AIMS Mathematics, 11(11): 38301–38334.
DOI: 10.3934/math.20261479
Received: 31 July 2026
Revised: 02 October 2026
Accepted: 09 November 2026
Published: 17 November 2026

https://www.aimspress.com/journal/Math

Research article

Measuring the returns to education in China and their urban–rural disparities from the perspective of nonignorable missingness

Chao Dai[1], **Tingfang Luo**[1], **Hanji He**[1,2,*] **and Jierui Du**[1,2]

[1] College of Mathematics and Statistics, Guilin University of Technology, Guilin 541000, China

[2] Guangxi Universities Key Laboratory of Applied Statistics, Guilin University of Technology, Guilin 541004, China

* **Correspondence:** Email: hhjtjedu@163.com.

Abstract: Equality of opportunity is an important pathway toward achieving common prosperity, and the return to education is a key indicator from the perspective of income distribution equity. Estimates of the return to education are mainly based on micro-level survey data, but such estimates may be biased due to missing data. Using data from the China Family Panel Studies from 2018 to 2022, this study estimated the returns to education in China under both the missing at random (MAR) and parametric not missing at random (NMAR) settings, employing fully parametric and semiparametric models, respectively. It further examined the effect of reclassifying migrant samples on the measurement of urban–rural differences. The results showed that: (1) Under the parametric NMAR specification adopted in this study, the estimated return to education for the national sample is lower than that obtained under the corresponding MAR specification, indicating that the measurement of returns to education is relatively sensitive to assumptions regarding the income missingness mechanism. (2) Across different missingness mechanisms, model specifications, and sample classifications, the estimated return to education is consistently higher in urban areas than in rural areas, while the rural estimates are more sensitive to the assumed missingness mechanism. (3) After treating the work-experience term nonparametrically, the estimated return to education remains higher in urban areas than in rural areas, suggesting that the main conclusion is robust to the functional-form specification of work experience. (4) Among the employed sample and under the reclassification scheme adopted in this study, reassigning employed individuals who currently reside in urban areas but hold rural hukou status to the rural sample further widens the estimated urban–rural gap in returns to education. Further comparisons showed that both assumptions about the missingness mechanism and the classification of migrant populations affect the measurement of urban–rural differences in returns to education. These findings highlight the importance of carefully accounting for uncertainty arising from income missingness mechanisms and sample-classification rules when estimating returns to education using micro-level survey data.

Keywords: returns to education; nonrandom missingness; semiparametric model; income distribution
Mathematics Subject Classification: 62D10, 62G08

1. Introduction

Education is an important channel for human capital accumulation, while the rate of return to education is a key indicator of the economic benefits derived from educational investment. It reflects how the labor market values human capital and influences individuals' decisions regarding educational investment and labor market participation. Identifying the returns to education and their heterogeneity across population groups not only contributes to understanding the formation of individual income disparities in the context of educational expansion, but also provides an important basis for optimizing the allocation of educational resources and improving income distribution. In particular, given that China's urban–rural dual structure has not yet been fully dismantled, urban–rural differences in the returns to education reflect not only disparities in the rewards to human capital across different groups, but also the persistent urban–rural income gap and labor market segmentation. From the perspective of income distribution, urban–rural disparities in the returns to education constitute an important dimension for understanding income inequality in China [1]. However, estimates of the returns to education typically rely on income data obtained from micro-level surveys, and may therefore be affected by income nonresponse, the classification of urban and rural samples, and model specification. Consequently, further research is needed to accurately estimate the returns to education and their urban–rural differences while adequately accounting for the characteristics of survey data.

Research on the returns to education is grounded in human capital theory. The earnings function proposed by Mincer [2] links individual income to years of schooling and work experience. Since then, the coefficient on years of schooling has gradually become a widely used empirical measure of the economic return to education in both the economics of education and labor economics. Existing studies generally show that education generates significant economic returns, although the magnitude of these returns varies across stages of economic development, labor market conditions, educational levels, and population groups [3]. Studies on China have found that the returns to education have generally increased with the advancement of market-oriented reforms and adjustments in the labor market [4]. This trend has also been closely associated with changes in the structure of the labor force and the evolution of income inequality [5]. Based on a meta-analysis of a large body of existing research, Ma and Iwasaki [6] showed that the returns to education in China increased overall during the process of economic transition, while exhibiting substantial heterogeneity across regions, hukou status, employment sectors, and gender. Against this background, urban–rural disparities in the returns to education have increasingly attracted scholarly attention. Gao and Li [7] examined the long-term evolution of the returns to education in urban and rural China, and found that the returns were generally higher in urban areas than in rural areas, with the two exhibiting different trajectories over time. Although the expansion of higher education has raised the overall level of educational attainment, it has not completely eliminated disparities in educational opportunities and economic returns between urban and rural residents or across hukou groups [8]. Using data from the 2012 and 2021 waves of the Chinese General Social Survey (CGSS), Hu et al. [9] found that the returns

to higher education remained relatively stable overall. However, the returns declined among urban residents but increased among rural residents, resulting in a substantial narrowing of the urban–rural gap. Meanwhile, the hukou system and the resulting segmentation of the labor market may further shape urban–rural income disparities and differences in human capital returns by affecting access to education, employment opportunities, and labor mobility [10]. These studies provide an important foundation for understanding the returns to education and their urban–rural heterogeneity in China. However, the existing literature has mainly focused on the magnitude of the returns to education and their variation across population groups, while paying relatively limited attention to the potential selection bias arising from income nonresponse in micro-level surveys.

