ability by Sector Choice

	Agri sector	NonAgri sector
Agri ability	0.063	-0.148
NonAgri ability	-0.002	0.005
NonAgri ability - Agri ability	-0.065	0.154

The average agricultural ability of agricultural workers is higher than that of non-agricultural workers (6.3 vs. -14.8 log points). This implies that there exists negative selection in agricultural ability for migrants in the sense that they have lower agricultural ability than non-migrants. Meanwhile, the average non-agricultural ability of non-agricultural workers is slightly lower than that of non-agricultural workers (-0.2 vs. 0.5 log points). This suggests that there exists positive selection in non-agricultural ability for migrants. The selection in non-agricultural ability is much weaker than the selection in agricultural ability.

In addition to the selection in absolute advantage, we further analyze the selection in terms of comparative advantage. We show the distribution of the relative ability (non-agricultural ability minus agricultural ability) in Figure 8 and Table 11. Overall, people with higher relative ability in non-agriculture work in non-agriculture and vice versa, suggesting that there exists positive selection in comparative advantage. Our estimated structural parameters in Table 9 suggest that absolute advantage and comparative advantage are positively correlated in both sectors. Specifically, $Cov(U_a, U_a - U_{na}) = \sigma_a^2 - \sigma_{na,a} = 0.08$ and $Cov(U_{na}, U_{na} - U_a) = \sigma_{na}^2 - \sigma_{na,a} = 0.01$. Since $Cov(U_a, U_a - U_{na}) > 0$ and $Cov(U_{na}, U_{na} - U_a) > 0$, comparative advantage aligns with absolute advantage. The covariance is higher for the agricultural sector, which echoes the finding that the selection on absolute productivity is stronger in agriculture.

Figure 8: Ability distribution for Agri and Non-agri workers



(a) Agricultural ability

(b) Non-agricultural ability

(c) Non-agri - Agri ability

5.6 Robustness

We also estimate a model that allows for differential returns to human capital. In other words, we allow β to vary across sectors. The results are shown in Appendix Tables D.5 and D.6. We find that the return to education is lower in the agricultural sector (2.0 vs. 3.0 log points), which is consistent with the findings of Herrendorf and Schoellman (2018) for other countries, but the difference is small. The average underlying APG is 48.5 log points and the average migration cost is 62.2 log points, similar to the estimates from the baseline model. Therefore, in the following analysis, we use the baseline model that assumes no differential returns to human capital between the two sectors.

6 General Equilibrium Analysis

We now embed our structural model into a three-sector general equilibrium model. The three sectors are rural agriculture, urban non-agriculture as described in Sections 3 and 5, and rural non-agriculture, indexed by $j = rn$. We introduce this third sector because, in the data, there is a non-trivial share of rural workers working in non-agricultural jobs in the rural areas. Let $L_{r,t}$ and $L_{u,t}$ be the total populations with rural and urban *hukou*, respectively.

6.1 Preferences and Consumption Demands

All individuals have identical Stone-Geary utility functions:

$$u\left(c_{a,t}, c_{na,t}\right) = \left(c_{a,t} - \bar{c}\right)^{a} \left(c_{na,t}\right)^{1-a},$$

where $c_{a,t}$ and $c_{na,t}$ represent consumption of agricultural and non-agricultural goods, respectively, and $0 < a < 1$. The consumption demands of an individual with nominal income I are

$$c_{a,t}(I) = \bar{c} + a\frac{(1-s_t)I - P_{a,t}\bar{c}}{P_{a,t}}, \tag{21}$$

$$c_{na,t}(I) = (1-a)\frac{(1-s_t)I - P_{a,t}\bar{c}}{P_{na,t}}. \tag{22}$$

Here, $P_{j,t}$ is the price of the sector-j good ($j = a, na$), and s_t is the savings rate. The worker's indirect utility is $V_t(I) = [(1-s_t)I - P_{a,t}\bar{c}]/P_t$, where P_t is the aggregate consumption price index, $P_t = \left(P_{a,t}/a\right)^{a}\left(P_{na,t}/(1-a)\right)^{1-a}$. Let $p_{a,t} = P_{a,t}/P_t$ and $p_{na,t} = P_{na,t}/P_t$ be the real prices of agricultural and non-agricultural goods, respectively. Then, we can write the indirect utility of a worker as a function of her real income $w = I/P$:

$$V_t(w) = (1-s_t)w - p_{a,t}\bar{c}. \tag{23}$$

6.2 Technologies and Real Wages

The technologies of the three sectors are:

$$Y_{a,t} = A_{a,t}H_{a,t}; Y_{na,t} = A_{na,t}H_{na,t}; Y_{rn,t} = A_{rn,t}H_{rn,t}.$$

Here $H_{j,t}$ is the efficiency units of labor in sector j. Thus, the real wages of the three sectors are

$$w_{a,t} = p_{a,t}A_{a,t}; w_{na,t} = p_{na,t}A_{na,t}; w_{rn,t} = p_{na,t}A_{rn,t}. \tag{24}$$

6.3 Human Capital, Migration Decisions, and Labor Allocation

Since it is very rare for a worker with urban *hukou* to work in agriculture in China, we assume in the model that every worker with urban *hukou* works in the non-agricultural sector and the average efficiency units of labor of these workers is ψ_t.

We also assume that a worker's human capital and wage per unit of human capital in rural non-agriculture are the same as those in agriculture. We make this assumption because our NFP data does not have good information about workers' earnings from rural non-agricultural jobs. For our assumption to be true, we assume $A_{rn,t} = p_{a,t}A_{a,t}/p_{na,t}$ in equilibrium so that, from (24), $w_{a,t} = w_{nr,t}$. Under this assumption, a worker who decides to stay in the rural areas will be indifferent between working in agriculture or non-agriculture. Therefore, we simply assume that an exogenous portion of workers with rural *hukou*, $\bar{\pi}_{rn,t}$, work in rural non-agriculture.

Human capital functions with rural *hukou* are the same as those described in Section 5. From equation (23), we can see that an individual's utility is linear in her real wage. Therefore, rural workers' migration decision rule is exactly the same as that in condition (20) of Section 5. We interpret the observable components of migration costs as taxes on migrant workers who work in the urban non-agricultural sector, and the tax revenues are transferred to workers with urban *hukou* as lump-sum transfers.

Let $\pi_{na,t}(\boldsymbol{X_t}, \boldsymbol{Z_t})$ be the proportion of workers with rural *hukou* and characteristics $(\boldsymbol{X_t}, \boldsymbol{Z_t})$ who migrate to the urban non-agricultural sector to work, and $\bar{\pi}_{na,t} = E\left[\pi_{na,t}(\boldsymbol{X_t}, \boldsymbol{Z_t})\right]$. Then, the employment in the three sectors are:

$$L_{a,t} = \left(1 - \bar{\pi}_{rn,t} - \bar{\pi}_{na,t}\right) L_{r,t}; L_{rn,t} = \bar{\pi}_{rn,t}L_{r,t}; L_{na,t} = L_{u,t} + \bar{\pi}_{na,t}L_{r,t}. \tag{25}$$

Let

$$\overline{h}_{a,t} = E\left[h_a(\boldsymbol{X_t}, \boldsymbol{U}) | V \geq R - m(\boldsymbol{X_t}, \boldsymbol{Z_t}) - U_{c,t}\right],$$

and let

$$\overline{h}_{na,t} = E\left[h_{na}(\boldsymbol{X_t}, \boldsymbol{U}) | V < R - m(\boldsymbol{X_t}, \boldsymbol{Z_t}) - U_{c,t}\right].$$

Then, the effective labor in the three sectors are:

$$H_{a,t} = \overline{h}_{a,t}L_{a,t}; H_{rn,t} = \overline{h}_{a,t}L_{rn,t}; H_{na,t} = \psi_t L_{u,t} + \overline{h}_{na,t}\bar{\pi}_{na,t}L_{r,t}. \tag{26}$$

6.4 Market Clearing Conditions

The aggregate nominal demand for the agricultural good is

$$P_{a,t}C_{a,t} = (1-a)\,L_t P_{a,t}\bar{c} + a(1-s_t)P_t Y_t,$$

where $P_t Y_t = P_{a,t}Y_{a,t} + P_{na,t}Y_{na,t} = P_{a,t}A_{a,t}H_{a,t} + P_{na,t}\left(A_{na,t}H_{na,t} + A_{rn,t}H_{rn,t}\right)$ is the aggregate nominal income. The aggregate nominal supply of the agricultural good is $P_{a,t}Y_{a,t} = P_{a,t}A_{a,t}H_{a,t}$. So, the market clearing condition is

$$P_{a,t}A_{a,t}H_{a,t} = (1-a)\,L_t P_{a,t}\bar{c} + a(1-s_t)\left[P_{a,t}A_{a,t}H_{a,t} + P_{na,t}\left(A_{na,t}H_{na,t} + A_{rn,t}H_{rn,t}\right)\right],$$

which is equivalent to:

$$\frac{H_{a,t}}{L_t} = \frac{1-a}{1-a(1-s_t)}\frac{\bar{c}}{A_{a,t}} + \frac{a(1-s_t)}{1-a(1-s_t)}\left[\frac{P_{na,t}A_{na,t}}{P_{a,t}A_{a,t}}\frac{H_{na,t}}{L_t} + \frac{P_{na,t}A_{rn,t}}{P_{a,t}A_{a,t}}\frac{H_{rn,t}}{L_t}\right].$$

Note that, by definition, $P_{na,t}A_{na,t} = P_{a,t}A_{a,t}e^{R_t}$, and by the labor market no-arbitrage condition in rural areas, $P_{na,t}A_{rn,t} = P_{a,t}A_{a,t}$. So, the equation above can be rewritten as

$$\frac{H_{a,t}}{L_t} = \frac{1-a}{1-a(1-s_t)}\frac{\bar{c}}{A_{a,t}} + \frac{a(1-s_t)}{1-a(1-s_t)}\left[e^{R_t}\frac{H_{na,t}}{L_t} + \frac{H_{rn,t}}{L_t}\right].$$

From (25) and (26), we then have

$$(1-\bar{\pi}_{rn,t}-\bar{\pi}_{na,t})\frac{L_{r,t}}{L_t} = \frac{1-a}{1-a(1-s_t)}\frac{\bar{c}}{A_{a,t}\bar{h}_{a,t}} + \frac{a(1-s_t)}{1-a(1-s_t)}\left[e^{R_t}\left(\frac{\psi_t}{\bar{h}_{a,t}}\frac{L_{u,t}}{L_t} + \frac{\bar{h}_{na,t}}{\bar{h}_{a,t}}\bar{\pi}_{na,t}\frac{L_{r,t}}{L_t}\right) + \bar{\pi}_{rn,t}\frac{L_{r,t}}{L_t}\right]. \tag{27}$$

6.5 Observed APG and Aggregate Real Income and Productivity

The observed APG is

$$APG_t^{\text{observed}} = \ln(P_{na,t}A_{na,t}\bar{h}_{na,t}) - \ln(P_{a,t}A_{a,t}\bar{h}_{a,t}) = R_t + \ln(\bar{h}_{na,t}/\bar{h}_{a,t}). \tag{28}$$

The aggregate real income valued in domestic prices is

$$Y_t = \frac{P_{a,t}A_{a,t}H_{a,t} + P_{na,t}\left(A_{rn,t}H_{rn,t} + A_{na,t}H_{na,t}\right)}{P_t} = \frac{P_{a,t}}{P_t}A_{a,t}\left(H_{a,t} + H_{rn,t} + e^{R_t}H_{na,t}\right).$$

Note that

$$\frac{P_{a,t}}{P_t} = a^a(1-a)^{1-a}\left(\frac{P_{na,t}}{P_{a,t}}\right)^{a-1} = a^a(1-a)^{1-a}\left(\frac{A_{a,t}}{A_{na,t}}\right)^{a-1}(e^{R_t})^{a-1}. \tag{29}$$

Thus, we have

$$Y_t = (aA_{a,t})^a\left((1-a)A_{na,t}\right)^{1-a}e^{-(1-a)R_t}\left(H_{a,t}+H_{rn,t}+e^{R_t}H_{na,t}\right). \tag{30}$$

The aggregate real GDP valued at base-year international PPP prices is

$$Y_t^{\text{PPP}} = P_a^{\text{PPP}}A_{a,t}H_{a,t} + P_{na}^{\text{PPP}}\left(A_{rn,t}H_{rn,t}+A_{na,t}H_{na,t}\right), \tag{31}$$

where P_j^{PPP} is the PPP price of sector j output, $j = a, na$.

