Income Inequality in Chinese Provinces: The Role of Human Capital

Preliminary Draft

Kristina Butaeva[1] and Albert F. Park[2,3]

[1]Harris School of Public Policy Stone Center for Research on Wealth Inequality and Mobility, University of Chicago
[2]Asian Development Bank (ADB)
[3]Department of Economics, Hong Kong University of Science and Technology

February 14, 2025

Abstract

In this paper, we conduct the first systematic empirical analysis of income inequality in China at the provincial level. Using data from the China Household Finance Survey (CHFS) and a semiparametric distribution model, we estimate Gini indices for Chinese provinces in 2012, 2014, 2016, and 2018. We find that differences in the "prices" and "quantities" of human capital are important factors that explain differences in inequality between provinces. Provinces with higher educational inequality, higher returns to schooling, and lower average years of schooling have higher levels of income inequality. Interprovincial differences in the returns to schooling may be attributable to the remaining barriers to labor mobility in China. Our findings suggest that poor provinces are highly disadvantaged compared to rich provinces given they face higher income and educational inequality, as well as higher returns to schooling, while at the same time having lower average educational attainment. We conclude that the reduction of existing interprovincial human capital gaps and the acceleration of labor market integration through appropriate government policies could substantially contribute to the reduction of income inequality and spatial disparities in inequality levels across regions in China.

Keywords: Income Inequality, Provinces, China, Human Capital.

We thank Grigory Kosenok, Shlomo Weber, Steven N. Durlauf, Li Gan, Jin Wang, Li Han, and Sujata Visaria for their insightful suggestions and indispensable feedback. Additionally, we appreciate the helpful comments from participants of the Applied Micro Seminar at HKUST. Butaeva gratefully acknowledges the financial support of the James M. and Catherine D. Stone Foundation and the Greater Eurasia Research Center (GEAR) at the Hong Kong University of Science and Technology (HKUST). She also recognizes Grace Kolavo and Joe Quinn for their instrumental support in preparation of the manuscript. Corresponding author: Kristina Butaeva (butaeva.k.o@gmail.com).

1 Introduction

Over the past four decades, the Chinese economy has experienced impressive economic growth. At the same time, the level of spatial and individual income inequality in the country has risen dramatically, and it seems that the fruits of economic success are still unevenly distributed.[1] Individual and household income inequality in China, as well as different types of spatial economic inequalities (rural-urban, regional, provincial, etc.), are well-studied in the existing literature.[2] However, to date there has been no systematic study of the nature of within-province income inequality and its variation across Chinese provinces.

Why is an analysis of provincial inequality important? First, intraprovincial income inequality substantially contributes to national-level inequality. The decomposition analysis from our study suggests that around 90% of the nationwide variation in annual household income per capita in 2012-2018 can be explained by intraprovincial income variation. Therefore, the analysis of the possible causes and the main contributing factors of intraprovincial income inequality, coupled with the estimation of each province's contribution to nationwide inequality has the potential to generate appropriate policies aimed at reducing nationwide income inequality. Second, income inequality is an important dimension of cross-regional economic disparities, though prior analyses have focused on the comparisons of GRP per capita or mean disposable income. Third, provincial income inequality has the capacity to impede provincial and national economic growth, discount social attitudes towards political legitimacy, and incite potentially adverse political instability. The negative impact of high income inequality on growth is well-documented in the literature. For example, high income inequality could create inequality of opportunities which limits the human capital development of disadvantaged social groups (Galor & Moav, 2006; Perotti, 1996). Another plausible mechanism is that high income inequality reduces social cohesion and may increase the risk of political conflicts.[3] Political instability creates uncertainty, which could adversely impact investments and economic growth (Alesina & Perotti, 1996; Esteban & Ray, 1999). Moreover, high income inequality may facilitate lobbying on behalf of financial elites. This could hurt political institutions, create inefficiency, and impede sustainable economic growth (Chong & Gradstein, 2017; Savoia, Easaw, & McKay, 2010). Finally, studying the causes of

[1] See, e.g., Naughton (2006), Ravallion and Chen (2009), Kanbur, Wang, and Zhang (2021), Jain-Chandra et al. (2018), Gradín and Wu (2020).

[2] Studies on China's inequality trends are summarized in Kanbur et al. (2021). A comprehensive overview of the trends in the Gini index in China is presented in Xie and Zhou (2014). A review of existing studies on inequality in China employing top-income corrections is presented in Q. Li, Li, and Wan (2020), whereas spatial inequality studies are described in C. Wang, Wan, and Yang (2014).

[3] Previous studies find that a notion of fairness is positively related to social tolerance towards income inequality (Lü, 2012). Florida, Mellander, and Rentfrow (2013) point out that human capital plays a key role in the happiness of a cities' residents.

provincial inequality will produce insights into the efficiency of educational investments and the performance of local labor markets.

In this paper, we aim to fill the gap in the existing literature on subnational-level income inequality in China. Our purpose is four-fold. First, we aim to provide Gini index estimates for 29 mainland Chinese provinces (excluding Tibet and Xinjiang), applying a new robust measurement approach developed in Butaeva (forthcoming) to the data from four waves of the China Household Finance Survey (CHFS 2013, 2015, 2017, and 2018). Second, we seek to identify the main contributing factors that explain the variation in income inequality across Chinese provinces, by examining the Mincer (1974) model. The Mincer model posits that education and the returns to education are key determinants of earnings, and therefore we assume that higher inequality in educational attainment and higher returns to education would result in higher income inequality. These two human capital factors are policy-relevant because they reflect local educational and labor market institutions. Many economic studies have noted substantial increases in educational attainment, including a remarkable expansion of higher education in China over the last several decades.[1] At the same time, previous studies have found that under a decentralized fiscal system, significant progress in educational development has advanced in the richer coastal provinces and large cities more than other regions (e.g., Xiang, Stillwell, Burns, & Heppenstall, 2020).

Many scholars also recognize the positive effects of Chinese institutional market reforms which encourage labor market liberalization. For example, Cai, Park, and Zhao (2008) emphasizes that internal migration in China has increased substantially over time, helping to promote local labor market integration, which has already led to progressive convergence of returns to education across regions over time. Significant progress in labor migration is also outlined by Cai et al. (2008). However, as pointed out by Fleisher and Yang (2008), despite the improvements, some serious challenges to the full integration of the labor market in China remain. In particular, the *hukou*[2] system still limits urbanization and industrialization in China (Ngai, Pissarides, & Wang, 2019). Therefore, our third goal is to contribute to the discussion of the unequal rewards of additional schooling across Chinese provinces. Unrestricted mobility implies free flows of skilled and unskilled labor among different locations which should equalize the returns to schooling across provinces. However, we continue to observe differences in returns to schooling across provinces, which suggests that despite the significant progress in labor market integration in China there still exist barriers to labor migration. Finally, we evaluate the potential effects of narrowing educational disparities and

[1] See, e.g., Chan and Ngok (2011), Lei and Shen (2015), Mendoza (2018).

[2] *Hukou* is a household registration system in mainland China introduced in the 1950s, which imposes formal restrictions on changes of permanent residence.

promoting labor market integration. To that end, we conduct a simulation analysis that examines how provincial and national income inequality would change in response to: (a) equalization of educational structure across provinces; (b) equalization of returns to schooling; and (c) both an equalization of educational structure and returns to schooling.

To summarize, in this study we find that there is a substantial disparity in levels of provincial income inequality in China. In 2018, the Gini index in the most unequal province, Sichuan, was 1.5 times larger than in the least unequal, Tianjin (0.667 versus 0.450, respectively). We also find that poor provinces – those with low GRP per capita – are severely disadvantaged compared to rich provinces. They not only have lower average levels of education and higher inequality in education but also face greater income inequality. Existing human capital disparities and differences in returns to schooling both seem to be the key contributing factors explaining the variation in income inequality across Chinese provinces. In turn, the factors spurring inequities in the returns to schooling include variation in the relative supply of more skilled workers, unemployment rates, and provincial industrial structure. We also find that current levels of interprovincial migration do not alter regional variation in human capital structure enough to substantially reduce the existing interprovincial differences in income inequality. Our simulation analysis reveals that keeping all other factors constant, equalization of educational structure across provinces could lead to an 18.1% decrease in nationwide income inequality. Our findings prompt an examination of regional educational policies together with policies on labor market integration as fundamental instruments for mitigating across-provincial income disparities in China.

The paper is organized as follows. In Section 2, we review the main results of previous studies on income and human capital disparities in China. In Section 3, we describe the data. In Section 4, we estimate provincial Gini indices in China, applying the semiparametric method introduced in Butaeva (forthcoming) to the data from CHFS. In Section 5, we test different patterns of nationwide income inequality in China, applying decomposition methods: (a) we decompose the country-level Theil T income inequality index by province, analyzing the relative contribution of inter- and intraprovincial variation in household income per capita to the nationwide variation in income; (b) we perform regression-based decomposition by applying the classical Mincer model. In Section 6, we estimate provincial educational Gini indices and provincial returns to schooling. In Section 7, we study the key factors of interprovincial variation in income inequality. Namely, we regress the provincial Gini indices on provincial educational Gini indices, provincial returns to schooling, and their interaction term to evaluate the role of human capital in the variation of income inequality across Chinese provinces. We also regress our provincial returns to schooling on different provincial labor supply and labor demand factors to identify the key factors of the exist-

ing interprovincial disparities in returns to schooling. Finally, in Section 8, we conduct a simulation analysis, which aims to study how income inequality in China would change in response to: (a) equalization of educational structure; (b) labor market liberalization (which implies equalization of returns to schooling); (c) both equalization of educational structure and labor market liberalization. In Section 9, we conclude.

2 Related literature

Provincial income inequality in China

The uneven impact of economic development since the beginning of market reforms in China has attracted the attention of many scholars. Current interprovincial economic inequality is considerable. In 2018, the richest province (Beijing) had 4.7 times higher GRP per capita than the poorest province (Gansu).[1] The coefficient of variation in GRP per capita among provinces is around 0.44-0.46. For comparison, the ratio of rich to poor GRP per capita among the US states in 2016 was 2.8, and the coefficient of variation was 0.18 (Fraumeni, He, Li, & Liu, 2019). Nevertheless, Hao, Sun, Tombe, and Zhu (2020) show that between 2000 and 2015, the variance in the cross-province (log) GDP per worker fell by one-third, which they claim was mostly explained by the enhancement of the internal migration.

Due to the data limitations, the number of studies on provincial-level income inequality in China is small. (Existing estimates of provincial-level Gini indices are summarized in Table A1, Appendix A). The earliest studies used survey data from the Chinese Academy of Social Sciences (CASS) (Hussain, Lanjouw, & Stern, 1994; Khan & Riskin, 2005). Hussain et al. (1994) estimate Gini indices for most Chinese provinces in 1986 separately for rural and urban populations. They find that intraprovincial income inequality was the main contributing factor to nationwide income inequality (explaining 94% of (log) income variance for urban households and 85% for rural households). They also find that, unlike most other developing countries, income inequality in China's urban areas appeared to be lower than in rural areas. Analyzing provincial income inequality in 1995 and 2002, Khan and Riskin (2005) find a decline in both urban and rural income inequality in most provinces and an increase in the rural-urban gap. Employing the Survey on Urban Household Income, Xu and Zou (2000) study the potential factors explaining provincial income inequality dynamics between 1985 and 1995. They find urban inequality increases with a reduction in the share of state-owned enterprises (SOEs) in GDP, with an increase in inflation and regional growth, and (less significantly) with greater exposure to foreign trade. At the same time, no statis-

[1] Provincial GRP per capita for 2012, 2014, 2016, and 2018 is presented in Table E1, Appendix E.

tically significant impact of schooling or urbanization was found. Using Urban Household Survey data for 1990-2009, Nie and Xing (2019) find the rise of urban income inequality is mainly attributed to the increase in returns to education. Nonetheless, the fragmented evidence from the existing studies on provincial-level income inequality in terms of their timing and geographic scope does not substitute for a comprehensive analysis of provincial income inequality in China. Our study aims to fill that gap by providing robust Gini index estimates for 29 mainland provinces (excluding Tibet and Xinjiang) for the years 2012, 2014, 2016, and 2018, and, furthermore, by identifying the key factors that shape inter- and intraprovincial income inequality in China.

Educational inequality in China

In this study, we find educational inequality and returns to schooling to be the main explanatory factors for the current variation in provincial income inequality levels in China. According to human capital theory, education is a fundamental determinant of earnings. Therefore, disparities in education could considerably widen disparities in income. In a multi-country study, Coady and Dizioli (2018) found a large, positive, and statistically significant relationship between inequality of schooling and income inequality, especially in emerging and developing countries.[1] Educational disparities in China have been studied quite extensively. One must note the remarkable progress made in furthering educational attainment since the launch of the government's higher education expansion policy in 1999. According to the Educational Statistical Yearbook of China, enrollment in higher education (undergraduate and graduate study, excluding adult higher education) increased eight times from 3.6 million individuals in 1998 to 28.9 million in 2016. However, recent studies continue to provide new evidence on educational inequalities.[2] Based on the 2010 China Family Panel Studies (CFPS) data, Lei and Shen (2015) underscore the widening gap in years of schooling between rural and urban residents. Introducing a new multidimensional index, Xiang et al. (2020) show that the northeastern regions of China provide better education compared to the southwest part of the country. J. Yang, Huang, and Liu (2014) estimate the educational Gini index in each province for several years from 1996 to 2008 and find that intraprovincial educational inequality has declined but regional variation remains large. estimating provincial-level educational Gini from the 2010-2012 CFPS data, Golley and Kong (2018) find the lack of equality in educational opportunities is primarily due to the individual's *hukou* status. Yu, Yu, de Jong, and Storm (2015) and Peng, Qi, Wan, Li,

[1]The paper also provides a theoretical framework and comprehensive literature review of the topic.

[2]For e.g., Heckman and Yi (2012) emphasize a substantial inequality in access to higher education in China.

and Hu (2020) contribute to the provincial educational Gini discussion by considering the regional economic development levels. Yu et al. (2015) show that the relationship between inequality in educational attainment and growth in China depends on the level of economic development of a particular area. More specifically, they claim that educational inequality has a larger impact on growth than educational attainment, and this effect is stronger in less developed western provinces than in more prosperous eastern provinces. Peng et al. (2020) conclude that poor provinces not only encounter lower educational resources but also experience a greater demand for education because they have more children to educate. In summary, there exists an extensive literature linking sub-national educational inequality and economic disparities, and as well as studies linking educational and income inequality in China (e.g., Chan & Ngok, 2011; Gustafsson & Wan, 2020; Sicular, Ximing, Gustafsson, & Shi, 2007). However, there are no provincial-level analyses of the combined impact of interrelated economic, educational, and income disparities. Our paper contributes to this literature by providing evidence that in China poor provinces exhibit higher levels of both income and educational inequality as compared to rich provinces.

Returns to schooling in China

Finally, our study provides an analysis of provincial differences in returns to schooling in China. We are not the first to find that returns to schooling per se, as well as in combination with educational inequality, are significant explanatory factors for cross-provincial income inequality variation in China. However, we also seek to explain provincial differences with respect to returns to schooling itself. To evaluate provincial returns to schooling we start with the classical Mincer human capital regression model, which has been used in many studies exploring differences in returns to schooling in China.[1] A recent paper by Asadullah and Xiao (2020) uses two waves (2010 and 2015) of the China General Social Survey (CGSS) to confirm the importance of the human capital factor in the Mincerian earnings equation in China. They also document the dynamics of greater declines in returns to schooling in urban locations, coastal regions, and among women. At the same time, S. Li, Whalley, and Xing (2014) document sharp increases in unemployment among young college graduates in response to China's educational expansion policy predominately in non-coastal regions. The authors suggest that encouraging interregional mobility in China could help to better match local labor supply and demand. Overall, most existing studies, including ours, attribute differences in returns to schooling across China to remaining labor market segmentation and labor mobility restrictions. Many scholars claim that the notorious *hukou* system, first intro-

[1] Zhang, Zhao, Park, and Song (2005) provide a comprehensive review of existing estimates of returns to schooling in China. Churchill and Mishra (2018) also present a broad literature review on the topic.

duced in 1951, which has become significantly less impactful in recent decades, still affects people's educational and employment opportunities, economic and social mobility, and labor market returns to education. Employing the data from the 2005 (1%) National Population Sample Survey, Fu and Ren (2010) demonstrate that *hukou* status still considerably shapes people's labor-market returns: an individual's human capital is largely determined by their birthplace (rural versus urban). A recent study by Ngai et al. (2019) also finds that the *hukou* system alters urbanization and industrialization in China. In addition to formal residential mobility restrictions, researchers also outline the presence of informal discrimination against labor migrants (Park, Song, Zhang, & Zhao, 2003).

3 Data

We use data for 29 mainland Chinese provinces (excluding Tibet and Xinjiang) from four waves of the China Household Finance Survey (CHFS 2013, CHFS 2015, CHFS 2017, CHFS 2019), a survey conducted by the Research Institute of Economics and Management at Southwestern University of Finance and Economics in Chengdu. The CHFS is a stratified multi-stage random sampling panel that contains micro-level information on household demographics and finance. The CHFS sample sizes are reported in Table B1, Appendix B.[1] All four aforementioned survey waves are nationally and provincially representative.

In this paper, we use the annual household disposable income per capita to measure income inequality. Household disposable income in CHFS includes income generated from labor, business, assets, transfers, and other sources received during the last twelve months.[2] Labor income includes total after-tax monetary wages and salaries, bonuses, and subsidies from employer (both monetary and in-kind). Business income is the sum of net income from agricultural production and business operations. Asset income includes net after-tax income from non-financial assets (rental income from lands, properties, and vehicles) and income from financial assets (interests on deposits, stocks, bonds, dividends from commercial life insurances, etc.). Transferred and other income include retirement benefits and pensions; unemployment insurance; other government compensation and subsidies; private transfers; net after-tax income from all other sources including lotteries, sale of properties, vehicles, intellectual property rights, etc.).

In the CHFS, some missing income values were recorded as zeros. Therefore, in our study, we do not use zero household incomes to avoid potential measurement bias. Negative incomes

[1] For more details on the CHFS dataset, see https://chfs.swufe.edu.cn/sjzx/sjsq.htm. A comprehensive review of the survey and key statistics of the data are presented in Gan et al. (2014).

[2] Therefore, we assign our inequality estimates to the year preceding each survey year.

are also excluded from the analysis. The restriction to positive income values reduces sample size across waves by 1-2.5%. All estimates are produced employing survey weights. It should especially be noted that CHFS contains a larger sample size of those living in high-income neighborhoods compared to other existing Chinese household surveys. In Tables B2 and B3 in the Appendix B we report summary statistics of all household- and provincial-level variables used in the current study.

4 Provincial-level income inequality

We begin by reporting the robust provincial Gini index estimates obtained using the new methodological approach introduced in Butaeva (forthcoming). In essence, this method allows one to address a well-known income inequality measurement issue related to the high sensitivity of the Gini index to the presence of top income outliers in the data.[1] On the one hand, sparse and incomplete top income data yield a high variance of Gini estimates across different years, because any random change in several top incomes may cause a sizable shift in inequality estimates for the whole geographic area under consideration. On the other hand, we can't just ignore or winsorize those observations because they represent an essential element of income inequality analysis, namely, the quantifiable difference between the rich and poor. The problem is especially striking in small samples, which is the case in the current study, where our provincial-level samples from CHFS range from 400 to 3000 observations in different survey years. The method we apply yields a semiparametric probability density function (PDF):

$$f(x) = \begin{cases} p_\tau \times \left[\dfrac{1}{Nh} \displaystyle\sum_{i=1}^{N} K\left(\dfrac{x_i - x}{h} \right) \right], & \text{if } x \leq \tau \\ (1 - p_\tau) \times \left[\dfrac{1}{\sigma} \left(1 + \xi \dfrac{x - \tau}{\sigma} \right)^{-\frac{1}{\xi} - 1} \right], & \text{if } x > \tau. \end{cases} \tag{1}$$

where τ is a threshold income, and p_τ is a corresponding threshold income percentile. The bottom p_τ% of incomes are represented by the nonparametric kernel density, where N is the number of observations, h is a smoothing parameter (bandwidth), and $K(\cdot)$ is a kernel function. The upper $(1 - p_\tau)$% of incomes are approximated by a GPD model where τ is a location parameter, ξ is a shape, and σ is a scale. The weights p_τ and $(1 - p_\tau)$ are introduced to guarantee the integral of the semiparametric PDF is equal to one.

The parameters of the distribution in (1) are estimated as follows. Common rules of

[1] (See, e.g., Cowell & Flachaire, 2007)

thumb are applied to the choice of bandwidth h and kernel function $K(\cdot)$. Namely, for $K(\cdot)$ we use a Gaussian kernel and h is estimated by cross-validation (CV). A threshold percentile p_τ was arbitrarily chosen at 90^{th} percentile, and this choice was governed by balancing a standard bias-variance trade-off: the fewer observations on the tail considered as a part of a GPD density, the more volatile the GPD parameters estimates are but the lower the potential bias, and vice versa (Charpentier & Flachaire, 2019). The parameters of the Generalized Pareto top income tail are estimated by applying a maximum penalized likelihood estimator (MPLE) originating from the extreme values literature (Coles & Dixon, 1999; Huang, Lin, & Ren, 2012; Martins & Stedinger, 2000, 2001; Pauli & Coles, 2001). Following Butaeva (forthcoming), beta distribution with parameters $\hat{\alpha} = 8$ and $\hat{\beta} = 8$ ($beta(8, 8)$) has been chosen as a penalty function. Butaeva (forthcoming) shows that unlike other estimators commonly used in income inequality literature (e.g. MLE, OBRE[1]), a MPLE demonstrates high efficiency and robustness in the presence of sparse high-income observations as well as potential top-income data contaminations. To compute confidence intervals, we use a semiparametric bootstrap algorithm proposed in Davidson and Flachaire (2007) and validated in Cowell and Flachaire (2007). Semiparametric bootstrap demonstrates better finite-sample performance due to the higher stability of the parametrically modeled upper tail. The robustness of the baseline results was checked against the threshold, bandwidth, and penalty function choices. For more details on the estimation procedures, see Butaeva (forthcoming).

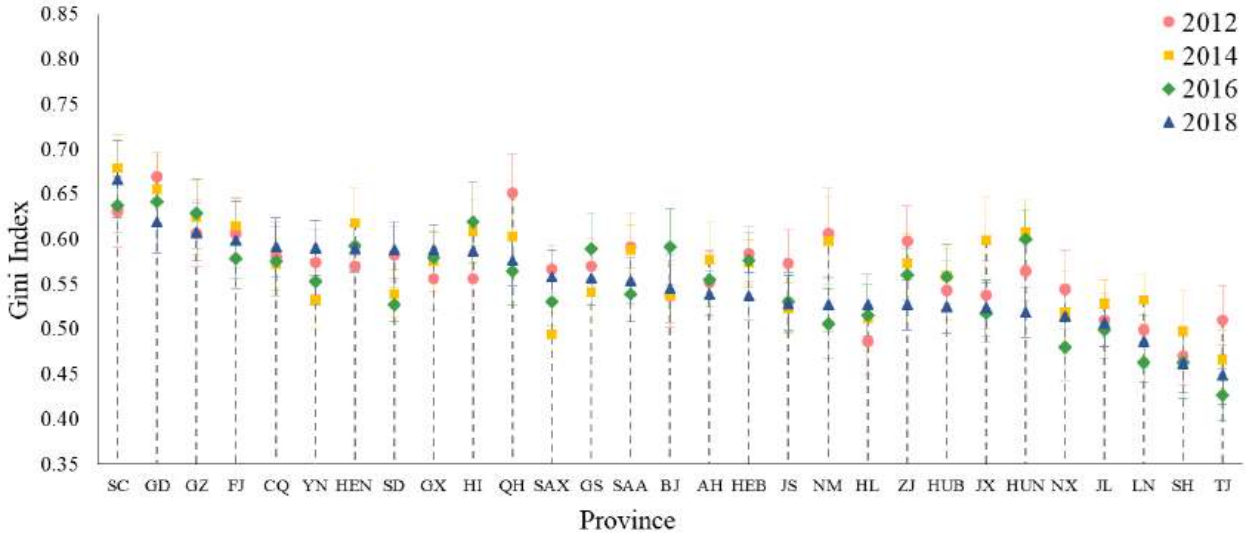
Figures 1 and 2 show the estimated provincial-level Gini index and mean household income per capita, respectively.[2] The spatial distribution of provincial inequality estimates is presented on China's maps in Figure C1, Appendix C. It can be seen from Figure 1 that there is a substantial variation in levels of income inequality among Chinese provinces. In 2018, the provincial Gini in China ranged from 0.45-0.46 in Tianjin and Shanghai to 0.62-0.67 in Guangdong and Sichuan. The largest increase (6-8%) in income inequality from 2012 to 2018 occurred in Heilongjiang, Guangxi, Sichuan, and Hainan. The largest decrease (12-13%) was in Tianjin, Zhejiang, and Inner Mongolia. Notably, over the period under study, the most unequal province, Sichuan, experienced the largest increase in inequality, and the least unequal, Tianjin, had the largest decline in inequality, which led to a widening gap in income inequality across Chinese provinces from 2012-2018.

Economic disparities across provinces are also substantial. Mean household income per capita in 2018 ranges from around 18K CN¥ in Guangxi and Hebei to 69K CN¥ in Shanghai

[1]OBRE was introduced by Victoria-Feser and Ronchetti (1994) and later used in many other inequality studies (e.g., Cowell & Victoria-Feser, 2007; Jenkins, 2017; Van Kerm, 2007; Victoria-Feser, 2000).

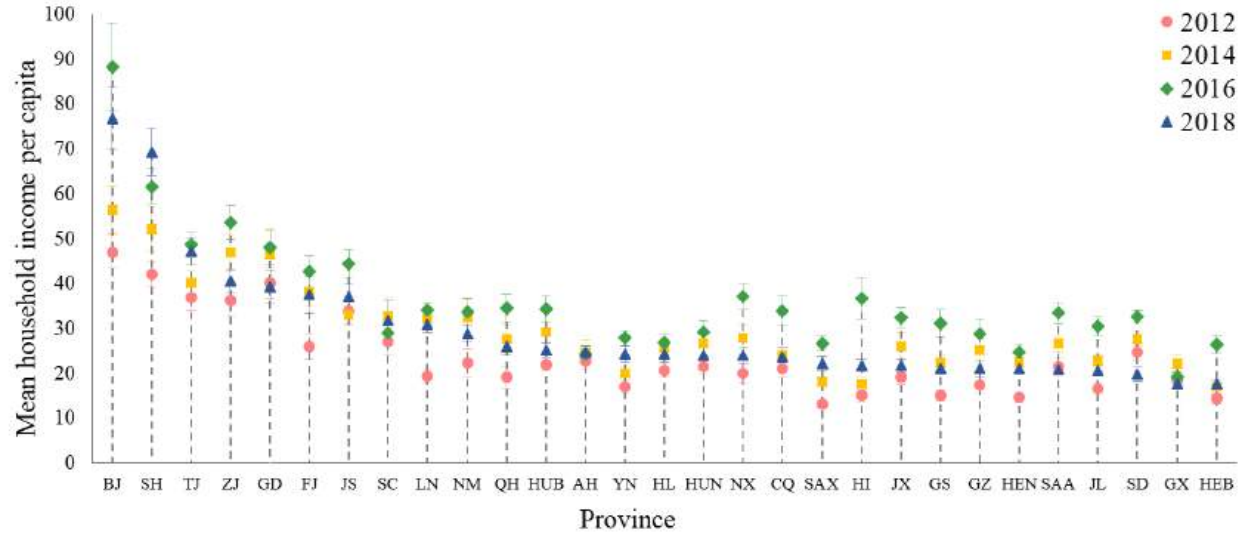
[2]The corresponding point estimates are reported in Table C1, Appendix C. Provincial abbreviations presented in Table C1 are used throughout the paper. The provincial order in all tables is similar: descending on average (among 2012, 2014, 2016, and 2018) GRP per capita.

Figure 1: Baseline provincial Gini index estimates for China.



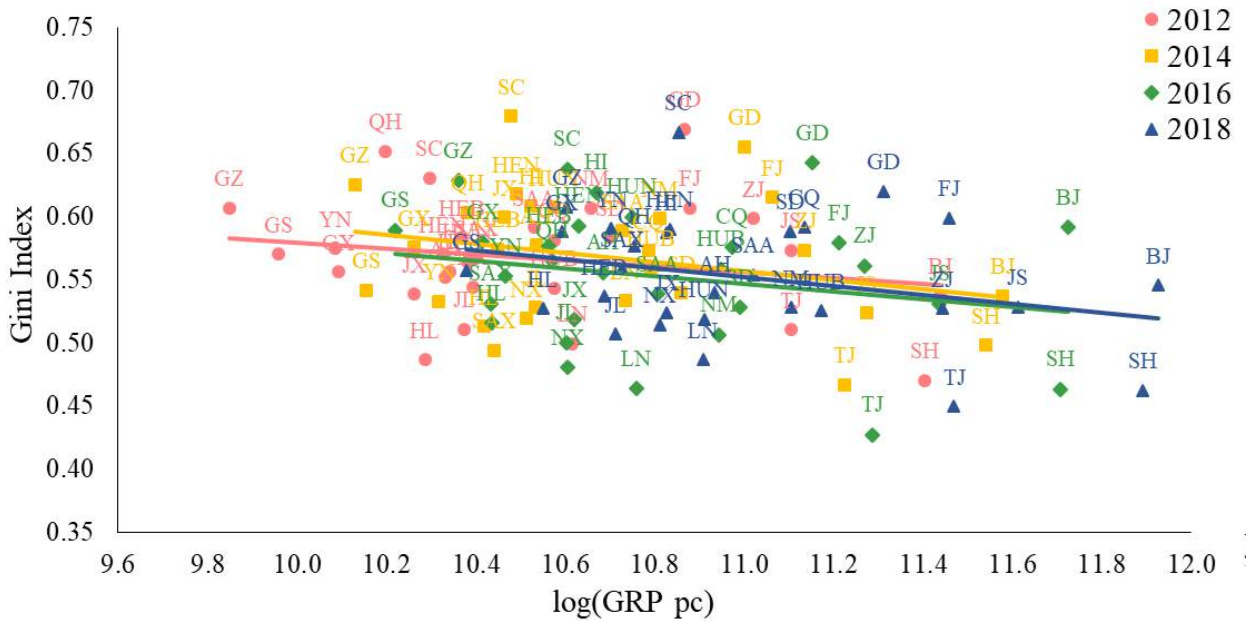
Notes: This figure shows our baseline provincial income Gini index estimates obtained from the CHFS. The provinces on the x-axis are arranged in descending order of the Gini index in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure 2: Provincial mean household income per capita (in 1000 CN¥).