In fact, household- and individual-level micro-surveys, on which research on the returns to education commonly relies, are often affected by missing data. Wages and income are particularly sensitive survey items and are therefore more prone to refusal, item nonresponse, and reporting errors. Using matched survey and administrative records, Bollinger et al. [11] found a systematic relationship between income nonresponse and the true income distribution, with this relationship being especially pronounced in the tails of the distribution. Income missingness, therefore, may not be completely random, and failure to account for the missingness mechanism may lead to biased estimates of both the income distribution and related regression parameters. Rubin [12] established the fundamental classification framework for modern missing-data mechanisms, distinguishing among missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Under the MAR framework, existing research has extended to a variety of models and statistical tasks, including weighted imputation and empirical likelihood inference with randomly missing responses [13], model averaging based on weighted generalized moments [14], and ultrahigh-dimensional group feature screening with randomly missing covariates [15]. However, when the probability of missingness depends on the unobserved outcome variable, the MAR assumption is no longer valid. In such cases, model identification and statistical inference must be further addressed under an NMAR mechanism.

For NMAR data, existing studies have developed a relatively systematic body of research on model identification, response probability modeling, and semiparametric estimation. Kim and Yu [16] investigated semiparametric estimation of mean functionals under nonignorable missingness. Zhao and Shao [17] further developed a semiparametric pseudo-likelihood approach for generalized linear models by incorporating instrumental variables. Subsequently, semiparametric inverse probability weighting methods [18] and studies on the identifiability of models with nonignorable missingness [19] further enriched the framework for statistical inference under NMAR mechanisms. In recent years, related research has gradually been extended to more flexible model specifications, including generalized additive partial linear models with nonignorable missing responses [20] and varying-coefficient partially nonlinear models [21]. Cheng et al. [22] combined a logistic response probability model with a semiparametric proportional likelihood ratio model, exploiting marginal information on covariates to improve identifiability and estimation efficiency under NMAR, and developed a semiparametrically efficient estimation procedure. Shetty et al. [23], without directly specifying an outcome regression model, proposed a robust estimation method under a semiparametric response probability model for NMAR data, further extending the toolkit for semiparametric inference under NMAR. Overall, the existing literature has developed a broad range of statistical inference methods for missing data. However, their application to research on the returns to education remains relatively

limited. In particular, systematic comparisons of how alternative missingness mechanism assumptions affect estimates of the returns to education and their urban–rural disparities are still lacking.

In addition to the income missingness mechanism, the definition of urban and rural samples may also affect comparisons of the returns to education. Zhao [24] examined the rural–urban migration decisions of China's rural labor force and showed that rural labor migration is characterized by pronounced selectivity. Meng and Zhang [25] further investigated the labor market outcomes of rural migrants after entering urban areas and found that substantial disparities remained between rural migrants and urban residents in terms of employment opportunities, occupational attainment, and wage levels. Yao et al. [26] further examined the returns to education among rural migrant workers in China from the perspective of the institutional environment. Gao and Li [27] redefined the urban and rural samples from the perspective of hukou conversion and found that reclassifying individuals who had changed from agricultural to non-agricultural hukou according to their rural origin significantly increased the estimated returns to education in rural areas. This finding suggests that sample classification itself may affect the measurement of urban–rural differences in returns to education. Wu and Amorim [28] further found that return migration and hukou status jointly influence household educational investment, indicating that migration processes and the hukou system may continuously reshape the human capital composition of urban and rural populations. These findings suggest that the criteria used to classify urban and rural status may directly affect the composition of urban and rural samples. When individuals are classified solely according to their current place of residence, some migrant workers with rural hukou who are employed in urban areas may be included in the urban sample. Such classification may alter the human capital structure and income distribution of both urban and rural samples, thereby affecting the estimated returns to education and the measured urban–rural gap in those returns. However, the existing literature has generally examined migration, labor market outcomes, and educational returns separately, while paying relatively limited attention to urban–rural disparities in the returns to education under the joint effects of migrant-sample reclassification and income missingness mechanisms.

Building on the above studies, the existing literature has provided extensive discussions of the returns to education and their urban–rural heterogeneity, statistical inference for missing data, and population migration. However, the measurement of the returns to education relies primarily on income data from micro-level surveys, and income nonresponse may be selective. Consequently, alternative assumptions regarding the missingness mechanism may affect estimates of both the overall returns to education and their urban–rural differences. At the same time, population migration may alter the composition of urban and rural samples, while different criteria for classifying urban and rural populations may further affect comparisons of the returns to education between the two groups. In addition, the conventional Mincer earnings equation generally requires the functional forms of variables such as work experience to be specified in advance, meaning that the resulting estimates may be sensitive to particular model specifications. It is therefore necessary to incorporate the income missingness mechanism, urban–rural sample classification, and model specification into a unified analytical framework, and to conduct a more systematic investigation of the returns to education and their urban–rural disparities in China. Using data from the 2018, 2020, and 2022 waves of the China Family Panel Studies (CFPS), this study investigates the returns to education and their urban–rural differences under alternative missingness mechanisms and sample classification criteria. First, the returns to education are estimated under both ignorable and nonignorable missingness mechanisms,

and the effects of alternative missingness assumptions on the estimated returns for the full sample and for urban and rural subsamples are compared. Second, migrant observations are reclassified to examine how the estimated returns to education and the corresponding urban–rural disparities change under different urban–rural classification criteria. Third, a semiparametric model is introduced alongside the parametric model, allowing the effect of work experience to be modeled nonparametrically so as to reduce the influence of restrictive functional-form assumptions on the estimation results. Finally, urban–rural disparities in the returns to education are further examined across different regions to assess regional heterogeneity in educational returns. Through these analyses, this study integrates nonrandom income missingness, migration-induced reconstruction of urban and rural samples, and differences in model specification into a unified analytical framework. This study provides a new statistical perspective and empirical evidence for examining returns to education in China and their urban–rural differences under alternative assumptions about the missingness mechanism.