6.6 Calibration of the Model

We calibrate the parameters of the model in two ways. First, for the parameters that affect individuals' migration behavior (i.e., migration cost parameters, distribution parameters for individual abilities, and idiosyncratic productivity and migration shocks), and the underlying productivity gap, we use the values we structurally estimated from the micro panel data in Section 5. Second, for other parameters in the general equilibrium model, we either take the values from the literature or set them to match some aggregate moments of the Chinese economy in 2012, the last year of our sample period. Our calibrations are summarized in Table 12. For notational simplicity, we suppress the time subscript. The values of all the time-dependent variables are those of 2012.

The first eight rows show parameters or moments calculated outside of the model. The savings rate (s) and agricultural employment share (L_a/L) are taken directly from the China Statistical Yearbook. The agricultural and non-agricultural GDP in 2005 international dollars ($P_a^{PPP}Y_a$ and $P_{na}^{PPP}Y_{na}$) are taken from the GGDC 10-Sector Database. The agricultural and non-agricultural GDP in 2012 domestic prices ($P_a^{2012}Y_a$ and $P_{na}^{2012}Y_{na}$) are taken from the China Statistical Yearbook. The ratio of consumption prices in the urban and rural areas are calculated based on the data posted by Carsten Holz on his webpage, which update the original series reported in Brandt and Holz (2006) to more recent years. The expenditure share on the agricultural good of rural workers (φ_a^{data}) and the proportion of rural-*hukou* workers in rural non-agriculture ($\bar{\pi}_{rn}$) are calculated from the NFP data.

The next four rows in Table 12 present the parameters estimated from our structural Roy

Table 12: Calibration of the General Equilibrium Model to the 2012 Data

Parameter	Meaning	Source	Value
s	Savings rate	China Statistical Yearbook	0.505
L_a/L	Share of workers working in the agricultural sector	China Statistical Yearbook	0.336
$P_a^{2012}Y_a$	Agricultural GDP in 2012 domestic prices (100 million RMB)	China Statistical Yearbook	49,085
$P_{na}^{2012}Y_{na}$	Non-agricultural GDP in 2012 domestic prices (100 million RMB)	China Statistical Yearbook	489,495
$P_a^{PPP}Y_a$	Agricultural GDP in 2005 international dollars	GGDC 10-Sector Database	937,360
$P_{na}^{PPP}Y_{na}$	Non-agricultural GDP in 2005 international dollars	GGDC 10-Sector Database	11,294,664
$P_{uc}^{2012}/P_{rc}^{2012}$	Ratio of consumption prices in the urban and rural areas	Brandt and Holz (2006)	1.311
φ^{data}	Expenditure share on agricultural good for rural households	National Fixed Point Survey	0.439
$\bar{\pi}_{rn}$	Share of workers with rural *hukou* working in the rural non-agricultural sector	National Fixed Point Survey	0.122
$\bar{\pi}_{na}$	Share of workers with rural *hukou* working in the urban non-agricultural sector	Estimated from the structural model	0.407
$\overline{h}_a$	Average human capital of workers with rural *hukou* working in the rural agricultural sector	Estimated from the structural model	1.886
$\overline{h}_{na}$	Average human capital of workers with rural *hukou* working in the urban non-agricultural sector	Estimated from the structural model	1.669
R	Underlying APG	Estimated from the structural model	0.703
L_r/L	Share of workers with rural *hukou*	Calibrated from Eqn (25)	0.667
ψ	Average human capital of workers with urban *hukou*	Calibrated from Eqn (33)	5.172
a	Preference weight on agricultural good	Calibrated from Eqn (35)	0.113
$\bar{c}/A_a$	Minimum agricultural consumption divided by agricultural productivity	Calibrated from Eqn (34)	0.343
$P_a^{PPP}A_a$	Real productivity level in agriculture (PPP)	Calibrated from Eqn (36)	59.953
$P_{na}^{PPP}A_{na}$	Real productivity level in non-agriculture (PPP)	Calibrated from Eqn (36)	190.037

model, including the share of workers with rural *hukou* working in the urban non-agricultural sector ($\bar{\pi}_{na}$), the average human capital of rural-*hukou* workers in either agriculture or rural non-agriculture ($\overline{h}_a$), and in the urban non-agricultural sector ($\overline{h}_{na}$), and the underlying APG (R).

The last five rows list the parameter values that we infer from the equilibrium conditions. From (25), we can calculate L_r/L from L_a/L, $\bar{\pi}_{rn}$, and $\bar{\pi}_{na}$.

In the model, we assume that the consumption price is the same in the two sectors. In the data, however, the rural and urban prices of consumption are different and we adjust for the price difference when estimating R from the data. Thus, we have

$$\frac{P_{na}A_{na}}{P_aA_a} = \frac{P_{uc}^{2012}}{P_{rc}^{2012}}e^R. \tag{32}$$

Note that

$$\frac{P_{na}^{2012}Y_{na}}{P_a^{2012}Y_a} = \frac{P_{na}\left(A_{rn}H_{rn} + A_{na}H_{na}\right)}{P_aA_aH_a}.$$

From the no-arbitrage condition for rural areas, we have $P_aA_a = P_{na}A_{rn}$, or $A_{rn} = P_aA_a/P_{na}$. Therefore, we have

$$\frac{P_{na}^{2012}Y_{na}}{P_a^{2012}Y_a} = \frac{P_aA_aH_{rn} + P_{na}A_{na}H_{na}}{P_aA_aH_a} = \frac{H_{rn} + \frac{P_{na}A_{na}}{P_aA_a}H_{na}}{H_a}.$$

Substituting (32) into the equation above yields:

$$\frac{P_{na}^{2012}Y_{na}}{P_a^{2012}Y_a} = \frac{H_{rn} + \frac{P_{uc}^{2012}}{P_{rc}^{2012}}e^RH_{na}}{H_a}. \tag{33}$$

Given the values of $P_{na}^{2012}Y_{na}/P_a^{2012}Y_a$ and $P_{uc}^{2012}/P_{rc}^{2012}$, we can recover ψ from equations (26) and (33), which equals 5.172.

The expenditure share on the agricultural good of rural workers is

$$\varphi_a^{data} = \frac{p_aC_a^r}{p_aC_a^r + p_{na}C_{na}^r} = a + \frac{(1-a)p_a\bar{c}}{(1-s)p_aA_a\overline{h}_a} = a + \frac{(1-a)\bar{c}}{(1-s)A_a\overline{h}_a}. \tag{34}$$

It implies that

$$\frac{\bar{c}}{A_a\overline{h}_a} = \frac{1-s}{1-a}\left(\varphi_a^{data} - a\right).$$

Substituting it into the the market clearing condition (27), we have

$$(1-\bar{\pi}_r-\bar{\pi}_{na})\frac{L_r}{L} = \frac{1-s}{1-a(1-s)}\left\{\varphi_a^{data} - a + a\left[e^R\left(\frac{\psi}{\overline{h}_a}\frac{L_u}{L} + \frac{\overline{h}_{na}}{\overline{h}_a}\bar{\pi}_{na}\frac{L_r}{L}\right) + \bar{\pi}_r\frac{L_r}{L}\right]\right\}, \tag{35}$$

which we use to solve for the value of a. The result is $a = 0.113$.

Given the value of a, we then choose the value of $\bar{c}/A_a$ such that the model's expenditure share on the agricultural good of rural workers matches that in the data (equation (34)). This results in a value of 0.343 for $\bar{c}/A_a$.

The real agricultural and non-agricultural productivity, $P_a^{PPP}A_a$ and $P_{na}^{\text{PPP}}A_{na}$, can be calculated from the following equation:

$$P_a^{\mathrm{PPP}} A_a = P_a^{\mathrm{PPP}} Y_a / H_a; P_{na}^{\mathrm{PPP}} A_{na} = \frac{P_{na}^{\mathrm{PPP}} Y_{na}}{e^{-R} H_{rn} + H_{na}}. \tag{36}$$

6.7 Counterfactual Experiments

We consider two types of counterfactual experiments. The first one is related to the NRPS and *hukou* policies. The second one is to hypothetically reduce migration costs or increase productivity.

For each experiment, the counterfactual underlying nominal productivity gap R' can be solved from equation (27). Given R', the observed APG, migrant share, and human capitals in the three sectors can all be directly simulated from the model. Then, from equation (30), we can calculate the relative change of the aggregate real income as follows:

$$\frac{Y'}{Y} = (\frac{A'_a}{A_a})^a (\frac{A'_{na}}{A_{na}})^{1-a} e^{-(1-a)(R'-R)} \frac{H'_a + H'_{r,na} + e^{R'} H'_{na}}{H_a + H_{r,na} + e^R H_{na}}. \tag{37}$$

For the real GDP measured in PPP terms, we have

$$\frac{Y^{\mathrm{PPP}\prime}}{Y^{\mathrm{PPP}}} = \omega_a \frac{A'_a H'_a}{A_a H_a} + \omega_{rn} \frac{A'_{rn} H'_{rn}}{A_{rn} H_{rn}} + \omega_{na} \frac{A'_{na} H'_{na}}{A_{na} H_{na}}, \tag{38}$$

where $\omega_a = P_a^{\mathrm{PPP}} Y_a / Y^{\mathrm{PPP}}$, $\omega_{rn} = P_{na}^{\mathrm{PPP}} A_{rn} H_{rn} / Y^{\mathrm{PPP}} = e^{-R} P_{na}^{\mathrm{PPP}} A_{na} H_{rn} / Y^{\mathrm{PPP}}$, and $\omega_{na} = P_{na}^{\mathrm{PPP}} A_{na} H_{na} / Y^{\mathrm{PPP}}$. Since we hold the total employment in the economy as exogenously given, equation (38) also gives us the change in real GDP per worker in PPP terms, which we will call the change in real productivity.