Notes: This figure shows our baseline provincial mean household disposable income per capita estimates obtained from the CHFS. The provinces on the x-axis are arranged in descending order of the indicator in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure 3: Comparison of provincial Gini index estimates and log GRP per capita.



Notes: This figure shows our baseline provincial Gini index estimates obtained from the CHFS plotted versus the log GRP per capita from China Statistical Yearbook. Correlations are: -20% in 2012, -28% in 2014, -22% in 2016, and -29% in 2018.

and 77K CN¥ in Beijing (Figure 2). The ratio of rich to poor GRP per capita (Table E1, Appendix E) among Chinese provinces in 2018 (4.7 times) is even larger than the ratio of rich to poor mean household income per capita (4.3 times). From Figure 3, we also observe that poor provinces (e.g., Guizhou, Guangxi, Yunnan, Qinghai) are also more unequal, and richer provinces (e.g., Beijing, Shanghai, Jiangsu, Tianjin) are less unequal. We find a negative correlation between log GRP per capita and estimated provincial Gini: -20% in 2012, -28% in 2014, -22% in 2016, and -29% in 2018. The exceptions are Fujian and Guangdong, which are both rich (by GRP per capita) and highly unequal and Heilongjiang and Shanxi, which are relatively poor and less unequal.

Several points should be added regarding the benefits of the estimation method used. First, while estimating Gini coefficients, we noticed that several observations of top incomes in each province have a large effect on provincial Gini index estimates. There could be different reasons for yearly shifts in the composition of top income outliers: for example, an increase in the sample size could add a new top income household or, conversely, some of the top income participants may quit the survey. Moreover, households may experience large swings in their household finance. All these non-systematic factors could contribute to large changes in the resulting Gini index estimates from year to year. The application of the estimation method from Butaeva (forthcoming), which is based on the semiparametric den-

sity function (1), helps us to address this issue by treating top incomes on the distributional tail as a continuous Generalized Pareto function in which parameters are estimated with the robust and efficient MPLE. This substantially improves the stability of the obtained income inequality estimates. It can be seen from Figure D1, Appendix D that the resulting provincial Gini's are much more stable over time than the Gini estimates from the raw data (the variation in Gini index across years, on average across provinces, drops for 68%). Second, this substantial improvement of stability in Gini estimates generates more stable correlations among the key provincial economic statistics reported by NBS. For example, Gini estimates from the raw data without any model assumptions are correlated with the log GRP per capita by -26% in 2012, -14% in 2014, -1% in 2016, and -25% in 2018, which is much less stable compared to the correlation between log GRP per capita and our robust baseline Gini reported in the previous paragraph. Finally, the considered estimation method doesn't upset the provincial income rankings. Our mean household income per capita estimates are correlated with the raw data mean incomes by 95% in 2012, 93% in 2014, 99% in 2016, and 97% in 2018. Our estimates are also substantially correlated with the per capita disposable income from the China Statistical Yearbook (CSY). The relevant correlations are: 89% in 2012 and 2014, 91% in 2016, and 95% in 2018.

Figures I4-I32 in Appendix I show the robustness check towards the choice of threshold, bandwidth, and the likelihood penalty function. Together with the baseline threshold of 90^{th} percentile, we consider lower (85^{th} percentile) and higher (95^{th} percentile) thresholds for separating the nonparametric kernel density approximating the distribution bulk from the Generalized Pareto density representing top income tail. Together with the default cross-validated bandwidths ($h = CV$), we consider $h = 0.5 \times CV$ and $h = 0.5 \times CV$ bandwidths. We also tried more, $beta(11, 11)$, and less, $beta(5, 5)$, restrictive penalties to the likelihood function, compared to the baseline $beta(8, 8)$. Conditioning on confidence intervals, all anticipated alternative approaches reveal no substantial changes in Gini estimates compared to our baseline results.

5 Inequality decomposition

5.1 Theil T index decomposition by population subgroups

To analyze the relative contribution of inter- and intraprovincial variation in incomes to nationwide income inequality in China, we perform a subgroup decomposition of Theil T

inequality index by province. The Theil T index formula is:

$$T = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\bar{y}}ln\left(\frac{y_i}{\bar{y}}\right), \tag{2}$$

where y_i is an annual income per capita of a household i, $\bar{y}$ is a sample mean of household income per capita, and N is the number of households in the sample.

We choose Theil T index because, in contrast to many other common inequality measures, it can be effectively decomposed into within and between population sub-group components without an overlap component (which is typically difficult to interpret). We decompose Theil T index into between-province and within-province shares applying the following formula:[1]

$$T = \sum_{g=1}^{h} s_g T_g + \sum_{g=1}^{h} s_g ln\left(\frac{s_g}{p_g}\right), \tag{3}$$

where the first term is the "within" component (T_g is each subgroup's Theil index, weighted by the subgroup's share in total income $s_g = p_g\bar{y}_g/\bar{y}$), and the second term is the "between" component. Note that the Theil T index is calculated by using the mean income of each subgroup, where $p_g = N_g/N$ is the share of each subgroup's population N_g in the total population N.

Table 1 presents the results of the decomposition of nationwide income inequality in China by province. We can see that within-provincial inequality explains around 87-92% of nationwide income inequality measured by the Theil T index, and the remaining 8-12% is attributed to the between-provincial income variation. It can be also observed that the largest share of nationwide inequality, around 9-11%, is explained by the income inequality in Guangdong province (due to the large share of the national population in Guangdong). The second most important province in explaining nationwide inequality are Sichuan and Zhejiang provinces (6.5%, on average). The correlation coefficient between the provincial Theil T indices and our baseline Gini indices are: 97% in 2012, 96% in 2014, 97% in 2016, and 96% in 2018.

We also apply eq. (3) to study the variation of incomes within and btween educational groups. In Table 2, we summarize the results of the inequality decomposition by educational level of working-age[2] household members. We can see that the dominant part, 76-81%, of nationwide income inequality in China is attributed to variation within educational groups, and the remaining 19-24% is explained by between-group variation.

[1] This formula goes back to (Bourguignon, 1979; F. A. Cowell, 1980; Shorrocks, 1984). More details can also be found in F. A. Cowell (2000) and Jain-Chandra et al. (2018).

[2] Here and below, we use 15 and 64 years-old as thresholds while referring to working-age individuals.

Table 1: Decomposition of nationwide income inequality by province.

	2012		2014		2016		2018	
	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beijing	0.379	4%	0.367	3%	0.496	5%	0.391	4%
Shanghai	0.292	3%	0.281	2%	0.264	2%	0.247	2%
Jiangsu	0.443	5%	0.401	6%	0.372	6%	0.388	7%
Tianjin	0.335	2%	0.286	1%	0.240	1%	0.269	1%
Zhejiang	0.487	9%	0.444	6%	0.435	6%	0.395	5%
Fujian	0.525	3%	0.535	3%	0.459	3%	0.491	3%
Guangdong	0.704	10%	0.571	11%	0.579	10%	0.530	9%
Shandong	0.502	4%	0.409	6%	0.419	5%	0.497	5%
Inner Mongolia	0.519	1%	0.542	2%	0.395	1%	0.396	2%
Hubei	0.412	4%	0.444	4%	0.447	4%	0.383	3%
Chongqing	0.473	3%	0.495	2%	0.454	2%	0.495	2%
Shaanxi	0.502	2%	0.543	3%	0.422	2%	0.485	2%
Liaoning	0.336	2%	0.410	4%	0.291	3%	0.344	3%
Hunan	0.480	4%	0.536	4%	0.495	4%	0.370	3%
Anhui	0.436	3%	0.456	3%	0.445	3%	0.420	3%
Hainan	0.497	1%	0.564	0%	0.543	1%	0.527	0%
Ningxia	0.443	1%	0.361	0%	0.341	0%	0.380	0%
Henan	0.481	2%	0.576	5%	0.517	4%	0.542	5%
Sichuan	0.562	6%	0.667	8%	0.588	5%	0.600	7%
Jilin	0.382	2%	0.411	1%	0.332	2%	0.379	1%
Jiangxi	0.432	2%	0.538	3%	0.403	2%	0.374	2%
Hebei	0.499	2%	0.509	3%	0.462	4%	0.424	3%
Shanxi	0.494	2%	0.365	1%	0.409	2%	0.457	2%
Qinghai	0.646	2%	0.571	0%	0.481	0%	0.515	0%
Heilongjiang	0.341	2%	0.353	2%	0.350	2%	0.373	2%
Yunnan	0.502	2%	0.429	2%	0.472	3%	0.509	3%
Guangxi	0.467	2%	0.516	2%	0.541	2%	0.526	2%
Guizhou	0.549	2%	0.595	2%	0.577	2%	0.565	2%
Gansu	0.488	2%	0.449	1%	0.488	2%	0.473	1%
Within	0.469	89%	0.474	92%	0.442	90%	0.436	87%
Between	0.057	11%	0.041	8%	0.050	10%	0.063	13%
Total	0.526	100%	0.515	100%	0.492	100%	0.499	100%

5.2 Regression-based decomposition of income inequality

To analyze the possible explanatory factors shaping the levels of income inequality in China, we turn to the regression-based decomposition. We apply the following variant of a Mincer

Table 2: Decomposition of nationwide income inequality by educational level of working-age (15-64) household members.

	2012		2014		2016		2018	
	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Primary school and below	0.592	10%	0.620	8%	0.615	9%	0.529	11%
Middle school	0.505	24%	0.501	21%	0.489	21%	0.448	22%
High school	0.419	24%	0.430	25%	0.407	26%	0.376	23%
College and above	0.325	23%	0.331	27%	0.303	25%	0.278	20%
Within	0.421	80%	0.415	80%	0.397	80%	0.374	75%
Between	0.105	20%	0.102	20%	0.099	20%	0.121	25%
Total	0.526	100%	0.517	100%	0.496	100%	0.496	100%

model (Mincer, 1974) for each year of the data:

$$\ln Y_i = \delta_0 + \delta_1 S_i + \delta_2 E_i + \delta_3 E_i^2 + \delta_4 \ln N_i + \sum_{l=1}^{3} \kappa_l K_{li} + \sum_{k=1}^{2} \eta_k X_{ki} + \sum_{q=1}^{Q-1} \gamma_q G_{qi} + \varepsilon_i, \qquad (4)$$

where $\ln Y_i$ is a logarithm of the annual income per capita of each household i; S_{ijt} is years of schooling per worker defined as average years of schooling among working-age (15-64) household members; E_i is average years of potential working experience of working-age household members (we follow the existing literature and use a standard measure of potential experience defined for each household member as his/her age minus years of schooling minus six); ${E_i}^2$ is average experience squared; and ε_i is an error term which represents other unobserved household-specific factors. In accordance with human capital theory, education enhances productivity and, therefore, more education is typically associated with higher earnings. Consequently, we expect to obtain a positive sign on the δ_1 coefficient estimate ($\hat{\delta_1} > 0$). The classical Mincer model uses individual earnings as the dependent variable. However, in the current study, we focus on household income per capita. Therefore, we must transform the regression equation accordingly by considering per-worker measures of education and experience, and by controlling for the log of household size $\ln N_i$. We also control for K_{li}, a set of factors shaping components of the household income other than labor income, namely: (a) to control for the agricultural business income endowment, we include the *Land dummy* variable, which indicates whether the household owns agricultural land; (b) we also include $Ln(land\ per\ capita)$, which is a log of the amount of land, owned by the household (in *mu*[1]) divided by the household size; (c) to consider other potential business

[1]One *mu* is equivalent to 0.06667 hectare.

income sources, we add $Ln(business\ assets\ per\ capita)$, a log of the total value of household assets from all business projects (in 100K CN¥) divided by the household size. We also control for the household-level covariates X_{ik} which, according to the existing literature, may also affect household income levels in China (namely, gender and Communist Party membership of the household head). To capture geographic factors, we also include G_{iq} spatial dummies represented in different specifications by rural and provincial dummies or rural and city dummies.

Following Fields (2003), we additionally estimate the share of the log-variance of household income per capita that can be attributed to each regressor in eq. (4):

$$s_z = \frac{\hat{\alpha}_z cov(V_z, \ln Y)}{\sigma^2(\ln Y)}, \tag{5}$$

where s_z is a resulting % of income inequality which can be attributed to the zth regressor in eq. (4), $\hat{\alpha_z}$ is estimated regression coefficient on each z^{th} independent variable, $cov(V_z, \ln Y)$ is the covariance between each z^{th} independent variable and the explanatory variable, and $\sigma^2(\ln Y)$ is the variance of the explanatory variable. The sign of the resulting s_z indicates whether the z^{th} regressor increases ($s_z > 0$) or decreases ($s_z < 0$) income inequality.

Table 3 reports the regression-based decomposition results. The OLS estimates of eq. (4) controlling for provincial dummies are reported in odd columns from (1) to (7) and those controlling for city dummies are reported in odd columns from (9) to (15). Each even column presents $\hat{s}_z$ estimates from eq. (5) reflecting the relative contribution of regressors from eq. (4) in nationwide income inequality.

Table 3 suggests that a one-year increase in the years of schooling per worker is associated in different years with a 10-12% increase in household income per capita. Variation in the years of schooling per worker explains the largest share of the variation in (log) household income per capita (10-14%). The positive sign of $\hat{s}_z$ estimates for schooling suggests that greater variation in the years of schooling per worker among households increases income inequality. Our results also confirm positive but diminishing returns to potential experience per worker. The inflection point in experience per worker is around 22-29 years old. Household size, naturally, is negatively correlated with (log) household income per capita, and it explains 4-6% of its variation. Households who own agricultural land receive on average lower incomes than those without land. A 1% increase in *mu* of owned agricultural land is associated with a 0.10-0.30% increase in the household income per capita. A 1% increase in the value of household business assets is associated with around 0.08-0.16% increase in income. Households whose heads are Communist Party members have 18-23% larger incomes per capita. Households with female heads receive 8-21% lower incomes than those

with male heads. Rural households have 18-53% lower incomes than urban households. The rural dummy explains 2-8% of inequality in China. Provincial and city dummies explain 3-5% and 7-12% of inequality in China, respectively.

To sum it up, Table 3 shows that the most important factor explaining the variation in household income is the educational attainment of working-age household members (explains around 10-15%). The second-largest contributors to income inequality are geographic factors: city dummies (7-12%), provincial dummies (4-5%), and the rural dummy (2-8%). The dominant role of education in explaining household income in China was also found by Goh, Xubei, and Nong (2009), who apply] the analogous Mincer regression decomposition framework to the 1989–2004 data from the China Health and Nutrition Survey (CHNS). However, in contrast to Wan and Zhou (2005)[1]; we haven't found any suggestive evidence on the prevalent impact of capital input in the variation of household income in China.

[1]Wan and Zhou (2005) uses a similar Mincer regression decomposition framework to study rural household incomes in China in 2002, employing the data from the Research Centre for Rural Economy (RCRE, Ministry of Agriculture).

Table 3: Mincer regression: explaining household income per capita variation in China. Heteroskedasticity-robust SEs are used.

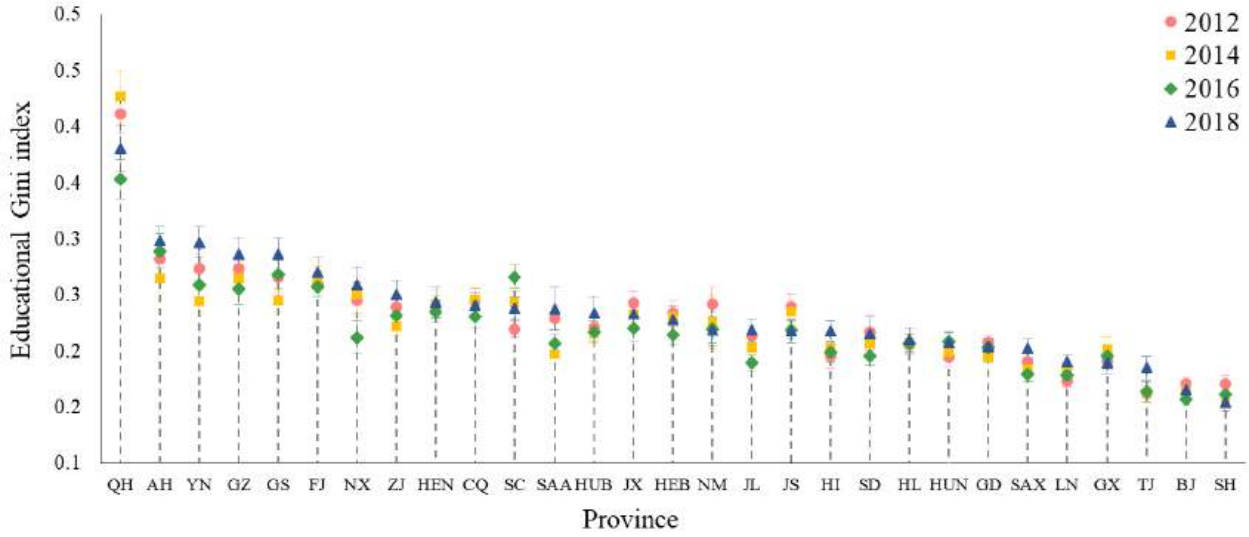
	Dependent variable: Ln(household income per capita)															
	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Schooling per worker	0.119***	12.7%	0.104***	11.0%	0.121***	14.4%	0.106***	12.2%	0.107***	11.4%	0.097***	10.2%	0.112***	13.2%	0.094***	10.9%
	(0.004)		(0.004)		(0.003)		(0.004)		(0.004)		(0.004)		(0.003)		(0.004)	
Experience per worker	0.015***	-2.4%	0.022***	-4.6%	0.029***	-6.3%	0.020***	-4.9%	0.017***	-2.6%	0.024***	-4.9%	0.033***	-6.9%	0.020***	-4.8%
	(0.004)		(0.004)		(0.003)		(0.003)		(0.004)		(0.004)		(0.003)		(0.003)	
Experience per worker^2 : 100	−0.026***	2.0%	−0.046***	5.3%	−0.052***	7.0%	−0.044***	6.5%	−0.030***	2.4%	−0.049***	5.6%	−0.057***	7.7%	−0.045***	6.6%
	(0.006)		(0.006)		(0.005)		(0.005)		(0.006)		(0.006)		(0.005)		(0.005)	
Ln(household size)	−0.897***	5.8%	−0.817***	4.6%	−0.833***	5.5%	−0.803***	3.8%	−0.889***	5.7%	−0.784***	4.4%	−0.806***	5.3%	−0.791***	3.7%
	(0.028)		(0.033)		(0.028)		(0.030)		(0.028)		(0.033)		(0.028)		(0.030)	
Land dummy	−0.063***	0.5%	−0.281***	3.5%	−0.274***	3.1%	−0.049**	0.4%	−0.042*	0.3%	−0.266***	3.3%	−0.215***	2.5%	−0.025	0.2%
	(0.021)		(0.035)		(0.030)		(0.024)		(0.022)		(0.035)		(0.031)		(0.025)	
Ln(land per capita)	0.107	-0.0%	0.198***	-2.0%	0.180***	-1.6%	0.314***	-0.1%	0.054	-0.0%	0.193***	-1.9%	0.150***	-1.3%	0.268***	-0.1%
	(0.080)		(0.026)		(0.023)		(0.060)		(0.083)		(0.027)		(0.024)		(0.060)	
Ln(business assets per capita)	0.153***	2.0%	0.159***	2.3%	0.118***	1.5%	0.075***	0.6%	0.145***	1.9%	0.156***	2.3%	0.116***	1.4%	0.074***	0.6%
	(0.007)		(0.007)		(0.008)		(0.010)		(0.007)		(0.007)		(0.008)		(0.009)	
Party member	0.223***	0.9%	0.201***	0.8%	0.179***	0.6%	0.178***	0.7%	0.235***	0.9%	0.218***	0.9%	0.183***	0.6%	0.193***	0.7%
	(0.024)		(0.023)		(0.022)		(0.024)		(0.023)		(0.023)		(0.022)		(0.024)	
Female	−0.083***	-0.2%	−0.093***	-0.2%	−0.203***	-0.2%	−0.078***	-0.1%	−0.087***	-0.2%	−0.096***	-0.2%	−0.211***	-0.2%	−0.096***	
	(0.020)		(0.022)		(0.021)		(0.022)		(0.020)		(0.021)		(0.021)		(0.022)	
Rural	−0.249***	2.9%	−0.527***	7.8%	−0.343***	4.2%	−0.399***	5.5%	−0.178***	2.0%	−0.484***	7.2%	−0.288***	3.6%	−0.299***	4.1%
	(0.022)		(0.029)		(0.023)		(0.023)		(0.023)		(0.031)		(0.024)		(0.023)	
Constant	9.404***		9.532***		9.540***		9.820***		9.081***		9.472***		9.411***		10.074***	
	(0.091)		(0.101)		(0.087)		(0.095)		(0.134)		(0.146)		(0.146)		(0.139)	
Provincial dummies	Yes	5.0%	Yes	3.4%	Yes	4.6%	Yes	5.2%	No		No		No		No	
City dummies	No	No	No	No	No	No	No	No	Yes	9.8%	Yes	6.8%	Yes	9.4%	Yes	11.9%
Residual		70.8%		68.1%		67.2%		70.2%		68.4%		66.3%		64.7%		66.3%
Observations	24,143		30,878		33,096		28,330		24,143		30,878		33,096		28,330	
R^2	0.317		0.343		0.351		0.326		0.341		0.361		0.375		0.365	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

6 Educational inequality and returns to schooling

To measure educational (human capital) inequality, we estimate educational Gini indices from two different data sources: CHFS for our baseline results and the China Population and Employment Statistical Yearbook (CPESY) for the robustness check. To calculate provincial-level educational Gini in China, we map years of schooling to educational attainment levels as follows: zero years for no schooling at all; six years for primary school; nine years for junior school; twelve years for regular or technical high school; fifteen years for vocational college; sixteen years for regular college and bachelor's; nineteen years for a master's degree; and twenty-two years for a doctoral degree.

Figure 4: Provincial-level educational Gini index in China.

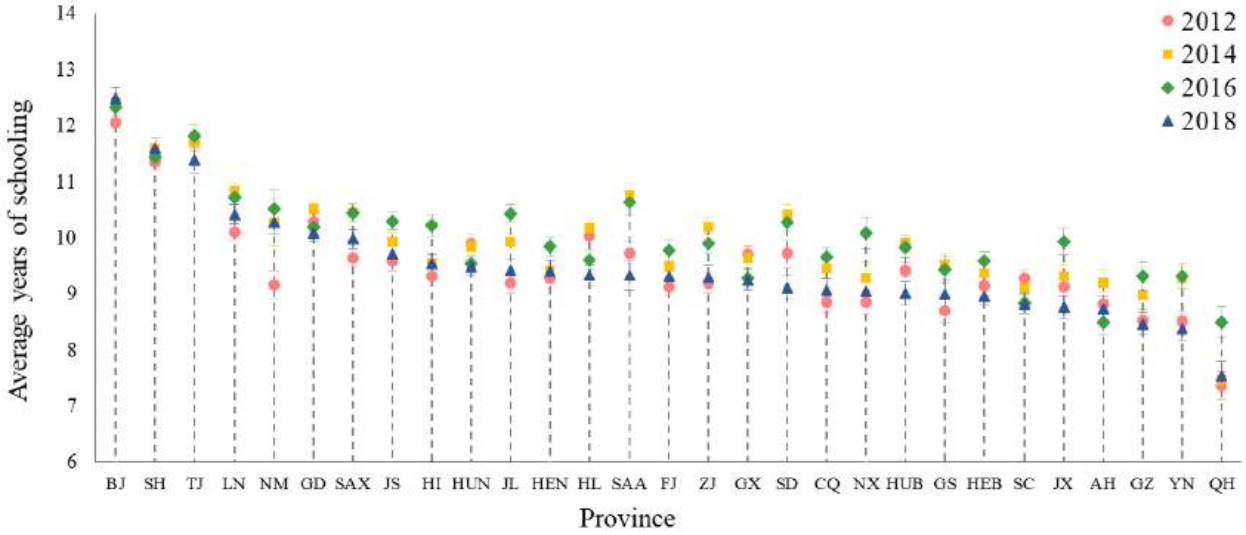


Notes: This figure plots the provincial-level educational Gini index estimated from the CHFS. The provinces on the x-axis are arranged in descending order of the indicator in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure 4 reports our baseline estimates of the provincial-level educational Gini indices from the CHFS[1]. The educational Gini index ranges from around 0.160 in Shanghai and Beijing to 0.39 in Qinghai. One notable observation is that provincial educational inequality has a strong negative correlation with average years of schooling (Figure 5): -83% in 2012, -86% in 2014, -83% in 2016, and -81% in 2018. The average number of years of schooling varies across provinces, from 7.7 in Qinghai to 12.3 in Beijing. Provincial educational patterns are also highly correlated with the level of economic development. Richer provinces, like Beijing, Shanghai, and Tianjin, exhibit lower levels of educational inequality (the cor-

[1]The corresponding point estimates of all educational variables presented in the current Section are reported in Table C2, Appendix C.

Figure 5: Provincial-level average years of schooling in China.

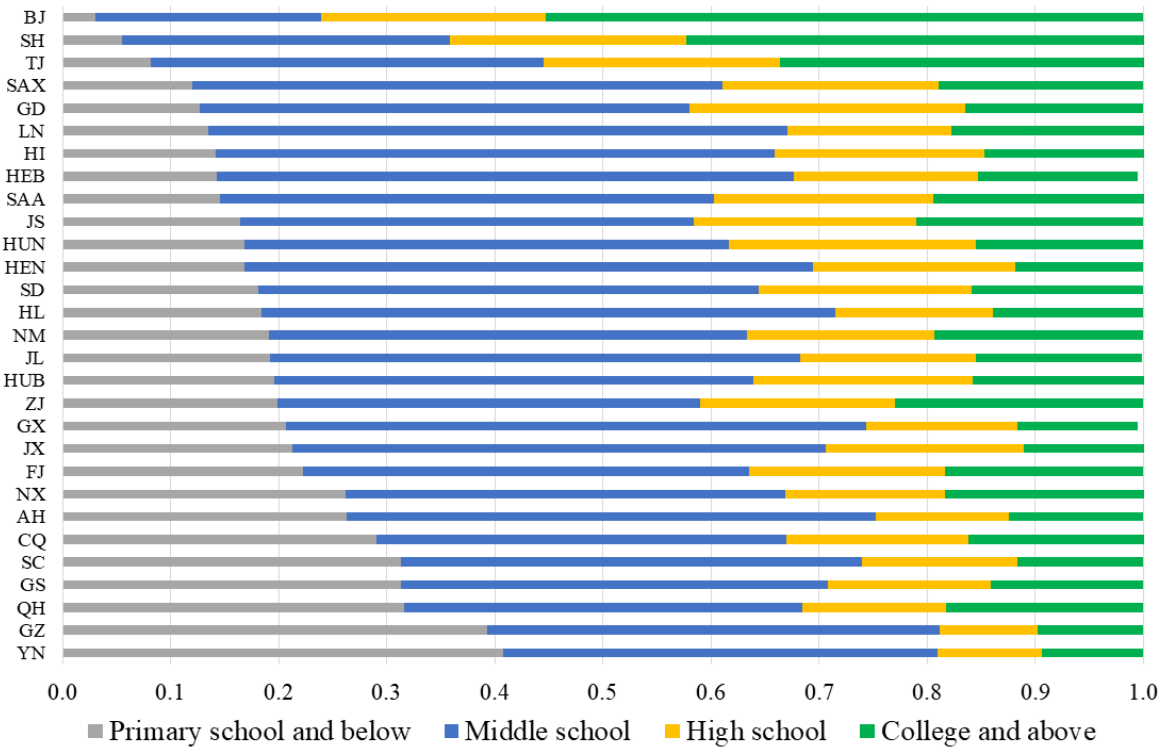


Notes: This figure plots the provincial-level average years of schooling estimated from the CHFS. The provinces on the x-axis are arranged in descending order of the indicator in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

relation between (log) GRP per capita and the educational Gini is −45% in 2012, −43% in 2014, −42% in 2016, and −44% in 2018), and their citizens have higher average levels of education (the correlation between (log) GRP per capita and average years of schooling Gini is 73% in 2012, 72% in 2014, 67% in 2016, and 69% in 2018). These patterns are likely to be attributed to provincial differences in educational attainment structure. Based on the data from the CPESY, in Figure 6, we summarize the provincial-level distributions of the employed population by their educational level averaged across four waves and grouped in four categories: at most primary school, middle school, high school, or "college degree and above"[1]. We can see that the richer provinces, like Beijing, Shanghai, and Tianjin, have a lower share of the employed population with primary school and below education and a higher share of employed individuals with college and above degrees. In the poorer provinces, like Yunnan, Guizhou, and Gansu, the distribution of education levels is reversed: they have the largest share of unskilled labor and the smallest share of high-skilled labor among all provinces. It can also be noticed that except in Beijing, Shanghai, and Tianjin, high school and college and above categories are more equally distributed across provinces compared to primary and middle school categories. Therefore, it seems that differences in educational inequality across provinces in China are mostly driven by the variation in the share of low-

[1]In Figure E1, Appendix C, we report the analogous educational structure reported in the China Statistical Yearbook (CSY). Figure E1 and Figure 6 demonstrate similar patterns for larger (six-years of education and above) and smaller (employed) subpopulations in Chinese provinces.

Figure 6: Average educational structure in Chinese provinces, 2012-2018.



Notes: This figure plots the average educational structure of the employed population in Chinese provinces for 2012-2018 estimated from the China Population & Employment Statistics Yearbook (CPESY). The estimates are ordered in the "Primary school and below" category.

and middle-skilled labor.