The remainder of this paper is organized as follows. Section 2 introduces the semiparametric specification and estimation under NMAR. Section 3 estimates the returns to education in China. Section 4 analyzes urban–rural disparities in the returns to education. Section 5 presents the main conclusions and policy implications.

2. Development of a semiparametric specification and estimation under NMAR

To examine the impact of assumptions regarding the income missingness mechanism on the estimated returns to education, this study specifies corresponding semiparametric models under both ignorable missingness and parametric nonignorable missingness frameworks. Let $Y_{it}^{(h)}$ denote the logarithm of labor income of individual i in year t, $X_{it}^{(h)}$ denote the fully observed covariates and let $R_{it}^{(h)}$ denote the income observation indicator. If $P(R_{it}^{(h)} = 1|Y_{it}^{(h)}, X_{it}^{(h)}) = P(R_{it}^{(h)} = 1)$, the missingness mechanism is missing completely at random (MCAR). If $P(R_{it}^{(h)} = 1|Y_{it}^{(h)}, X_{it}^{(h)}) = P(R_{it}^{(h)} = 1|X_{it}^{(h)})$, the missingness mechanism is missing at random (MAR). If, conditional on $X_{it}^{(h)}$, the probability that income is observed still depends on the latent income $y_{it}^{(h)}$, the missingness mechanism is nonignorable (NMAR), where $R_{it}^{(h)} = 1$ if income is observed and $R_{it}^{(h)} = 0$ otherwise. Here, $h \in \{u, r\}$ indexes the urban and rural subsamples, with u denoting the urban sample and r the rural sample.

2.1. Semiparametric outcome model and the nonignorable missingness mechanism

Under the semiparametric specification, years of schooling and year effects enter the income equation linearly, whereas work experience is modeled through an unknown smooth function. For each urban–rural subsample, the outcome equation is specified as follows:

$$Y_{it}^{(h)} = \beta_0^{(h)} + \beta_1^{(h)} eduy_{it} + \beta_2^{(h)} D_{2020,t} + \beta_3^{(h)} D_{2022,t} + s_h(exper_{it}) + \varepsilon_{it}^{(h)}, \tag{2.1}$$

where $eduy_{it}$ denotes years of schooling, $D_{2020,t}$ and $D_{2022,t}$ are year dummy variables, $exper_{it}$ denotes work experience, $s_h(\cdot)$ is an unknown smooth function, $\varepsilon_{it}^{(h)} \sim N(0, \sigma_g^2)$, and $\beta_1^{(h)}$ represents the change in log income associated with an additional year of schooling.

Under the nonignorable missingness mechanism, the probability that income is observed depends not only on observed covariates but may also depend on the underlying income level itself. The income observation probability is specified as follows:

$$\Pr\left(R_{it}=1 \mid Y_{it}^{(h)}, eduy_{it}, D_{2020,t}, D_{2022,t}, exper_{it}\right) = \frac{\exp\left(\eta_0+\alpha^{(h)}Y_{it}^{(h)}+\eta_1 eduy_{it}+\eta_2 D_{2020,t}+\eta_3 D_{2022,t}+\eta_4 exper_{it}\right)}{1+\exp\left(\eta_0+\alpha^{(h)}Y_{it}^{(h)}+\eta_1 eduy_{it}+\eta_2 D_{2020,t}+\eta_3 D_{2022,t}+\eta_4 exper_{it}\right)}, \tag{2.2}$$

where $\eta = (\eta_0, \eta_1, \eta_2, \eta_3, \eta_4)^{'}$ denotes the parameter vector in the missingness mechanism, with $\alpha^{(h)}$ characterizing the effect of latent income on the probability of income being observed. When $\alpha^{(h)} = 0$, the probability of income being observed no longer depends directly on latent income, and the missingness mechanism is therefore ignorable. When $\alpha^{(h)} \neq 0$, the missingness mechanism is nonignorable.

2.2. Identification conditions and interpretation boundaries under NMAR

Unlike under MAR, the probability of observing income under the NMAR model depends directly on the latent income $Y_{it}^{(h)}$. For individuals with $R_{it}^{(h)} = 0$, however, the true income is unobserved. Estimation of the model therefore requires explicit identification conditions. This study does not introduce external instrumental variables, exclusion restrictions, or other auxiliary information. Identification of the NMAR model is instead mainly derived from the joint restrictions imposed by the income outcome equation and the response mechanism. Specifically, the Gaussian outcome distribution, the logistic response mechanism, and the associated parameter restrictions jointly provide the information required for identification. The model can be viewed as a specific application, in the context of the income data considered in this study, of the generalized additive partial linear model with nonignorable response missingness studied by Du et al. [29].

Although the CFPS has a longitudinal survey structure, both the outcome equation and the response equation in this study are specified at the individual-year level. The response mechanism depends on contemporaneous latent income and contemporaneous covariates, without incorporating dynamic variables such as lagged income or historical income-missingness indicators. In addition, the main model parameters are assumed to remain constant across survey years. Accordingly, identification of these common parameters can be reduced to the identification problem associated with the corresponding individual-year cross-sectional selection model.

For a continuous outcome combined with a logistic missingness mechanism, Du et al. [29] showed that identification can be established when the direction of the effect of the latent outcome in the response mechanism is known. Building on this identification condition, and in light of empirical evidence on income nonresponse among high-income households, this study assumes that the true nonignorability parameter satisfies $\alpha^{(h)} < 0$; that is, the higher the latent income, the lower the probability that income is observed. Flachaire et al. [30] found that high-income households are relatively more difficult to contact in surveys and are more likely to refuse to report their income or to underreport income at the upper end of the income distribution, providing empirical support for this directional assumption.