6.7.1 Effects of NRPS and *hukou* Policies

Table 13 reports the counterfactual results of the policy experiments. As comparison, the baseline results for 2012 are reported in the first row. The top panel reports the results under the partial equilibrium assumption that R remains the same under counterfactual polices, and the bottom panel reports the results under general equilibrium, in which R responds to policy changes to clear the goods market. Our structural estimation suggests that the NRPS policy increases the migration rate of those with an elderly at home. The first counterfactual experiment is to eliminate this policy for households with elderly. The quantitative effects of this policy change on the migrant share, observed APG, aggregate real income, and aggregate productivity are all small in both the partial and general equilibrium cases. The results suggest that the NRPS policy only has a marginal effect on migration, APG, and aggregate productivity. Setting the Hukou Index to the 2003 level for all villages in the data also produces relatively small effects.

However, a counterfactual *hukou* policy reform that sets the Hukou Index for all villages to that of the highest observed (most liberal) level in 2012, i.e. that of Heyuan prefecture in Guangdong province, yields more significant quantitative effects. The average migrant share increases from 40.7% to 62.3% under partial equilibrium and to 44.3% under general equilibrium. The observed APG declines by 6 log points under partial equilibrium and by 20 log points under general equilibrium. The aggregate real income and real productivity increase by 4.7% and 6.8%, respectively, under partial equilibrium, and by 2.0% and 1.1%, respectively, under general equilibrium. The hypothetical migration policy reform has a larger effect on the migrant share under partial equilibrium because the underlying APG does not respond to the policy change. Under general equilibrium, however, the policy induced reduction in migration costs also results in lower underlying APG due to the changes in relative prices and therefore the net migration cost reduction is much smaller. As a result, the increases in the migrant share and the aggregate real income and productivity are smaller. This counterfactual experiment suggests that there is significant heterogeneity in *hukou* policies across regions in China and that setting them on par with the most liberal policy will have a significant effect on rural-urban migration, APG, and aggregate productivity.

Table 13: Counterfactual Experiments: the NRPS and Hukou Policies

	Underlying APG	$\overline{\pi}_{na}$	$\overline{h}_a$	$\overline{h}_{na}$	Observed APG	Real income	Real productivity
Baseline	0.703	0.407	1.886	1.669	0.614	1.000	1.000
Partial equilibrium:							
Without NRPS for those with elderly	0.703	0.397	1.883	1.669	0.615	0.998	0.997
2003 *hukou* policy	0.703	0.369	1.867	1.676	0.625	0.992	0.988
Most liberal *hukou* policy for all regions	0.703	0.623	1.981	1.673	0.556	1.047	1.068
General equilibrium:							
Without NRPS for those with elderly	0.712	0.404	1.881	1.677	0.626	0.999	1.000
2003 *hukou* policy	0.735	0.400	1.884	1.671	0.646	0.997	0.998
Most liberal *hukou* policy for all regions	0.523	0.443	1.905	1.667	0.419	1.020	1.011

Notes: All the rows use sample year 2012 to analyze the effects of different policies on the share of workers with rural *hukou* who work in the urban non-agricultural sector ($\overline{\pi}_{na}$) and their average human capital ($\overline{h}_{na}$), the average human capital of workers with rural *hukou* who work in the rural agricultural sector ($\overline{h}_a$), the underlying and observed APG, and the aggregate real income and productivity in the partial equilibrium and general equilibrium. The first row is the baseline model. The second row eliminates the NRPS policy for individuals with elderly aged 60 or above. The third row sets the Hukou Index to the 2003 level for each village. The fourth row sets the Hukou Index of all villages to the highest observed (most liberal) level in 2012 (Heyuan prefecture in Guangdong province).

6.7.2 Effects of Lowering Average Migration Cost

In the next set of counterfactual experiments, we reduce the constant term of the migration cost function and keep all other parameters of the function unchanged. These experiments reduce the average migration cost and keep the distribution of relative migration costs across villages and households constant. Table 14 reports the counterfactual results when we reduce the average migration cost proportionally by 10% to 100%. Again, we report the results for both the partial equilibrium and general equilibrium cases.

Table 14: Counterfactual Experiments: Reducing Migration Costs

	Underlying APG	$\overline{\pi}_{na}$	$\overline{h}_a$	$\overline{h}_{na}$	Observed APG	Real income	Real productivity
Benchmark	0.703	0.407	1.886	1.669	0.614	1.000	1.000
Partical equilibrium, reducing average migration cost by							
10%	0.703	0.464	1.913	1.666	0.601	1.012	1.018
50%	0.703	0.684	2.013	1.668	0.545	1.058	1.085
100%	0.703	0.881	2.206	1.649	0.427	1.092	1.140
General equilibrium, reducing average migration cost by							
10%	0.655	0.416	1.887	1.674	0.566	1.005	1.003
50%	0.464	0.451	1.901	1.674	0.372	1.027	1.014
100%	0.221	0.489	1.923	1.666	0.112	1.055	1.025

Notes: All the rows use sample year 2012 to analyze the effects of reducing migration costs on the share of workers with rural *hukou* who work in the urban non-agricultural sector ($\overline{\pi}_{na}$) and their average human capital ($\overline{h}_{na}$), the average human capital of workers with rural *hukou* who work in the rural agricultural sector ($\overline{h}_a$), the underlying and observed APG, and the aggregate real income and productivity in the partial equilibrium and general equilibrium. We reduce the average migration cost proportionally from 10% to 100%.

In the partial equilibrium, as the average migration cost declines from the 2012 benchmark level (56.6 log points) to zero, the share of migrant workers increases from 40.7% to 88.1%, the observed APG declines from 61 log points to 43 log points, and aggregate real income and productivity increase by 9.2% and 14.0%, respectively. In the general equilibrium case, the reductions in the average migration cost result in lower underlying APG. As a result, the share of migrant workers only increases by 8.2 pp when the average migration cost drops to zero, and is much smaller than that in the partial equilibrium. Consequently, the aggregate real income and productivity increase by only 5.5% and 2.5%, respectively. In contrast, the observed APG declines much more significantly in general equilibrium, from 61 log points to only 11 log points. So, migration costs have a larger effect on APG, but a smaller effect on migration and aggregate productivity in general equilibrium than in partial equilibrium.

6.7.3 Effects of Increasing Sectoral Productivity

In our final set of counterfactual experiments, we examine the effects of either increasing agricultural productivity only or economy-wide productivity. For these experiments, we examine the general equilibrium effects only. The results are reported in Table 15.

The market clearing condition (27) implies that the equilibrium underlying APG, R, depends only on the level of agricultural productivity A_a and is independent of the non-agricultural productivity A_{na}. When A_a increases, household income rises due to the productivity increase. Since the income elasticity of the demand for the agricultural good is less than one under the Stone-Geary utility function, the increase in the demand for the agricultural good due to the income increase is less than the increase in the supply of the agricultural good due to the productivity increase in agriculture. For the agricultural good market to clear, employment in the agricultural sector has to decline, or the migration share has to increase. Without any change in the migration costs faced by individuals, this is only possible if the relative price of the agricultural good declines so that the underlying APG, R, increases. Thus, regardless of we increase the agricultural productivity only or increase both sectors' productivity by the same proportion, the effects on the underlying APG, migration share, and observed APG are the same: they all increases as a result of the productivity increases. The effects on aggregate income and productivity, however, are different depending on whether we increase agricultural productivity or both sectors' productivity. For example, a 50% increase in agricultural productivity only results in a 4.6% increase in aggregate income and a 5.3% increase in aggregate productivity. In contrast, a 50% increase in both sectors' productivity results in a 52% increase in aggregate income and 53% increase in aggregate productivity. The increase in non-agricultural productivity is much more important quantitatively because the GDP share of the non-agricultural sector is much larger than that of the agricultural sector.

Table 15: Counterfactual Experiments: Increasing Sectoral Productivity

	Underlying APG	Migrant worker share	$\overline{h}_a$	$\overline{h}_{na}$	Observed APG	Real income	Real productivity
Benchmark	0.703	0.407	1.886	1.669	0.614	1.000	1.000
Increasing agricultural productivity by							
10%	0.721	0.426	1.891	1.674	0.630	1.011	1.013
30%	0.751	0.455	1.906	1.670	0.656	1.030	1.035
50%	0.772	0.476	1.915	1.670	0.669	1.046	1.053
Increasing agricultural and non-agricultural productivity by							
10%	0.721	0.426	1.891	1.674	0.630	1.104	1.106
30%	0.751	0.455	1.906	1.670	0.656	1.311	1.318
50%	0.772	0.476	1.915	1.670	0.669	1.518	1.530

Notes: The upper panel analyzes the impact of increasing agricultural productivity and the bottom panel analyzes the impact of increasing both agricultural and non-agricultural productivity by the same proportion. All the rows use sample year 2012 to analyze the effects of changing productivity on the share of rural-*hukou* workers who work in the urban non-agricultural sector ($\overline{\pi}_{na}$) and their average human capital ($\overline{h}_{na}$), the average human capital of workers with rural *hukou* who work in the rural agricultural sector ($\overline{h}_a$), underlying APG, observed APG, aggregate real income, and aggregate real productivity in the general equilibrium. The productivity increases in the counterfactual experiment range from 10% to 50%.

Overall, the counterfactual experiments suggest a strong price effect. The change in the underlying productivity gap in the general equilibrium weakens the effect of a reduction in migration costs on the share of migrant workers and amplifies the effect on the raw APG. Therefore, it is important to allow for changes in prices using a general equilibrium framework.

7 Conclusion

In this paper, we use a nationally representative long-term panel data, the National Fixed Point Survey, to analyze the impact of migration costs and sorting on the agricultural productivity gap in China. Based on insights from labor economics, we use a policy experiment, the gradual implementation of the New Rural Pension Scheme, as an exogenous instrument to control for selection bias and estimate both the average migration cost and the underlying sectoral productivity difference in China. Our estimation results reveal that there are substantial migration costs and a large underlying sectoral productivity difference. For the observed agricultural productivity gap in China, we find that more than half of it can be attributed to the underlying productivity difference, with less than half of it accounted for by sorting of workers. This result is in contrast to several recent studies suggesting that sorting accounts for most of the observed agricultural productivity gap in several other countries, but it is consistent with the fact that China has an institutional arrangement, the *hukou* system, that explicitly restricts rural-to-urban migration.

We then extend our analysis by structurally estimating a general equilibrium Roy model so that we can conduct counterfactual analysis. If we implement a *hukou* reform by setting the *hukou* liberalization index in all regions of China to the level of the most liberal region, the observed agricultural productivity gap would decrease by more than 30%, the migrant share would increase by about 9%, and the aggregate productivity would increase by 1.1%. Even when we reduce the average migration cost for all migrants to zero, the migration share would increase by only 20% and the aggregate productivity would increase by only 2.6%. The modest gains in aggregate productivity is partly due to the fact that most migrant workers in China work in low-skill manufacturing and service industries. This suggests that there may be additional barriers to moving into more productive skill-intensive industries in China. Understanding what is behind these barriers is an interesting and important question for future research.

References

Adamopoulos, T., L. Brandt, J. Leight, and D. Restuccia (2017). Misallocation, selection and productivity: A quantitative analysis with panel data from China. Working Paper, National Bureau of Economic Research.