Next, we evaluate differences in returns to human capital across provinces. To do so, we estimate provincial-level returns to schooling and returns to high school for each province and year. For returns to schooling, we apply the following provincial-level analog of eq. (4)[1] to each year of the CHFS household-level data:

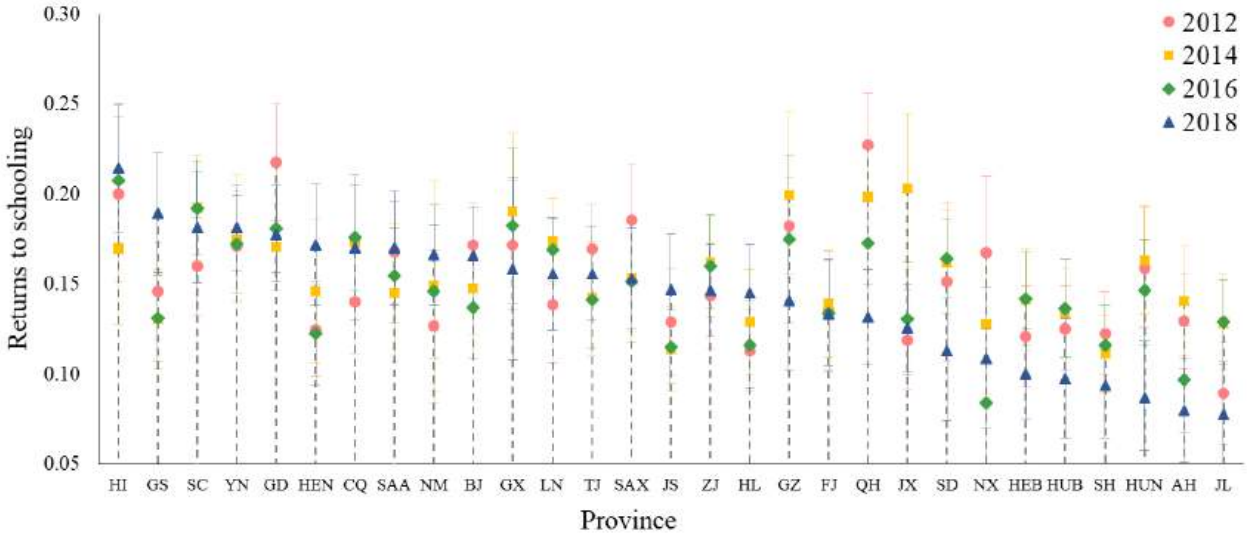
$$\ln Y_{ijt} = \delta_{0jt} + \delta_{1jt} Z_{ijt} + \delta_{2jt} E_{ijt} + \delta_{3jt} E_{ijt}^2 + \delta_{4jt} \ln N_{ijt} + \varepsilon_{ijt}, \tag{6}$$

where for each household i in province j and year t, $Z_{ijt} = \{S_{ijt}; H_{ijt}\}$ denotes either years of schooling per worker S_{ijt} or a dummy variable indicating whether working-age household members obtained, on average, high school or above educational level ($H_{ijt} = 1$), or not ($H_{ijt} = 0$). In case of $Z_{ijt} = S_{ijt}$, we call the estimate of $\hat{\delta}_{1jt}$ *"returns to schooling"*, which defines

[1]To evaluate provincial-level returns to schooling and returns to high school, we will focus on the human capital explanation of the variation in the household income per capita and ignore non-labor income covariates X_{ik}, the household head characteristics K_{li}, and spatial dummies G_{iq} from eq. (4).

an increase in household income per capita in response to an increase in average years of schooling among working-age household members[1]. In case of $Z_{ijt} = H_{ijt}$ the estimate of $\hat{\delta}_{1jt}$ is called *"returns to high school"*.

Figure 7: Provincial-level returns to schooling in China.



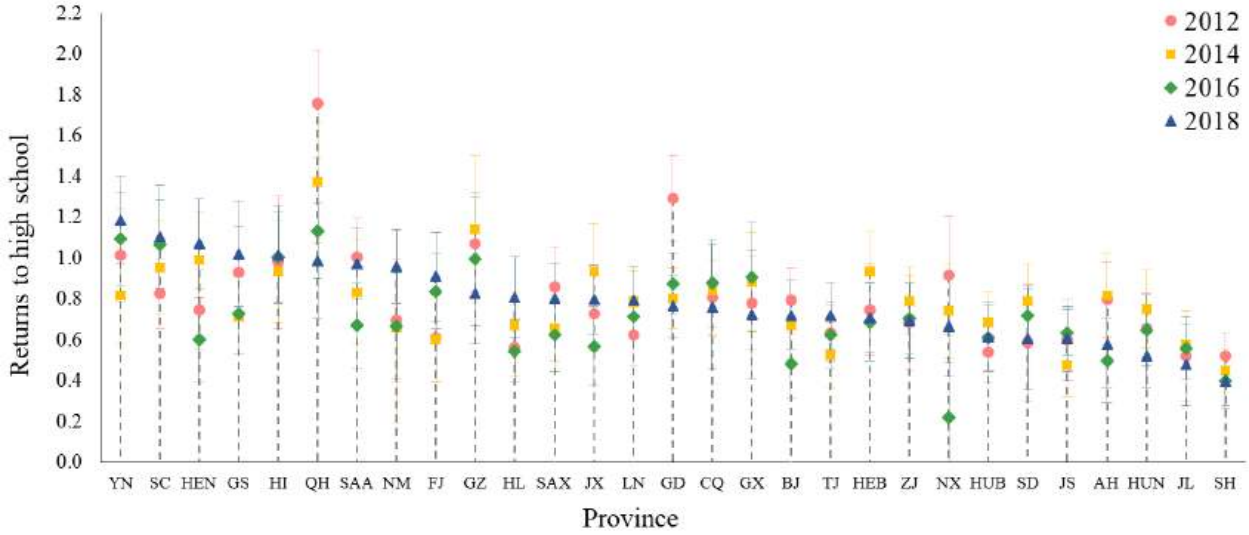
Notes: This figure plots the provincial-level returns to schooling estimated from the CHFS. The provinces on the x-axis are arranged in descending order of the indicator in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figures 7 and 8 report estimated provincial returns to schooling and returns to high school (respectively). On average in 2012-2018, an additional year of schooling of working-age household members varied from 10.5% (Jilin) to 19.9% (Hainan) increases in the household income per capita. The returns to high school vary from 43.9% in Shanghai to 101.2% in Yunnan. It should also be noted that richer provinces exhibit lower returns to schooling (the correlation between (log) GRP per capita and returns to schooling is -11% in 2012, -43% in 2014, -16% in 2016, and -8% in 2018) and lower returns to high school (the correlation between (log) GRP per capita and returns to high school is -36% in 2012, -60% in 2014, -30% in 2016, and -39% in 2018). These negative correlations could be explained by the fact that richer provinces have a higher relative supply of skilled labor than poor provinces.

In Figures C2-C5 (Appendix C) we also show the interrelations between all of the estimated human capital indicators and the provincial-level income Gini estimates introduced in Section 4. In short, this demonstrates that provincial income inequality is positively correlated with provincial educational inequality, provincial returns to schooling, and provincial returns to high school, and negatively correlated with provincial average years of schooling.

[1]Our definition of returns to schooling is close to the "ex-post returns to schooling" from Heckman, Lochner, and Todd (2006).

Figure 8: Provincial-level returns to high school in China.



Notes: This figure plots the provincial-level returns to high school estimated from the CHFS. The provinces on the x-axis are arranged in descending order of the indicator in 2018. Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

7 Interprovincial variation in income inequality

7.1 Provincial income inequality determinants

To evaluate the empirical relationship between the level of human capital development and variation in income inequality across provinces, we regress provincial income Gini, introduced in Section 4, on provincial educational Gini and returns to schooling from Section 6:

$$I_{jt} = \beta_0 + \beta_1 I^E_{jt} + \beta_2 \hat{\delta}_{1jt} + \beta_3 I^E_{jt} \times \hat{\delta}_{1jt} + \sum_{m=1}^{4} \tau_m T_{mjt} + \nu_t + \varepsilon_{jt}, \tag{7}$$

where I_{jt} is the income inequality Gini index in province j at year t, I^E_{jt} is the educational Gini index, $\hat{\delta}_{1jt}$ is the estimated returns to schooling, ν_t is year fixed effect, and ε_{jt} captures other unobserved factors that can vary across time and provinces.

To motivate our expectation toward the signs of $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$, let us consider a simplified version of the initial Mincer model from eq. (4):

$$\ln Y_i = \delta_0 + \delta_1 S_i + u_i, \tag{8}$$

where Y_i is annual household income per capita, S_i is average years of schooling among working-age household members, and u_i captures other factors that may influence household

income per capita. Then the variance of household income per capita can be expressed by the formula:

$$\text{Var}(\ln Y_i) = \delta_1^2 \text{Var}(S_i) + \text{Var}(u_i) + 2\text{Cov}(\delta_1 S_i, u_i), \tag{9}$$

where the variance in household income per capita $\text{Var}(\ln Y)$ might be considered as a proxy measure for income inequality, and $\text{Var}(S)$ is educational inequality. Therefore, if we believe that the income-generating process is defined by our Mincer model introduced in (4), then from (9) we can see that an increase in educational inequality or an increase in returns to schooling, and especially their interaction, would unambiguously lead to an increase in income inequality among households within the same province. Therefore, if we estimate eq. (7) with provincial fixed effects, we expect to obtain positive signs in the educational Gini, returns to schooling, and their interaction term.[1]

To capture other provincial-specific factors, which could be important for explaining variation in provincial income inequality, we further control eq. (7) for four T_{jtm} provincial-level covariates collected from the China Statistical Yearbook. First, we include the *Social spending* variable which measures local governments' expenditures on social insurance per working-age adult (in 1,000 CN¥).[2] We use this variable as a proxy for social public spending which aims at reducing income inequality, and, therefore, we expect to obtain a negative coefficient on this variable. Second, we include *SOE assets share* - the share of the total current assets of state-holding industrial enterprises in provincial GRP. This variable captures the size of the government sector in the provincial economy. Usually, the public sector exhibits a more egalitarian wage structure than the private sector. Therefore, we expect to find a negative relationship between *SOE assets share* and provincial-level income inequality. Third, we include *Foreign trade share* - the total volume of foreign trade (imports and exports) as a share of provincial GRP. This variable is used as a proxy for the degree of openness of the provincial economy. We expect to obtain a negative coefficient on this variable because in China greater trade openness is likely associated with increased demand and prices for unskilled labor, negatively impacting income inequality. Finally, we include the *Ln(GRP per capita)* to control for the differences in levels of economic development across provinces. We documented in Section 4 the negative relationship between log GRP per capita and income inequality in Chinese provinces, and we expect to confirm this correlation with a negative *Ln(GRP per capita)* coefficient estimate.

[1] For more details refer to Coady and Dizioli (2018).

[2] We use an indicator called "Social Safety Net and Employment Effort" which reflects the government's social security and employment spending. Related matters include social security and employment management services, civil administration affairs, fiscal subsidies to social insurance funds, employment subsidies, retirement resettlement, social welfare, disability services, urban minimum living security, etc.

Table 4: Regression estimates of Gini index determinants using educational Gini index calculated from CHFS. Heteroskedasticity-robust standard errors are used.

	Dependent variable: Gini index															
	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Educational Gini	0.455***		0.386***	0.471***	0.444***		0.390***	0.402***	0.595***		0.500**	0.494**	0.658***		0.574***	0.555**
	(0.093)		(0.099)	(0.093)	(0.083)		(0.073)	(0.071)	(0.177)		(0.199)	(0.216)	(0.191)		(0.210)	(0.225)
Returns to schooling		0.796***	0.711***	0.762***		0.763***	0.709***	0.716***		0.360***	0.316***	0.312**		0.385***	0.344***	0.333***
		(0.111)	(0.107)	(0.106)		(0.107)	(0.089)	(0.092)		(0.119)	(0.119)	(0.135)		(0.114)	(0.112)	(0.126)
Educational Gini × Returns to schooling				−5.349***				−0.698				0.168				0.584
				(1.589)				(1.450)				(1.346)				(1.294)
Social spending CSY					−0.012**	−0.010*	−0.015***	−0.015***					0.007	0.004	0.007	0.008
					(0.005)	(0.005)	(0.004)	(0.005)					(0.011)	(0.008)	(0.011)	(0.011)
SOE assets share CSY					−0.109***	−0.110***	−0.078***	−0.077***					0.056	0.064	0.108	0.113
					(0.034)	(0.031)	(0.028)	(0.028)					(0.086)	(0.075)	(0.085)	(0.087)
Foreign trade share CSY					0.032	−0.030	−0.002	−0.002					−0.038	−0.041	−0.041	−0.038
					(0.029)	(0.025)	(0.024)	(0.024)					(0.046)	(0.047)	(0.048)	(0.049)
Ln(GRP per capita) CSY					−0.030	−0.003	0.003	0.003					0.077	0.073	0.087*	0.092*
					(0.020)	(0.019)	(0.018)	(0.018)					(0.048)	(0.048)	(0.047)	(0.049)
2014.year	0.003	−0.003	−0.001	−0.001	0.012	−0.001	0.003	0.003	0.004	−0.001	0.002	0.002	−0.011	−0.015	−0.015	−0.016
	(0.011)	(0.011)	(0.010)	(0.009)	(0.011)	(0.011)	(0.010)	(0.010)	(0.006)	(0.006)	(0.006)	(0.006)	(0.012)	(0.010)	(0.011)	(0.012)
2016.year	−0.011	−0.011	−0.008	−0.008	0.011	−0.004	0.005	0.005	−0.009	−0.013*	−0.009	−0.009	−0.041**	−0.041**	−0.043**	−0.044**
	(0.012)	(0.011)	(0.010)	(0.010)	(0.014)	(0.014)	(0.012)	(0.012)	(0.007)	(0.008)	(0.007)	(0.007)	(0.020)	(0.019)	(0.020)	(0.021)
2018.year	−0.017	−0.007	−0.009	−0.011	0.013	0.003	0.009	0.008	−0.018***	−0.011*	−0.014**	−0.014**	−0.067**	−0.053*	−0.066**	−0.068**
	(0.011)	(0.011)	(0.010)	(0.010)	(0.018)	(0.017)	(0.016)	(0.016)	(0.006)	(0.006)	(0.006)	(0.006)	(0.032)	(0.030)	(0.033)	(0.034)
Constant	0.566***	0.565***	0.564***	0.566***	0.911***	0.645***	0.577***	0.575***								
	(0.007)	(0.007)	(0.006)	(0.006)	(0.205)	(0.198)	(0.189)	(0.189)								
Observations	116	116	116	116	116	116	116	116	116	116	116	116	116	116	116	116
R^2	0.203	0.265	0.391	0.428	0.396	0.473	0.564	0.565	0.183	0.200	0.248	0.248	0.221	0.234	0.295	0.296
Number of provinces									29	29	29	29	29	29	29	29

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4 reports our baseline estimates of eq. (7). It should be noted that we centered our key educational Gini I_{jt}^{E} and returns to schooling $\hat{\delta}_{1jt}$ variables by subtracting the mean in each case to make main effects in the specification with interaction terms in columns (5), (8), (12), and (16) more interpretable. The OLS results (columns 1-8) suggest a strong positive relationship between educational and income inequality in Chinese provinces. A 10% increase in the educational Gini is associated with about 3.8-4.7 percentage point increase in the income Gini index. The coefficient is significant at the 1% level in all OLS specifications. The mean provincial income Gini is 0.552-0.568 in different years, and the standard deviation is around 0.05 (Table C1, Appendix C). The same significant and positive relationship is found between provincial income Gini indices and returns to schooling. A 10% increase in returns to schooling is associated with a 7.1-8.0 percentage point increase in the income Gini. A negative sign on the interaction term in columns (5) and (8) suggests educational inequality has a lower effect on income inequality in places with higher returns to schooling. And vice versa: returns to schooling have a lower effect on income inequality in areas with higher educational inequality. Since we centered I_{jt}^{E} and $\hat{\delta}_{1jt}$, the main effects in columns (5) and (8) indicate that a 10% increase in educational Gini is associated with 4-4.7 percentage point increase in the income Gini for the place with average returns to schooling. At the same time, a 10% increase in returns to schooling is associated with a 7.2-7.6 percentage point increase in the income Gini for the place with average returns to schooling.

The fixed effects results (columns 9-16) indicate positive effects of educational inequality, returns to schooling, and their interaction with income inequality within each province, as we expected from eq. (9). Educational inequality has a larger effect on income inequality in the fixed-effects specification: a 10% increase in the educational Gini within a province is associated, on average, with a 4.9-6.5 percentage point increase in the income Gini index. Returns to schooling have a lower impact on income inequality in the fixed-effects model: a 10% increase in returns to schooling in a province is associated with a 3.1-3.9 percentage point increase in income Gini. The main effects are significant at the 1-5% level, and an interaction term was found to be insignificant.

For other covariates, in the OLS specification, we find a significant negative impact of social public spending (*Social spending*) and government share of the provincial economy (*SOE assets share*) on income inequality (as we suggested above). A 1,000 CN¥increase in local governments' expenditures on social insurance per working-age adult is associated with a 1-1.5 percentage point increase in the income Gini. A 10% increase in the share of the total assets of state-holding industrial enterprises in provincial GRP is associated with a 7.7-11 percentage point increase in the income Gini. Economic openness proxied by the share of imports and exports in GRP (*Foreign trade*) was found to be negatively related to income

inequality, but the coefficient is not statistically significant in most of the specifications. The impact of the $Ln(GRP\ per\ capita)$is also insignificant.

To examine the robustness of these results, we also estimate eq. (7) using the educational Gini calculated from the Employment Statistics Yearbook (CPESY). The resulting estimates are reported in Table I1, Appendix I. Table I1 suggests that the positive impact of educational inequality on income inequality remains significant and of a similar magnitude, as in our baseline results: a 10% increase in the educational Gini is associated with about 3.7-4.9 percentage point increase in the income Gini index. However, in the fixed effects specification, we find effects which are half the size (compared to our baseline results) and insignificant. The main effects of returns to schooling on income inequality do not change their magnitude and remain significant in all specifications. Table I1 confirms a statistically significant negative effect of government presence in the provincial economy and government social spending on income inequality in the OLS specification.

We also trace the possible impact of migration on our results in an additional robustness check. Employing the individual-level CHFS, we assign individuals to provinces based on their *hukou* instead of residence. We then recalculate the educational Gini, assuming no migration.[1] We then re-estimate our baseline income inequality regression from eq. (7) using the adjusted educational Gini measures. Figure I1 in Appendix I shows the comparison between predicted income Gini from our baseline regression (vertical axes) with predicted values employing educational Gini based on *hukou* (horizontal axes). The predicted income Gini values are correlated at 99%, which is indicative that migration has not altered our results.

The impact of educational policies

Our key result on the negative impact of provincial educational inequality on income inequality suggests that targeted educational policies are promising instruments by which income inequality may be reduced. What are the main characteristics of educational inequality in China? In Table H1 (Appendix H), we report the results of educational inequality decomposition by province and birth cohorts. To perform this decomposition, we employ individual-level data from the CHFS. In each available data year in every province, we define 50 birth cohorts for every birth year of working-age adults (from 15 to 64 years old). Then we apply our standard formula from eq. (3), performing the decomposition by population subgroups. First, we decompose the educational Theil T index by province, and second, we decompose it by birth cohort within each province. From columns (2), (6), (10), and (14) in Table

[1]The percentage of those whose *hukou* location differs from their residence is around 3-6% of the CHFS samples.

H1, we can see that the within-province component explains the dominant share (98%) of variation in years of schooling in China. In columns (3)-(4), (7)-(8), (11)-(12), and (15)-(16) we report how the within-province Theil T is attributed to within- and between-birth cohort sub-components. Again, the within-birth cohort component describes the most significant part (80%) of the variation in years of schooling in each province.

In order to relate current provincial inequality levels to the historical data on provincial educational expenditures, we also perform the following additional analysis. First, for every birth cohort in every province/year, we estimate average government educational expenditures, employing the data from the China Compendium of Statistics 1949-2008 (CCS), and the China Statistical Yearbook (CSY). For every birth cohort in every province we calculate average government expenditures[1] on education at age 12 starting from the year individuals from each birth cohort were 6 years old. Second, we estimate the educational Gini index for every province/birth cohort. Third, we create an aggregated data set of year/province/birth cohort level (4 years $\times$ 29 provinces $\times$ 50 birth cohorts = 5,800 observations) and run OLS to identify the impact of educational expenditures on educational inequality.

Table 5: OLS estimates of government educational expenditures' impact on educational inequality. Data sources: CHFS, The China Compendium of Statistics 1949-2008, China Statistical Yearbook.

	Dependent variable:			
	Educational Gini by province/birth cohort			
	(1)	(2)	(3)	(4)
Average educational expenditures		-0.890^{***}		-0.396^{***}
		(0.025)		(0.027)
Constant	0.248^{***}	0.260^{***}	0.304^{***}	0.306^{***}
	(0.008)	(0.008)	(0.016)	(0.018)
Province dummies	Yes	Yes	No	No
Birth cohort dummies	No	No	Yes	Yes
Observations	5,800	5,584	5,800	5,584
R^2	0.336	0.539	0.441	0.442

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 5 reports the obtained results. From column (2), we can see that an increase in average annual educational expenditures per capita in 1000 CN¥ is associated with an 8.9 percentage point reduction in the educational Gini index, controlling for the provincial fixed effects (the coefficient is significant at the 1% level). Adding annual educational

[1]Government expenditures are measured in 1000 CN¥ of annual spending per capita, deflated using the retail price index.

expenditures per capita in column (2) to provincial dummies, considered in column (1), significantly increases the explanatory power of the regression (R squared increases from 0.336 to 0.539). Overall, the results from column (2) suggest that the difference in educational spending within-province across the cohorts matters for the explanation of the variation in educational inequality. In columns (3) and (4), we report the results of a similar analysis that employs birth-cohort dummies instead of provincial dummies. We can see that the impact of educational spending controlling for the birth-cohort dummies is also significant, but twice as small: an increase in average annual educational expenditures per capita in 1000 CN¥ is associated with a 4 percentage point reduction in the educational Gini. R squared has not increased in column (4) compared to column (3). Therefore, the variation in government educational spending across provinces within birth cohorts plays a smaller role in explaining educational inequality compared to the variation within provinces. These results are consistent with the decomposition results, reported in Table H1 (Appendix H): within-province variation share is 98%, and within-birth cohort share in each province is around $0.8 \times 98\% = 78.4\%$. To summarize, the variation in government educational expenditures within provinces over time (across birth cohorts), but not across provinces, explains the largest part of the variation in educational inequality in China. In particular, lower government spending on education for older cohorts compared to the higher spending for younger cohorts is the key factor shaping educational inequality. Since we observe that increased educational government spending over time helped to reduce educational inequality, we should expect that educational inequality will proceed to decline in the future as older cohorts are replaced by younger cohorts in the labor market.

7.2 Provincial returns to schooling determinants

Next, we study the key factors contributing to interprovincial variation in returns to schooling and returns to high school. We estimate the following model involving labor supply and labor demand factors:[1]

$$R_{jt} = \beta_0 + \beta_1 L_{jt} + \sum_{p=2}^{4} \beta_p D_{pjt} + \nu_t + \varepsilon_{jt}, \tag{10}$$

where R_{jt} represents in province j at year t either returns to schooling or returns to high school, a $\hat{\delta}_{1jt}$ estimate from eq. (6). L_{jt} is the relative supply of skilled labor in province j at year t. Following D. T. Yang (2005), this variable is calculated as the share of workers with a middle-school and above education in the working-age population. A larger supply of skilled labor is likely to shrink the earnings of skilled labor, and we expect to obtain a negative $\hat{\beta}_1$

[1] This approach is close to those discussed in D. T. Yang (2005), and Zhang et al. (2005).

estimate. D_{jt} are provincial labor demand factors. Including *Foreign trade share* again to capture the openness of the local economy. As we already mentioned, in China greater trade openness is likely to increase the demand for unskilled labor. Therefore, we expect places with higher foreign trade shares in GRP to have higher returns on education. We also include *SOE assets share* again to control for the size of the government sector in the provincial economy. A larger state sector implies higher control of the local labor market, and therefore we expect it will have a negative impact on returns to schooling (returns to high school). Finally, we control for *Secondary employment* - the share of employment in the secondary industry.[1]

Table 6: OLS estimates of returns to schooling and returns to high school determinants. Heteroskedasticity-robust standard errors are used.

	Dependent variable:					
	Returns to schooling			Returns to high school		
	(1)	(2)	(3)	(4)	(5)	(6)
Skilled labor share	−0.107**			−1.760***		
	(0.047)			(0.360)		
Skilled labor share CPESY		−0.084**			−1.074***	
		(0.033)			(0.245)	
Skilled labor share CSY			−0.136***			−1.595***
			(0.042)			(0.329)
Secondary employment CSY	−0.066**	−0.059**	−0.071**	−0.511***	−0.376***	−0.509***
	(0.028)	(0.028)	(0.028)	(0.140)	(0.136)	(0.145)
SOE assets share CSY	−0.016	−0.025	−0.015	0.077	−0.087	0.018
	(0.027)	(0.026)	(0.027)	(0.148)	(0.141)	(0.147)
Foreign trade share CSY	0.019	0.014	0.025*	0.144	0.005	0.118
	(0.013)	(0.013)	(0.014)	(0.092)	(0.083)	(0.091)
2014.year	0.009	0.007	0.007	0.052	0.003	0.010
	(0.007)	(0.008)	(0.007)	(0.043)	(0.053)	(0.047)
2016.year	−0.0004	−0.004	−0.002	−0.024	−0.092*	−0.071
	(0.008)	(0.008)	(0.007)	(0.045)	(0.054)	(0.046)
2018.year	−0.015*	−0.012	−0.010	−0.064	−0.025	−0.011
	(0.009)	(0.009)	(0.008)	(0.052)	(0.053)	(0.048)
Constant	0.257***	0.246***	0.271***	2.247***	1.837***	2.065***
	(0.036)	(0.027)	(0.030)	(0.270)	(0.207)	(0.239)
Observations	116	116	116	116	116	116
R^2	0.117	0.112	0.154	0.371	0.259	0.358

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 6 reports the OLS estimates of eq. (10). Columns (1) and (4) show the results from our baseline specification, where the *Skilled labor share* is estimated from the CHFS. To check the robustness, in columns (2) and (4) we use skilled labor share estimated from

[1]Secondary industry refers to mining, manufacturing, construction, production, and supply of electricity, water, and gas.

the China Employment Statistics Yearbook (CPESY), and in columns (3) and (6) we used skilled labor share computed from the China Statistical Yearbook (CSY). *Skilled labor share CPESY*, therefore, measures the share of individuals with nine or more years of schooling among the employed population, and *Skilled labor share CSY* - among those age 6 and above in the employed population. The signs, magnitude, and significance level of the skilled labor variables from the NBS statistical yearbooks are close to our baseline result. We find that provinces with a larger skilled labor share and a higher share of employment in the secondary industry have lower returns to schooling and lower returns to high school in China. A 10% increase in the provincial share of skilled labor is associated with a 0.8-1.3 percentage point decrease in the returns to a year of schooling. The coefficients are statistically significant at the 1-5% level. Mean provincial returns to schooling are 0.141-0.157 in different years, and the standard deviation is around 0.03 (Table C2, Appendix C). At the same time, a 10% increase in the skilled labor share is associated with a 10.7-17.6 percentage point decrease in the returns to high school. The coefficients are significant at the 1% level. Mean provincial returns to high school are 0.706-0.798, and the standard deviation is 0.2 (Table C2, Appendix C). In all specifications, the coefficients on secondary employment share are significant at the 1-5% level but have a lower magnitude than the skilled labor coefficients: a 10% increase in secondary industry employment share is associated with a 0.6-0.7 percentage point decrease in returns to a year of schooling and to a 3.8-5.1 decrease in the returns to high school. We find *SOE assets share* and *Foreign trade share* to be insignificant but of the expected signs in most of the specifications.

It should also be noted that our independent variables here reflect the level of provincial economic development. Considering possible factors correlated both with the level of economic development and with the returns to schooling, we should describe the potential bias which might be caused by omitting the educational quality variable from our regression model.[1] If we assume that educational quality is higher in rich Chinese provinces compared to the poor provinces and that higher educational quality leads to higher returns to schooling then our main results could be upwardly biased. Therefore, the provincial level of economic development might be less negatively related to the returns to schooling.

We also repeat the robustness check that examines the potential role of the migration. We re-estimate skilled shares based on *hukou* province instead of the province of residence, and then we plugged them into eq. 10. Figures I2 and I3 in Appendix I plot predicted returns to schooling (returns to high school) from 10 based on the province of residence versus based on the province of *hukou*. Predicted returns to schooling are 98% correlated.

[1]Previous studies (e.g., Wei & Hao, 2011) emphasize the significant role of an educational quality factor in the analysis of human capital in China.

Predicted returns to high school are correlated at 97%. Therefore, we conclude that our results with and without migration, again, are very similar.

8 Simulation study

Finally, we conduct Monte Carlo simulations aimed at studying how income inequality in Chinese provinces could be impacted by: (a) equalization of educational structure; (b) labor market liberalization (equalization of returns to schooling); and (c) both equalization of educational structure and labor market liberalization. For this analysis, we again employ Theil T inequality index.

We can reformulate eq. (3) for a decomposition by province:

$$T = \sum_{j=1}^{29} \frac{N_j \bar{y}_j}{N \bar{y}} T_j + \sum_{j=1}^{29} \frac{N_j \bar{y}_j}{N \bar{y}} \log\left(\frac{\bar{y}_j}{\bar{y}}\right), \tag{11}$$

where T is the national Theil T index, T_j is a Theil T index of each province j (in our case $j = 1, ..., 29$), N and $\bar{y}$ are the total number of households and the nationwide mean annual household income per capita, and N_j and $\bar{y}_j$ are the number of households and the mean household income per capita in each province j.

Each provincial Theil T index is estimated using the standard formula:

$$T_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{y_{ij}}{\bar{y}_j} \log\left(\frac{y_{ij}}{\bar{y}_j}\right), \tag{12}$$

where y_{ij} is log income of household i from province j.

Incorporating (12) in (11) we obtain:

$$\begin{aligned} T &= \sum_{j=1}^{29} \frac{\bar{y}_j}{N \bar{y}} \left(\sum_{i=1}^{N_j} \frac{y_{ij}}{\bar{y}_j} \log\left(\frac{y_{ij}}{\bar{y}_j}\right) \right) + \sum_{j=1}^{29} \frac{N_j \bar{y}_j}{N \bar{y}} \log\left(\frac{\bar{y}_j}{\bar{y}}\right) = \\ &= \frac{1}{N \bar{y}} \left(\sum_{j=1}^{29} \sum_{i=1}^{N_j} y_{ij} \log\left(\frac{y_{ij}}{\bar{y}_j}\right) + \sum_{j=1}^{29} N_j \bar{y}_j \log\left(\frac{\bar{y}_j}{\bar{y}}\right) \right) \end{aligned} \tag{13}$$

At the same time, each y_{ij} could be derived from the eq. (6) Mincer-type regression. All of the above allow us to perform the following policy simulations employing the most recent (2019) wave of the CHFS.