It should be emphasized that $\alpha^{(h)} < 0$ restricts only the direction of the nonignorability effect rather than fixing its numerical value a priori. In the model considered in this study, the magnitude of $\alpha^{(h)}$ is estimated jointly with the other model parameters through the iterative MCEM procedure and estimating equations. Thus, $\alpha^{(h)}$ is an unknown parameter to be estimated rather than an externally specified sensitivity parameter.

Accordingly, the NMAR estimates reported in this study should be interpreted as model-dependent results obtained under the assumptions of a Gaussian outcome distribution, a logistic response mechanism, and the directional restriction on the nonignorability parameter. This study does not attempt to establish, solely on the basis of the observed data, that the true income-missingness mechanism must be NMAR. Instead, returns to education and their urban–rural differences are examined under both MAR and NMAR specifications in order to assess their sensitivity to alternative assumptions about the income-missingness mechanism. Because no external instrumental variables or exclusion restrictions are introduced, the NMAR results remain dependent on the aforementioned distributional, functional-form, and directional assumptions. These assumptions therefore constitute an important boundary for the interpretation of the empirical findings.

In addition, to examine the local numerical behavior of the theoretical identification conditions in the empirical sample, this study further evaluates the Jacobian matrix of the estimating equations for the response mechanism in a neighborhood of the baseline NMAR estimates. This diagnostic is intended primarily to assess the local degree of dependence between the latent-income coefficient $\alpha^{(h)}$ and the other response-mechanism parameters, rather than to serve as an independent formal test of identification. The results show that, even after standardizing the parameter scales, the Jacobian matrix remains moderately ill-conditioned, indicating relatively strong local dependence between $\alpha^{(h)}$ and the other response-mechanism parameters. Accordingly, the parameter should not be interpreted as being strongly empirically identified by the observed data alone. In light of this finding, the interpretation of the NMAR parameter is kept deliberately cautious. The stability of the main empirical conclusions is further assessed by jointly considering the repeated-sampling uncertainty reflected in the cluster bootstrap based on the personal identification number (PID) and sensitivity analyses under alternative response link functions.

2.3. Observed-data likelihood and MCEM estimation

Based on the above semiparametric outcome equation and nonignorable missingness mechanism, we first define the conditional mean of income for individual i in year t to construct the observed-data likelihood function as follows:

$$\mu_{it}^{(h)} = \beta_0^{(h)} + \beta_1^{(h)} eduy_{it} + \beta_2^{(h)} D_{2020,t} + \beta_3^{(h)} D_{2022,t} + s_h(exper_{it}). \tag{2.3}$$

Let $X_{it}^{(h)} = (eduy_{it}, exper_{it}, D_{2020,t}, D_{2022,t})$ denote the covariate set in the outcome model, and $Z_{it}^{(h)}$ denote the covariate set in the missingness mechanism model. Let

$$\theta^{(h)} = \left(\beta^{(h)}, \sigma_h^2, \eta^{(h)}, \alpha^{(h)}\right) \tag{2.4}$$

denote the set of model parameters, further defining the response probability function:

$$\pi^{(h)}(y, z) = P\left(R_{it}^{(h)} = 1 \mid Y_{it}^{(h)} = y,\ Z_{it}^{(h)} = z\right),$$

that is, the conditional probability that income is observed in the urban–rural subsample h, given a latent income value y and response-mechanism covariates z. For the specific observation of individual i in period t, let

$$\pi_{it}^{(h)} = \pi^{(h)}\left(Y_{it}^{(h)}, Z_{it}^{(h)}\right).$$

Therefore, $\pi^{(h)}(y,z)$ denotes the response probability function itself, whereas $\pi_{it}^{(h)}$ denotes the value of this function for a specific individual-year observation and $f(\cdot)$ denotes the normal density function corresponding to the outcome equation. The observed-data likelihood function can then be expressed as follows:

$$L\left(\theta^{(h)}\right) = \prod_{i,t} \left[\pi_{it}^{(h)} f\left(y_{it}^{(h)} \mid X_{it}^{(h)}; \beta^{(h)}, \sigma_h^2\right)\right]^{R_{it}^{(h)}} \times \left[\int \left\{1 - \pi_{it}^{(h)}\left(u, Z_{it}^{(h)}\right)\right\} f\left(u \mid X_{it}^{(h)}; \beta^{(h)}, \sigma_h^2\right) du\right]^{1-R_{it}^{(h)}}. \tag{2.5}$$

For observations with observed income, the likelihood contribution is jointly determined by the income density and the probability of income being observed. For observations with missing income, the likelihood contribution is obtained by integrating over the unobserved income.

Since the likelihood function above involves integration over the missing income values, a closed-form solution is generally difficult to obtain directly. Therefore, this study employs the Monte Carlo EM (MCEM) algorithm for numerical estimation. The basic idea is that, at each iteration, Monte Carlo candidate values for the missing income are first generated from the current outcome model. These candidate values are then reweighted according to the current response mechanism, taking into account the fact that the individual is actually in the income-missing state. The resulting weighted samples are used to approximate the first- and second-order conditional moments required for the missing income, after which the parameters of the outcome model and the response mechanism are updated sequentially.