Alvarez, J. (2020). The agricultural wage gap: Evidence from Brazilian micro-data. *American Economic Journal: Macroeconomics 12*(1), 153–173.

Beegle, K., J. De Weerdt, and S. Dercon (2011). Migration and economic mobility in Tanzania: Evidence from a tracking survey. *Review of Economics and Statistics 93*(3), 1010–1033.

Benjamin, D., L. Brandt, and J. Giles (2005). The evolution of income inequality in rural China. *Economic Development and Cultural Change 53*(4), 769–824.

Brandt, L. and C. A. Holz (2006). Spatial price differences in China: Estimates and implications. *Economic Development and Cultural Change 55*(1), 43–86.

Brandt, L., T. Tombe, and X. Zhu (2013). Factor market distortions across time, space and sectors in China. *Review of Economic Dynamics 16*(1), 39–58.

Bryan, G., S. Chowdhury, and A. M. Mobarak (2014). Underinvestment in a profitable technology: The case of seasonal migration in Bangladesh. *Econometrica 82*(5), 1671–1748.

Bryan, G. and M. Morten (2019). The aggregate productivity effects of internal migration: Evidence from Indonesia. *Journal of Political Economy*.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica 69*(5), 1127–1160.

Caselli, F. (2005). Accounting for cross-country income differences. *Handbook of Economic Growth 1*, 679–741.

Chan, K. W. (2019). China's hukou system at 60: Continuity and reform. *in Ray Yep, Jun Wang, Thomas Johnson (eds.) Edward Elgar Handbook on Urban Development in China, Edward Elgar, pp.59-79.*.

Chari, A., E. Liu, S.-Y. Wang, and Y. Wang (2020). Property rights, land misallocation and agricultural efficiency in china. *Review of Economic Studies*.

Chen, X., K. Eggleston, and A. Sun (2018). The impact of social pensions on intergenerational relationships: Comparative evidence from china. *The Journal of the Economics of Ageing 12*, 225–235.

Cornelissen, T., C. Dustmann, A. Raute, and U. Schönberg (2016). From LATE to MTE: Alternative methods for the evaluation of policy interventions. *Labour Economics 41*, 47–60.

Donovan, K. and T. Schoellman (2020). The role of labor market frictions in structural transformation. Working Paper.

Eggleston, K., A. Sun, and Z. Zhan (2016). The impact of rural pensions in China on labor migration. *The World Bank Economic Review 32*(1), 64–84.

Eisenhauer, P., J. J. Heckman, and E. Vytlacil (2015). The generalized roy model and the cost-benefit analysis of social programs. *Journal of Political Economy 123*(2), 413–443.

Fan, J. (2019). Internal geography, labor mobility, and the distributional impacts of trade. *American Economic Journal: Macroeconomics 11*(3), 252–88.

Gollin, D., D. Lagakos, and M. E. Waugh (2014). Agricultural productivity differences across countries. *American Economic Review 104*(5), 165–70.

Gollin, D., S. Parente, and R. Rogerson (2002). The role of agriculture in development. *American Economic Review 92*(2), 160–164.

Hamory, J., M. Kleemans, N. Y. Li, and E. Miguel (2021). Reevaluating agricultural productivity gaps with longitudinal microdata. *Journal of the European Economic Association*.

Hao, T., R. Sun, T. Tombe, and X. Zhu (2020). The effect of migration policy on growth, structural change, and regional inequality in China. *Journal of Monetary Economics 113*, 112–134.

Heckman, J. J. and B. E. Honore (1990). The empirical content of the Roy model. *Econometrica 58*(5), 1121–1149.

Herrendorf, B. and T. Schoellman (2018). Wages, human capital, and barriers to structural transformation. *American Economic Journal: Macroeconomics 10*(2), 1–23.

Hsieh, C.-T. and P. J. Klenow (2009). Misallocation and manufacturing tfp in china and india. *Quarterly Journal of Economics 124*(4), 1403–1448.

Huang, W. and C. Zhang (2020). The power of social pensions: Evidence from China's new rural pension scheme. *American Economic Journal: Applied Economics*.

Imbert, C. and J. Papp (2020). Costs and benefits of rural-urban migration: Evidence from india. *Journal of Development Economics 146*, 102473.

Jiao, N. (2016). Does public pension affect intergenerational support in rural China? *Population Research (in Chinese) 4*, 88–102.

Kim, R. and J. Vogel (2020). Trade and welfare (across local labor markets). Technical report, National Bureau of Economic Research.

Kinnan, C., S.-Y. Wang, and Y. Wang (2018). Access to migration for rural households. *American Economic Journal: Applied Economics 10*(4), 1–43.

Lagakos, D. (2020). Urban-rural gaps in the developing world: Does internal migration offer opportunities. *Journal of Economic Perspectives 34*(3), 174–192.

Lagakos, D., S. Marshall, M. Mobarak, C. Vernot, and M. E. Waugh (2020). Migration costs and observational returns to migration in the developing world. *Journal of Monetary Economics 113*, 138–154.

Lagakos, D., M. Mobarak, and M. E. Waugh (2018). The welfare effects of encouraging rural-urban migration. Working Paper, National Bureau of Economic Research.

Lagakos, D. and M. E. Waugh (2013). Selection, agriculture, and cross-country productivity differences. *American Economic Review 103*(2), 948–80.

Munshi, K. and M. Rosenzweig (2016). Networks and misallocation: Insurance, migration, and the rural-urban wage gap. *American Economic Review 106*(1), 46–98.

Nakamura, E., J. Sigurdsson, and J. Steinsson (2016). The gift of moving: Intergenerational consequences of a mobility shock. Technical report, National Bureau of Economic Research.

Ngai, L. R., C. A. Pissarides, and J. Wang (2019). China's mobility barriers and employment allocations. *Journal of the European Economic Association 17*(5), 1617–1653.

Pulido, J. and T. Świecki (2018). Barriers to mobility or sorting? sources and aggregate implications of income gaps across sectors and locations in indonesia.

Restuccia, D., D. T. Yang, and X. Zhu (2008). Agriculture and aggregate productivity: A quantitative cross-country analysis. *Journal of Monetary Economics 55*(2), 234–250.

Schoellman, T. (2020). Comment on "migration costs and observational returns to migration in the developing world". *Journal of Monetary Economics 113*, 155–157.

Song, Z., K. Storeslletten, and F. Zilibotti (2011). Growing like China. *American Economic Review 101*(1), 196–233.

Tian, Y., J. Xia, and R. Yang (2020). Trade-induced urbanization and the making of modern agriculture. Technical report.

Tombe, T. and X. Zhu (2019). Trade, migration, and productivity: A quantitative analysis of China. *American Economic Review 109*(5), 1843–72.

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources 50*(2), 420–445.

Young, A. (2013). Inequality, the urban-rural gap, and migration. *The Quarterly Journal of Economics 128*(4), 1727–1785.

Zhang, C. and B. Chen (2014). Can "public pension system" substitutes "family mutual insurance"? *Economic Research Journal (in Chinese) 11*, 102–115.

APPENDIX

A Data Appendix

A.1 Sector of Employment and Migration

The NFP provides the following information, which can be used to infer sector of employment and earnings for each sector: (i) number of working days in each of within-town agricultural and non-agricultural sectors, (ii) number of working days out of town, (iii) net income from agricultural production at the household level, and (iv) income earned out of town at the individual level. Table A.3 shows that out-migration status and non-agricultural employment are highly correlated. On the one hand, those who work more than 180 days out of town only spend 3.6% of working days in agricultural production on average, and 91.9% of these workers report non-agriculture as their sector of employment. On the other hand, for those who spend less than 180 working days out of town, the share of working days allocated to agricultural production is 78.1% (i.e, the weighted average of the statistics in columns (1) and (2)), and the share of workers reporting non-agriculture as their sector of employment is only 20.4%. Column (2) of Table A.3 shows that the majority of workers with out-of-town working days within the range $(0, 180]$ still report agriculture as their sector of employment.

Based on these observations, this paper does not distinguish between sector choice and location choice. We loosely define sector of employment as follows: an individual is affiliated with the na sector if she works out of town for more than 180 days, and in the a sector otherwise.[19] Panel A of Appendix Figure A.1 shows the distributions of working days allocated to within-town agriculture, within-town non-agriculture, and out of town for workers who are grouped into the a sector. We find that for workers in the a sector, 65.4% have zero working day in the within-town na sector and 90.6% spend zero working day out of town. Analogously, Panel B reveals that, for workers in the na sector, 72.3% have zero working days in the within-town a sector and 94.8% have zero working days in the within-town na sector.

A.2 Hukou Index

We extend the prefecture-level *hukou* policy liberalization index constructed by Fan (2019) to 2012. Specifically, we search and review all *hukou*-related official news articles, and laws and regulations at the prefecture level from Peking University's Law Information Database and Baidu. Following the narrative approach by Fan (2019), we rate each document describing

[19] The National Bureau of Statistics of China adopts a cutoff of 180 days to define migrant workers.

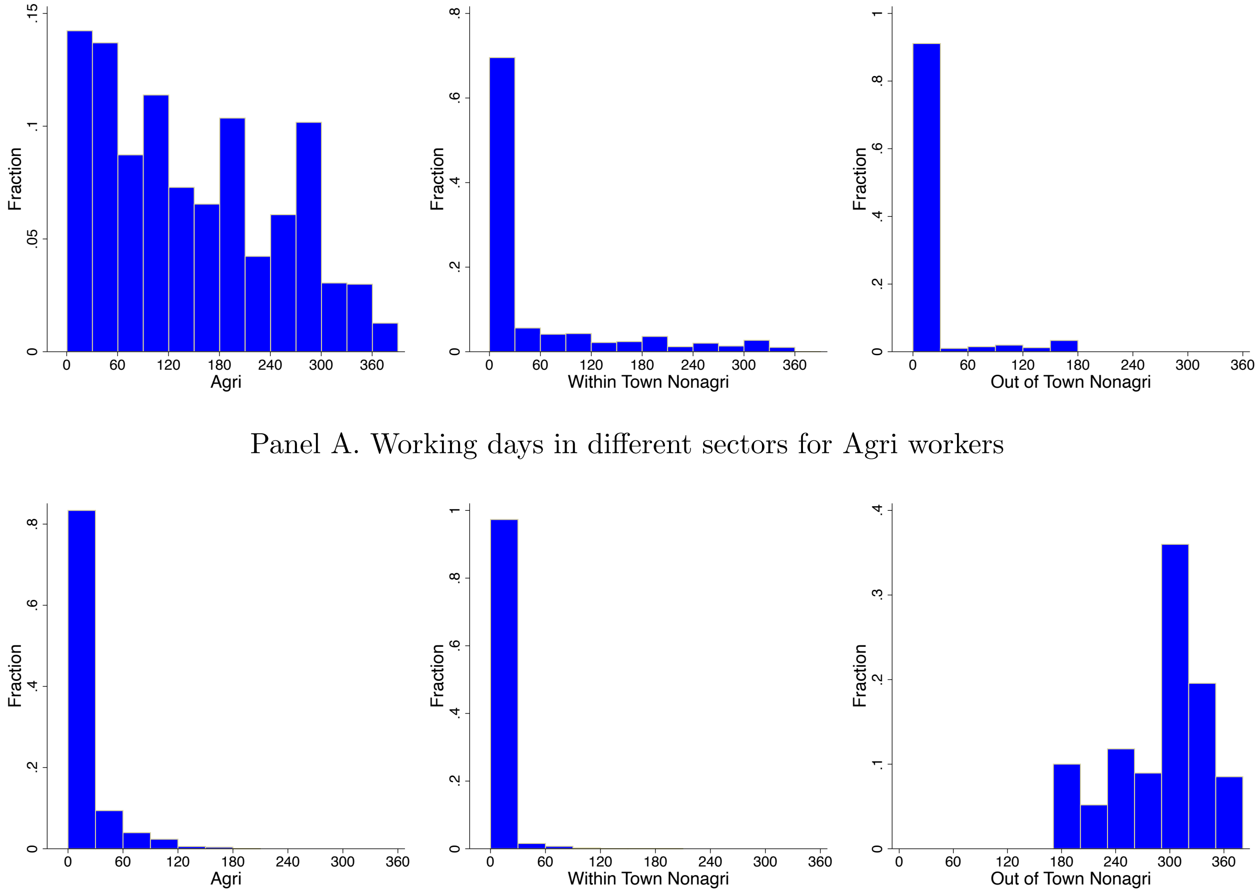
hukou policies on a score of 0 to 6, with 0 being the most stringent and 6 being completely open.[20] The average policy liberalization index increased from 2.04 in 2003 to 3.31 in 2010, and to 3.58 in 2012. This *hukou* index in general captures a migrant's job stability and the prospect of long-term settlement.