First, to study the impact of equalization of educational structure on income inequality we randomly reassign years of schooling to individuals adjusting the educational structure in each province in accordance with the national educational structure reported in the China

Employment Statistics Yearbook (CPESY). Then we aggregate the individual-level simulated data to households. This will result in adjusting the schooling variable $\tilde{S}_{ijt}$. Then, using the updated $\tilde{S}_{ijt}$ and the prior estimated regression coefficients from eq. (6), we predict adjusted income variable $\tilde{y}_{ij}$. Incorporating the new $\tilde{y}_{ij}$ into (13) we are getting a new predicted income inequality Theil's $\tilde{T}$ index.

Second, since equalization of education leads to different skilled labor shares in each province, after obtaining new $\tilde{S}_{ijt}$ we can re-estimate skilled shares $\tilde{L}_{jt}$. Again, keeping regression coefficients and the error term estimated from the initial data constant, and using adjusted $\tilde{L}_{jt}$, we can predict new returns to schooling $\tilde{R}_{jt}$ from (10). Then we incorporate both adjusted schooling $\tilde{S}_{ijt}$ and adjusted returns to schooling $\tilde{R}_{jt}$ into (6). This again gives us new $\tilde{y}_{ij}$, from which we can re-estimate $\tilde{T}$ from eq. (13).

Third, we can consider labor market integration only. This scenario implies an exogenous equalization of returns to schooling to the national-level returns to schooling, without any adjustments to the provincial educational structure. National-level returns to schooling are obtained by running the (6) OLS at the household-level data from the CHFS 2019.

Finally, we consider both educational equalization and labor market integration. For that, we repeat the same educational equalization from (1) and perform an exogenous equalizing change in the returns to schooling to the national-level returns to schooling. All of the subsequent procedures remain the same as in (2). Through this simulation, we are trying to answer the question: what would happen to the Theil T index if both educational structure and returns to schooling become similar across provinces?

The results of the policy simulation study are reported in Table 7. Columns (1) and (2) report our baseline results of income inequality decomposition in 2018 (from Table 1). Columns (3) and (4) show the results from our first simulation scenario, where we equalize the distribution of education in each province to the nationwide educational distribution reported in CPESY. In comparison to the initial Theil T decomposition results, equalization of educational structure leads to an 18.1% decrease in the nationwide Theil T index from 0.499 to 0.409. The within-province component of nationwide household income per capita variation increases from 87% to 91%. The poor to rich province income inequality ratio declines narrowly from 1.21 (0.473 in Gansu divided by 0.391 in Beijing) to 1.19 (0.394 in Gansu divided by 0.333 in Beijing).

The results from our second simulation scenario are presented in columns (5) and (6). Equalization of educational structure among Chinese provinces, together with adjusting returns to schooling, leads to an almost similar decrease in nationwide income inequality (-18.4%) and an increase in the within-province variation component (from 87% to 91%). However, as opposed to the first simulation, the poor-to-rich province inequality ratio slightly

Table 7: Theil T indices resulting from policy simulations.

	Baseline Estimates 2018		(a) Uniform educational structure		(b) Uniform educational structure and adjusted RS		(c) Uniform RS		(d) Uniform educational structure and uniform RS	
	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Beijing	0.391	4%	0.333	2%	0.341	2%	0.397	2%	0.330	1%
Shanghai	0.247	2%	0.248	2%	0.217	2%	0.315	4%	0.257	3%
Jiangsu	0.388	7%	0.342	6%	0.339	6%	0.385	5%	0.344	5%
Tianjin	0.269	1%	0.258	1%	0.261	1%	0.286	1%	0.260	1%
Zhejiang	0.395	5%	0.319	6%	0.312	6%	0.371	4%	0.319	5%
Fujian	0.491	3%	0.410	4%	0.389	3%	0.502	3%	0.408	3%
Guangdong	0.530	9%	0.390	7%	0.415	9%	0.478	5%	0.395	4%
Shandong	0.497	5%	0.410	6%	0.382	5%	0.490	6%	0.408	6%
Inner Mongolia	0.396	2%	0.339	2%	0.368	2%	0.362	1%	0.348	1%
Hubei	0.383	3%	0.372	4%	0.317	3%	0.409	4%	0.360	5%
Chongqing	0.495	2%	0.426	2%	0.444	2%	0.445	1%	0.430	1%
Shaanxi	0.485	2%	0.387	2%	0.402	2%	0.444	1%	0.387	1%
Liaoning	0.344	3%	0.325	3%	0.334	3%	0.357	2%	0.327	2%
Hunan	0.370	3%	0.360	4%	0.303	3%	0.413	6%	0.353	6%
Anhui	0.420	3%	0.348	4%	0.291	3%	0.495	7%	0.332	6%
Hainan	0.527	0%	0.434	0%	0.489	1%	0.434	0%	0.435	0%
Ningxia	0.380	0%	0.358	0%	0.316	0%	0.430	0%	0.348	0%
Henan	0.542	5%	0.443	6%	0.472	6%	0.481	3%	0.451	3%
Sichuan	0.600	7%	0.469	8%	0.497	9%	0.546	4%	0.473	4%
Jilin	0.379	1%	0.376	2%	0.329	1%	0.458	3%	0.372	2%
Jiangxi	0.374	2%	0.317	3%	0.292	2%	0.386	2%	0.316	3%
Hebei	0.424	3%	0.341	3%	0.309	3%	0.465	4%	0.345	4%
Shanxi	0.457	2%	0.385	2%	0.389	2%	0.436	1%	0.387	1%
Qinghai	0.515	0%	0.445	1%	0.425	1%	0.529	0%	0.443	0%
Heilongjiang	0.373	2%	0.362	3%	0.355	3%	0.374	2%	0.364	2%
Yunnan	0.509	3%	0.372	2%	0.395	3%	0.446	1%	0.376	1%
Guangxi	0.526	2%	0.451	2%	0.455	2%	0.526	1%	0.452	2%
Guizhou	0.565	2%	0.495	3%	0.483	3%	0.546	2%	0.494	2%
Gansu	0.473	1%	0.394	1%	0.447	2%	0.413	1%	0.405	1%
Within	0.436	87%	0.374	91%	0.369	91%	0.430	76%	0.366	78%
Between	0.063	13%	0.035	9%	0.038	9%	0.132	24%	0.100	22%
Total	0.499	100%	0.409	100%	0.407	100%	0.562	100%	0.466	100%

Notes: This table reports the results of the policy simulation study based on the data from the CHFS 2019. Monte Carlo Simulations were repeated 1000 times.

increased from 1.21 to 1.31.

The results of our third simulation study are reported in columns (7) and (8). Exogenous equalization of the provincial returns to schooling to the national-level returns leads to a 12.5% increase in inequality, and almost twice an increase in the between-province component explaining national-level inequality from 13% to 24%. This happens because increase in income inequality in provinces with low returns to schooling (27% increase of Theil T in Shanghai, 21% increase in Jilin, 18% increase in Anhui, etc.) overcome inequality decline in provinces with high returns to schooling (-18% in Theil T in Hainan, -13% in Gansu, -12% in Yunnan, etc.). The weight of each rise and fall of provincial Theil T in the nationwide increase in inequality is mechanically attributed to the initial role of each province in explaining the nationwide inequality, reported in column (2). We already mentioned, that returns to schooling in China are negatively associated with GRP per capita. Therefore, naturally, equalization of returns to schooling to the average national level leads to richer provinces mostly getting inequality rise, and poorer provinces getting inequality reduction. This intervention reduces by 15% the poor-to-rich province income inequality ratio from 1.21 to 1.04.

The last two columns in Table 7 report the result of our fourth simulation. Equalization of the educational structure together with equalization of the returns to schooling leads to a 6.7% decline in income inequality. The between-province component of nationwide inequality again increases to 22%, which is similar to the result from the previous simulation. The poor-to-rich province Theil T index ratio, however, remains almost the same compared to the baseline result (1.22 and 1.21, respectively).

Instead of equalizing provincial educational structure and returns to schooling at the national level, we could consider a policy that targets returns to schooling and educational structure to make them similar to the current levels in one of the leading and highly developed provinces. For example, Shanghai is one of the top provinces by the level of GRP per capita with high levels of average schooling among its population (Figure 5), low educational inequality (Figure 4) and low returns to schooling (Figure 7). All of these are desirable properties for the potential policy intervention. The results of this additional simulation analysis are reported in Table J1, Appendix J. If all provinces had an educational structure similar to that of Shanghai, it would reduce income inequality by 17.5% (Scenario "a"). If the equalization of educational structure is accompanied by adjusting returns to schooling (Scenario "b"), inequality would drop by 15.6%. If returns to schooling in all provinces reach the same low as in Shanghai (Scenario "c"), income inequality would decline by 3.1%. Finally, if both returns to schooling and educational structure in all provinces come to resemble that of Shanghai (Scenario "d"), inequality would decline by 14.4%.

9 Conclusion

In this paper, we introduce robust Gini index estimates for 29 mainland Chinese provinces (excluding Xinjiang and Tibet), evaluated from four waves of the China Household Finance Survey (CHFS 2013, 2015, 2017, and 2019). We also perform a novel empirical analysis of the provincial-level income inequality in China. With the support of the evidence from the first provincial-level inequality study by Hussain et al. (1994), we found that intraprovincial inequality is the main contributing factor to the national inequality, explaining around 90% of the variation in annual household income per capita.

Our key empirical results suggest that human capital disparities play a significant role in explaining variation in the levels of income inequality across Chinese provinces. Educational inequality and returns to schooling are positively and statistically significantly related to the current provincial income inequality. We found, that in 2012-2018, a 10% increase in provincial-level educational Gini and provincial returns to schooling are associated with increases of 3.8-4.7 and 7.1-8.0 percentage point in income Gini index, respectively. This result is consistent with the results from the recent study by Nie and Xing (2019), where returns to education is identified as a key explanatory factor of income inequality in China (but only for urban households). We also find that the government's presence in the local economy and the governmental social spending have a significant negative impact on provincial income inequality.

We argue that the existing labor market disparities are mostly responsible for the current interprovincial differences in returns to schooling and returns to high school in China. In a fully integrated labor market, returns to schooling across provinces would be similar. However, in China, there still exists, albeit diminishing, labor market segmentation. We found a statistically significant negative relationship between the share of skilled labor and provincial returns to schooling (as well as returns to high school). A 10% increase in a skilled labor share is associated with a decrease of 0.8-1.3 percentage point in returns to schooling, and a decrease of 10.7-17.6 percentage point in returns to high school. We also find that secondary employment share has a significant negative impact on returns to schooling and returns to high school. These results align with the prior findings of Meng and Wu (1998), that the important determinant of regional income disparities is the degree of regional marketization in China, are still in force.

Lastly, we find that poor provinces are severely disadvantaged compared to rich provinces in China. Poor provinces experience higher income and educational inequality, and at the same time, their citizens have comparatively lower levels of education. The results from our simulation analysis suggest that equalization of educational structure across provinces to the

national level could lead to an 18% decrease in income inequality. Equalization of returns to schooling in provinces to the national level, accompanied by the equalization of educational structure, could lead to a 6% decrease in income inequality. Equalization of returns to schooling to the low levels of highly-developed Shanghai, together with the equalization of educational structure, could lead to a 14.4% decrease in national income inequality. These findings support results from previous studies showing that human capital could be an important factor in policy design for regional development in China (Fraumeni et al., 2019). The benefits of targeting educational policies would also emerge from the premise that promoting education is associated with the equalization of opportunities. Government policies that foster educational development in disadvantaged provinces further reduce barriers to labor mobility, support labor market integration, reduce income disparities and enhance sustainable and balanced regional economic development in China.

References

Alesina, A., & Perotti, R. (1996). Income distribution, political instability, and investment. *European Economic Review*, *40*(6), 1203–1228.

Asadullah, M. N., & Xiao, S. (2020). The changing pattern of wage returns to education in post-reform China. *Structural Change and Economic Dynamics*, *53*, 137–148.

Bourguignon, F. (1979). Decomposable income inequality measures. *Econometrica: Journal of the Econometric Society*, 901–920.

Cai, F., Park, A., & Zhao, Y. (2008). The Chinese labor market in the reform era. *China's great economic transformation, Cambridge University Press Cambridge*, 167–214.

Chan, W. K., & Ngok, K. (2011). Accumulating human capital while increasing educational inequality: a study on higher education policy in China. *Asia Pacific Journal of Education*, *31*(3), 293–310.

Charpentier, A., & Flachaire, E. (2019). Pareto models for top incomes.

Chong, A., & Gradstein, M. (2017). Political and economic inequities and the shaping of institutions and redistribution. *Southern Economic Journal*, *83*(4), 952–971.

Churchill, S. A., & Mishra, V. (2018). Returns to education in China: a meta-analysis. *Applied Economics*, *50*(54), 5903–5919.

Coady, D., & Dizioli, A. (2018). Income inequality and education revisited: persistence, endogeneity and heterogeneity. *Applied Economics*, *50*(25), 2747–2761.

Coles, S. G., & Dixon, M. J. (1999). Likelihood-based inference for extreme value models. *Extremes*, *2*, 5–23.

Cowell, & Flachaire, E. (2007). Income distribution and inequality measurement: The problem of extreme values. *Journal of Econometrics*, *141*(2), 1044–1072.

Cowell, & Victoria-Feser, M.-P. (2007). Robust stochastic dominance: A semi-parametric approach. *The Journal of Economic Inequality*, *5*(1), 21–37.

Cowell, F. A. (1980). On the structure of additive inequality measures. *The Review of Economic Studies*, *47*(3), 521–531.

Cowell, F. A. (2000). Measurement of inequality. *Handbook of income distribution*, *1*, 87–166.

Davidson, R., & Flachaire, E. (2007). Asymptotic and bootstrap inference for inequality and poverty measures. *Journal of Econometrics*, *141*(1), 141–166.

Esteban, J., & Ray, D. (1999). Conflict and distribution. *Journal of Economic Theory*, *87*(2), 379–415.

Fields, G. S. (2003). Accounting for income inequality and its change: A new method, with application to the distribution of earnings in the United States. In *Worker well-being and public policy*. Emerald Group Publishing Limited.

Fleisher, B., & Yang, D. (2008). China's evolving labor market. *Policy reform and Chinese markets: Progress and challenges, Edward Elgar Publishing*, 125–158.

Florida, R., Mellander, C., & Rentfrow, P. J. (2013). The happiness of cities. *Regional studies*, *47*(4), 613–627.

Fraumeni, B. M., He, J., Li, H., & Liu, Q. (2019). Regional distribution and dynamics of human capital in China 1985–2014. *Journal of Comparative Economics*, *47*(4), 853–866.

Fu, Q., & Ren, Q. (2010). Educational inequality under China's rural–urban divide: The hukou system and return to education. *Environment and Planning A*, *42*(3), 592–610.

Galor, O., & Moav, O. (2006). Das human-kapital: A theory of the demise of the class structure. *The Review of Economic Studies*, *73*(1), 85–117.

Gan, L., Yin, Z., Jia, N., Xu, S., Ma, S., Zheng, L., et al. (2014). Data you need to know about China. *Springer Berlin Heidelberg. https://doi*, *10*, 978–3.

Goh, C.-c., Xubei, L., & Nong, Z. (2009). Income growth, inequality and poverty reduction: A case study of eight provinces in China. *China Economic Review*, *20*(3), 485–496.

Golley, J., & Kong, S. T. (2018). Inequality of opportunity in China's educational outcomes. *China Economic Review*, *51*, 116–128.

Gradín, C., & Wu, B. (2020). Income and consumption inequality in China: A comparative approach with India. *China Economic Review*, *62*, 101463.

Gustafsson, B., & Wan, H. (2020). Wage growth and inequality in urban China: 1988–2013. *China Economic Review*, *62*, 101462.

Hao, T., Sun, R., Tombe, T., & Zhu, X. (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., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. *Handbook of the Economics of Education*, *1*, 307–458.

Heckman, J. J., & Yi, J. (2012). *Human capital, economic growth, and inequality in China* (Tech. Rep.). National Bureau of Economic Research.

Huang, C., Lin, J.-G., & Ren, Y.-Y. (2012). Statistical inferences for generalized pareto distribution based on interior penalty function algorithm and bootstrap methods and applications in analyzing stock data. *Computational Economics*, *39*, 173–193.

Hussain, A., Lanjouw, P., & Stern, N. (1994). Income inequalities in China: Evidence from household survey data. *World Development*, *22*(12), 1947–1957.

Jain-Chandra, M. S., Khor, N., Mano, R., Schauer, J., Wingender, M. P., & Zhuang, J. (2018). *Inequality in China–trends, drivers and policy remedies*. International Monetary Fund.

Jenkins, S. P. (2017). Pareto models, top incomes and recent trends in uk income inequality. *Economica*, *84*(334), 261–289.

Kanbur, R., Wang, Y., & Zhang, X. (2021). The great Chinese inequality turnaround. *Journal of Comparative Economics*, *49*(2), 467–482.

Khan, A. R., & Riskin, C. (2005). China's household income and its distribution, 1995 and 2002. *The China Quarterly*, 356–384.

Lei, X., & Shen, Y. (2015). Inequality in educational attainment and expectation: evidence from the China Family Panel Studies. *China Economic Journal*, *8*(3), 252–263.

Li, Q., Li, S., & Wan, H. (2020). Top incomes in China: Data collection and the impact on income inequality. *China Economic Review*, *62*, 101495.

Li, S., Whalley, J., & Xing, C. (2014). China's higher education expansion and unemployment of college graduates. *China Economic Review*, *30*, 567–582.

Lü, X. (2012). Equality of educational opportunity and attitudes toward income inequality: Evidence from China. *Quarterly Journal of Political Science*.

Martins, E. S., & Stedinger, J. R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, *36*(3), 737–744.

Martins, E. S., & Stedinger, J. R. (2001). Generalized maximum likelihood pareto-poisson estimators for partial duration series. *Water Resources Research*, *37*(10), 2551–2557.

Mendoza, O. M. V. (2018). Heterogeneous determinants of educational achievement and

inequality across urban China. *China Economic Review*, *51*, 129–148.

Meng, X., & Wu, H. X. (1998). Household income determination and regional income differential in rural China. *Asian Economic Journal*, *12*(1), 35–63.

Mincer, J. (1974). Schooling, Experience, and Earnings. *Human Behavior & Social Institutions*, *2*.

Naughton, B. J. (2006). *The Chinese economy: Transitions and growth*. MIT press.

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

Nie, H., & Xing, C. (2019). Education expansion, assortative marriage, and income inequality in China. *China Economic Review*, *55*, 37–51.

Park, A., Song, X., Zhang, J., & Zhao, Y. (2003). The growth of wage inequality in urban China, 1988 to 1999. *Unpublished Paper, Chinese University of Hong Kong*.

Pauli, F., & Coles, S. (2001). Penalized likelihood inference in extreme value analyses. *Journal of Applied Statistics*, *28*(5), 547–560.

Peng, H., Qi, L., Wan, G., Li, B., & Hu, B. (2020). Child population, economic development and regional inequality of education resources in China. *Children and Youth Services Review*, *110*, 104819.

Perotti, R. (1996). Growth, income distribution, and democracy: What the data say. *Journal of Economic growth*, *1*(2), 149–187.

Ravallion, M., & Chen, S. (2009). *China's (uneven) progress against poverty*. Routledge.

Savoia, A., Easaw, J., & McKay, A. (2010). Inequality, democracy, and institutions: A critical review of recent research. *World Development*, *38*(2), 142–154.

Shorrocks, A. F. (1984). Inequality decomposition by population subgroups. *Econometrica: Journal of the Econometric Society*, 1369–1385.

Sicular, T., Ximing, Y., Gustafsson, B., & Shi, L. (2007). The Urban–Rural Income Gap and Inequality in China. *Review of Income and Wealth*, *53*(1), 93–126.

Sun, W., Lu, J., & Bai, C. (2007). Analysis of income flow in rural China (in Chinese). *Economic Research*, *08*, 43–57.

Van Kerm, P. (2007). Extreme incomes and the estimation of poverty and inequality indicators from eu-silc.

Victoria-Feser, M.-P. (2000). Robust methods for the analysis of income distribution, inequality and poverty. *International Statistical Review*, *68*(3), 277–293.

Victoria-Feser, M.-P., & Ronchetti, E. (1994). Robust methods for personal-income distribution models. *Canadian Journal of Statistics*, *22*(2), 247–258.

Wan, G., & Zhou, Z. (2005). Income inequality in rural China: Regression-based decomposition using household data. *Review of development economics*, *9*(1), 107–120.

Wang, C., Wan, G., & Yang, D. (2014). Income inequality in the People's Republic of China: Trends, determinants, and proposed remedies. *Journal of Economic Surveys*, *28*(4), 686–708.

Wang, Z. (2006). Study on the estimation of Gini coefficient in the six provinces of Central China (in Chinese). *Zhongguo She Hui Ke Xue*, *4*, 77–87.

Wei, Z., & Hao, R. (2011). The role of human capital in China's total factor productivity growth: A cross-Province analysis. *The Developing Economies*, *49*(1), 1–35.

Weimin, T. (2012). The Gini coefficient of provincial residents' income and its change trend analysis (in Chinese). *Economic Science*, *02*, 48–59.

Xiang, L., Stillwell, J., Burns, L., & Heppenstall, A. (2020). Measuring and assessing regional education inequalities in China under changing policy regimes. *Applied Spatial Analysis and Policy*, *13*(1), 91–112.

Xie, Y., & Zhou, X. (2014). Income inequality in today's China. *Proceedings of the National Academy of Sciences*, *111*(19), 6928–6933.

Xu, L. C., & Zou, H.-f. (2000). Explaining the changes of income distribution in China. *China Economic Review*, *11*(2), 149–170.

Yang, D. T. (2005). Determinants of schooling returns during transition: Evidence from Chinese cities. *Journal of Comparative Economics*, *33*(2), 244–264.

Yang, J., Huang, X., & Liu, X. (2014). An analysis of education inequality in China. *International Journal of Educational Development*, *37*, 2–10.

Yu, N., Yu, B., de Jong, M., & Storm, S. (2015). Does inequality in educational attainment matter for China's economic growth? *International Journal of Educational Development*, *41*, 164–173.

Zhang, J., Zhao, Y., Park, A., & Song, X. (2005). Economic returns to schooling in urban China, 1988 to 2001. *Journal of comparative economics*, *33*(4), 730–752.

Appendix

A Empirical evidence on provincial income inequality

Table A1: Summary of provincial-level Gini index estimates in China.

Authors & year	Data source	Year	Rural/ Urban	Gini Summary Statistics				
				Min	Max	Mean	SD	#provinces
Hussain et al. (1994)	Based on the Chinese Academy of Social Sciences (CASS) survey of households 1986.	1986	Urban	0.180	0.243	0.212	0.017	23
		1986	Rural	0.195	0.281	0.241	0.026	10
Xu and Zou (2000)	Based on the published results of average incomes of different percentiles for urban residents in each province (World Bank, 1996).	Mean 1985-1995	Urban	0.173	0.279	0.208	0.023	29
Khan and Riskin (2005)	Based on the Chinese Academy of Social Sciences (CASS) survey of households.	1995	Urban	0.272	0.432	0.331	0.045	19
		2002	Urban	0.235	0.409	0.309	0.041	21
		1995	Rural	0.223	0.482	0.280	0.074	11
		2002	Rural	0.250	0.328	0.282	0.026	11
Sun, Lu, and Bai (2007)	Based on an annual fixed-point survey of farmers conducted jointly by the Central Policy Research Office and the Ministry of Agriculture.	1986	Rural	0.223	0.413	0.322	0.074	6
		1987	Rural	0.185	0.386	0.288	0.070	6
		1988	Rural	0.213	0.363	0.296	0.057	6
		1989	Rural	0.224	0.370	0.295	0.059	6
		1990	Rural	0.234	0.388	0.294	0.057	6
		1991	Rural	0.234	0.420	0.321	0.064	6
		1995	Rural	0.345	0.369	0.360	0.011	6
		1996	Rural	0.363	0.420	0.385	0.021	6
		1997	Rural	0.367	0.394	0.380	0.011	6
		1998	Rural	0.356	0.393	0.378	0.013	6
		1999	Rural	0.366	0.402	0.379	0.014	6
		2000	Rural	0.356	0.421	0.390	0.021	6
		2001	Rural	0.395	0.416	0.408	0.008	6
Z. Wang (2006)	Based on grouped income distribution data from Statistical Yearbooks of each province.	1995	Rural and urban	0.271	0.379	0.331	0.037	6
		2000	Rural and urban	0.338	0.378	0.354	0.014	6
		2003	Rural and urban	0.385	0.428	0.407	0.015	6
Weimin (2012)	Based on grouped income distribution data from Statistical Yearbooks of each province.	1995	Rural and urban	0.241	0.438	0.348	0.056	27
		1996	Rural and urban	0.231	0.437	0.328	0.054	27
		1997	Rural and urban	0.228	0.418	0.325	0.050	27
		1998	Rural and urban	0.229	0.410	0.328	0.046	27
		1999	Rural and urban	0.248	0.421	0.343	0.043	27
		2000	Rural and urban	0.256	0.456	0.368	0.053	27
		2001	Rural and urban	0.270	0.466	0.380	0.054	27
		2002	Rural and urban	0.265	0.478	0.396	0.053	26
		2003	Rural and urban	0.267	0.483	0.406	0.052	26
		2004	Rural and urban	0.289	0.483	0.399	0.050	26
		2005	Rural and urban	0.280	0.478	0.405	0.046	26
		2006	Rural and urban	0.276	0.490	0.407	0.049	25
		2007	Rural and urban	0.280	0.491	0.406	0.052	24
		2008	Rural and urban	0.296	0.486	0.408	0.049	24
		2009	Rural and urban	0.290	0.484	0.408	0.049	24
		2010	Rural and urban	0.274	0.476	0.394	0.049	24
Nie and Xing (2019)	1% National Population Sample Survey 2005	2005	Rural and urban	0.335	0.544	0.421	0.045	31

B Data and summary statistics: detailed

Table B1: CHFS sample size.

	CHFS 2013	CHFS 2015	CHFS 2017	CHFS 2019
Number of provinces	29	29	29	29
Number of counties	267	351	355	343
Number of communities (villages)	1,048	1,396	1,439	1,360
Number of households	28,143	37,289	40,011	34,643

Table B2: Summary statistics, household-level variables. Data source: CHFS.

	2012				2014				2016				2018			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Ln(household income per capita)	9.36	1.29	4.80	12.15	9.46	1.35	5.48	12.23	9.73	1.28	5.13	13.06	9.62	1.26	5.46	12.53
Schooling per worker	9.91	3.49	0	22.00	10.01	3.48	0	22.00	9.89	3.59	0	22.00	9.79	3.59	0	22.00
Experience per worker	25.32	10.94	0	58.00	26.35	10.79	0	58.00	28.21	12.13	0	58.00	29.30	12.37	0	58.00
Experience per worker^2 : 100	7.61	6.48	0	33.64	8.11	6.60	0	33.64	9.43	7.61	0	33.64	10.12	7.89	0	33.64
Household size	3.07	1.47	1	17	3.15	1.49	1	18	3.08	1.49	1	18	3.07	1.51	1	16
Land dummy	0.39	0.49	0	1	0.43	0.50	0	1	0.33	0.47	0	1	0.44	0.50	0	1
Ln(land per capita)	0.01	0.09	0	2.71	0.43	0.65	0	3.80	0.32	0.58	0	3.85	0.06	0.18	0	2.40
Ln(business assets per capita)	0.36	1.09	0	7.43	0.37	1.12	0	7.58	0.35	1.10	0	7.13	0.30	0.97	0	6.22
Party member	0.18	0.38	0	1	0.18	0.39	0	1	0.13	0.33	0	1	0.18	0.39	0	1
Female	0.24	0.43	0	1	0.25	0.43	0	1	0.21	0.40	0	1	0.25	0.43	0	1
Rural	0.32	0.47	0	1	0.30	0.46	0	1	0.32	0.46	0	1	0.35	0.48	0	1

Table B3: Summary statistics, provincial-level variables.