Let k denote the kth outer iteration of the MCEM algorithm. Given the outcome-model parameters and missingness-mechanism parameters obtained from the $(k-1)$th iteration, M Monte Carlo candidate income values are generated for each observation with missing income from the conditional distribution implied by the current outcome model:

$$Y_{it}^{(m,k)} \sim N\left(X_{it}^{(h)\prime}, \beta^{(h,k-1)}, \sigma_h^{2(k-1)}\right), \qquad m = 1, \ldots, M.$$

The candidate values above are generated solely from the current outcome model and do not yet incorporate the information that the individual actually satisfies $R_{it}^{(h)} = 0$. Therefore, importance weights are constructed according to the current missingness mechanism as follows:

$$\omega_{it}^{(m,k)} = 1 - \pi_{it}^{(h)}\left(Y_{it}^{(m,k)}, Z_{it}^{(h)}; \gamma^{(h,k-1)}, \alpha^{(h,k-1)}\right). \tag{2.6}$$

Here, the weights reflect the degree of compatibility between each candidate income value and the fact that the individual's actual income is unobserved. The higher the missingness probability associated with a candidate income value, the greater the weight assigned to that value in the conditional distribution of the missing income. After normalizing the weights, the first conditional moment of the missing income can be approximated by the weighted Monte Carlo average:

$$\widetilde{Y}_{it}^{(h,k)} = R_{it}^{(h)} Y_{it}^{(h)} + (1 - R_{it}^{(h)}) \frac{\sum_{m=1}^{M} \omega_{it}^{(m,k)} Y_{it}^{(m,k)}}{\sum_{m=1}^{M} \omega_{it}^{(m,k)}}. \tag{2.7}$$

When $R_{it}^{(h)} = 1$, Eq (2.7) is simply equal to the observed income. When $R_{it}^{(h)} = 0$, it is approximated by the weighted mean of the Monte Carlo candidate income values, which serves as an approximation to the corresponding conditional expectation. Since updating the error variance in the Gaussian outcome model also requires the conditional moment of squared income, the second conditional moment is further computed as follows:

$$\widetilde{Y}_{it}^{2(h,k)} = R_{it}^{(h)} \left(Y_{it}^{(h)}\right)^2 + (1 - R_{it}^{(h)}) \frac{\sum_{m=1}^{M} \omega_{it}^{(m,k)} \left(Y_{it}^{(m,k)}\right)^2}{\sum_{m=1}^{M} \omega_{it}^{(m,k)}}. \tag{2.8}$$

In the M-step, the outcome equation is first updated using the first- and second-order conditional moments described above. For the fully parametric model, these moments are used to update the regression coefficients and error variance in the Mincer earnings equation. For the semiparametric model, a Gaussian generalized additive model (GAM) is refitted using the completed conditional income responses, thereby updating the coefficient on years of schooling, the year effects, the error variance, and the smooth function of work experience.

$$\left(\widehat{\beta}^{(h)}, \widehat{s_h}\right) = \arg\min_{\beta,g} \sum_{i,t} \left[\widetilde{Y}_{it}^{(h,s)} - \beta_0^{(h)} - \beta_1^{(h)} edu y_{it} - \beta_2^{(h)} D_{2020,t} - \beta_3^{(h)} D_{2022,t} - s_h(exper_{it})\right]^2. \tag{2.9}$$

After updating the outcome equation, the response-mechanism parameters are updated using the Newton method based on the estimating equations obtained from the current Monte Carlo approximation. The updated outcome equation and response-mechanism parameters are then used to initiate the next round of Monte Carlo sampling and parameter updating. This procedure is repeated until the change in parameter estimates between two successive iterations satisfies a prespecified stopping criterion or the maximum number of iterations is reached.

In the numerical implementation, the initial conditional means for the MCEM algorithm are first constructed using observations with observed income. For the fully parametric model, the fitted values from an ordinary least squares regression based on the complete income observations are used. For the semiparametric model, the fitted values from a Gaussian GAM estimated using the complete income observations are employed, with the work-experience term initially specified as $s_h(exper, k = 6)$. These fitted values are used to initialize the conditional mean of the candidate distribution for the missing income values. At the same time, the starting values for the MCEM parameter iterations are set as follows: The regression coefficients in the outcome equation are initialized at 0, the error variance at 1, the coefficient on latent income in the response mechanism at -0.2, and all remaining response-mechanism parameters at 0.

The MCEM parameter iterations are initialized from a single set of starting values. In each Monte Carlo step, $M = 500$ candidate income values are generated for each individual with missing income. The parameters of the response mechanism are updated using the Newton method. To prevent excessively large Newton steps, when the largest absolute component of the update vector exceeds 5, the entire update vector is rescaled proportionally so that its largest absolute component does not exceed 5. For the response and missingness probabilities involved in the MCEM algorithm,

numerical calibration is applied whenever the corresponding probability falls below $\varepsilon = 0.01$, thereby avoiding numerical instability caused by probabilities approaching the boundary too closely. The stopping criterion is based on the maximum absolute change in the outcome-equation parameters and response-mechanism parameters between two successive iterations. The algorithm is terminated when this change does not exceed 0.05, with the maximum number of iterations set to 10. In addition, the Monte Carlo sample size is varied and the estimation procedure is repeated to assess the numerical stability with respect to the choice of M; the corresponding results are reported in Appendix A.

2.4. Ignorable missingness mechanism and the fully parametric model

When $\alpha^{(h)} = 0$, the probability of income being observed does not depend on latent income, corresponding to an ignorable missingness mechanism. The probability of income being observed can then be estimated using a logistic regression model, and the inverse probability weights are constructed as follows:

$$\omega_{it}^{(h)} = \frac{1}{\widehat{\pi}_{it}^{(h)}}. \tag{2.10}$$

For observations with observed income, the corresponding weighted semiparametric estimator can be expressed as follows:

$$\left(\widehat{\beta}^{(h)}, \widehat{s_h}\right) = \arg\min_{\beta,g} \sum_{R_{it}^{(h)}=1} \omega_{it}^{(h)} \left[Y_{it}^{(h)} - \beta_0^{(h)} - \beta_1^{(h)} eduy_{it} - \beta_2^{(h)} D_{2020,t} - \beta_3^{(h)} D_{2022,t} - s_h(exper_{it})\right]^2. \tag{2.11}$$

When $s_h(exper_{it}) = \delta_1^{(h)} exper_{it} + \delta_2^{(h)} exper_{it}^2$, the semiparametric outcome equation reduces to the fully parametric Mincer earnings equation. Accordingly, the semiparametric nonignorable missingness model is adopted as the general formulation, while the fully parametric model and the ignorable missingness model are treated as special cases.