To construct the *hukou* index faced by potential out-migrants from different localities, we proceed as follows. First, for each prefecture, we use the 2000 Population Census to calculate the shares of out-migrants to different destination prefectures. Second, employing the predetermined migration shares as weights, we calculate the average of *hukou* policy liberalization indices across different destination prefectures. This measure is negatively related to the migration barriers faced by potential out-migrants in different origins, and is named Hukou Index in the paper. Lastly, with the mapping of villages and prefectures, we assign the prefecture-level HuKou Index measures to the villages. The indices for 2012 across prefectures are displayed in Figure 1. Table 1 presents the summary statistics.[21]

[20] See the details of the rating criteria in the appendix of Fan (2019). In the data, for each prefecture-year observation, there is at most one document of *hukou* policy. If such a document exists, the score of the document is the *hukou* index for the prefecture in a given year. If there is no new document introducing new *hukou* reforms, we adopt the measure from the preceding year.

[21] Note that the average of the destination-based index increases after 2010, while the average of the origin-based index decreases (see Table 1). This is because many first-tier cities tightened their *hukou* policy restrictions after 2010, and they constitute large weights in out-migration flows.

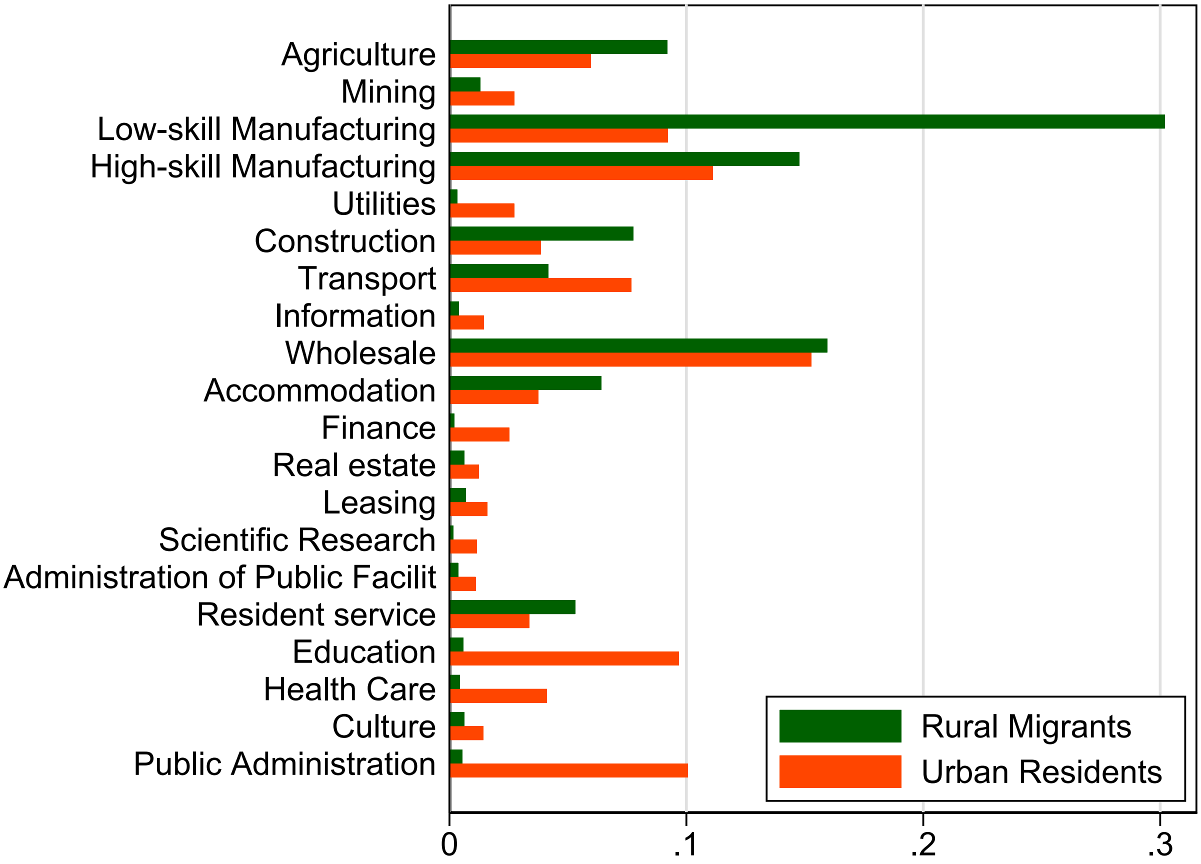
Figure A.1: Distribution of Working Days for Agri/NonAgri Workers



Panel A. Working days in different sectors for Agri workers

Panel B. Working days in different sectors for NonAgri workers

Figure A.2: Sectoral Distribution across Rural Migrants and Urban Residents



Notes: We disaggregate the manufacturing sector into high- and low-skill manufacturing. We define high–skilled workers having a college degree or above, and low–skilled workers as the rest. High-skill manufacturings are the manufacturing industries that have a higher share of high-skilled workers than the median manufacturing industry.

Table A.1: Summary Statistics: The NFP and the 2005 China 1% Population Sampling Survey

	NFP	Census,2005	
		Rural Hukou	Urban Hukou
Age	36.937	33.933	36.873
	(17.848)	(20.443)	(19.444)
Female	0.467	0.502	0.487
	(0.499)	(0.500)	(0.500)
Years of Schooling	6.779	6.509	9.728
	(3.080)	(3.559)	(4.251)
Poor Health Status	0.040	0.031	0.018
	(0.195)	(0.174)	(0.133)
Share of Workers	0.726	0.613	0.653
	(0.446)	(0.487)	(0.476)
Share of Elders	0.098	0.113	0.126
	(0.297)	(0.317)	(0.332)
Share of Workers Working in Non-agriculture	0.486	0.374	0.960
	(0.500)	(0.484)	(0.197)
Share of Workers to Migrate	0.166	0.123	0.198
	(0.373)	(0.329)	(0.399)
Rural Migrant/Urban Resident's Annual Earnings (log)	8.715	9.076	9.355
	(0.640)	(0.590)	(0.617)
Share of Migrants Working in:			
Agriculture	0.096	0.094	0.037
	(0.294)	(0.292)	(0.188)
Industry	0.256	0.463	0.266
	(0.436)	(0.499)	(0.442)
Construction	0.132	0.077	0.042
	(0.338)	(0.267)	(0.201)
Service	0.517	0.365	0.655
	(0.500)	(0.482)	(0.475)

Notes: Standard deviation in parentheses.

Table A.2: Distribution of the Number of Years Individuals Can be Tracked

	Raw Data		Our Sample	
Years	Observations	Share(%)	Observations	Share(%)
1	51,414	30.46		
2	24,416	14.47	13,148	25.45
3	16,507	9.78	9,372	18.14
4	13,707	8.12	7,131	13.80
5	10,087	5.98	5,444	10.54
6	8,604	5.10	4,544	8.79
7	7,458	4.42	3,641	7.05
8	8,766	5.19	3,394	6.57
9	11,867	7.03	3,059	5.92
10	15,946	9.45	1,935	3.75
Total	168,772	100	51,688	100

Table A.3: Summary Statistics
Labor Allocation and Sector of Employment by Out-of-town Labor Supply

	Agri Sector		Non-Agri Sector
Sample: Number of working days out of town	0 day	$(0, 180]$ days	$>$ 180 days
	(1)	(2)	(3)
Total working days	205.757	234.989	303.440
	(107.945)	(77.007)	(44.074)
Share of working days in:			
Within-town agri production	0.817	0.428	0.036
	(0.303)	(0.228)	(0.078)
Within-town non-agri production	0.183	0.071	0.006
	(0.303)	(0.150)	(0.032)
Out-of-town	0.000	0.502	0.958
	(0.000)	(0.236)	(0.086)
(Self-reported) Non-agricultural sector	0.186	0.386	0.919
	(0.389)	(0.487)	(0.273)
ln Daily wage in Non-agricultural sector	0.000	3.531	3.458
	(0.000)	(0.682)	(0.660)
ln Daily wage in agricultural sector	3.001	2.894	2.993
	(1.009)	(0.997)	(1.032)
Number of observations	147,571	15,243	71,217

Notes: Standard deviation in parentheses.

B Proofs of Propositions

B.1 Proof of Proposition 1

Note that

$$E\left[U_{na}|V > m(\boldsymbol{X},\boldsymbol{Z}) - R\right] = E\left[U_{na}|U_{na} - U_a > m(\boldsymbol{X},\boldsymbol{Z}) - R\right],$$

and

$$E\left[U_a|V \leq m(\boldsymbol{X},\boldsymbol{Z}) - R\right] = E\left[U_a|U_a - U_{na} > -\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right].$$

Under the assumption of the proposition, it is obvious that the first conditional expectation increases with $m(\boldsymbol{X},\boldsymbol{Z})-R$ and the second conditional expectations decreases with $m(\boldsymbol{X},\boldsymbol{Z})-R$. So,

$$R_{OLS} - R = E\left[U_{na}|U_{na} - U_a > m(\boldsymbol{X},\boldsymbol{Z}) - R\right] - E\left[U_a|U_a - U_{na} > -\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right]$$

is increasing with $m(\boldsymbol{X},\boldsymbol{Z}) - R$.