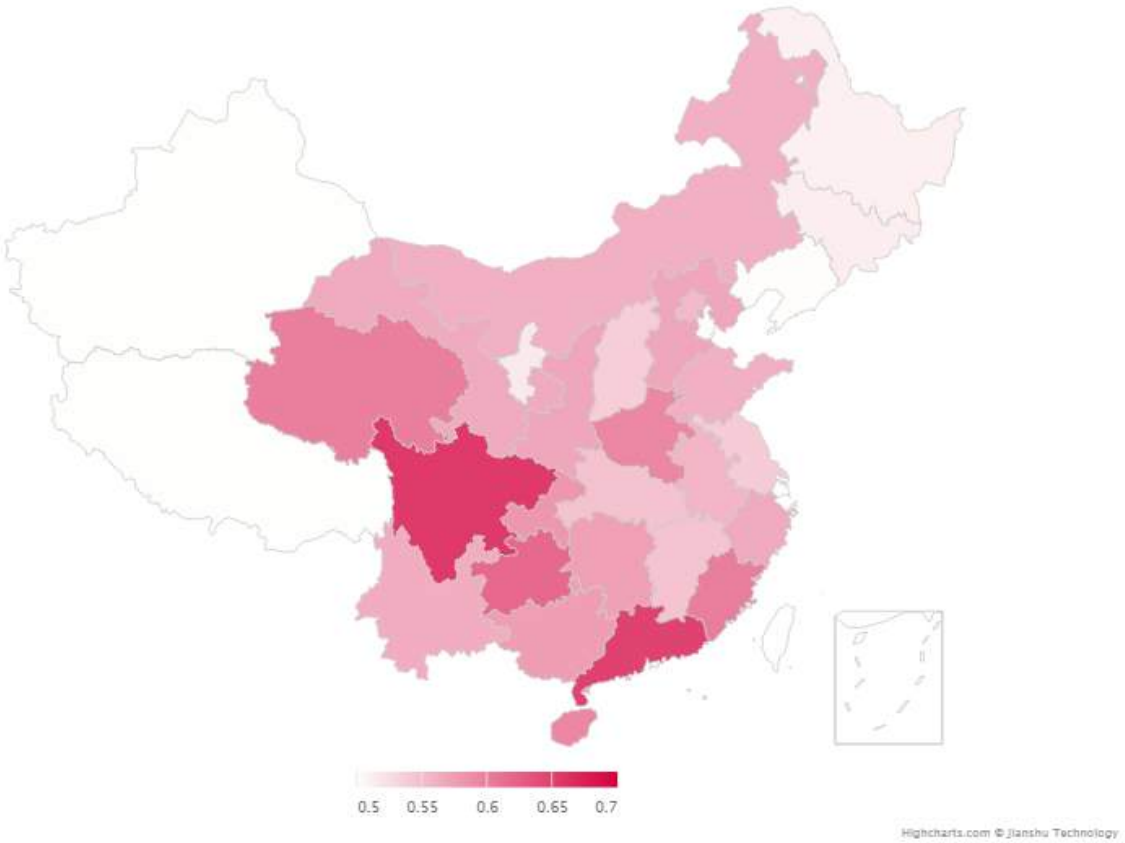
	2012				2014				2016				2018			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Data: individual-level CHFS, working-age population [15-64].																
Educational Gini	0.230	0.048	0.162	0.412	0.225	0.050	0.154	0.427	0.221	0.042	0.157	0.353	0.234	0.046	0.155	0.380
Average years of schooling	9.510	0.959	7.366	12.064	9.934	0.954	7.467	12.421	10.009	0.869	8.487	12.335	9.489	1.005	7.531	12.482
Skilled labor share	0.369	0.091	0.278	0.640	0.425	0.088	0.322	0.690	0.423	0.086	0.279	0.672	0.368	0.095	0.271	0.664
Data: household-level CHFS.																
Income Gini index	0.567	0.046	0.470	0.670	0.568	0.050	0.467	0.680	0.552	0.054	0.427	0.642	0.552	0.048	0.450	0.667
Returns to schooling	0.151	0.032	0.085	0.228	0.157	0.026	0.110	0.205	0.147	0.029	0.085	0.208	0.141	0.036	0.078	0.215
Returns to high School	0.798	0.259	0.492	1.749	0.790	0.191	0.440	1.378	0.706	0.210	0.227	1.112	0.772	0.193	0.390	1.158
Data: China Statistical Yearbook (CSY).																
Skilled labor share CSY	0.279	0.089	0.168	0.596	0.294	0.086	0.167	0.600	0.312	0.098	0.179	0.646	0.331	0.097	0.213	0.686
Social spending CSY	1.464	0.749	0.716	4.344	1.784	0.677	0.877	3.445	2.423	1.033	1.211	5.181	2.992	1.077	1.546	5.396
SOE assets share CSY	0.257	0.127	0.041	0.490	0.252	0.131	0.046	0.542	0.248	0.136	0.049	0.638	0.216	0.113	0.051	0.519
Ln(GRP per capita) CSY	10.531	0.400	9.849	11.438	10.696	0.383	10.131	11.578	10.822	0.392	10.218	11.723	11.009	0.392	10.379	11.925
Foreign trade share CSY	0.332	0.371	0.048	1.354	0.312	0.317	0.057	1.137	0.239	0.244	0.032	0.965	0.250	0.249	0.018	0.949
Secondary employment CSY	0.448	0.104	0.206	0.672	0.458	0.113	0.197	0.682	0.431	0.113	0.178	0.659	0.398	0.118	0.157	0.647
Data: China Population & Employment Statistics Yearbook (CPESY).																
Educational Gini CPESY	0.176	0.025	0.140	0.238	0.180	0.026	0.142	0.243	0.191	0.029	0.138	0.258	0.187	0.027	0.136	0.249
Skilled labor share CPESY	0.323	0.121	0.166	0.761	0.348	0.126	0.185	0.758	0.376	0.118	0.192	0.754	0.392	0.124	0.203	0.771

C Provincial-level variables estimated from the CHFS

Table C1: Provincial Gini indices and mean household annual income per capita. Data source: CHFS.

		Household income per capita (CN¥)				Gini index			
		2012	2014	2016	2018	2012	2014	2016	2018
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beijing	BJ	47,146	56,506	88,358	76,913	0.537	0.537	0.592	0.546
Shanghai	SH	42,169	52,226	61,711	69,358	0.470	0.498	0.463	0.463
Jiangsu	JS	34,030	33,057	44,381	37,191	0.573	0.524	0.531	0.529
Tianjin	TJ	37,076	40,251	48,689	47,279	0.510	0.467	0.427	0.450
Zhejiang	ZJ	36,444	47,144	53,725	40,572	0.598	0.574	0.561	0.527
Fujian	FJ	26,000	38,347	42,832	37,611	0.607	0.615	0.579	0.599
Guangdong	GD	40,170	46,578	48,125	39,359	0.670	0.656	0.642	0.620
Shandong	SD	24,765	27,869	32,698	19,770	0.584	0.540	0.528	0.588
Inner Mongolia	NM	22,372	32,463	33,762	28,880	0.607	0.599	0.506	0.528
Hubei	HUB	21,950	29,334	34,412	25,230	0.544	0.560	0.559	0.525
Chongqing	CQ	21,069	24,162	34,096	23,692	0.581	0.573	0.576	0.592
Shaanxi	SAA	21,583	26,647	33,504	20,993	0.592	0.589	0.539	0.554
Liaoning	LN	19,442	32,329	34,178	31,030	0.500	0.533	0.464	0.487
Hunan	HUN	21,478	26,802	29,281	24,249	0.565	0.608	0.600	0.519
Anhui	AH	22,853	24,790	24,380	24,682	0.552	0.577	0.556	0.539
Hainan	HI	15,072	17,617	36,738	21,760	0.557	0.609	0.619	0.587
Ningxia	NX	20,178	27,931	37,207	24,098	0.544	0.519	0.480	0.514
Henan	HEN	14,624	22,571	24,829	21,062	0.570	0.618	0.593	0.590
Sichuan	SC	27,197	32,919	29,060	31,932	0.631	0.680	0.638	0.667
Jilin	JL	16,724	22,783	30,490	20,639	0.511	0.528	0.500	0.507
Jiangxi	JX	19,282	26,064	32,504	21,723	0.538	0.600	0.518	0.524
Hebei	HEB	14,541	17,065	26,435	17,703	0.584	0.574	0.577	0.537
Shanxi	SAX	13,235	18,251	26,752	22,258	0.568	0.494	0.531	0.559
Qinghai	QH	19,189	27,689	34,535	25,978	0.652	0.603	0.565	0.577
Heilongjiang	HL	20,820	25,820	26,833	24,340	0.487	0.513	0.516	0.527
Yunnan	YN	17,178	20,005	27,939	24,388	0.575	0.533	0.553	0.591
Guangxi	GX	18,647	22,134	19,265	17,758	0.556	0.576	0.580	0.588
Guizhou	GZ	17,503	25,132	28,919	21,204	0.607	0.625	0.629	0.608
Gansu	GS	15,132	22,408	31,277	21,217	0.570	0.542	0.589	0.557
Mean		23,720	29,893	36,445	29,754	0.567	0.568	0.552	0.552
SD		9,079	10,192	13,687	14,149	0.046	0.050	0.054	0.048
CV		0.383	0.341	0.376	0.476	0.081	0.088	0.097	0.087

Figure C1: Provincial-level Gini index estimates on China's map (average for 2012-2018).

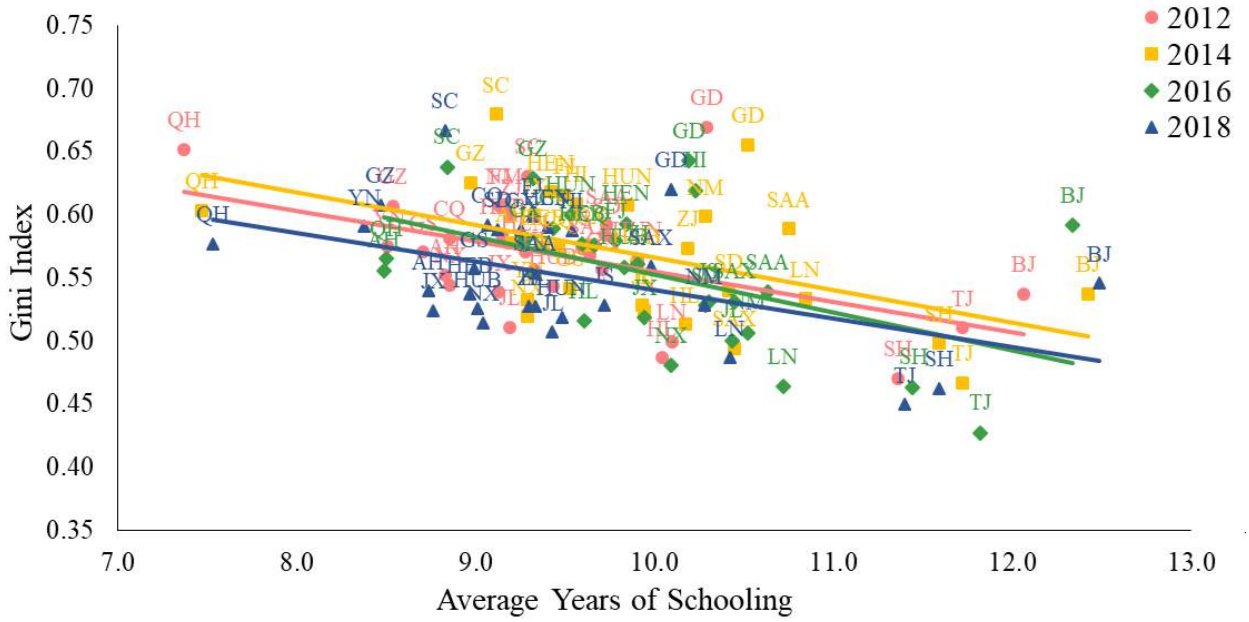


Notes: This figure plots the average provincial-level Gini index estimates for 2012-2018 calculated from the CHFS.

Table C2: Provincial-level estimates of average years of schooling, educational Gini, and returns to high school in China. Data source: China Statistical Yearbook for average years of schooling and educational Gini, CHFS for returns to schooling and returns to high school.

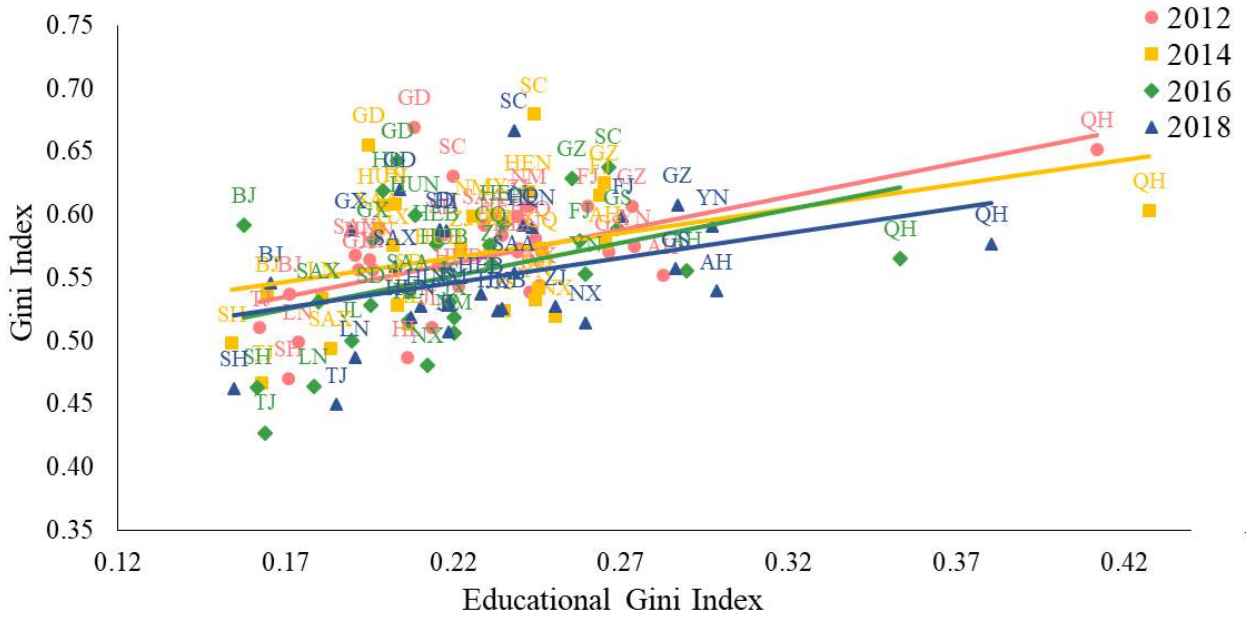
	Average years of schooling				Educational Gini index				Returns to schooling				Returns to high school			
	2012	2014	2016	2018	2012	2014	2016	2018	2012	2014	2016	2018	2012	2014	2016	2018
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Beijing	12.1	12.4	12.3	12.5	0.171	0.164	0.157	0.165	0.169	0.148	0.135	0.163	0.789	0.674	0.476	0.716
Shanghai	11.4	11.6	11.4	11.6	0.171	0.154	0.162	0.155	0.124	0.110	0.115	0.092	0.530	0.440	0.397	0.390
Jiangsu	9.6	9.9	10.3	9.7	0.239	0.235	0.219	0.218	0.126	0.115	0.114	0.148	0.581	0.483	0.634	0.605
Tianjin	11.7	11.7	11.8	11.4	0.162	0.163	0.164	0.185	0.171	0.147	0.137	0.152	0.636	0.543	0.603	0.705
Zhejiang	9.2	10.2	9.9	9.3	0.239	0.222	0.232	0.250	0.143	0.162	0.163	0.144	0.673	0.799	0.702	0.673
Fujian	9.1	9.5	9.8	9.3	0.260	0.263	0.258	0.270	0.135	0.141	0.132	0.129	0.607	0.600	0.833	0.873
Guangdong	10.3	10.5	10.2	10.1	0.208	0.194	0.203	0.204	0.217	0.171	0.179	0.174	1.294	0.802	0.868	0.756
Shandong	9.7	10.4	10.3	9.1	0.217	0.207	0.195	0.216	0.151	0.161	0.162	0.113	0.578	0.797	0.707	0.599
Inner Mongolia	9.2	10.3	10.5	10.3	0.242	0.226	0.220	0.219	0.127	0.155	0.145	0.168	0.692	0.713	0.653	0.969
Hubei	9.4	9.9	9.8	9.0	0.222	0.216	0.217	0.234	0.124	0.133	0.139	0.096	0.530	0.688	0.622	0.598
Chongqing	8.9	9.5	9.7	9.1	0.244	0.246	0.231	0.241	0.140	0.174	0.174	0.165	0.813	0.836	0.865	0.734
Shaanxi	9.7	10.8	10.6	9.3	0.229	0.198	0.206	0.238	0.166	0.148	0.153	0.168	0.994	0.851	0.666	0.952
Liaoning	10.1	10.8	10.7	10.4	0.174	0.181	0.178	0.191	0.141	0.174	0.169	0.158	0.642	0.783	0.708	0.813
Hunan	9.9	9.8	9.5	9.5	0.195	0.199	0.209	0.207	0.159	0.164	0.147	0.084	0.654	0.750	0.655	0.509
Anhui	8.8	9.2	8.5	8.7	0.283	0.265	0.289	0.298	0.130	0.142	0.097	0.079	0.802	0.811	0.495	0.573
Hainan	9.3	9.6	10.2	9.5	0.196	0.203	0.199	0.218	0.200	0.173	0.208	0.215	0.947	0.935	0.997	0.996
Ningxia	8.9	9.3	10.1	9.0	0.246	0.250	0.212	0.259	0.170	0.127	0.085	0.106	0.923	0.757	0.227	0.631
Henan	9.3	9.4	9.8	9.4	0.239	0.242	0.235	0.243	0.124	0.144	0.120	0.173	0.742	0.975	0.581	1.085
Sichuan	9.3	9.1	8.8	8.8	0.220	0.244	0.266	0.238	0.158	0.193	0.192	0.177	0.816	0.944	1.058	1.066
Jilin	9.2	9.9	10.4	9.4	0.213	0.203	0.190	0.219	0.085	0.132	0.127	0.078	0.492	0.595	0.551	0.476
Jiangxi	9.1	9.3	9.9	8.8	0.243	0.232	0.220	0.233	0.118	0.205	0.132	0.121	0.724	0.932	0.578	0.761
Hebei	9.1	9.4	9.6	9.0	0.234	0.231	0.215	0.228	0.121	0.140	0.137	0.098	0.746	0.920	0.644	0.705
Shanxi	9.6	10.4	10.4	10.0	0.190	0.183	0.180	0.202	0.184	0.154	0.149	0.151	0.850	0.655	0.606	0.796
Qinghai	7.4	7.5	8.5	7.5	0.412	0.427	0.353	0.380	0.228	0.197	0.171	0.131	1.749	1.378	1.112	0.976
Heilongjiang	10.0	10.2	9.6	9.3	0.206	0.207	0.207	0.210	0.114	0.131	0.115	0.142	0.575	0.687	0.535	0.776
Yunnan	8.5	9.3	9.3	8.4	0.274	0.244	0.259	0.297	0.169	0.179	0.170	0.177	0.994	0.823	1.075	1.158
Guangxi	9.7	9.6	9.3	9.3	0.191	0.202	0.196	0.189	0.172	0.191	0.182	0.154	0.783	0.884	0.903	0.697
Guizhou	8.5	9.0	9.3	8.5	0.273	0.265	0.255	0.287	0.179	0.201	0.172	0.137	1.054	1.145	0.974	0.791
Gansu	8.7	9.5	9.4	9.0	0.266	0.245	0.269	0.286	0.144	0.130	0.131	0.188	0.924	0.706	0.737	1.007
Mean	9.5	9.9	10.0	9.5	0.230	0.225	0.221	0.234	0.151	0.157	0.147	0.141	0.798	0.790	0.706	0.772
SD	1.0	1.0	0.9	1.0	0.048	0.050	0.042	0.046	0.032	0.026	0.029	0.036	0.259	0.191	0.210	0.193
CV	0.101	0.096	0.087	0.106	0.209	0.221	0.190	0.197	0.213	0.165	0.196	0.254	0.325	0.242	0.298	0.250

Figure C2: Gini index and average years of schooling



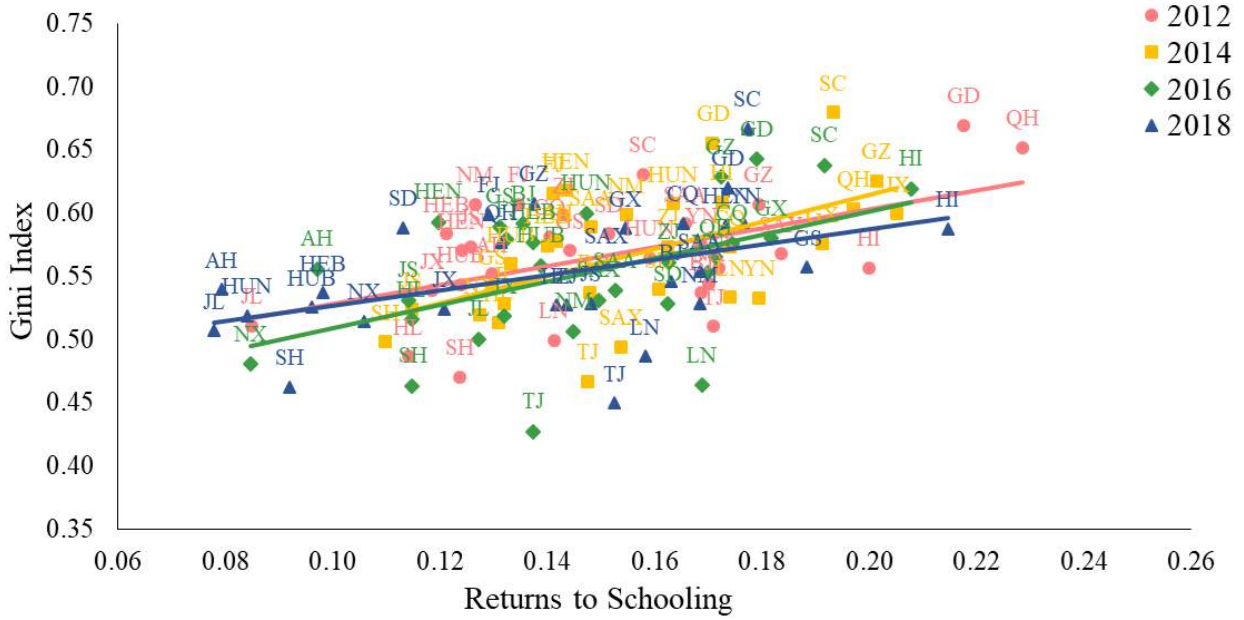
Notes: This figure plots provincial-level Gini index and average years of schooling estimated from the CHFS. Correlations are: -50% in 2012, -49% in 2014, -48% in 2016, and -48% in 2018.

Figure C3: Gini index and educational Gini index.



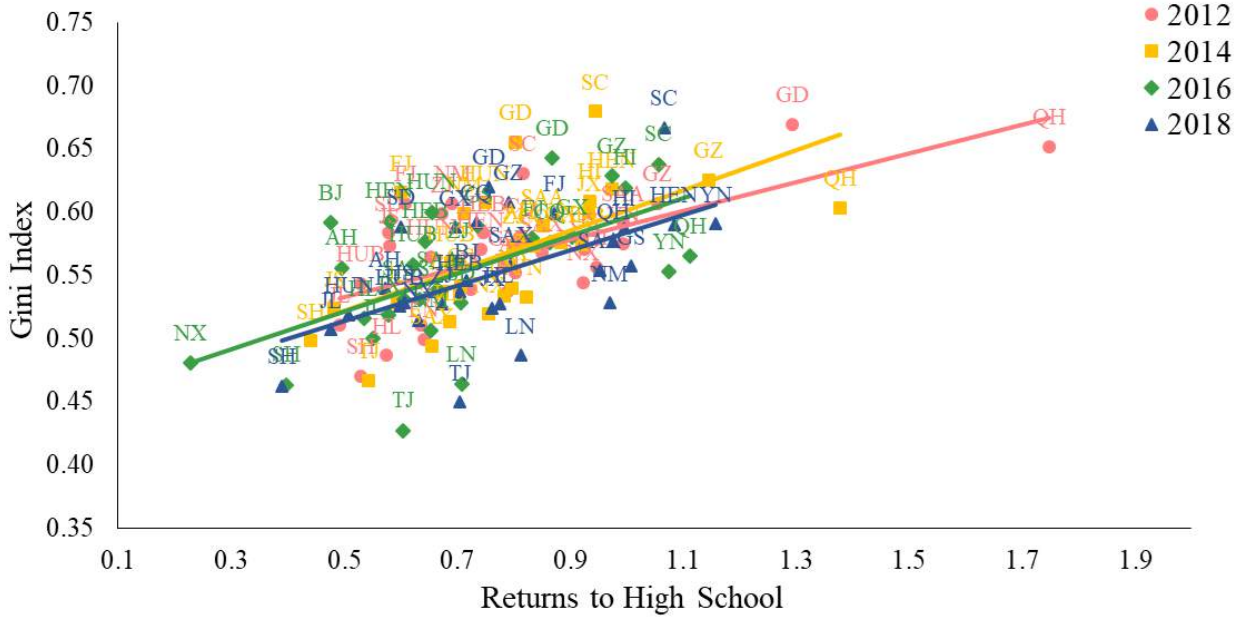
Notes: This figure plots provincial-level Gini index and educational Gini index estimated from the CHFS. Correlations are: 56% in 2012, 39% in 2014, 42% in 2016, and 39% in 2018.

Figure C4: Gini index and returns to schooling.



Notes: This figure plots provincial-level Gini index and returns to schooling estimated from the CHFS. Correlations are: 52% in 2012, 55% in 2014, 49% in 2016, and 45% in 2018.

Figure C5: Gini index and returns to high school.

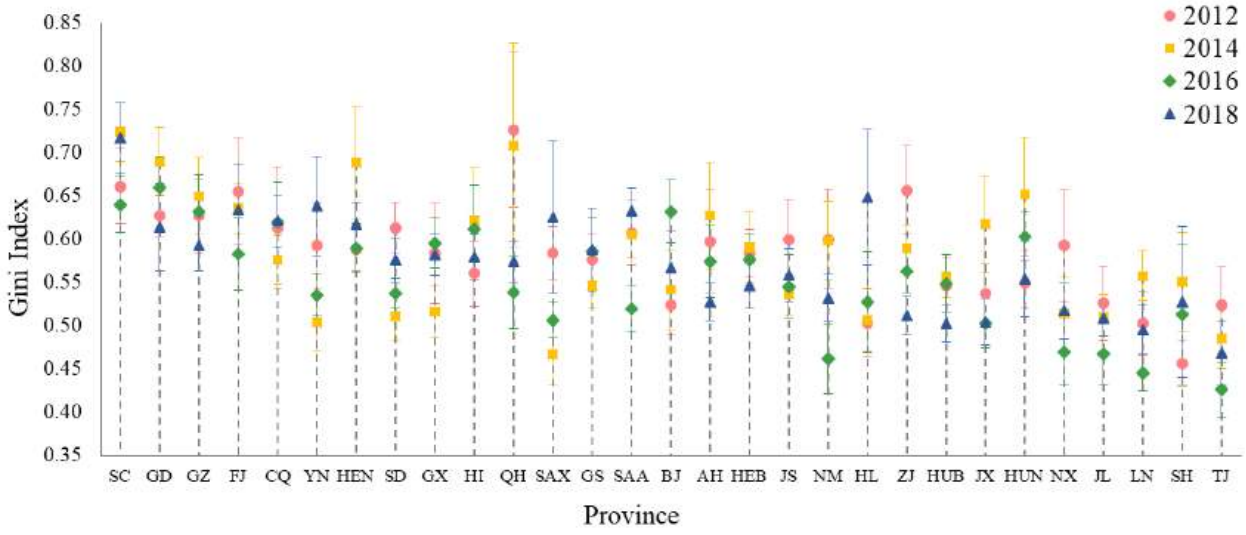


Notes: This figure plots provincial-level Gini index and returns to high school estimated from the CHFS. Correlations are: 64% in 2012, 61% in 2014, 58% in 2016, and 56% in 2018.

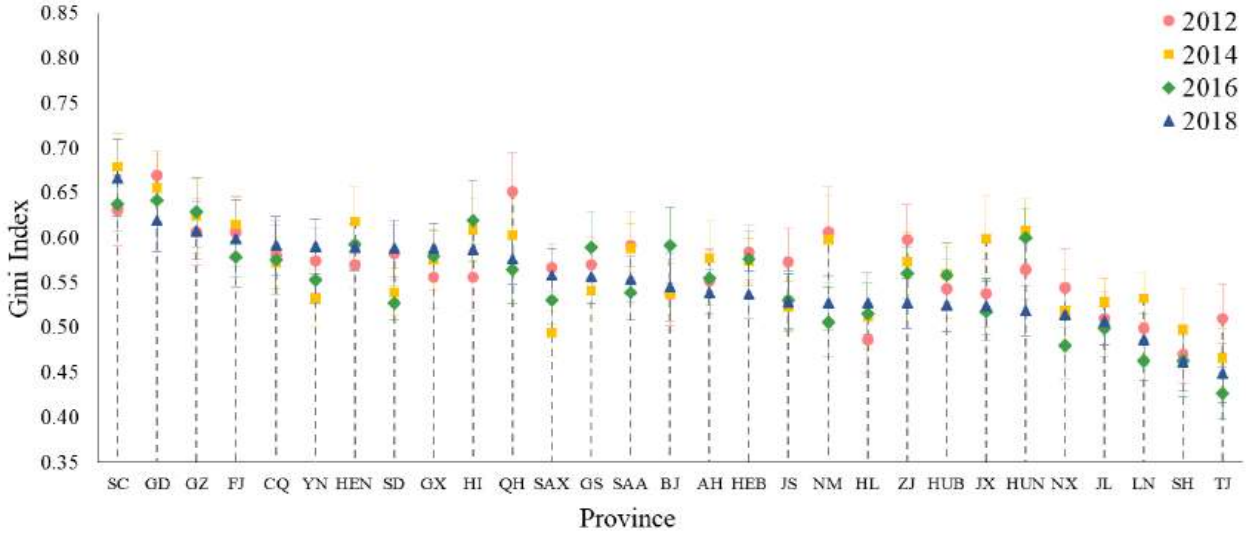
D Raw Gini estimates

Figure D1: Baseline versus raw data estimates of provincial Gini index.

(a) Raw.



(b) Baseline.



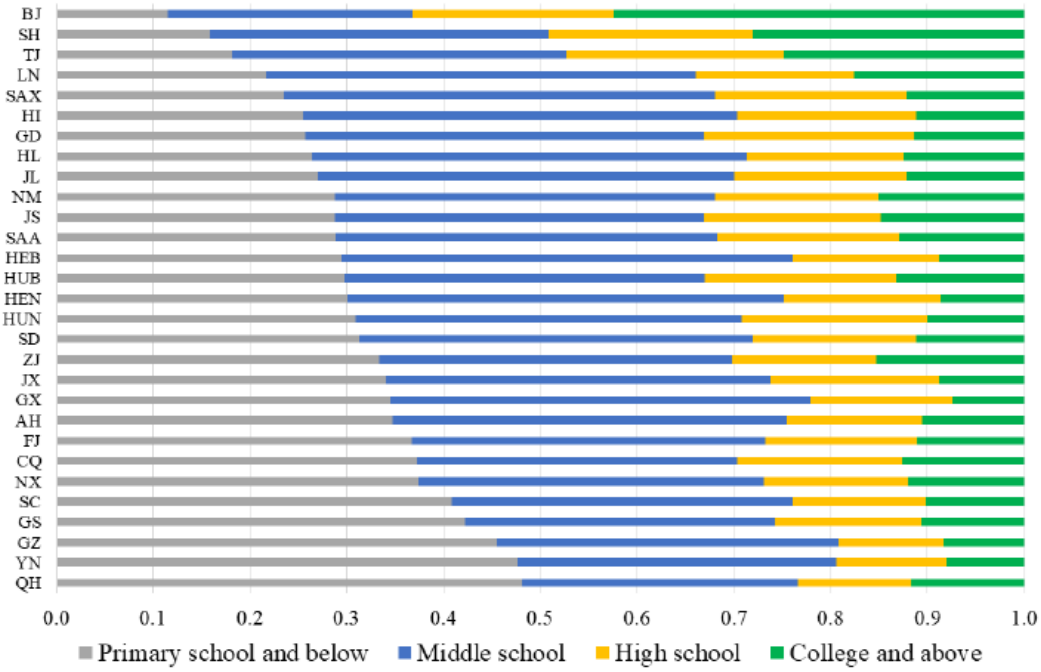
Notes: These figures illustrate the substantial reduction of variation in our baseline estimated provincial Gini indices (b) compared to the raw weighted data estimates (a).Error bars represent 95% confidence intervals computed from 1000 bootstrap samples. Semiparametric bootstrap, as in Davidson and Flachaire (2007), is used to produce our estimates, and conventional bootstrap is applied to the raw weighted data results reported here.