2.5. Statistical inference and the treatment of repeated observations

Because the CFPS is a longitudinal survey, the same individual may appear repeatedly across multiple survey waves, including 2018, 2020, and 2022. Consequently, individual-year observations are not fully independent of one another. Treating all individual-year observations as mutually independent may therefore fail to adequately account for within-individual correlation across survey waves.

To address this issue, this study implements a cluster bootstrap using individual PID as the clustering unit. In each bootstrap replication, individuals rather than individual-year observations are sampled. When an individual is selected, all observations for that individual across the survey waves are included jointly in the corresponding bootstrap sample, and both the response mechanism and the outcome equation are re-estimated using that sample. Specifically, each bootstrap replication fully repeats the estimation of the response mechanism, the MCEM iterations, and the GAM fitting procedure in the semiparametric model. The resulting bootstrap standard errors therefore reflect uncertainty arising from both repeated sampling and re-estimation of the corresponding models. This procedure preserves the within-individual dependence structure across survey waves during resampling.

A total of $B = 500$ PID-cluster bootstrap replications are conducted, and the standard deviation of the estimated returns to education across replications is used as the PID-cluster bootstrap standard error. The standard errors reported in parentheses in the main result tables below are all calculated using this procedure. It should be noted that the cluster bootstrap is used primarily to quantify uncertainty in the estimated returns to education, whereas the point estimates reported in the main text are obtained from the corresponding models fitted to the original sample.

3. Measurement of returns to education in China

3.1. Data sources and variable definitions

This study uses three waves of data from the China Family Panel Studies (CFPS) conducted in 2018, 2020, and 2022. Individuals are matched across waves using their personal identification numbers (PIDs) to construct an individual-level unbalanced panel dataset. In accordance with the research objectives, the analysis focuses primarily on employed individuals and involves the data processing and cleaning of key variables, including labor income, years of schooling, work experience, urban–rural status, and household registration status. On this basis, the sample is further divided into an original sample and an adjusted sample to examine how the reclassification of migrant workers affects the estimation of urban–rural differences in returns to education. The urban–rural classifications described above are determined separately for each survey year based on the contemporaneous urban–rural status of the household residence and the individual's hukou status. Accordingly, the classification of the same individual may change across survey waves as their place of residence or hukou status changes. The CFPS is conducted by the Institute of Social Science Survey at Peking University. Its baseline sample covers 25 provinces, municipalities, and autonomous regions across China and is broadly nationally representative.

With respect to variable processing, the analysis mainly retains information from the individual questionnaire, including age, gender, annual income, household registration status, years of schooling, and the urban–rural status of household residence. The original sample is processed as follows:

(1) Observations with missing information on age or years of schooling are excluded. The sample is restricted to the working-age population, and individuals who may still be enrolled in school are removed.

(2) Observations with missing, inapplicable, or no household registration information, as well as those with missing urban–rural status, are excluded. Structurally unobserved cases are also removed according to income observation status.

After sample processing, the final sample contains 22,952 observations, of which 5.1% have unobserved annual income.

For sample classification, this study constructs two alternative urban–rural classification schemes: the original classification and an adjusted classification. Under the original classification, individuals are assigned to the urban or rural sample according to the urban–rural status of their household residence recorded in the CFPS survey: Individuals currently residing in urban areas are classified as urban, whereas those currently residing in rural areas are classified as rural. The adjusted classification further incorporates hukou status. Individuals currently residing in rural areas remain in the rural sample, individuals currently residing in urban areas but holding rural hukou are reassigned to the rural sample, and individuals currently residing in urban areas with urban hukou remain in the urban sample.

By comparing the estimated returns to education under the original and adjusted classifications, this study examines how the treatment of migrant populations affects the measurement of urban and rural returns to education and the corresponding urban–rural gap.

Under this reclassification scheme, 2760, 2249, and 2247 individual-year observations were reassigned from the original urban sample to the rural sample in 2018, 2020, and 2022, respectively, yielding a total of 7256 observations across the three survey waves. To further illustrate the compositional changes induced by the reclassification, Appendix B further divides the sample into urban residents, rural residents, and rural migrants, and reports the corresponding sample composition, income missingness patterns, and supplementary estimation results. The baseline analysis in the main text continues to focus on the original and adjusted urban–rural classification schemes.

The dependent variable is the logarithm of labor income. The key explanatory variable is years of schooling, which measures the extent of an individual's formal education. The control variables include work experience and its squared term. Year effects are controlled for using dummy variables for 2020 and 2022, with 2018 serving as the reference year. In addition, an income observation indicator is constructed to indicate whether an individual's income is observed.

Before conducting the econometric estimation, descriptive statistical analyses are first performed on the main sample variables to characterize the basic features of the sample and provide an intuitive basis for the subsequent estimation of returns to education. Since this study adopts both the original and adjusted urban–rural classification schemes, Table 1 reports not only the overall distribution of the main variables but also the changes in the mean values of urban variables under the two classification schemes, thereby illustrating the changes in sample composition resulting from sample reclassification.

Table 1. Descriptive statistics of the main variables.