B.2 Proof of Proposition 2

If the distribution of U is symmetric with respect to U_a and U_{na}, then we have

$$E\left[U_a|U_a - U_{na} > -\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right] = E\left[U_{na}|U_{na} - U_a > -\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right].$$

So,

$$R_{OLS} - R = E\left[U_{na}|U_{na} - U_a > m(\boldsymbol{X},\boldsymbol{Z}) - R\right] - E\left[U_{na}|U_{na} - U_a > -\left(m(\boldsymbol{X},\boldsymbol{Z}) - R\right)\right],$$

which is zero if $m(\boldsymbol{X},\boldsymbol{Z}) - R = 0$. If, in addition, the assumption of Proposition 1 holds, then $E\left[U_{na}|U_{na} - U_a > x\right]$ is increasing with x, and the above equation implies that $R_{OLS} - R$ is increasing in the net migration cost, $m(\boldsymbol{X},\boldsymbol{Z})-R$. It is positive, zero, or negative if $m(\boldsymbol{X},\boldsymbol{Z})-R$ is positive, zero, or negative, respectively.

B.3 Proof of Proposition 3

Let

$$G(x) = \int_{-\infty}^{x} v f(v) dv,$$

and let $p(\boldsymbol{X}, \boldsymbol{Z})$ be the PDF of $(\boldsymbol{X}, \boldsymbol{Z})$. Then, we have

$$\begin{aligned}&E\left[V|m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m<V<m(\boldsymbol{X},\boldsymbol{Z})-R\right]\\&=\frac{\int\left[G(m(\boldsymbol{X},\boldsymbol{Z})-R)-G(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)\right]p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}{\int\left[F(m(\boldsymbol{X},\boldsymbol{Z})-R)-F(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)\right]p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}.\end{aligned}$$

Note that

$$\lim_{\Delta m\to 0}\frac{G(m(\boldsymbol{X},\boldsymbol{Z})-R)-G(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)}{\Delta m}$$

$$=G'(m(\boldsymbol{X},\boldsymbol{Z})-R)=(m(\boldsymbol{X},\boldsymbol{Z})-R)f((m(\boldsymbol{X},\boldsymbol{Z})-R)),$$

and

$$\lim_{\Delta m\to 0}\frac{F(m(\boldsymbol{X},\boldsymbol{Z})-R)-F(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)}{\Delta m}$$

$$=F'(m(\boldsymbol{X},\boldsymbol{Z})-R)=f((m(\boldsymbol{X},\boldsymbol{Z})-R)).$$

Thus, we have

$$\lim_{\Delta m\to 0}R_{\text{LATE}}=R+\lim_{\Delta m\to 0}E\left[V|m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m<V<m(\boldsymbol{X},\boldsymbol{Z})-R\right]$$

$$=R+\lim_{\Delta m\to 0}\frac{\int\frac{G(m(\boldsymbol{X},\boldsymbol{Z})-R)-G(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)}{\Delta m}p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}{\int\frac{F(m(\boldsymbol{X},\boldsymbol{Z})-R)-F(m(\boldsymbol{X},\boldsymbol{Z})-R-\Delta m)}{\Delta m}p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}$$

$$=R+\frac{\int(m(\boldsymbol{X},\boldsymbol{Z})-R)f(m(\boldsymbol{X},\boldsymbol{Z})-R)p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}{\int f(m(\boldsymbol{X},\boldsymbol{Z})-R)p(\boldsymbol{X},\boldsymbol{Z})d(\boldsymbol{X},\boldsymbol{Z})}$$

$$=R+\frac{E\left[(m(\boldsymbol{X},\boldsymbol{Z})-R)f(m(\boldsymbol{X},\boldsymbol{Z})-R)\right]}{E\left[f(m(\boldsymbol{X},\boldsymbol{Z})-R)\right]}$$

$$=\frac{E\left[m(\boldsymbol{X},\boldsymbol{Z})f(m(\boldsymbol{X},\boldsymbol{Z})-R)\right]}{E\left[f(m(\boldsymbol{X},\boldsymbol{Z})-R)\right]}.$$

C OLS versus IV Estimation

This appendix illustrates the possibility that the IV estimate can be larger than the OLS estimate, based on a simple model that belongs to the general Roy model in Section 3.[22] For clarity and without loss of generality, we drop observed characteristics $\mathbf{X}$ in the following discussions, and make the simplifying assumption that (U_a, U_{na}) follows a joint normal distribution.

Consider an exogenous policy shock that reduces migration cost from M_c to M_c'. As is shown in Appendix B, when $\Delta = M_c - M_c'$ is sufficiently small, IV estimation yields a consistent estimate for the baseline migration cost M_c. Hence, the difference between the OLS estimate and the IV estimate is given by:

$$\beta^{OLS} - \beta^{IV} = R - M_c + \sigma_a \rho_{a,v} \frac{\phi(\frac{R-M_c}{\sigma_v})}{1 - \Phi(\frac{R-M_c}{\sigma_v})} - \sigma_{na}\rho_{na,v}\left(-\frac{\phi(\frac{R-M_c}{\sigma_v})}{\Phi(\frac{R-M_c}{\sigma_v})}\right). \tag{C.1}$$

In the following, we show that $\beta^{IV} > \beta^{OLS}$ when $\sigma_a^2 > \sigma_{na,a}$, and migration cost M_c is sufficiently large.[23] To see this, equation (C.1) can be rewritten as

$$\beta^{OLS} - \beta^{IV} = \sigma_v x - \frac{\sigma_a^2 - \sigma_{na,a}}{\sigma_v} \frac{\phi(x)}{1 - \Phi(x)} - \frac{\sigma_{na}^2 - \sigma_{na,a}}{\sigma_v}\left(-\frac{\phi(x)}{\Phi(x)}\right), \tag{C.2}$$

where $x = \frac{R-M_c}{\sigma_v}$. Denote $f(x) = \sigma_v x$ and $g(x) = \frac{\sigma_a^2 - \sigma_{na,a}}{\sigma_v}\frac{\phi(x)}{1-\Phi(x)} + \frac{\sigma_{na}^2 - \sigma_{na,a}}{\sigma_v}\left(-\frac{\phi(x)}{\Phi(x)}\right)$.

We now prove that $\lim_{x\to-\infty} f(x) < \lim_{x\to-\infty} g(x)$ when $\sigma_a^2 > \sigma_{na,a}$. There are two cases to consider: (i) when $\sigma_{na}^2 < \sigma_{na,a}$, the relation trivially holds; (ii) when $\sigma_{na}^2 > \sigma_{na,a}$, both functions approach $-\infty$ when x goes to $-\infty$.

For case (ii), given the properties of the standard normal distribution:

$$\lim_{x\to-\infty} \frac{d\left(\frac{\phi(x)}{1-\Phi(x)}\right)}{dx} = 0 \quad \text{and} \quad \lim_{x\to-\infty} \frac{d\left(-\frac{\phi(x)}{\Phi(x)}\right)}{dx} = 1,$$

it is straightforward to show that

$$\lim_{x\to-\infty} \frac{g'(x)}{f'(x)} = \frac{\sigma_{na}^2 - \sigma_{na,a}}{\sigma_v^2} < 1. \tag{C.3}$$

The inequality follows because $\sigma_a^2 - \sigma_{na,a} > 0$ and $\sigma_{na}^2 - \sigma_{na,a} > 0$ imply $\frac{\sigma_{na}^2 - \sigma_{na,a}}{\sigma_v^2} < 1$. By

[22] When the sectoral switches are induced by exogenous migration costs, the following discussion also applies to the comparison of the OLS and FE estimates.

[23] In our data, $\sigma_a > \sigma_{na}$ which implies that $\sigma_a^2 > \sigma_{na,a}$. Our data also supports that $\sigma_{na}^2 > \sigma_{na,a}$.

L'Hôpital's rule,

$$\lim_{x\to-\infty}\frac{g(x)}{f(x)}=\frac{\sigma_{na}^2-\sigma_{na,a}}{\sigma_v^2}<1\implies\lim_{x\to-\infty}f(x)<\lim_{x\to-\infty}g(x).$$

Therefore, when M_c is sufficiently large, $\beta^{OLS}<\beta^{IV}$.[24]

In the following section, we conduct a series of Monte Carlo simulations based on data generation process described in the simple model, and confirm the above analytical results.

C.1 Monte Carlo Simulations

We simulate the data in a way that is analogous to the rollout of the NRPS. Specifically, we randomly assign a reduction in migration cost Δ to 1/5 of workers from year 6 (segment 1), 1/5 of workers from year 7 (segment 2), and so forth. Therefore, the treatment takes five phases to roll out. By year 10,the migration cost of all workers is reduced by Δ. Worker i migrates to the na sector in year t if

$$U_{i,n}-U_{i,na}+R-M_c+\Delta*I_{it}-\varepsilon_{it}>0,$$

where I_{it} is a binary variable that equals 1 if the treatment of a reduction in migration is turned on; ε_{it} is the idiosyncratic migration cost that follows the normal distribution $N(0,\sigma_\varepsilon^2)$.

We set $\sigma_{na}=0.52$, $R=0.59$, $\sigma_\varepsilon=0.1$,[25] $\Delta=0.05$ and consider two cases: (i) low migration cost $M_c=0.3$, and (ii) high migration cost $M_c=0.7$. With different parameterizations of $\{\sigma_a,\rho_{na,a}\}$, we simulate panel datasets, each with 2,000 workers and 10 years, and estimate the pooled cross-sectional OLS model and the IV/2SLS model as follows:

$$OLS:\ y_{it}=\beta d_{it}+D_t+u_{it};\qquad IV:\ y_{it}=\beta\hat{d}_{it}+D_j+D_t+v_{it},$$

where D_j is the segment fixed effect (analogous to village fixed effect in our regression analysis), and $\hat{d}_{it}$ is the fitted value of migration status from the first-stage regression: $d_{it}=\gamma I_{it}+D_j+D_t+\omega_{it}$. Denote the estimates of β from these two models by $\hat{\beta}^{OLS}$ and $\hat{\beta}^{IV}$, respectively.

Figure C.1 shows the differences between the OLS and IV estimates across grids of $\{\sigma_a,\rho_{na,a}\}$ for the low-cost case (left panel) and the high-cost case (right panel). With the negative selection of compilers relative to U_a, i.e., when $\sigma_a>\sigma_{na}$, the IV estimate can be larger than the OLS

[24] Note that with an additional parameter restriction of $\sigma_{na}^2>\sigma_{na,a}$, we can show that $\lim_{x\to+\infty}f(x)>\lim_{x\to+\infty}g(x)$. That is, when migration cost is small and APG is sufficiently large, the OLS estimate is larger than the FE estimate. Since both $f(x)$ and $g(x)$ are continuous, there would be a value of x such that $\beta^{OLS}=\beta^{IV}$.

[25] The parameters σ_{na}, R, and $\sigma_\varepsilon=0.1$ are taken from the estimates in Tables 9 and 10.

estimate. Moreover, when M_c is larger, the selection force is stronger, and hence we observe more cases with $\hat{\beta}^{OLS} - \hat{\beta}^{IV} < 0$. This is indeed the case: When $\sigma_a = 0.58$, $\rho_{na,a} = 0.85$, and $M_c = 0.7$ (as in the ballpark of our baseline estimates), $\hat{\beta}^{OLS} < \hat{\beta}^{IV}$, and the difference is similar to the difference between the corresponding estimates in Tables 3 and 5.