E Comparison to the NBS official statistics

Table E1: Demographic and economic statistics on Chinese provinces. Data source: China Statistical Yearbook.

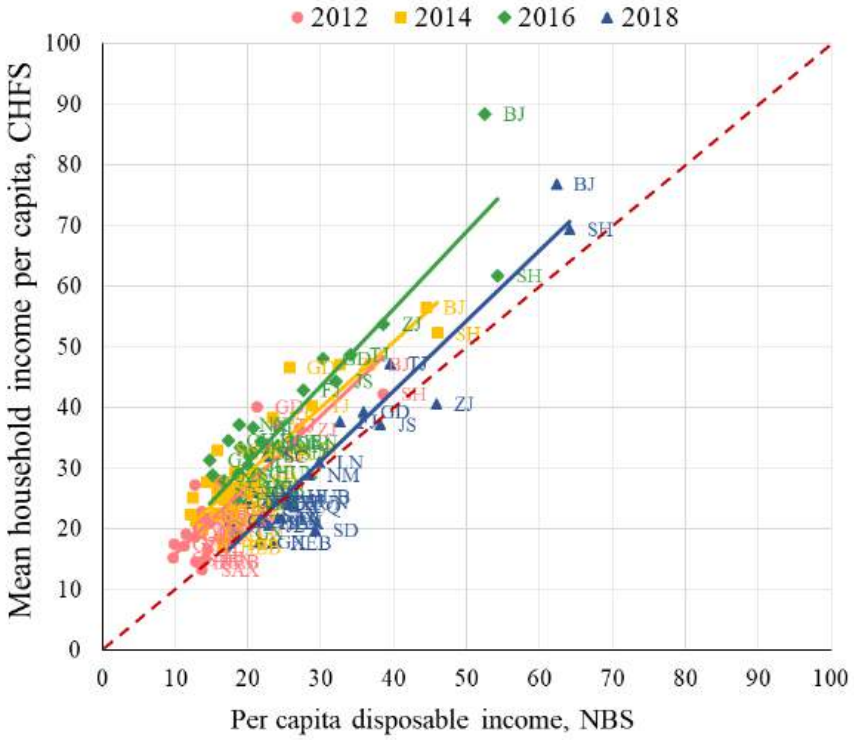
	Population (10,000 persons)				Rural popilation share (%)				GRP per capita (CN¥)				Per capita disposable income (CN¥)			
	2012	2014	2016	2018	2012	2014	2016	2018	2012	2014	2016	2018	2012	2014	2016	2018
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Beijing	2,078	2,171	2,195	2,192	0.137	0.135	0.133	0.129	92,758	106,732	123,391	150,962	36,817	44,489	52,530	62,361
Shanghai	2,399	2,467	2,467	2,475	0.107	0.107	0.110	0.109	89,613	102,827	121,369	145,767	38,550	45,966	54,305	64,183
Jiangsu	8,120	8,281	8,381	8,446	0.370	0.343	0.311	0.288	66,533	78,711	92,658	110,508	22,432	27,173	32,070	38,096
Tianjin	1,378	1,429	1,443	1,383	0.184	0.174	0.167	0.161	66,517	74,960	79,647	95,689	24,030	28,832	34,074	39,506
Zhejiang	5,685	5,890	6,072	6,273	0.371	0.350	0.323	0.300	61,097	68,569	78,384	93,230	27,020	32,658	38,529	45,840
Fujian	3,841	3,945	4,016	4,104	0.407	0.380	0.356	0.330	52,959	63,709	74,024	94,719	19,141	23,331	27,608	32,644
Guangdong	11,041	11,489	11,908	12,348	0.329	0.314	0.299	0.282	52,308	59,909	69,671	81,625	21,268	25,685	30,296	35,810
Shandong	9,708	9,808	9,973	10,077	0.480	0.452	0.409	0.385	44,348	51,933	59,239	66,284	17,127	20,864	24,685	29,205
Inner Mongolia	2,464	2,449	2,436	2,422	0.416	0.390	0.366	0.345	42,441	49,585	56,560	66,491	16,800	20,559	24,127	28,376
Hubei	5,781	5,816	5,885	5,917	0.468	0.443	0.414	0.390	39,149	48,635	56,844	71,097	14,809	18,283	21,787	25,815
Chongqing	2,975	3,043	3,110	3,163	0.434	0.403	0.367	0.334	39,180	48,311	58,327	68,460	14,924	18,352	22,034	26,386
Shaanxi	3,787	3,827	3,874	3,931	0.503	0.470	0.436	0.403	37,453	45,610	49,341	61,115	12,885	15,837	18,874	22,528
Liaoning	4,375	4,358	4,327	4,291	0.344	0.330	0.311	0.297	40,778	45,915	47,069	54,657	18,761	22,820	26,040	29,701
Hunan	6,590	6,611	6,625	6,635	0.538	0.510	0.473	0.439	32,203	39,181	46,606	54,763	14,391	17,622	21,115	25,241
Anhui	5,978	5,997	6,033	6,076	0.537	0.507	0.474	0.444	30,697	37,580	43,686	56,063	13,593	16,796	19,998	23,984
Hainan	910	936	957	982	0.490	0.467	0.433	0.408	30,993	37,166	43,009	50,263	14,180	17,476	20,653	24,579
Ningxia	659	678	695	710	0.489	0.451	0.413	0.379	32,609	36,815	40,339	49,614	13,104	15,907	18,832	22,400
Henan	9,532	9,645	9,778	9,864	0.580	0.550	0.512	0.478	30,497	35,982	41,326	50,714	12,772	15,695	18,443	21,964
Sichuan	8,085	8,139	8,251	8,321	0.566	0.535	0.500	0.465	29,627	35,563	40,297	51,658	12,753	15,749	18,808	22,461
Jilin	2,698	2,642	2,567	2,484	0.455	0.432	0.413	0.391	32,005	37,539	40,259	44,925	14,395	17,520	19,967	22,798
Jiangxi	4,475	4,480	4,496	4,513	0.526	0.494	0.460	0.427	28,624	34,988	40,950	50,347	13,567	16,734	20,110	24,080
Hebei	7,262	7,323	7,375	7,426	0.534	0.506	0.461	0.427	31,844	34,507	38,688	43,808	13,647	16,647	19,725	23,446
Shanxi	3,548	3,528	3,514	3,502	0.487	0.457	0.427	0.401	32,864	34,248	33,972	45,517	13,592	16,538	19,049	21,990
Qinghai	571	576	582	587	0.522	0.491	0.464	0.428	26,839	32,218	38,968	46,854	11,470	14,374	17,302	20,757
Heilongjiang	3,724	3,608	3,463	3,327	0.431	0.408	0.389	0.365	29,352	33,464	34,025	38,199	14,302	17,404	19,838	22,726
Yunnan	4,631	4,653	4,677	4,703	0.615	0.588	0.554	0.526	23,992	30,217	35,051	44,446	11,233	13,772	16,720	20,084
Guangxi	4,694	4,770	4,857	4,947	0.565	0.535	0.508	0.482	24,181	28,603	33,340	39,837	12,644	15,557	18,305	21,485
Guizhou	3,587	3,677	3,758	3,822	0.637	0.597	0.544	0.505	18,947	25,101	31,589	40,271	9,850	12,371	15,121	18,430
Gansu	2,550	2,531	2,520	2,515	0.612	0.577	0.539	0.503	21,141	25,724	27,396	32,178	9,768	12,185	14,670	17,488
Mean	4,591	4,647	4,698	4,739	0.453	0.427	0.399	0.373	40,743	47,735	54,346	65,519	16,891	20,593	24,332	28,771
SD	2,791	2,850	2,919	2,986	0.134	0.126	0.115	0.106	18,725	21,141	24,861	30,064	7,072	8,368	9,832	11,604
CV	0.608	0.613	0.621	0.630	0.296	0.294	0.288	0.285	0.460	0.443	0.457	0.459	0.419	0.406	0.404	0.403

Figure E1: Average educational structure in Chinese provinces, 2012-2018.



Notes: This figure plots the average educational structure of the six-years and above population in Chinese provinces for 2012-2018 estimated from the China Statistical Yearbook (CSY). The estimates are ordered in the "Primary school and below" category.

Figure E2: Mean household income per capita versus the NBS statistics (in 1000 CN¥).



Notes: This figure plots provincial-level mean household income per capita estimated from the CHFS versus per capita disposable income from the China Statistical Yearbook (CSY). Correlations are: 89% in 2012 and 2014, 91% in 2016, and 95% in 2018.

F Mincer regression for the returns to schooling

Table F1: Mincer regression: explaining household income per capita variation in China. Heteroskedasticity-robust SEs are used.

	Dependent variable: Ln(household income per capita)															
	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Schooling per worker	0.151***	16.0%	0.156***	16.7%	0.147***	17.5%	0.138***	15.9%	0.129***	13.6%	0.137***	14.6%	0.130***	15.3%	0.114***	13.1%
	(0.003)		(0.003)		(0.003)		(0.003)		(0.003)		(0.004)		(0.003)		(0.004)	
Experience per worker	0.012***	-2.0%	0.021***	-4.2%	0.030***	-6.3%	0.018***	-4.4%	0.015***	-2.3%	0.024***	-4.6%	0.034***	-7.0%	0.020***	-4.7%
	(0.004)		(0.004)		(0.003)		(0.004)		(0.004)		(0.004)		(0.003)		(0.003)	
Experience per worker^2 : 100	−0.018***	1.5%	−0.038***	4.4%	−0.053***	6.9%	−0.039***	5.9%	−0.026***	2.0%	−0.044***	4.9%	−0.060***	7.8%	−0.043***	6.4%
	(0.006)		(0.007)		(0.005)		(0.006)		(0.006)		(0.006)		(0.005)		(0.005)	
Ln(household size)	−0.854***	5.5%	−0.909***	5.1%	−0.899***	5.9%	−0.852***	4.1%	−0.848***	5.4%	−0.848***	4.7%	−0.846***	5.5%	−0.814***	3.9%
	(0.028)		(0.032)		(0.027)		(0.030)		(0.028)		(0.032)		(0.027)		(0.029)	
Constant	9.011***		8.975***		9.199***		9.480***		8.731***		8.817***		9.033***		9.743***	
	(0.086)		(0.099)		(0.085)		(0.091)		(0.130)		(0.141)		(0.144)		(0.139)	
Provincial dummies	Yes	5.4%	Yes	4.0%	Yes	5.1%	Yes	5.8%	No		No		No		No	
City dummies	No	No	No	No	No	No	No	No	Yes	10.8%	Yes	9.3%	Yes	10.6%	Yes	13.6%
Residual		73.7%		74.0%		70.9%		72.8%		70.5%		71.0%		67.7%		67.8%
Observations	24,918		32,156		34,940		28,908		24,918		32,156		34,940		28,908	
R^2	0.285		0.282		0.311		0.298		0.317		0.311		0.343		0.348	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

G Mincer regression for the returns to high school

Table G1: Mincer regression: explaining household income per capita variation in China. Heteroskedasticity-robust SEs are used.

	Dependent variable: Ln(household income per capita)															
	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)	(2012)	($\hat{s}_z$)	(2014)	($\hat{s}_z$)	(2016)	($\hat{s}_z$)	(2018)	($\hat{s}_z$)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
High school per worker	0.768***	8.5%	0.770***	9.2%	0.696***	8.5%	0.751***	8.6%	0.616***	6.8%	0.647***	7.6%	0.578***	7.0%	0.575***	6.5%
	(0.020)		(0.022)		(0.018)		(0.022)		(0.020)		(0.022)		(0.019)		(0.022)	
Experience per worker	0.016***	-2.4%	0.026***	-5.1%	0.037***	-7.5%	0.031***	-7.0%	0.018***	-2.7%	0.028***	-5.3%	0.039***	-7.9%	0.030***	-6.7%
	(0.004)		(0.004)		(0.003)		(0.004)		(0.004)		(0.004)		(0.003)		(0.003)	
Experience per worker^2 : 100	−0.053***	3.9%	−0.076***	8.0%	−0.090***	11.3%	−0.084***	11.8%	−0.056***	4.1%	−0.077***	8.1%	−0.092***	11.6%	−0.079***	11.1%
	(0.006)		(0.007)		(0.005)		(0.005)		(0.006)		(0.007)		(0.005)		(0.005)	
Ln(household size)	−0.951***	6.1%	−1.012***	5.7%	−0.993***	6.5%	−0.934***	4.5%	−0.919***	5.9%	−0.920***	5.1%	−0.908***	5.9%	−0.868***	4.2%
	(0.029)		(0.032)		(0.027)		(0.030)		(0.028)		(0.032)		(0.027)		(0.029)	
Constant	10.471***		10.521***		10.614***		10.704***		9.898***		10.084***		10.248***		10.661***	
	(0.077)		(0.083)		(0.074)		(0.077)		(0.124)		(0.136)		(0.141)		(0.127)	
Provincial dummies	Yes	6.4%	Yes	4.6%	Yes	5.9%	Yes	6.4%	No		No		No		No	
City dummies	No	No	No	No	No	No	No	No	Yes	13.0%	Yes	10.9%	Yes	12.7%	Yes	15.1%
Residual		77.5%		77.6%		75.3%		75.8%		72.9%		73.5%		70.8%		69.8%
Observations	24,918		32,156		34,940		28,908		24,918		32,156		34,940		28,908	
R^2	0.246		0.244		0.267		0.267		0.292		0.285		0.311		0.328	

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

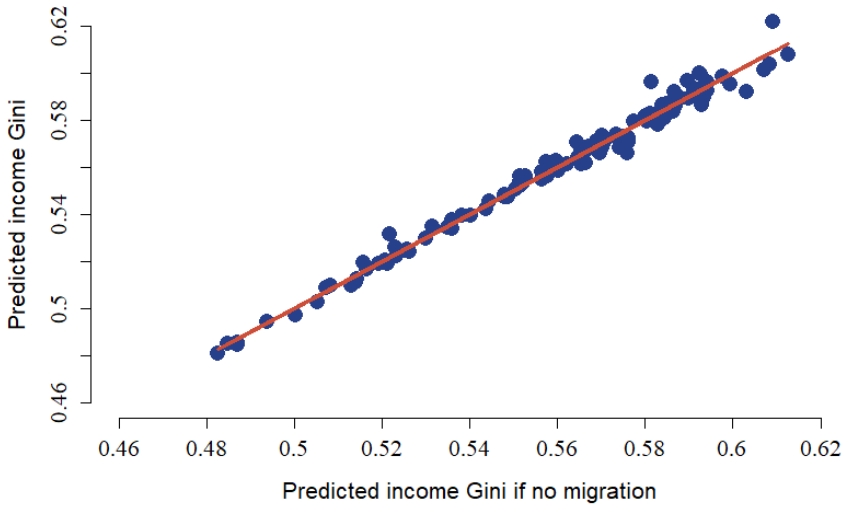
H Educational Theil T index decomposition

Table H1: Decomposition of educational Theil T index by province and birth cohorts.

	2012				2014				2016				2018			
	Theil T index	Share (%)	Within birth cohort (%)	Between birth cohort (%)	Theil T index	Share (%)	Within birth cohort (%)	Between birth cohort (%)	Theil T index	Share (%)	Within birth cohort (%)	Between birth cohort (%)	Theil T index	Share (%)	Within birth cohort (%)	Between birth cohort (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Beijing	0.088	2.0%	1.6%	0.4%	0.084	1.1%	0.8%	0.3%	0.075	1.1%	0.8%	0.3%	0.076	1.0%	0.8%	0.3%
Shanghai	0.088	1.4%	1.0%	0.3%	0.094	1.3%	1.0%	0.3%	0.093	1.4%	1.0%	0.4%	0.092	1.1%	0.8%	0.3%
Jiangsu	0.166	3.6%	2.7%	0.9%	0.203	7.4%	5.6%	1.8%	0.165	6.5%	5.2%	1.3%	0.157	5.7%	4.6%	1.1%
Tianjin	0.088	1.5%	1.2%	0.3%	0.091	0.8%	0.6%	0.1%	0.082	0.8%	0.7%	0.1%	0.095	0.8%	0.7%	0.1%
Zhejiang	0.183	5.0%	3.5%	1.5%	0.156	3.9%	2.7%	1.1%	0.174	4.2%	3.0%	1.2%	0.208	4.9%	3.4%	1.5%
Fujian	0.212	6.5%	5.2%	1.3%	0.213	3.5%	2.8%	0.7%	0.191	3.7%	2.9%	0.8%	0.204	3.3%	2.6%	0.7%
Guangdong	0.121	3.3%	2.9%	0.5%	0.125	6.0%	5.1%	0.9%	0.134	6.6%	5.6%	1.0%	0.137	6.1%	5.1%	1.0%
Shandong	0.158	2.5%	1.9%	0.6%	0.155	6.9%	5.3%	1.6%	0.137	4.7%	3.7%	1.0%	0.194	7.5%	6.0%	1.5%
Inner Mongolia	0.187	2.2%	1.7%	0.5%	0.153	1.9%	1.5%	0.3%	0.156	1.5%	1.2%	0.2%	0.144	1.7%	1.3%	0.4%
Hubei	0.157	4.4%	3.9%	0.5%	0.168	4.6%	4.0%	0.6%	0.160	4.3%	3.5%	0.7%	0.196	4.5%	3.7%	0.8%
Chongqing	0.193	4.6%	3.5%	1.1%	0.207	2.5%	1.9%	0.6%	0.166	2.4%	1.7%	0.6%	0.218	2.6%	1.8%	0.8%
Shaanxi	0.165	3.6%	3.1%	0.5%	0.145	2.6%	2.2%	0.4%	0.131	2.5%	2.2%	0.3%	0.184	3.1%	2.6%	0.5%
Liaoning	0.110	2.2%	1.8%	0.4%	0.096	2.1%	1.7%	0.4%	0.085	2.2%	1.8%	0.4%	0.091	2.0%	1.6%	0.5%
Hunan	0.143	3.7%	3.1%	0.6%	0.149	4.4%	3.6%	0.7%	0.157	4.9%	4.1%	0.8%	0.151	4.4%	3.6%	0.8%
Anhui	0.260	3.3%	2.6%	0.7%	0.261	6.2%	5.1%	1.2%	0.301	6.8%	5.5%	1.3%	0.290	6.8%	5.0%	1.8%
Hainan	0.139	2.1%	1.9%	0.2%	0.150	0.5%	0.5%	0.0%	0.122	0.6%	0.5%	0.1%	0.141	0.5%	0.4%	0.1%
Ningxia	0.216	3.1%	2.5%	0.6%	0.215	0.6%	0.5%	0.1%	0.163	0.6%	0.5%	0.1%	0.218	0.6%	0.4%	0.1%
Henan	0.199	2.7%	2.3%	0.5%	0.227	8.5%	7.2%	1.4%	0.201	7.7%	6.4%	1.3%	0.209	7.8%	6.4%	1.3%
Sichuan	0.138	5.0%	3.8%	1.2%	0.195	6.4%	5.0%	1.4%	0.219	6.9%	5.7%	1.2%	0.194	6.2%	4.9%	1.3%
Jilin	0.130	2.4%	2.1%	0.4%	0.121	1.5%	1.3%	0.2%	0.103	1.6%	1.4%	0.2%	0.130	1.6%	1.3%	0.3%
Jiangxi	0.196	3.5%	2.8%	0.7%	0.186	3.4%	2.7%	0.7%	0.155	3.3%	2.7%	0.6%	0.175	3.2%	2.5%	0.7%
Hebei	0.169	2.9%	2.4%	0.5%	0.180	5.2%	4.4%	0.8%	0.154	5.2%	4.2%	1.0%	0.181	5.1%	4.4%	0.8%
Shanxi	0.130	2.8%	2.4%	0.4%	0.116	2.0%	1.6%	0.4%	0.103	2.1%	1.7%	0.4%	0.129	2.2%	1.8%	0.4%
Qinghai	0.460	6.1%	5.5%	0.6%	0.447	0.8%	0.7%	0.1%	0.317	0.9%	0.8%	0.1%	0.370	0.8%	0.7%	0.1%
Heilongjiang	0.139	2.2%	2.0%	0.2%	0.119	2.2%	1.9%	0.3%	0.125	2.4%	2.1%	0.3%	0.122	2.1%	1.9%	0.3%
Yunnan	0.220	3.6%	3.0%	0.6%	0.205	3.9%	3.4%	0.5%	0.197	4.2%	3.7%	0.6%	0.240	4.2%	3.6%	0.6%
Guangxi	0.108	2.8%	2.5%	0.3%	0.140	2.7%	2.3%	0.4%	0.136	2.7%	2.3%	0.5%	0.118	2.1%	1.7%	0.4%
Guizhou	0.227	3.9%	3.2%	0.7%	0.224	3.0%	2.4%	0.6%	0.207	3.2%	2.6%	0.6%	0.255	3.3%	2.5%	0.8%
Gansu	0.232	4.3%	3.8%	0.6%	0.190	2.1%	1.9%	0.2%	0.219	2.8%	2.4%	0.3%	0.257	2.6%	2.2%	0.4%
Within	0.166	97.3%	79.7%	17.6%	0.167	98.0%	79.9%	18.1%	0.159	97.9%	80.1%	17.8%	0.175	97.8%	78.2%	19.6%
Between	0.005	2.7%			0.003	2.0%			0.003	2.1%			0.004	2.2%		
Total	0.171	100%			0.170	100%			0.162	100%			0.179	100%		

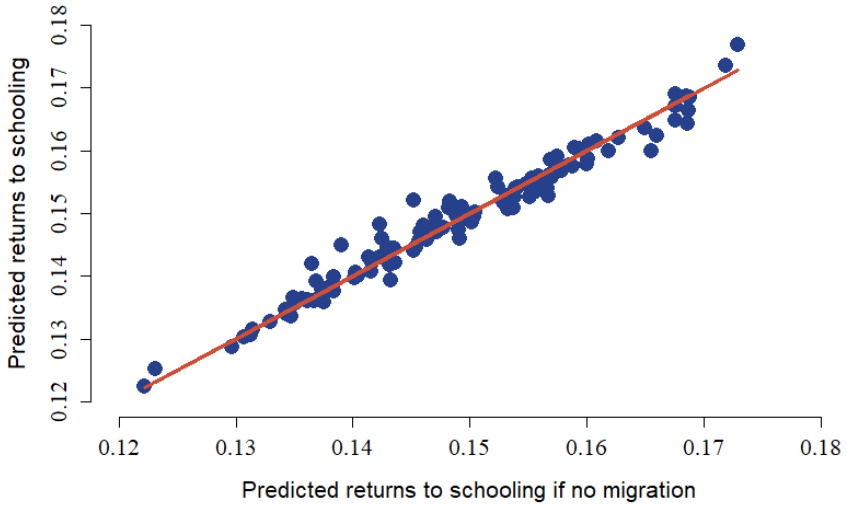
I Robustness checks

Figure I1: Predicted Gini index with/without migration: pooled OLS.



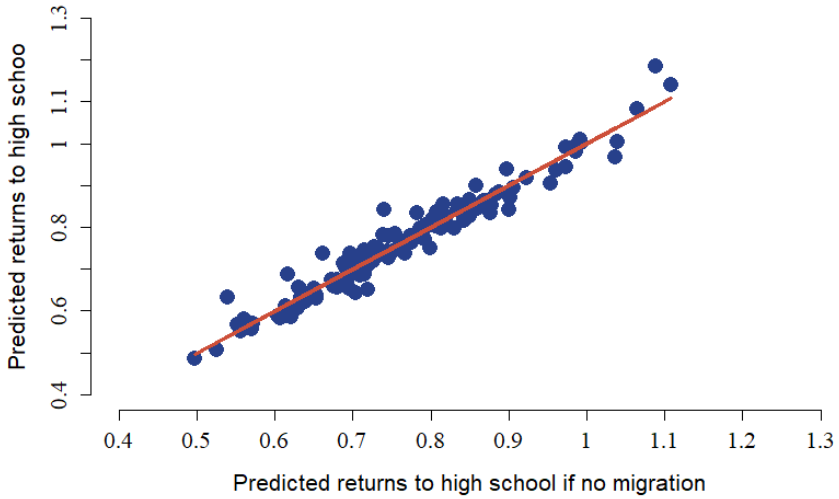
Notes: This figure plots predicted provincial-level Gini index with and without migration estimated from the CHFS. Each dot represents an estimate for each province/year. Correlation is 99%.

Figure I2: Predicted returns to schooling with/without migration: pooled OLS.



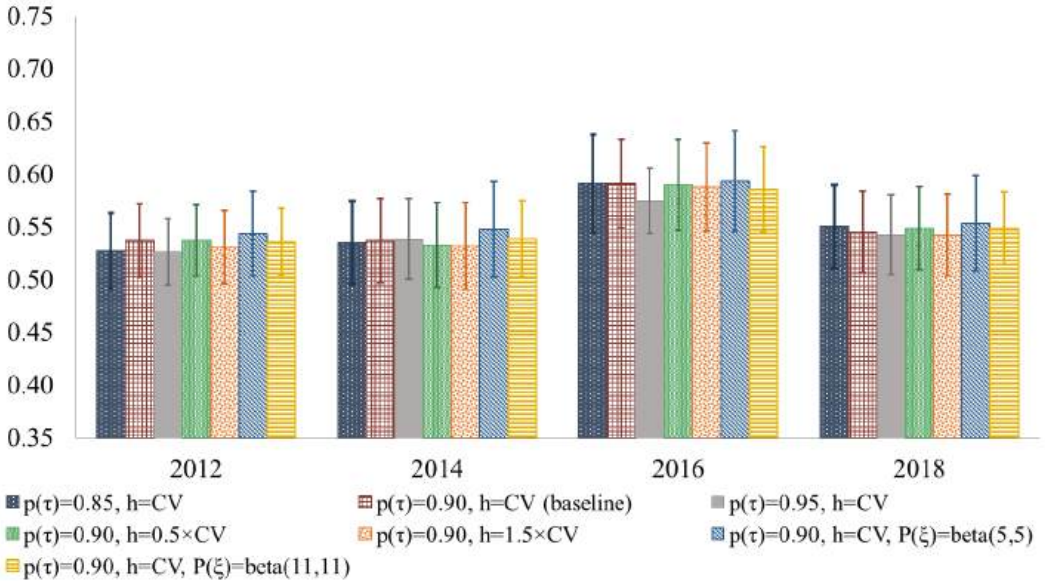
Notes: This figure plots predicted provincial-level returns to schooling with and without migration estimated from the CHFS. Each dot represents an estimate for each province/year. Correlation is 98%.

Figure I3: Predicted returns to high school with/without migration: pooled OLS.



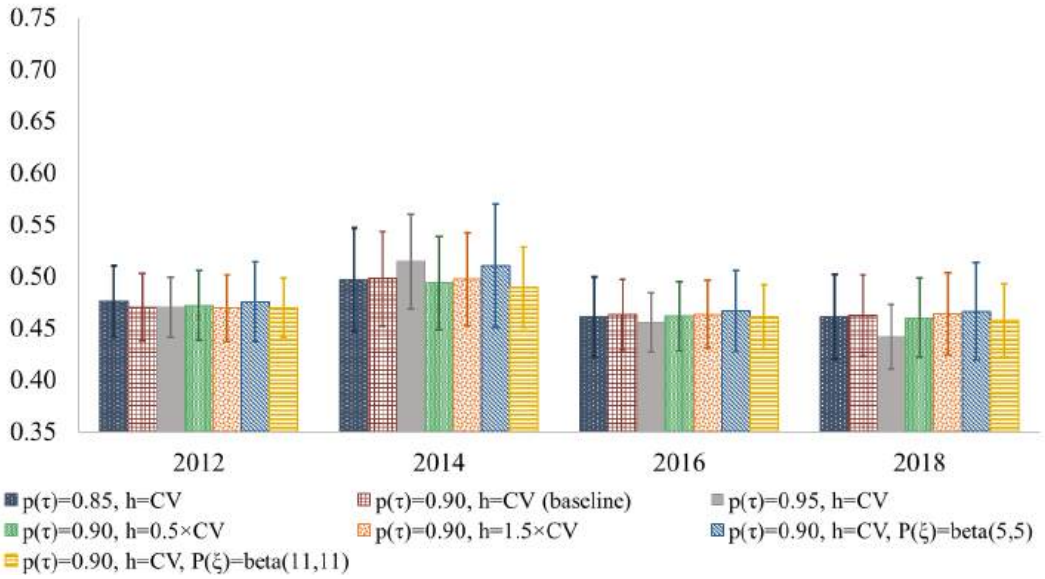
Notes: This figure plots predicted provincial-level returns to high school with and without migration estimated from the CHFS. Each dot represents an estimate for each province/year. Correlation is 97%.

Figure I4: Beijing Gini estimates under the alternative parameter choices.



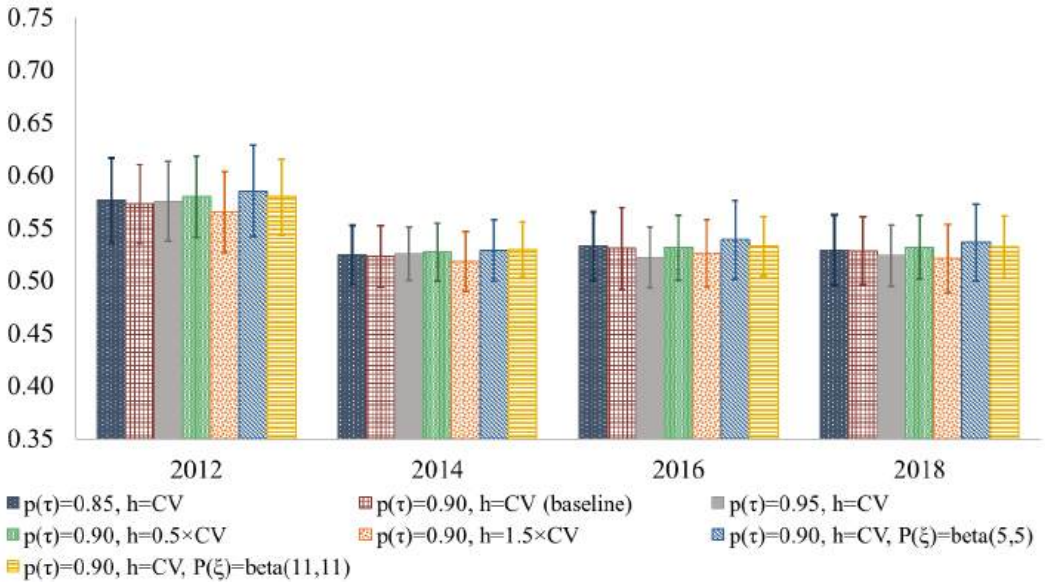
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I5: Shanghai Gini estimates under the alternative parameter choices.