Variable	Sample size	Mean	Std. Dev.	Min	Max
Annual income (10,000 yuan)	21,799	4.8093	4.4989	0	100.000
Log income	21,799	10.310	1.514	0	13.816
Years of schooling (years)	22,952	10.874	4.026	0	23.000
Work experience (years)	22,952	21.130	11.029	0	44.000
Age (years)	22,952	39.225	10.056	16.000	60.000
Gender	22,952	0.589	0.492	0	1.000
Urban (original sample)	22,952	0.633	0.482	0	1.000
Urban (adjusted sample)	22,952	0.317	0.465	0	1.000

Table 1 shows that the average annual income of individuals in the sample is RMB 48,093, with a mean log income of 10.310. The average years of schooling, work experience, and age are 10.874 years, 21.130 years, and 39.225 years, respectively, and males account for 58.9% of the sample. In terms of the urban–rural composition, the mean of the urban indicator declines from 0.633 to 0.317 under the adjusted classification, indicating a substantial change in the urban–rural composition of the sample after rural migrants are reclassified. This change provides the basis for the subsequent comparison of estimated returns to education and urban–rural differences under alternative sample-classification schemes. To further illustrate the reclassification of migrant populations and its effect

on sample composition, Appendix B reports, for 2018, 2020, and 2022, the number of reclassified observations, urban and rural sample sizes before and after reclassification, the composition of the three population groups, and income missingness rates. It also provides supplementary estimation results for urban residents, rural residents, and rural migrants.

3.2. Empirical results

In the empirical estimation, this study estimates the returns to education under both MAR and parametric NMAR specifications using fully parametric and semiparametric models. Given that years of schooling vary little over the sample period among employed adults, a fixed-effects model cannot effectively identify the returns to education. Therefore, this study adopts a pooled estimation framework while controlling for year effects. Table 2 reports the estimated returns to education for the national sample and the eastern, central, and western regions under different missingness mechanisms and model specifications. In this study, the estimated coefficient on years of schooling in the log-income equation is multiplied by 100 and reported as the return to education. Accordingly, the reported return to education represents the conditional association between an additional year of schooling and the approximate percentage change in labor income.

Table 2. Estimated returns to education by region under different missingness mechanisms and model specifications (%).

Region	Fully Parametric Model		Semiparametric Model	
Missingness mechanism	MAR	NMAR	MAR	NMAR
National	10.74***(0.35)	9.31***(0.34)	10.66***(0.35)	9.25***(0.34)
Eastern region	12.25***(0.63)	11.80***(0.63)	12.22***(0.64)	11.75***(0.63)
Central region	10.26***(0.69)	8.85***(0.66)	10.22***(0.69)	8.80***(0.67)
Western region	10.04***(0.64)	7.62***(0.60)	9.93***(0.65)	7.44***(0.61)
Northeastern region	8.21***(0.90)	7.65***(0.89)	8.20***(0.90)	7.64***(0.90)

Note: Standard errors in parentheses are calculated using an individual-level PID cluster bootstrap ($B = 500$). *** indicates statistical significance at the 1% level. The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

At the national level, the coefficients on years of schooling are significantly positive across all model specifications, indicating a significant positive association between educational attainment and individual labor income. Specifically, under the fully parametric model, the estimated return to education is 10.74% under MAR and 9.31% under NMAR; under the semiparametric model, the corresponding estimates are 10.66% and 9.25%, respectively. These results show that, under the model specifications adopted in this study, the NMAR estimates are lower than the corresponding MAR estimates, suggesting that the measurement of returns to education is clearly sensitive to assumptions regarding the income missingness mechanism.

In terms of model specification, the fully parametric and semiparametric models yield broadly similar estimates. Under both the MAR and NMAR mechanisms, the estimated returns to education

remain significantly positive, while the estimates under the NMAR mechanism are consistently lower than those under the MAR mechanism. Compared with the fully parametric model, the semiparametric model produces slightly lower estimates of the returns to education, although the differences are small. This suggests that allowing greater flexibility in the functional form of work experience does not alter the main empirical findings. Therefore, the core conclusions of this study exhibit a certain degree of robustness and do not depend on a specific functional-form assumption for the experience term.

With respect to regional variation*, the returns to education exhibit considerable heterogeneity across regions, with the eastern region showing the highest returns. Under the fully parametric model, the estimated returns to education in the eastern region are 12.25% and 11.80% under the MAR and NMAR mechanisms, respectively; under the semiparametric model, the corresponding estimates are 12.22% and 11.75%. These results suggest that the estimated association between years of schooling and labor income is relatively stronger in the eastern region. The central region has the second-highest returns to education. Under the fully parametric model, the estimates are 10.26% under MAR and 8.85% under NMAR, while the corresponding semiparametric estimates are 10.22% and 8.80%. This indicates that years of schooling are likewise significantly positively associated with labor income in the central region, although the estimated return to education is lower than that in the eastern region. In the western region, the estimates under the MAR mechanism are relatively close to those in the central region, reaching 10.04% and 9.93% under the fully parametric and semiparametric models, respectively. However, under the NMAR mechanism, the corresponding estimates decline markedly to 7.62% and 7.44%, indicating that the estimated returns to education in the western region are more sensitive to the specification of the income missingness mechanism. By contrast, the northeastern region exhibits relatively lower returns to education. Under the fully parametric model, the estimates are 8.21% under MAR and 7.65% under NMAR, while under the semiparametric model they are 8.20% and 7.64%, respectively. Such regional differences may be related to factors such as industrial structure and local labor market conditions. Overall, after moving from the MAR specification to the parametric NMAR specification, the estimated returns to education decline across all regions, with a particularly pronounced decrease in the western region. This indicates that the effect of the income missingness mechanism on the estimation of returns to education varies across regions.

Overall, the estimated returns to education are significantly positive in both the national and regional samples, although their magnitudes vary with the assumed missingness mechanism. In all regions, the NMAR estimates are lower than the corresponding MAR estimates, with a larger difference observed in the western region. This indicates that the sensitivity of estimated returns to education to assumptions about the income missingness mechanism varies across regions.