Figure C.2 shows the differences between the OLS estimates and the underlying APG, and the differences between the IV estimates and migration cost. In general, $\hat{\beta}^{OLS}$ is different from R, and the bias is more positive when M_c is larger and when $\sigma_{na} > \sigma_a$. As discussed, the IV estimate captures migration cost when Δ is sufficiently small, which is reflected in our simulations.

Figure C.1: $\hat{\beta}^{OLS} - \hat{\beta}^{IV}$

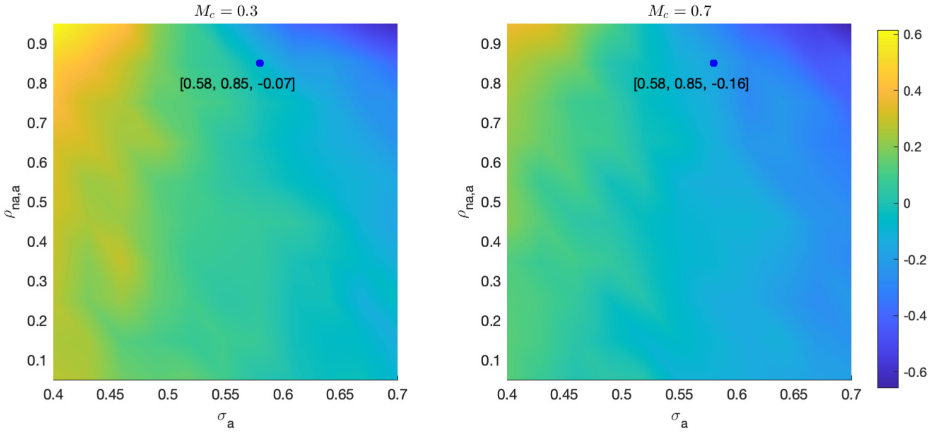
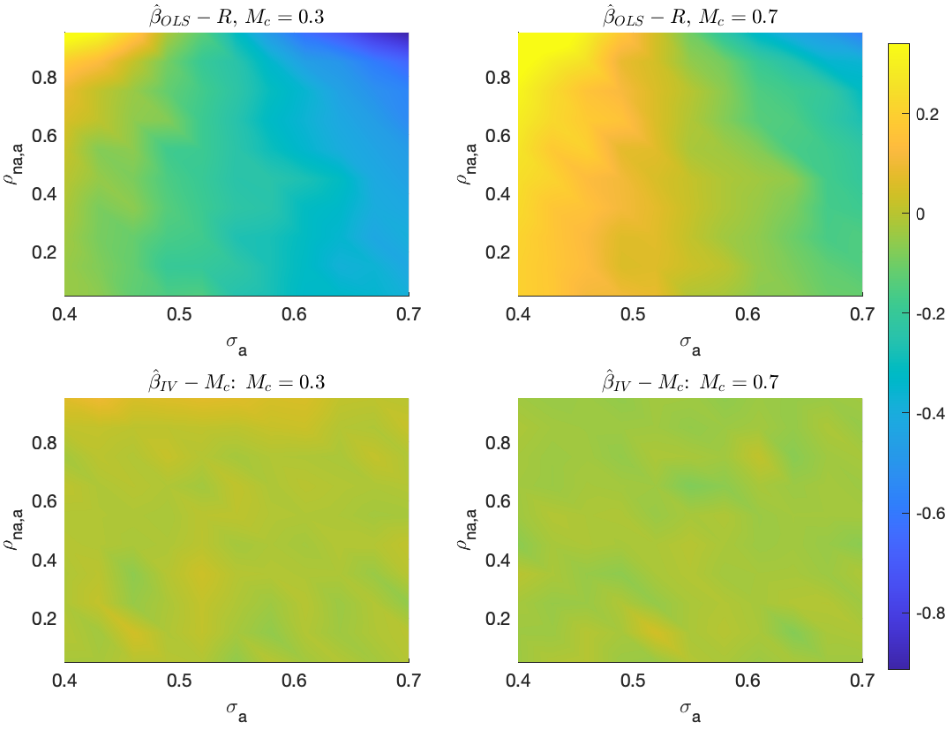


Figure C.2: $\hat{\beta}^{OLS} - R$ and $\hat{\beta}^{IV} - M_c$



D Additional Results

D.1 Estimation Results Based on the CFPS Data

In this section, we also estimate the observational returns to migration based on the China Family Panel Study (CFPS) data. The CFPS is a nationally representative, annual longitudinal survey of Chinese communities, families, and individuals. It focuses on both the economic and non-economic well-being of the Chinese people, covering substantive areas such as economic activities, educational attainment, and population migration. In 2010, The baseline national survey was launched in 25 provinces, representing 95% of the Chinese population. A total of 14,960 households and 42,590 individuals were successfully interviewed in the baseline survey. We use 2010, 2012, 2014, and 2016 waves for analysis.

Consumption and earnings are widely used to measure the observational gains from migration. We follow Lagakos et al. (2020) and construct the measure of per capita consumption by dividing a household's total consumption by the number of adult members in the household. The CFPS contains information on the consumption of 31 items of consumer goods and services in the CPI basket. The surveys also provide information on household agricultural production, including output and intermediate inputs. We subtract expenditures on intermediate inputs from the gross output to obtain the agriculture value-added at the household level. The information on individuals' working months in agricultural production is only available in the 2010 wave. Therefore, we construct individual agricultural income in the following way: (i) for the 2010 observations, we apportion the agricultural household agriculture value added to each member based on their working months in that year; and (ii) for the following waves, we equally allocate the agricultural household value added to each member who engages in agricultural production. The CFPS records individual employment earnings, including wages and subsidies from all jobs in the non-agricultural sector. We compute individual annual income by adding up their incomes from the agricultural and non-agricultural sectors, and deflate it using province-level spatial price deflators constructed by Brandt and Holz (2006), so that the measure reflects the real annual earnings.

The information on residential location (i.e., rural or urban) comes from the geographic identifier variables provided by the surveys. For the sector of employment, the CFPS asks the question *"Is your job an agricultural job or a non-agricultural job?"* or *"Are you currently engaging in any paid non-agricultural work?"* With the information, we identify if an individual works in the non-agricultural sector. To be consistent with our analysis based on the NFP, we follow the same sample selection criteria and focus on individuals aged between 20 and 54 with no more than 12 years of schooling, who appear at least two times in the whole sample period. Note that the NFP is an origin-based survey and almost all observations have rural *hukou*. Therefore, for consistency, we only keep individuals with rural *hukou* in the CFPS. The final panel data for regression analysis contains 44,665 individual-year observations and covers 15,021 individuals.

Estimates of observational migration returns are presented in Appendix Table D.1. In all regressions, we include individual and household control variables, namely: four age group dummies (20-29,30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, a dummy indicating whether there is an elderly aged 60 or above residing in the household and a province×year fixed effect. Columns (1)-(4) report the estimates of consumption gains from migration. Column (1) shows the cross-sectional gap, captured by the coefficient on an urban indicator. Column (2) reports the estimated coefficient from the specification augmented with individual fixed effects. We find that the cross-sectional consumption gap is 0.14, while the estimate declines to 0.07 when all the time-invariant individual characteristics are controlled for. Similar patterns hold when we substitute the urban indicator with the non-agricultural sector indicator in the regressions. These results confirm the findings of Lagakos et al. (2020).

The estimated return to migration based on consumption may be smaller than that based on income because income elasticity of consumption is generally less than 1. This conjecture is supported by the empirical findings in the following columns, where we employ the measure of individual real earnings as the dependent variable to estimate the income gap between sector na and sector a. Column (5) finds that the cross-sectional income gap between the non-agricultural and agricultural sectors is on average 109 log points. Moreover, the estimated gap increases to 129 log points when individual fixed effects are included in Column (6). These results align with our previous findings using the NFP data.

Table D.1: Observational Returns to Migration in CFPS

	(1) Consumption	(2) Consumption	(3) Consumption	(4) Consumption	(5) Annual Earning	(6) Annual Earning
Urban	0.1435***	0.0706*				
	(0.0156)	(0.0372)				
NonAgri			0.1919***	0.0198	1.0926***	1.2855***
			(0.0127)	(0.0220)	(0.0229)	(0.0464)
Individual and Household Controls	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Individual FE	N	Y	N	Y	N	Y
Observations	41046	39875	31143	27705	24504	19641
R-squared	0.2106	0.6266	0.2042	0.6357	0.2756	0.6889

Notes: Individual and household controls include four age group dummies (20-29,30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, a dummy indicating whether there is an elderly aged 60 or above residing in the household. Urban is a binary indicators for rural or urban status, and NonAgri is for employment in the non-agricultural sector. The dependent variable in columns (1)-(4) is log consumption per adult, and individual's real annual earnings in columns (5) and (6), respectively. Columns (1),(3), and (5) report the OLS regression results with individual and household controls, and province-year fixed effects. Columns (2),(4), and (6) add individual fixed effects to the regressions. All regressions are clustered at the household level. Standard errors are given in parentheses *** p<0.01, ** p<0.05, * p<0.1

D.2 The NRPS and Sector of Employment: Mechanisms

We propose two mechanisms through which the NRPS changes the sector switching cost of the young workers: (i) with the new pension plan, the elderly increase healthcare service consumption and rely less on the eldercare provided by their children; and (ii) the elderly reduce their labor supply and may allocate more time to home production and looking after their grandchildren. Both these channels reduce the shadow price of home production, which in effect lower the sector switching cost for young workers associated with being geographically distant from family.

We employ the information from the NFP on medical expenditure at the household level to test the first mechanism. Specifically, we restrict the sample to the individuals aged 60 or above and estimate the following regression:

$$\ln(1 + \text{Medical Expenditure}_{ihjt}) = \delta_1 NRPS_{jt} + X'_{ihjt}\delta_2 + D_j + D_{pt} + u_{ihjt},$$

where Medical Expenditure$_{ihjt}$ denotes the expenditure on medical services of individual i's household. The vector X_{ihjt} contains individual and household controls including dummies for age groups (60-64, 65-69, 70-74, and > 75), dummies for educational attainment (illiterate, primary school, middle school, high school, and college), gender, dummies for health status, arable land per capita, and type of Hukou.

Column (1) of Table D.2 confirms that households with an elderly spend more on medical services following the introduction of the NRPS. The NFP data contains a categorical variable indicating individual health status on a 5-point scale: 1 for “very good”, 2 for “good”, 3 for “medium”, 4 for “poor”, 5 for “disabled.” In column (2), we consider an individual to be in poor (respectively, good) health if her health status is “poor” or worse (respectively, “medium” or better), and find that the effect of the NRPS on medical expenditure is more pronounced if the elderly is in relatively poor health. Column (3) estimates a flexible specification, allowing for heterogeneous effects for each health status category. We find a stronger effect of the NRPS when the elderly is of health status “medium” or worse.