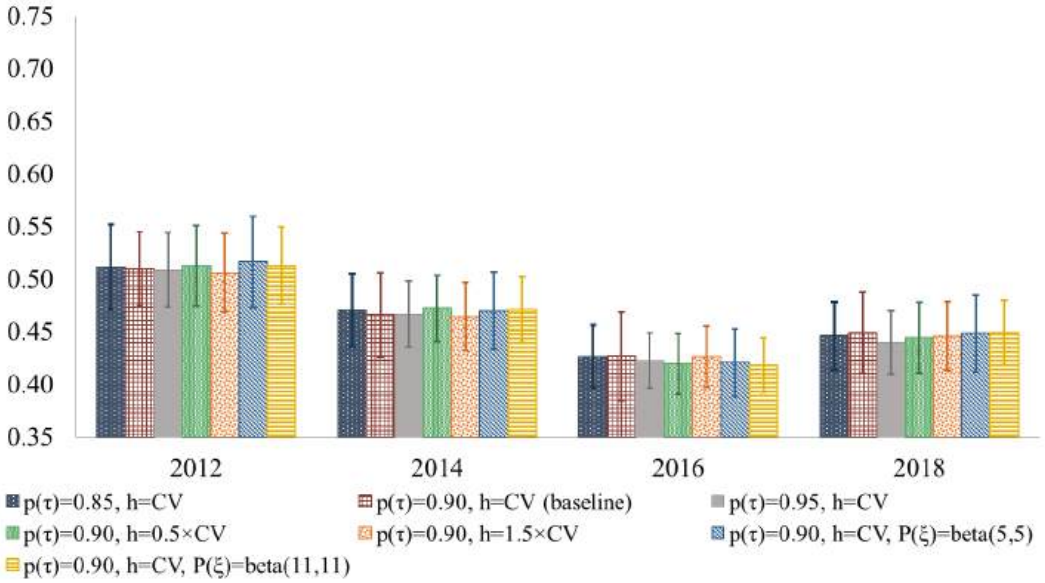
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I6: Jiangsu Gini estimates under the alternative parameter choices.



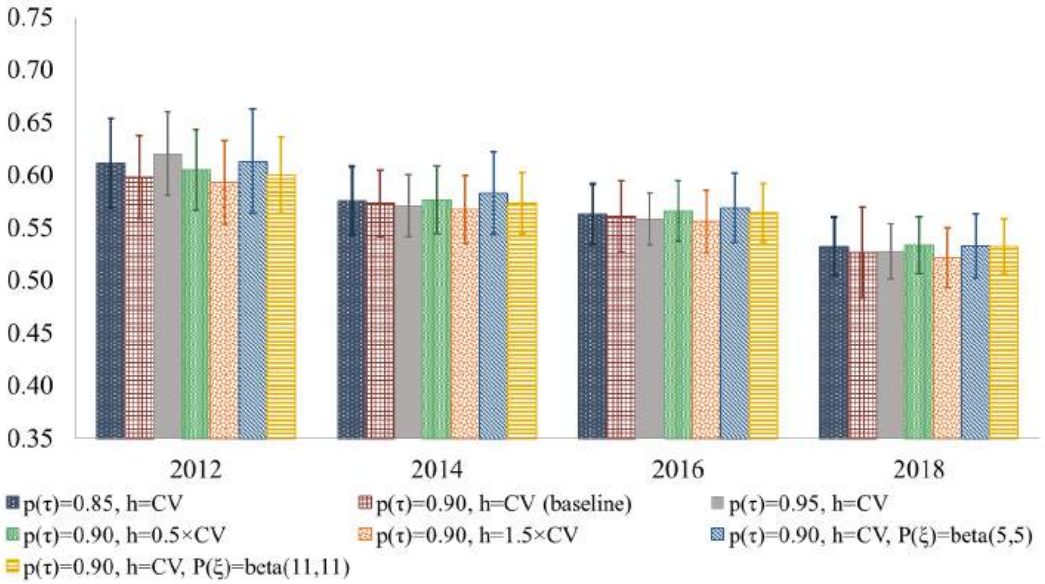
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I7: Tianjin Gini estimates under the alternative parameter choices.



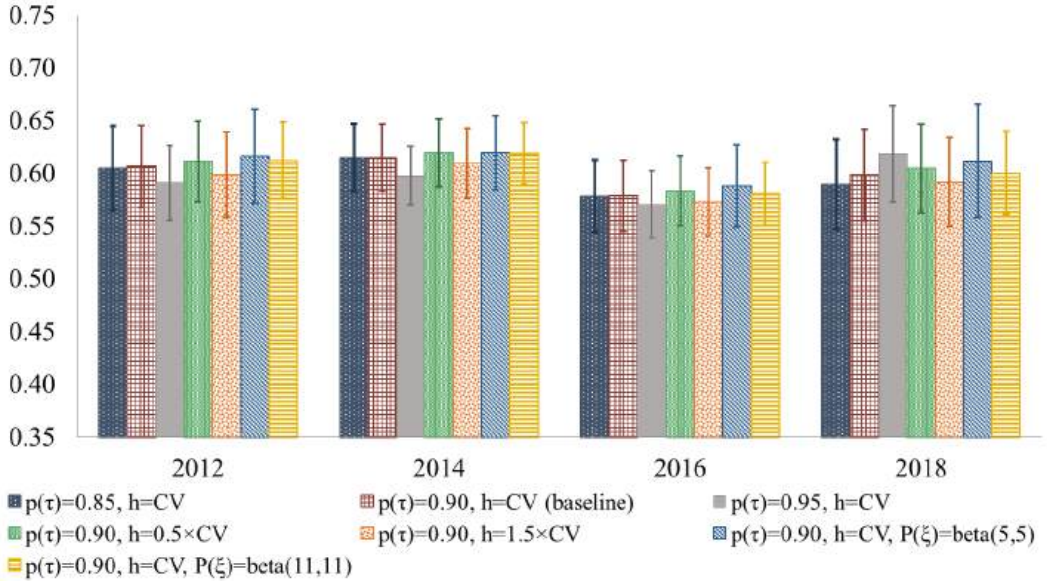
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I8: Zhejiang Gini estimates under the alternative parameter choices.



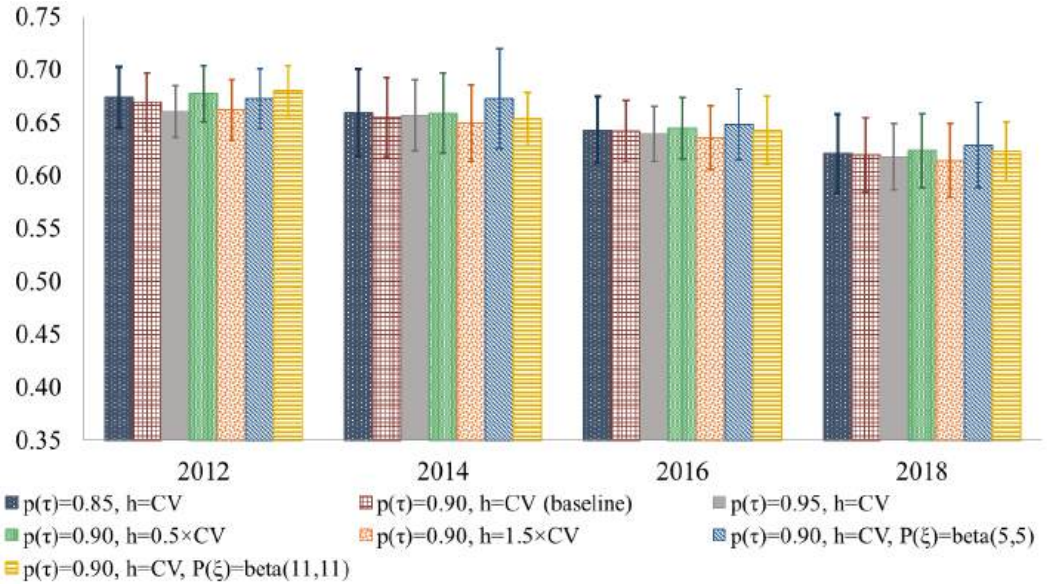
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I9: Fujian Gini estimates under the alternative parameter choices.



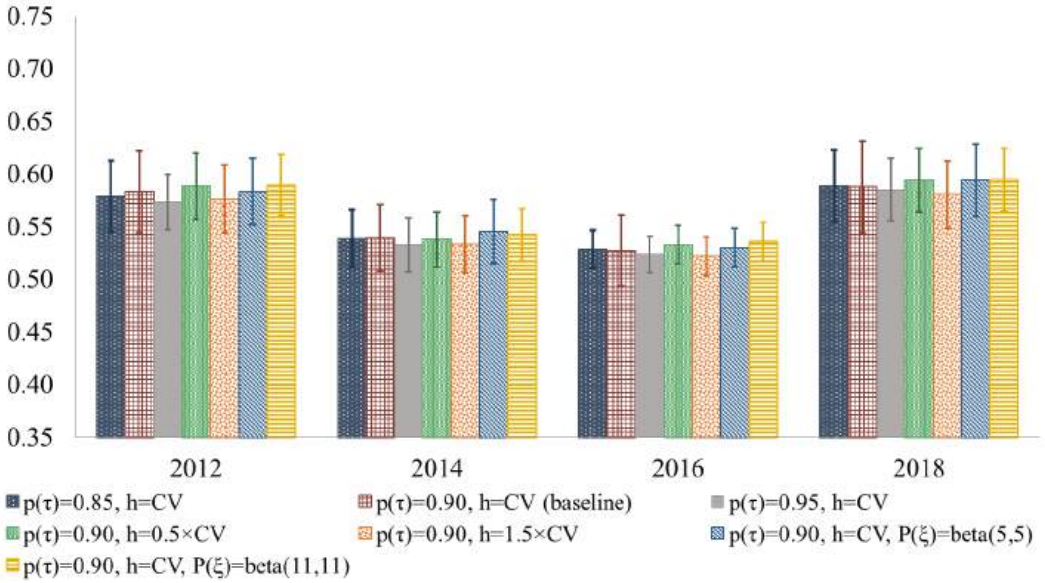
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I10: Guangdong Gini estimates under the alternative parameter choices.



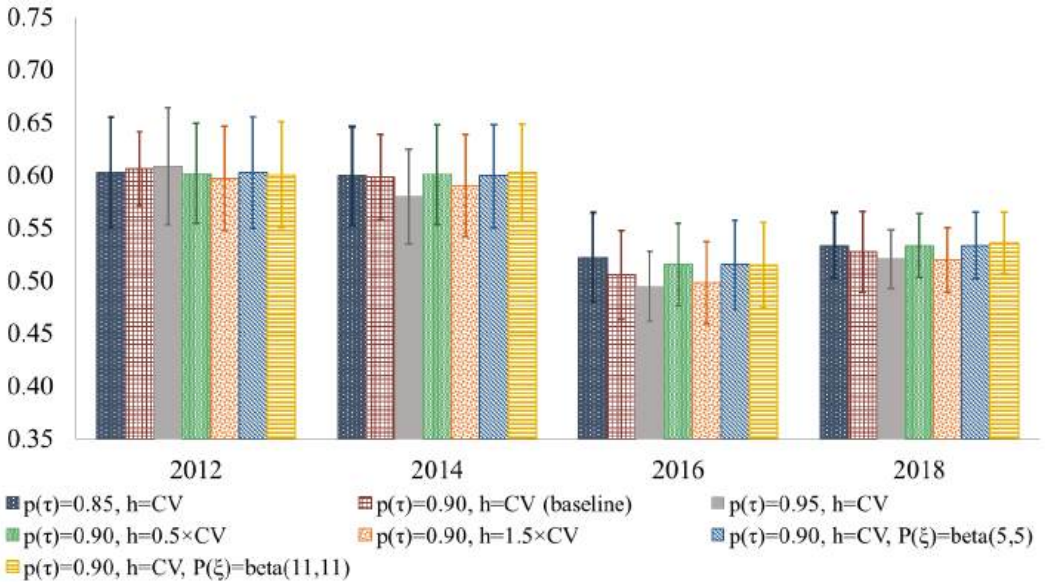
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I11: Shandong Gini estimates under the alternative parameter choices.



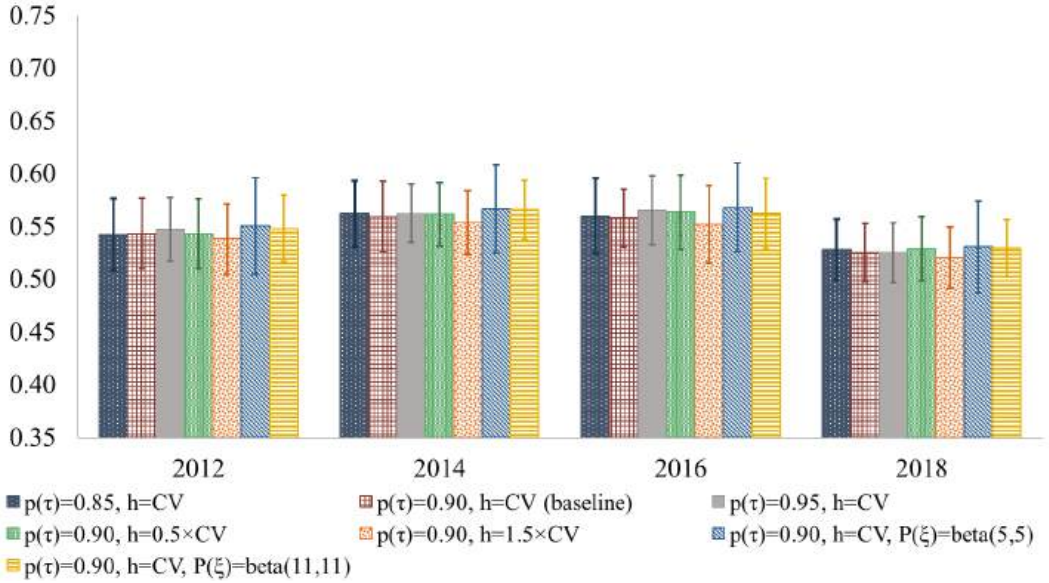
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I12: Inner Mongolia Gini estimates under the alternative parameter choices.



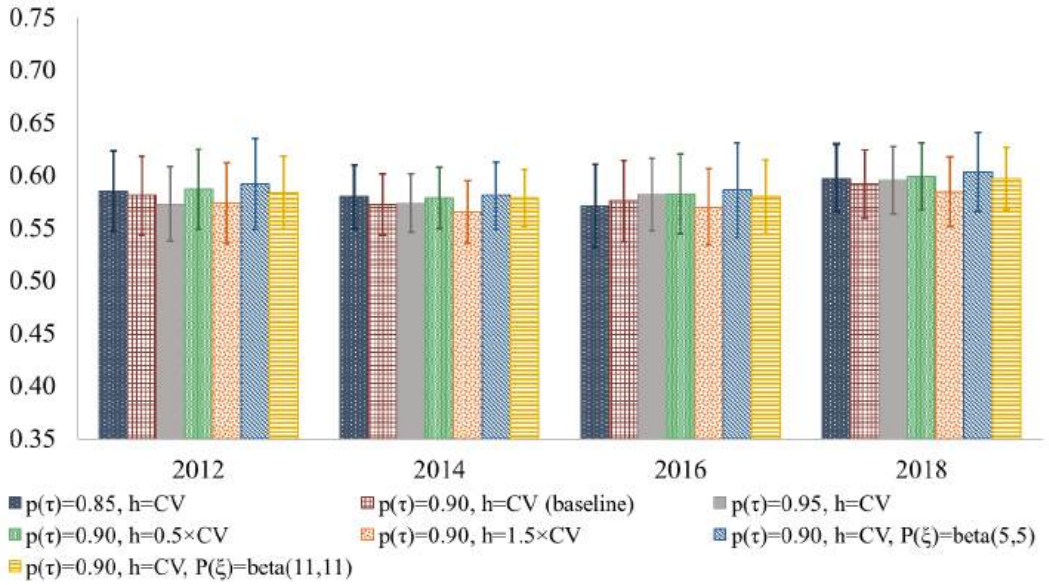
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I13: Hubei Gini estimates under the alternative parameter choices.



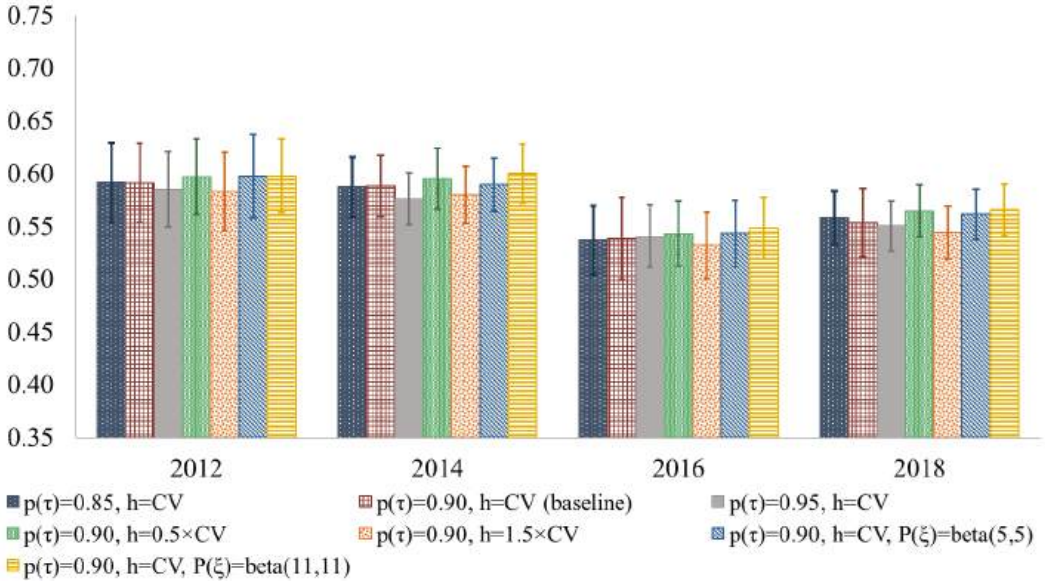
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I14: Chongqing Gini estimates under the alternative parameter choices.



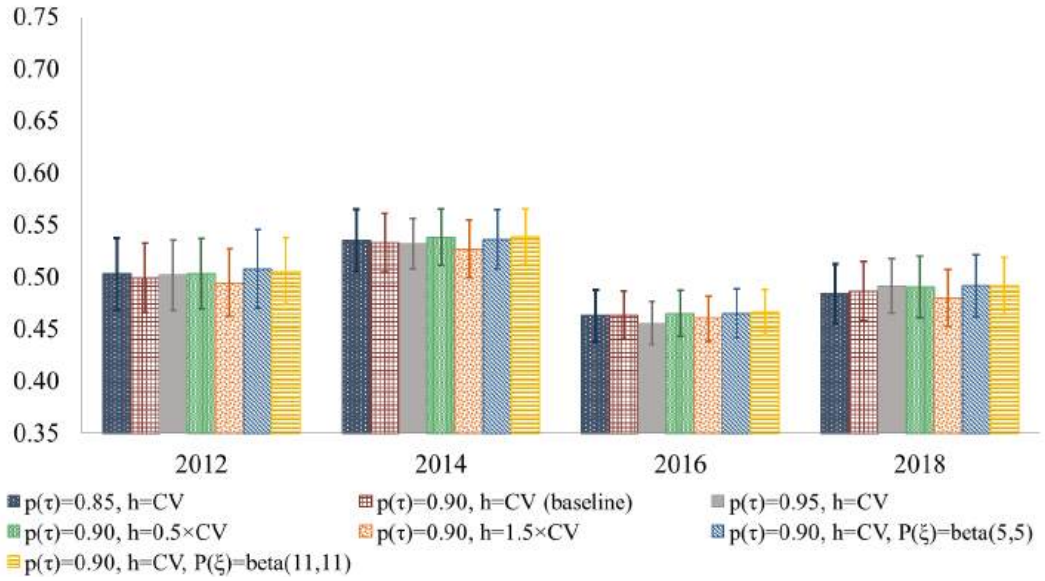
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I15: Shaanxi Gini estimates under the alternative parameter choices.



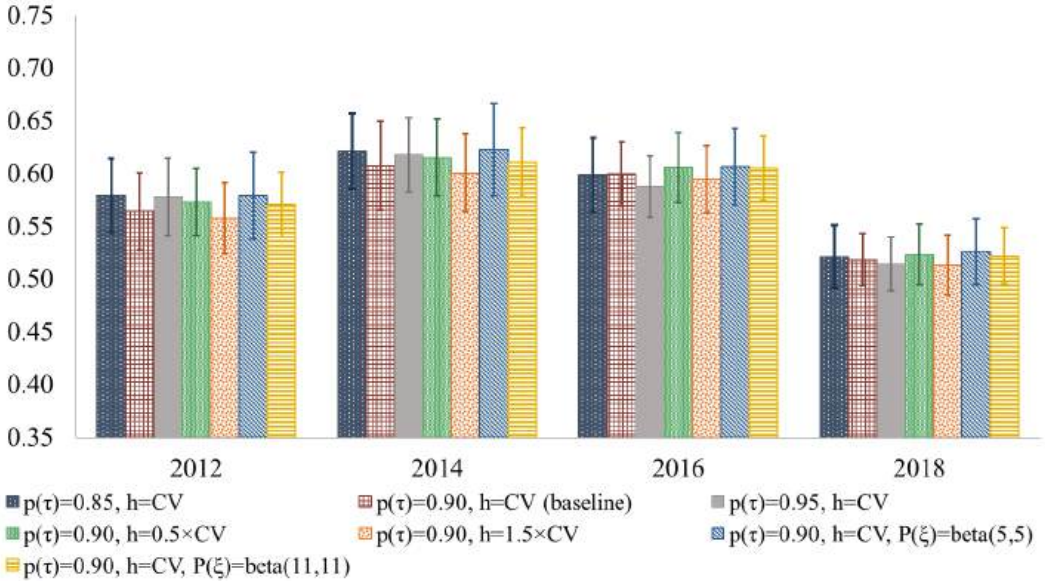
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I16: Liaoning Gini estimates under the alternative parameter choices.



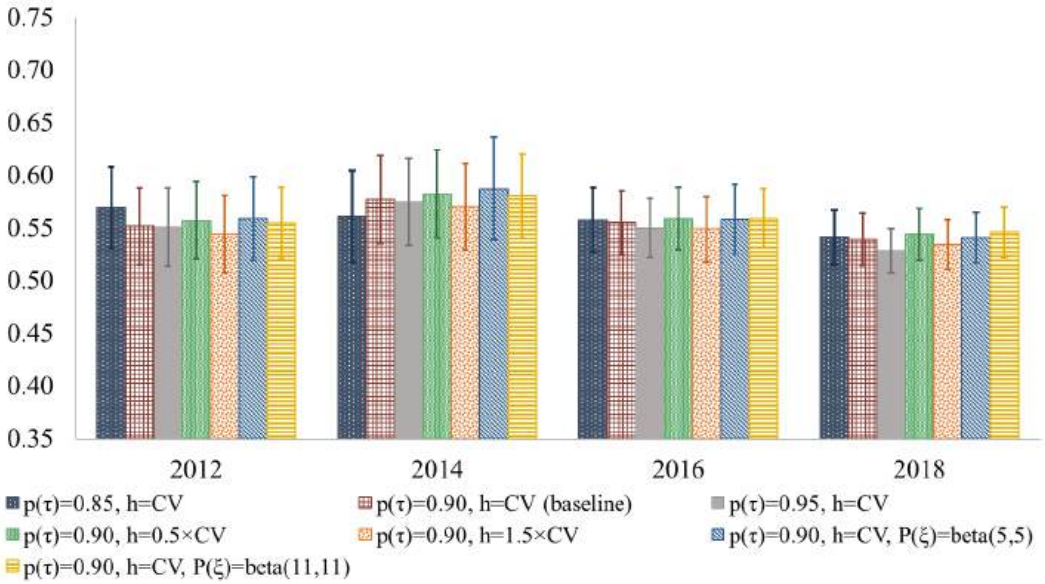
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I17: Hunan Gini estimates under the alternative parameter choices.



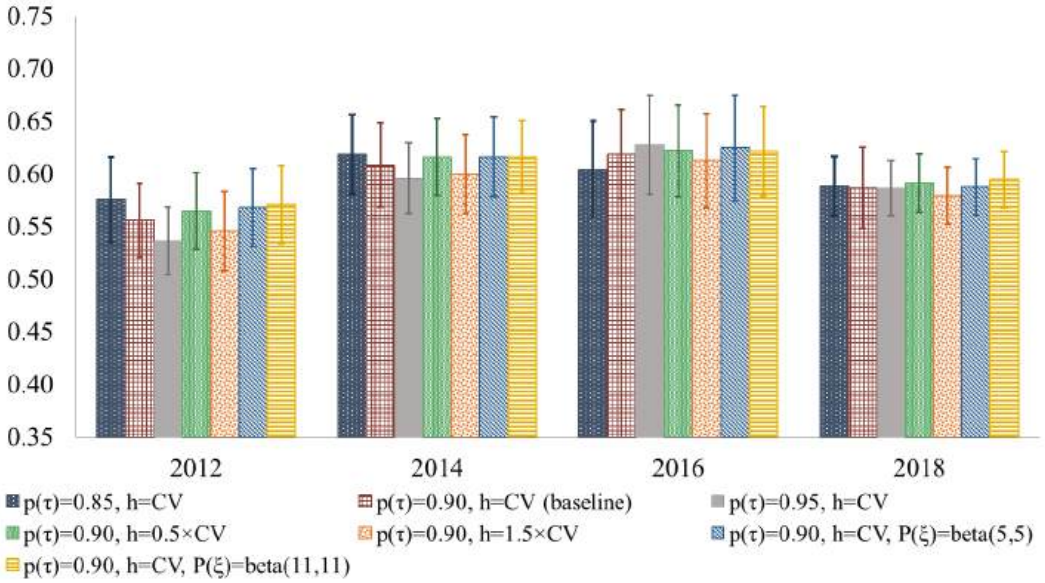
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I18: Anhui Gini estimates under the alternative parameter choices.



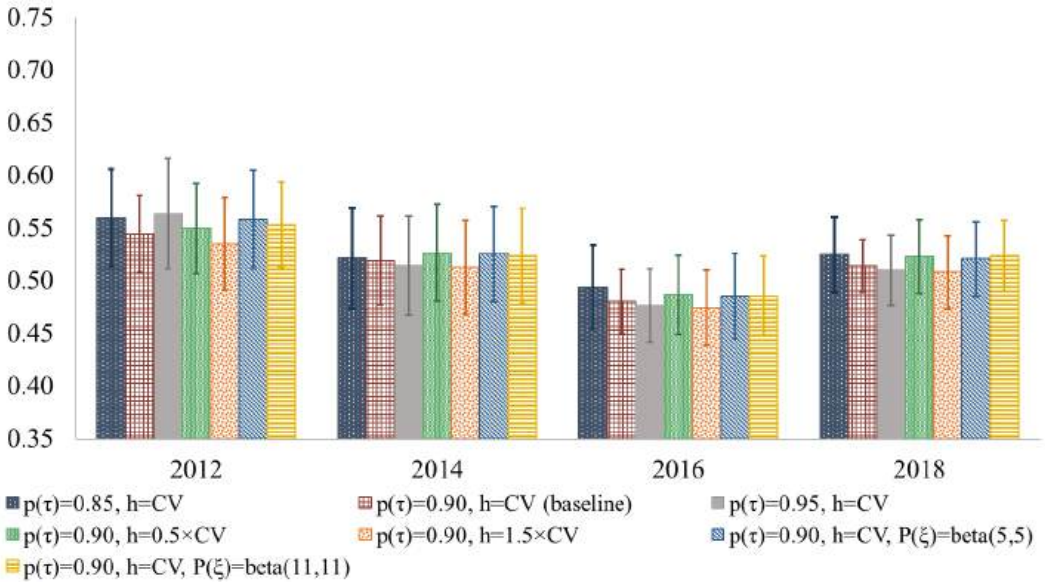
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I19: Hainan Gini estimates under the alternative parameter choices.



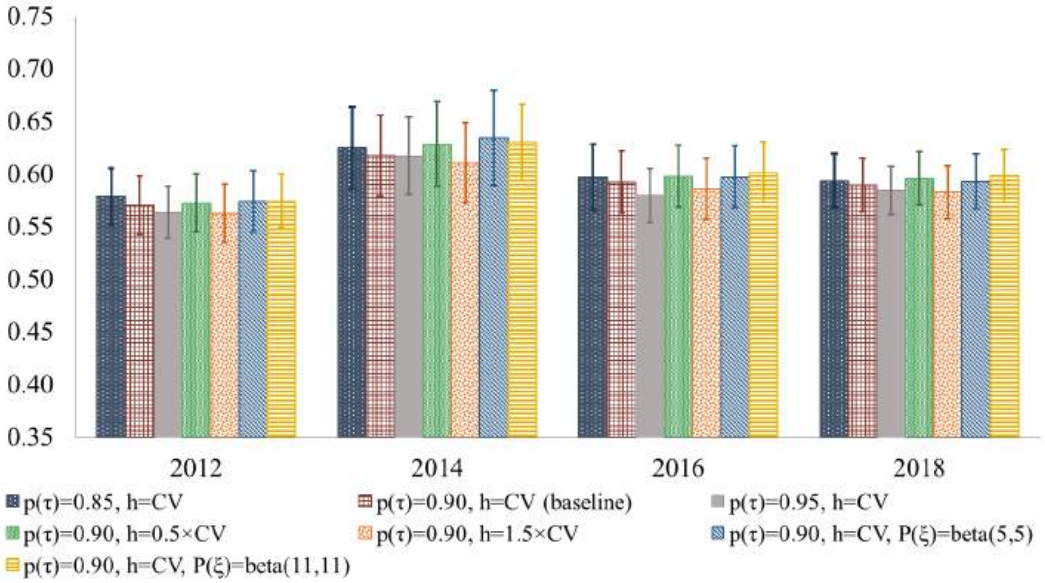
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I20: Ningxia Gini estimates under the alternative parameter choices.



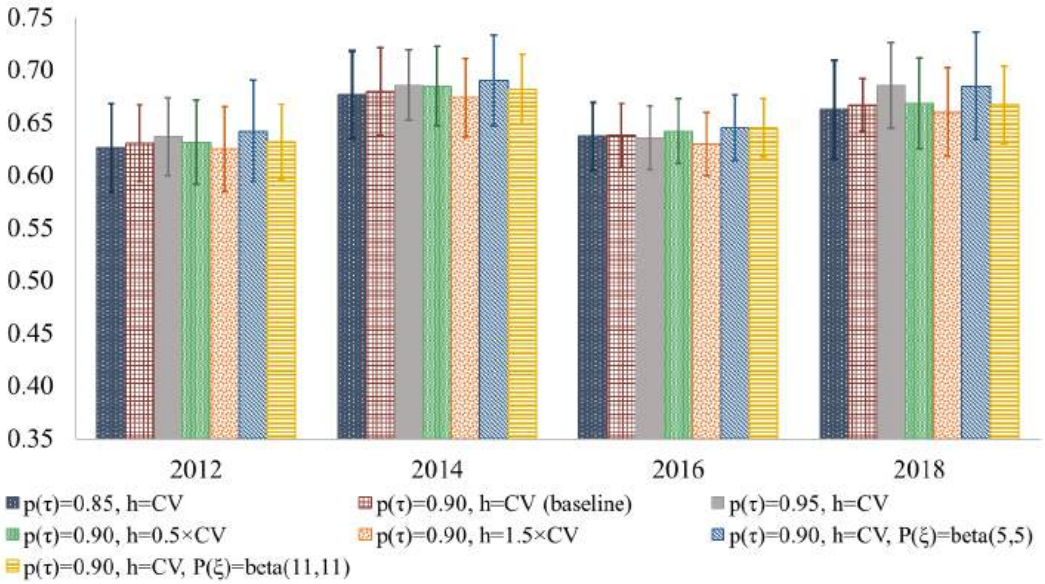
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I21: Henan Gini estimates under the alternative parameter choices.



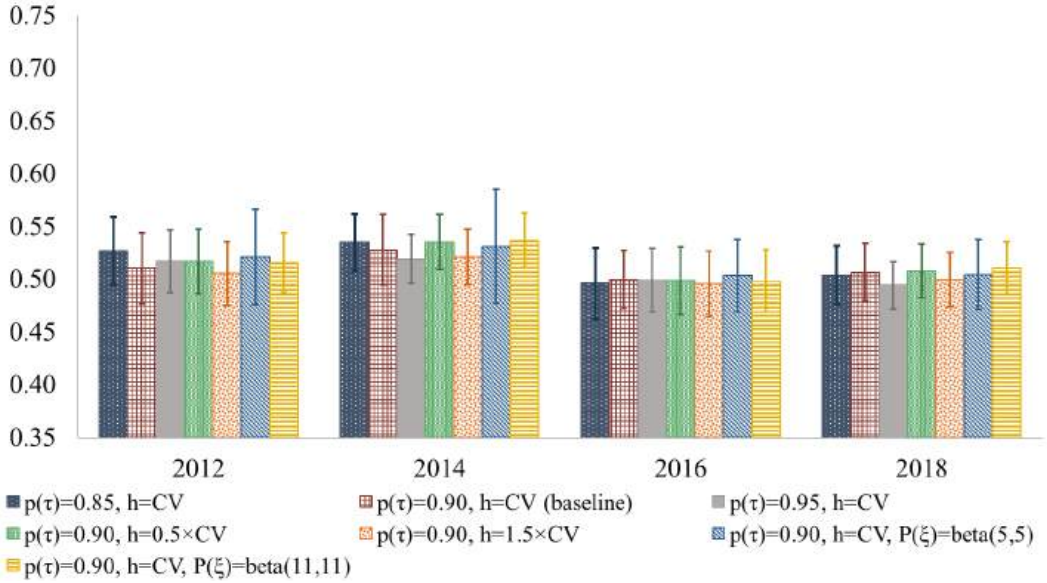
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I22: Sichuan Gini estimates under the alternative parameter choices.



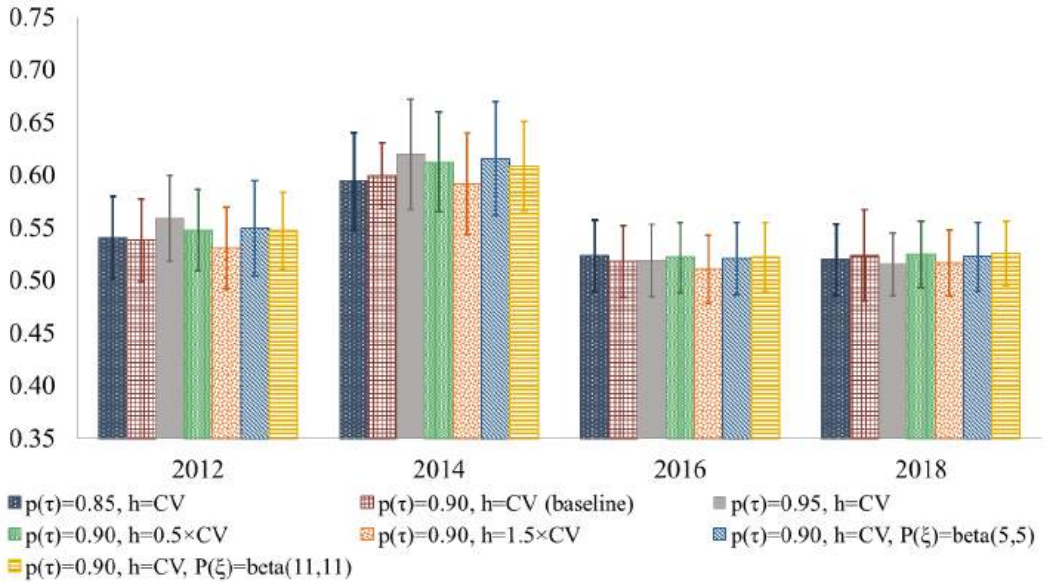
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I23: Jilin Gini estimates under the alternative parameter choices.



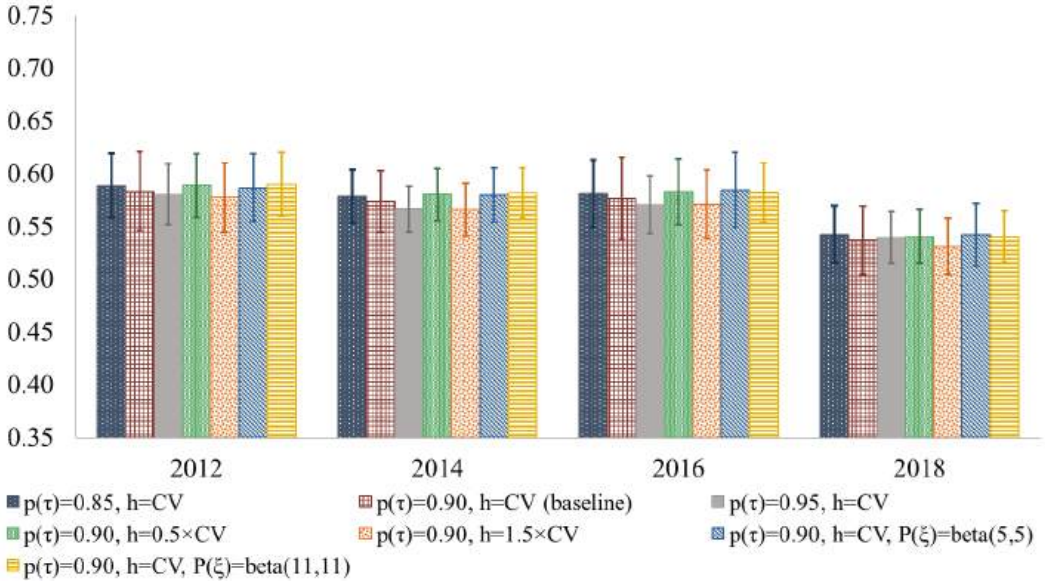
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I24: Jiangxi Gini estimates under the alternative parameter choices.



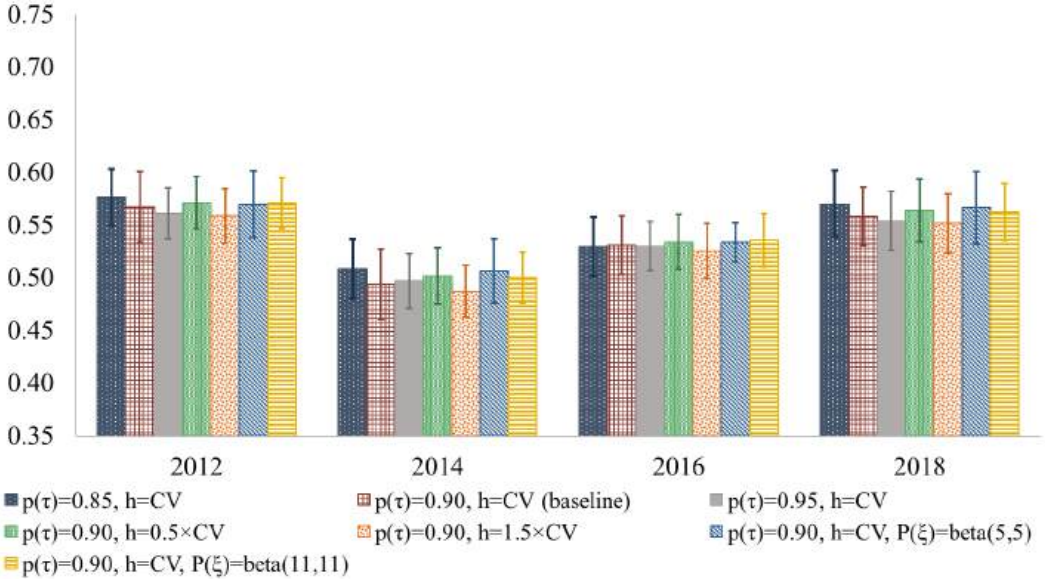
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I25: Hebei Gini estimates under the alternative parameter choices.



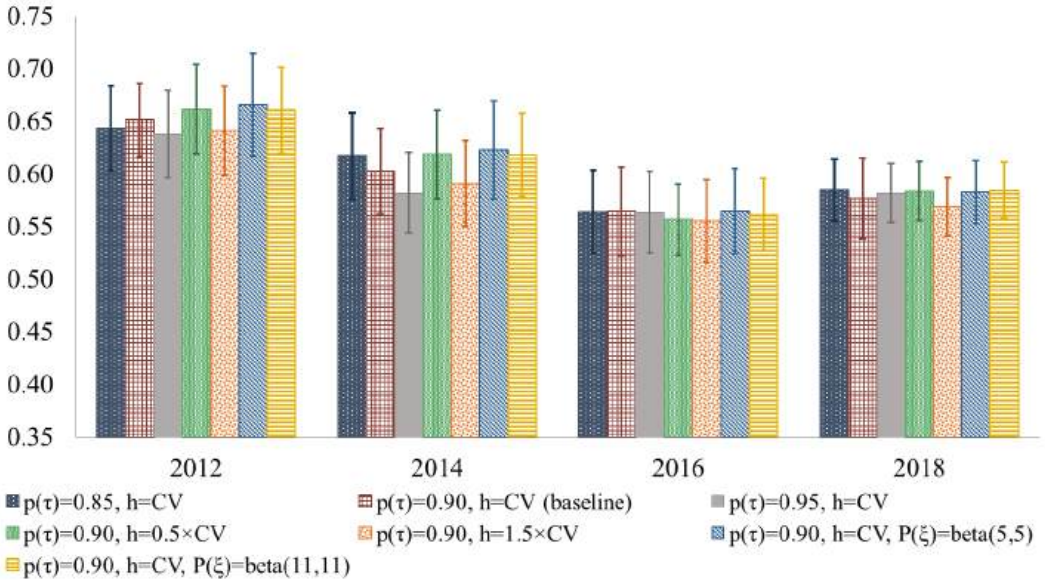
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I26: Shanxi Gini estimates under the alternative parameter choices.



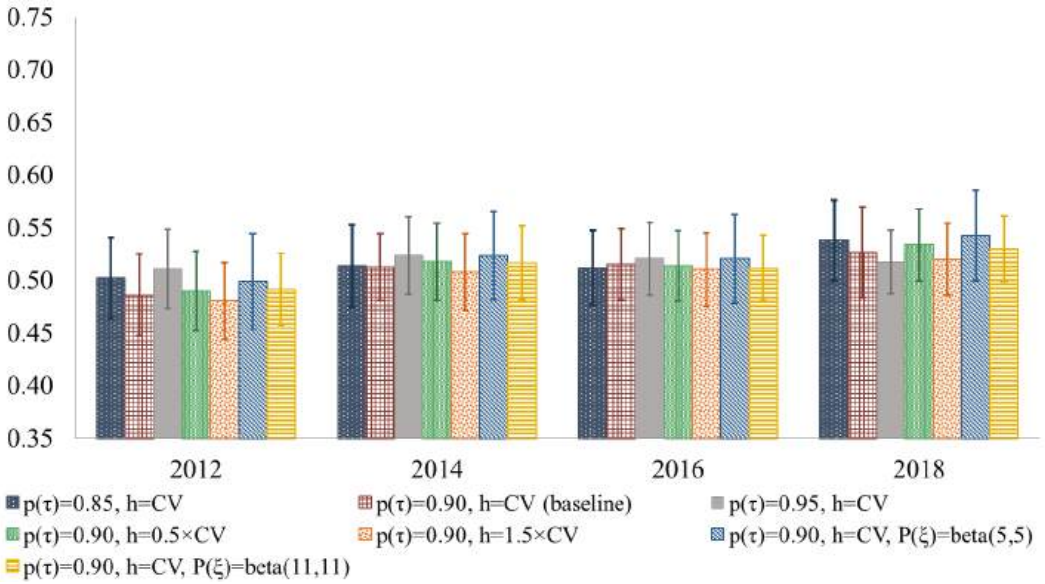
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I27: Qinghai Gini estimates under the alternative parameter choices.



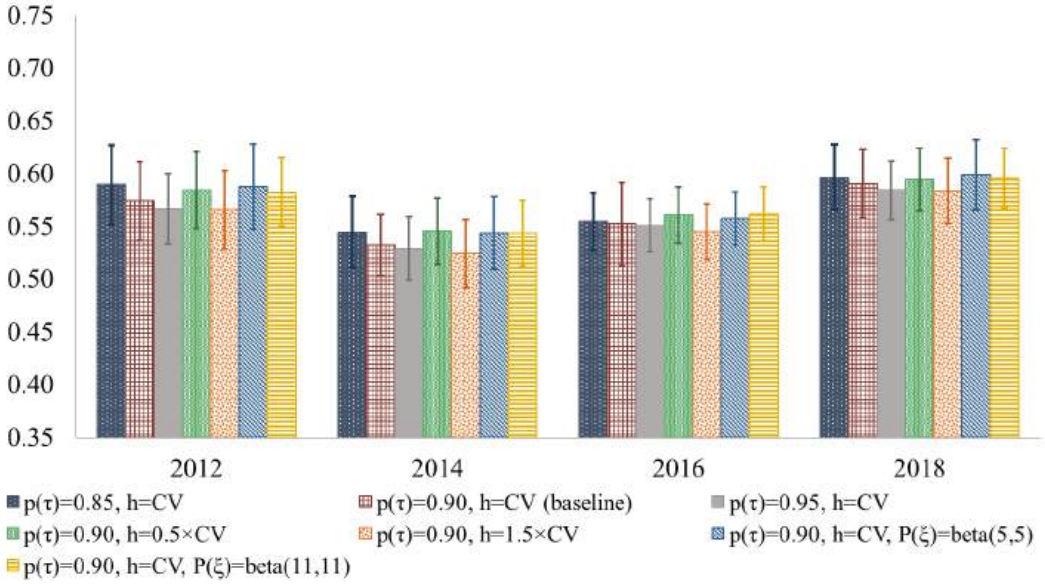
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I28: Heilongjiang Gini estimates under the alternative parameter choices.



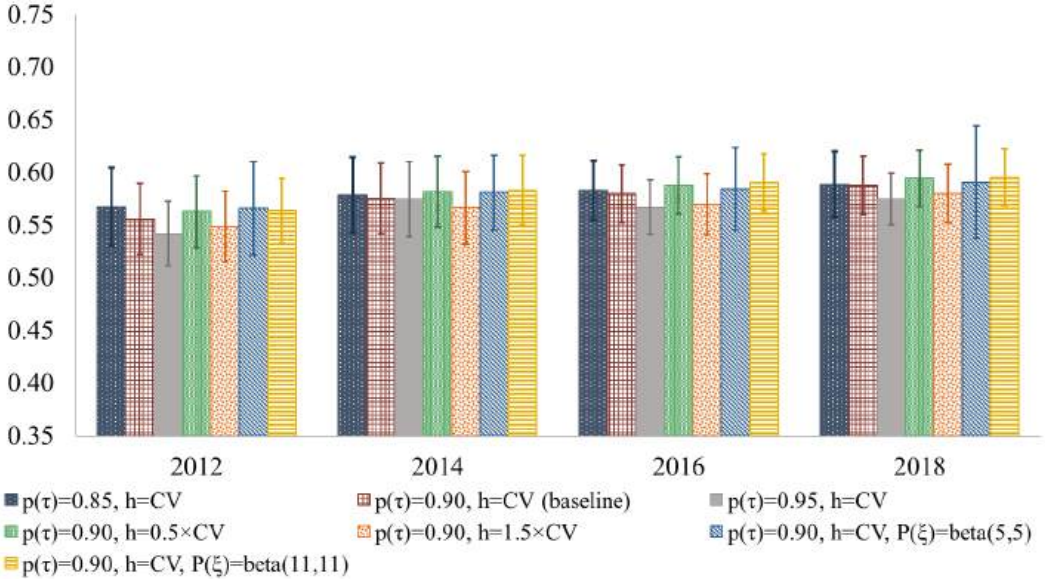
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I29: Yunnan Gini estimates under the alternative parameter choices.



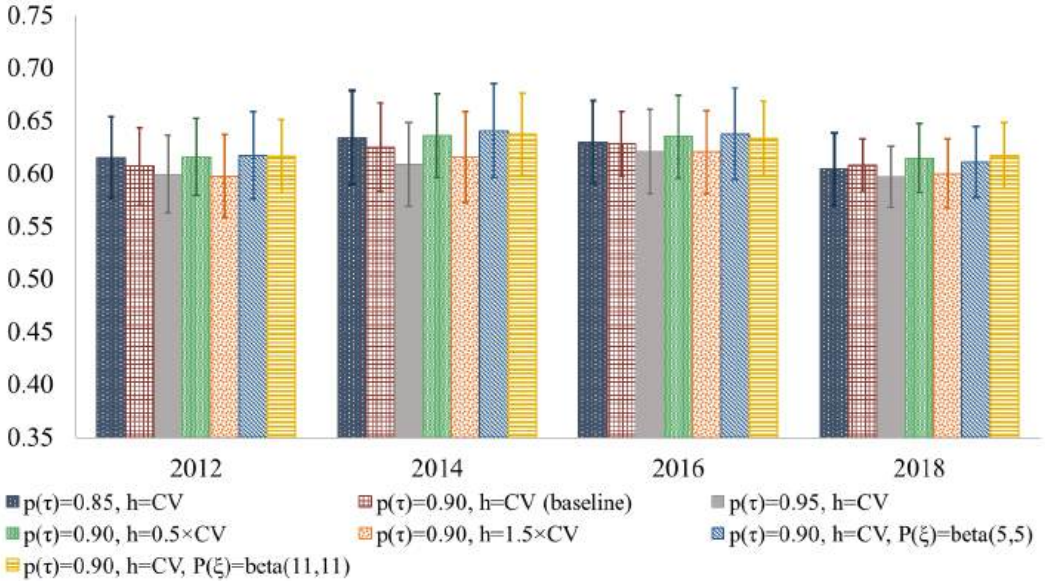
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I30: Guangxi Gini estimates under the alternative parameter choices.



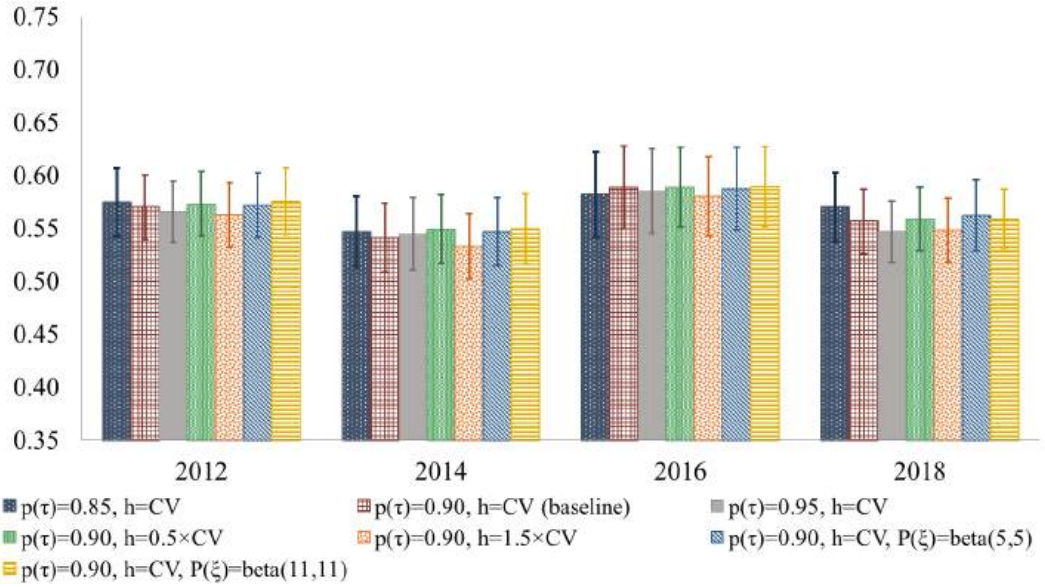
Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I31: Guizhou Gini estimates under the alternative parameter choices.



Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Figure I32: Gansu Gini estimates under the alternative parameter choices.



Notes: This figure presents the results of the robustness check on different chosen threshold percentiles ($p(\tau)$ =0.85, 0.90, and 0.95), bandwidths (h=0.5×CV, CV, and 1.5×CV, where CV is optimal bandwidth estimated using cross-validation), and alternative penalty functions ($P(\xi) = beta(5,5)$, or $beta(11,11)$). Error bars represent 95% semiparametric bootstrap confidence intervals computed from 1000 bootstrap samples.

Table I1: Regression estimates of Gini index determinants using educational Gini index calculated from CPESY. Heteroskedasticity-robust standard errors are used.

	Dependent variable: Gini index															
	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(OLS)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)	(FE)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Educational Gini CPESY	0.491***		0.429***	0.437***	0.413**		0.378***	0.370***	0.223		0.298	0.305	0.330		0.376	0.362
	(0.175)		(0.144)	(0.147)	(0.177)		(0.143)	(0.141)	(0.303)		(0.324)	(0.322)	(0.349)		(0.336)	(0.344)
Returns to schooling		0.796***	0.771***	0.779***		0.763***	0.752***	0.738***		0.360***	0.367***	0.355***		0.385***	0.389***	0.380***
		(0.111)	(0.110)	(0.113)		(0.107)	(0.102)	(0.099)		(0.119)	(0.118)	(0.135)		(0.114)	(0.110)	(0.124)
Educational Gini CPESY × Returns to schooling				−2.233				3.284				1.381				0.995
				(3.531)				(3.452)				(3.018)				(2.969)
Social spending CSY					−0.008	−0.010*	−0.012**	−0.013**					0.003	0.004	0.004	0.004
					(0.007)	(0.005)	(0.005)	(0.005)					(0.008)	(0.008)	(0.008)	(0.009)
SOE assets share CSY					−0.142***	−0.110***	−0.104***	−0.105***					−0.020	0.064	0.046	0.049
					(0.035)	(0.031)	(0.030)	(0.030)					(0.089)	(0.075)	(0.072)	(0.073)
Foreign trade share CSY					0.018	−0.030	−0.016	−0.014					−0.052	−0.041	−0.056	−0.053
					(0.031)	(0.025)	(0.026)	(0.026)					(0.051)	(0.047)	(0.052)	(0.054)
Ln(GRP per capita) CSY					−0.036*	−0.003	−0.001	−0.001					0.043	0.073	0.055	0.058
					(0.021)	(0.019)	(0.019)	(0.019)					(0.056)	(0.048)	(0.050)	(0.053)
2014.year	−0.001	−0.003	−0.005	−0.005	0.007	−0.001	−0.001	−0.001	0.00000	−0.001	−0.002	−0.002	−0.010	−0.015	−0.014	−0.014
	(0.012)	(0.011)	(0.011)	(0.011)	(0.012)	(0.011)	(0.011)	(0.010)	(0.007)	(0.006)	(0.006)	(0.006)	(0.011)	(0.010)	(0.010)	(0.010)
2016.year	−0.022*	−0.011	−0.017	−0.018	−0.003	−0.004	−0.008	−0.006	−0.018*	−0.013*	−0.017*	−0.017*	−0.040**	−0.041**	−0.043**	−0.043**
	(0.013)	(0.011)	(0.012)	(0.012)	(0.016)	(0.014)	(0.013)	(0.013)	(0.010)	(0.008)	(0.010)	(0.010)	(0.019)	(0.019)	(0.018)	(0.019)
2018.year	−0.020*	−0.007	−0.012	−0.012	0.005	0.003	0.002	0.004	−0.018**	−0.011*	−0.015*	−0.014*	−0.049*	−0.053*	−0.049*	−0.051*
	(0.012)	(0.011)	(0.011)	(0.011)	(0.020)	(0.017)	(0.016)	(0.016)	(0.008)	(0.006)	(0.008)	(0.008)	(0.029)	(0.030)	(0.029)	(0.031)
Constant	0.570***	0.565***	0.568***	0.568***	0.990***	0.645***	0.624***	0.631***								
	(0.008)	(0.007)	(0.007)	(0.007)	(0.215)	(0.198)	(0.199)	(0.200)								
Observations	116	116	116	116	116	116	116	116	116	116	116	116	116	116	116	116
R^2	0.093	0.265	0.317	0.319	0.315	0.473	0.504	0.508	0.118	0.200	0.207	0.209	0.146	0.234	0.242	0.243
Number of provinces									29	29	29	29	29	29	29	29

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

J Simulation study: additional results

Table J1: Theil T indices resulting from policy simulations equalizing educational structure and returns to schooling to Shanghai.

	Baseline Estimates 2018		(a) Uniform educational structure		(b) Uniform educational structure and adjusted RS		(c) Uniform RS		(d) Uniform educational structure and uniform RS	
	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)	Theil T index	Share (%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Beijing	0.391	4%	0.329	2%	0.342	3%	0.313	2%	0.283	1%
Shanghai	0.247	2%	0.247	2%	0.218	2%	0.267	4%	0.218	3%
Jiangsu	0.388	7%	0.340	6%	0.339	6%	0.325	4%	0.305	5%
Tianjin	0.269	1%	0.262	1%	0.265	1%	0.223	1%	0.215	1%
Zhejiang	0.395	5%	0.323	6%	0.316	6%	0.292	4%	0.284	5%
Fujian	0.491	3%	0.413	4%	0.392	3%	0.426	3%	0.370	3%
Guangdong	0.530	9%	0.395	8%	0.421	9%	0.390	4%	0.355	4%
Shandong	0.497	5%	0.415	6%	0.387	5%	0.429	6%	0.376	6%
Inner Mongolia	0.396	2%	0.340	2%	0.370	2%	0.300	1%	0.313	1%
Hubei	0.383	3%	0.377	4%	0.319	3%	0.362	5%	0.318	5%
Chongqing	0.495	2%	0.428	2%	0.448	3%	0.389	1%	0.386	1%
Shaanxi	0.485	2%	0.391	2%	0.407	2%	0.380	1%	0.353	1%
Liaoning	0.344	3%	0.321	3%	0.334	3%	0.295	2%	0.287	2%
Hunan	0.370	3%	0.362	4%	0.303	3%	0.365	6%	0.309	6%
Anhui	0.420	3%	0.354	4%	0.291	3%	0.405	6%	0.298	6%
Hainan	0.527	0%	0.434	0%	0.501	1%	0.400	0%	0.404	0%
Ningxia	0.380	0%	0.364	0%	0.320	0%	0.364	0%	0.310	0%
Henan	0.542	5%	0.443	6%	0.475	7%	0.404	3%	0.411	3%
Sichuan	0.600	7%	0.474	8%	0.508	9%	0.465	4%	0.430	4%
Jilin	0.379	1%	0.375	2%	0.328	1%	0.394	3%	0.334	2%
Jiangxi	0.374	2%	0.320	3%	0.294	2%	0.315	2%	0.274	2%
Hebei	0.424	3%	0.347	3%	0.308	2%	0.396	4%	0.306	4%
Shanxi	0.457	2%	0.385	2%	0.387	2%	0.372	1%	0.354	1%
Qinghai	0.515	0%	0.448	1%	0.431	1%	0.430	0%	0.407	0%
Heilongjiang	0.373	2%	0.372	3%	0.361	3%	0.313	2%	0.321	2%
Yunnan	0.509	3%	0.376	3%	0.401	3%	0.352	1%	0.342	1%
Guangxi	0.526	2%	0.457	2%	0.463	2%	0.462	1%	0.417	2%
Guizhou	0.565	2%	0.503	3%	0.490	3%	0.473	2%	0.455	2%
Gansu	0.473	1%	0.393	1%	0.448	2%	0.351	1%	0.354	1%
Within	0.436	87%	0.377	92%	0.377	89%	0.364	75%	0.327	76%
Between	0.063	13%	0.035	8%	0.044	11%	0.120	25%	0.101	24%
Total	0.499	100%	0.412	100%	0.422	100%	0.484	100%	0.428	100%

Notes: This table reports the results of the policy simulation study based on the data from the CHFS 2019. Monte Carlo Simulations were repeated 1000 times.