To explore the second mechanism, Table D.3 explores the effect of the NRPS on elderly labor supply, with the sample restricted to individuals aged 60 and above. Column (1) reports the result of the following regression:

$$E(WorkingDays_{ihjt}) = \exp(\rho_1 NRPS_{jt} + X'_{ihjt}\rho_2 + D_j + D_{pt} + u_{ihjt}).$$

We employ the poisson regression due to a large number of observations with zero working days in the data. Column (1) finds that there is no significant effect of the NRPS on elderly labor supply. However, as shown in column (2), the estimated effect becomes larger in magnitude when we restrict the sample to younger individuals (i.e., between 60-69) albeit it is statistically insignificant. In columns (3) and (4), we estimate the effect of the NRPS on the extensive margin of elderly labor supply using an OLS specification. We find that for individuals aged between 60-69, the NRPS lowers the probability of working more than 120 days annually by 3.2 percentage points. Columns (5) and (6) find that the NRPS lowers the income from labor supply, especially for the younger individuals. Due to the data constraint, we are not able to further study whether the elderly allocate more time to home production or to leisure. Therefore, we only consider the above findings as suggestive evidence for the second mechanism.

Table D.4 presents the heterogeneous effects of the NRPS on sector of employment, which also provides indirect evidence for the two mechanisms discussed above. Columns (1) to (3) explore the effect of $Elder60_{hjt} \times NRPS_{jt}$ by location of non-agricultural employment. The effect only reveals when the *na* employment is outside the county of the registered Hukou. Columns (4) and (5) show that the effect is stronger for female workers. These findings align with the proposed mechanisms. Specifically, sector switching costs associated with caring for dependants and home production increase with migration distance. In addition, female workers are more likely to be the main caregivers looking after seniors and children in China's context. Therefore, if the NRPS reduces the related costs, we should expect that its effect is more pronounced for *na* employment in more distant locations and for female workers. In column (6), we allow the effect of the NRPS to vary by the age of the elderly. Relative to households without an elderly, *na* employment probability increases by 2.1%, 5.3%, and 3.9% following the introduction of the NRPS, for workers from households with an elderly aged 55-59, 60-69, and 70 or above, respectively. Individuals aged 55-59 are not entitled to NRPS transfers, and hence the significantly positive estimate of Elder55-59×NRPS suggests anticipatory responses to the NRPS. More importantly, the effect is the most pronounced for the households with elderly aged 60-69, which is consistent with the finding that the labor supply channel is more pronounced for this age group.

Table D.2: NRPS and Medical Expenditure

Dep. Var.: ln (1+Medical Expenditure)	(1) OLS	(2) OLS	(3) OLS
NRPS	0.3349**		
	(0.1464)		
$\mathbf{1}$(HealthStatus≤ 3)$\times$ NRPS		0.3158**	
		(0.1487)	
$\mathbf{1}$(HealthStatus≥ 4)$\times$ NRPS		0.4072**	
		(0.1688)	
$\mathbf{1}$(HealthStatus$= 1$)$\times$ NRPS			0.1488
			(0.1768)
$\mathbf{1}$(HealthStatus$= 2$)$\times$ NRPS			0.3202**
			(0.1599)
$\mathbf{1}$(HealthStatus$= 3$)$\times$ NRPS			0.4755***
			(0.1655)
$\mathbf{1}$(HealthStatus$= 4$)$\times$ NRPS			0.3943**
			(0.1786)
$\mathbf{1}$(HealthStatus$= 5$)$\times$ NRPS			0.4473**
			(0.2006)
Individual and household controls	Y	Y	Y
Province × Year FE	Y	Y	Y
Village FE	Y	Y	Y
Observations	74,951	74,951	74,951
R-squared	0.2602	0.2603	0.2604

Notes: Individual-level and household-level controls include dummies for age groups (60-64, 65-69, 70-74, and > 75), dummies for educational attainment (illiterate, primary school, middle school, high school and college), gender, dummies for health status, arable land per capita, and type of Hukou. HeathStatus is a categorical variable with 1 for "very good", 2 for "normal", 3 for "medium", 4 for "poor", 5 for "disabled". In column (2), we consider an individual to be in poor health if her health status is "poor" or worse. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

Table D.3: NRPS and Elderly Labor Supply

	(1)	(2)	(3)	(4)	(5)	(6)
Sample:	All	Age<70	All	Age<70	All	Age<70
Dep. Var.:	Working days	Working days	Working days> 120	Working days> 120	Annual income	Annual income
	Poisson	Poisson	OLS	OLS	Poisson	Poisson
NRPS	-0.0063	-0.0380	-0.0123	-0.0318**	-0.0446	-0.0740*
	(0.0286)	(0.0279)	(0.0130)	(0.0146)	(0.0379)	(0.0391)
Individual controls	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Village FE	Y	Y	Y	Y	Y	Y
Observations	58,785	45,825	58,813	45,835	58,756	45,803
R-squared	–	–	0.3151	0.2976	–	–

Notes: All columns restrict the sample to the elderly with medium or better health status (i.e. $HealthStatus \leq 3$). Individual-level controls include dummies for age groups (60-64, 65-69, 70-74, and > 75), dummies for educational attainment (illiterate, primary school, middle school, high school and college), gender, dummies for health status, arable land per capita, and type of Hukou. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1.

Table D.4: NRPS and Sector of Employment

	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	NonAgri	NonAgri	NonAgri	NonAgri	NonAgri	NonAgri
	within County	outside County	outside Province	Male	Female	All
		within Province				
Elder60 × NRPS	-0.0002	0.0194***	0.0233***	0.0208**	0.0631***	
	(0.0051)	(0.0056)	(0.0069)	(0.0100)	(0.0110)	
NRPS	0.0072	-0.0019	-0.0040	0.0055	-0.0041	-0.0030
	(0.0071)	(0.0061)	(0.0060)	(0.0115)	(0.0099)	(0.0099)
Elder60	0.0007	0.0075***	0.0129***	0.0177***	0.0230***	
	(0.0015)	(0.0018)	(0.0020)	(0.0032)	(0.0032)	
Elder55-59×NRPS						0.0214**
						(0.0099)
Elder60-69×NRPS						0.0529***
						(0.0119)
Elder≥70×NRPS						0.0389***
						(0.0099)
Elder55-59						0.0468***
						(0.0035)
Elder60-69						0.0500***
						(0.0037)
Elder≥70						0.0130***
						(0.0031)
Individual and household controls	Y	Y	Y	Y	Y	Y
Province × Year FE	Y	Y	Y	Y	Y	Y
Village FE	Y	Y	Y	Y	Y	Y
Observations	234,031	234,031	234,031	124,185	109,846	234,031
R-squared	0.1535	0.1637	0.2967	0.3454	0.3443	0.3501

Notes: Individual controls include four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), a dummy for gender, a dummy for poor health, arable land per capita, and type of Hukou. Robust standard errors are clustered at the village×year level. *** p<0.01, ** p<0.05, * p<0.1

D.3 Differential Productivity Growth Across Sectors

In this section, we leverage the panel data on individual earnings and estimate the differential sectoral productivity growth based on a sample of workers who stay in the same sector over time.[26] The earnings of a stayer i in the a sector is determined by $y_{a,it} = w_{a,t} \exp(\mathbf{X_i}\beta_t + U_{i,a} + \epsilon_{it})$ for all t, where ϵ_i is the time-varying shock that affects i's productivity in a common way across sectors. Here, we allow for the return to individual characteristics, $\mathbf{X_i}$, to vary over time in a linear manner, i.e., $\beta = \beta_t - \beta_{t-1}$. The corresponding annual growth in earnings is $\Delta \ln y_{a,it} = \Delta \ln w_{a,t} + \mathbf{X_i}\beta + \Delta\epsilon_{it}$. Analogously, the annual growth in earnings for stayers in the na sector is given by $\Delta \ln y_{na,it} = \Delta \ln w_{na,t} + \mathbf{X_i}\beta + \Delta\epsilon_{it}$. We can therefore employ the sample of stayers to estimate the change in the underlying APG, $\Delta R = \Delta \ln w_{na,t} - \Delta \ln w_{a,t}$, by estimating the following equation:

$$\Delta \ln y_{j,it} = \alpha NonAgri_i + \mathbf{X_i}\beta + \Delta \ln \epsilon_{it}, \tag{D.1}$$

where the vector $\mathbf{X_i}$ contains four age group dummies (20-29, 30-39, 40-49, and 50-54), four educational attainment group dummies (illiterate, primary school, middle school, and high school), and a dummy for gender. The coefficient α captures ΔR. By focusing on stayers in both sectors, the selection on unobserved abilities $\{U_{i,a}, U_{i,na}\}$ is necessarily accounted for. Then, the identification relies on the assumption that the change in individual productivity $\Delta \ln \epsilon_{it}$ is uncorrelated with the sector of employment conditional on observed worker characteristics.

We estimate equation (D.1) year by year over the period 2004 to 2012. For each estimation, the sample is restricted to workers who remain in the same sector in both period $t-1$ and period t. We find that, on average, the annual productivity growth of the non-agricultural sector is 1.3 log points higher than that of the agricultural sector. The difference is statistically significant at the 1% level and aligns with the structural estimate in Table 9.

With the presence of differential growth rates across sectors, through the lens of the simple model in Appendix C, the FE and IV estimates obtained in Section 4 measure $M_c + \Delta R$. We find that the estimate of ΔR is much smaller in magnitude than the FE and IV estimates in Tables 4 and 5. It suggests that ignoring ΔR does not have a big impact on the interpretation of migration cost M_c.

[26] Kim and Vogel (2020) adopts a similar strategy to identify the change per efficient unit of labor across industries in the US.

D.4 Robustness: Differential Returns to Observable Characteristics

This section considers an augmented model where we allow for differential returns to observable characteristics across sectors. Specifically, equation (19) is modified to

$$h_{j,it} = \exp(\mathbf{X}_{it}\beta_j + u_{j,i} + \epsilon_{j,it}) \quad \text{for } j \in \{a, na\}.$$

The estimation results are reported in Table D.5 and D.6. The results remain robust.

Table D.5: Parameter Estimates of the Model with Differential Returns to Human Capital (Human Capital)

	Agri	NonAgri	NonAgri - Agri
Real wage level	8.489	8.973	0.485
	(1.433)	(1.536)	
Real wage growth rate	0.064	0.114	0.049
	(0.003)	(0.003)	
Human capital – male	0.221	0.293	0.072
	(0.013)	(0.031)	
Human capital – years of education	0.022	0.030	0.008
	(0.001)	(0.001)	
Human capital – age	0.103	0.057	-0.045
	(0.006)	(0.002)	
Human capital – age squared	-0.001	-0.001	0.000
	(0.000)	(0.000)	
SD of wage shock	0.774	0.387	
	(0.079)	(0.016)	
SD of ability	0.623	0.501	
	(0.077)	(0.031)	
Correlation between abilities	0.569		
	(0.048)		

Notes: Standard errors in parentheses.

Table D.6: Parameter Estimates of the Model with Differential Returns to Human Capital (Migration Costs)

	Migration Cost
Constant	0.622
	(0.039)
Male	-0.224
	(0.005)
Years of education	-0.020
	(0.001)
Age	-0.001
	(0.000)
Age squared	0.000
	(0.000)
Hukou Index	-1.247
	(0.049)
NRPS	0.012
	(0.001)
Elderly above 60	-0.036
	(0.002)
NRPS × elderly	-0.028
	(0.001)
SD of migration cost shock	0.278
	(0.009)

Notes: Standard errors in parentheses.