To further characterize the magnitude and uncertainty of the estimated nonignorability parameter in the NMAR response mechanism, this study additionally reports the estimated latent-income coefficient $\alpha^{(h)}$ for the main samples. It should be emphasized that only the negative direction of $\alpha^{(h)}$ is imposed a priori according to the identification condition, whereas its magnitude is jointly determined with the other model parameters during estimation. Because the same individual may appear repeatedly across multiple waves of the CFPS, uncertainty in $\alpha^{(h)}$ is likewise assessed using a bootstrap procedure

*The regional classification follows the criteria of the National Bureau of Statistics of China for the eastern, central, western, and northeastern regions. Specifically, the eastern region includes Beijing, Tianjin, Hebei, Shanghai, Jiangsu, Zhejiang, Fujian, Shandong, Guangdong, and Hainan; the central region includes Shanxi, Anhui, Jiangxi, Henan, Hubei, and Hunan; the western region includes Inner Mongolia, Guangxi, Chongqing, Sichuan, Guizhou, Yunnan, Tibet, Shaanxi, Gansu, Qinghai, Ningxia, and Xinjiang; and the northeastern region includes Liaoning, Jilin, and Heilongjiang. The regions are classified accordingly to meet the needs of this study.

clustered at the individual PID level. In each bootstrap replication, the corresponding NMAR estimation procedure is re-estimated in full, with a total of $B = 500$ replications. The results are reported in Table 3.

Table 3. Estimates and uncertainty of the nonignorability parameter $\alpha^{(h)}$ in the main samples.

Sample	Full Parameter $\alpha^{(h)}$	Bootstrap SE	95% CI	Semi-parameter $\alpha^{(h)}$	Bootstrap SE	95% CI
National	-4.193	0.026	[-4.244, -4.142]	-4.092	0.022	[-4.130, -4.045]
East	-4.112	0.053	[-4.207, -4.007]	-4.006	0.043	[-4.085, -3.917]
Central	-4.123	0.059	[-4.225, -3.996]	-4.024	0.047	[-4.111, -3.918]
West	-4.225	0.050	[-4.329, -4.136]	-4.110	0.043	[-4.190, -4.023]
Northeast	-4.128	0.133	[-4.275, -3.916]	-4.036	0.126	[-4.153, -3.831]
Original Rural	-4.153	0.045	[-4.239, -4.056]	-4.036	0.036	[-4.104, -3.959]
Original Urban	-4.137	0.038	[-4.212, -4.064]	-4.045	0.032	[-4.106, -3.984]
Adjusted Rural	-4.159	0.033	[-4.223, -4.094]	-4.057	0.027	[-4.103, -4.000]
Adjusted Urban	-4.128	0.066	[-4.260, -4.006]	-4.048	0.053	[-4.159, -3.955]

Note: $\alpha^{(h)}$ denotes the estimated nonignorability parameter associated with latent income in the NMAR response mechanism. The bootstrap standard error and the 95% confidence interval are both calculated using the individual PID cluster bootstrap with $B = 500$ replications; the 95% confidence interval is constructed using the percentile bootstrap method. Because the identification strategy adopted in this study imposes $\alpha^{(h)} < 0$ a priori, a confidence interval lying entirely below zero should not be interpreted as an independent statistical test that the true income missingness mechanism is NMAR.

Table 3 shows that, under the maintained assumptions of a Gaussian outcome distribution, the specified response mechanism, and the directional identification restriction on $\alpha^{(h)}$, the estimated nonignorability parameters are negative across all major samples. In the national sample, the estimates of $\widehat{\alpha}^{(h)}$ are -4.193 for the fully parametric model and -4.092 for the semiparametric model, with corresponding PID-cluster bootstrap standard errors of 0.026 and 0.022, respectively. The estimated magnitudes are quite similar across the two model classes, suggesting that estimation of the nonignorability parameter does not depend strongly on whether work experience is modeled parametrically or through a smooth function.

Across different subsamples, the estimates of $\widehat{\alpha}^{(h)}$ for the eastern, central, western, and northeastern

regions, as well as for the original and adjusted urban–rural samples, are mostly concentrated in the range of approximately −4.0 to −4.2. The fully parametric and semiparametric models also yield highly consistent estimates in both direction and magnitude. The bootstrap standard errors are relatively larger for the northeastern region and for some more finely disaggregated subsamples, indicating that uncertainty in the estimated nonignorability parameter increases as the sample is further subdivided.

It should be emphasized that these results represent parameter estimates and repeated-sampling uncertainty obtained under the identification conditions imposed in this study, and the location of the confidence intervals alone cannot establish that the true income missingness mechanism in the CFPS is necessarily NMAR. Their primary purpose is to make the estimated magnitude and uncertainty of the nonignorable selection effect more transparent and to provide a benchmark for subsequent sensitivity analyses under alternative response-mechanism functional forms. Although the estimates of $\widehat{\alpha}^{(h)}$ are broadly consistent in direction and magnitude across the major samples and across the fully parametric and semiparametric models, and the PID-cluster bootstrap results indicate good repeated-sampling stability, the preceding local Jacobian diagnostics caution against interpreting the nonignorability parameter as being strongly empirically identified. Rather, these findings suggest that, under the maintained parametric identification assumptions, the estimated NMAR parameter exhibits a certain degree of sample stability.

3.3. Sensitivity analysis of the response-mechanism link function

To examine whether the estimated returns to education depend on a particular functional form of the response probability, this study further re-estimates the returns to education for the national sample using Probit and complementary log-log (Cloglog) link functions in addition to the baseline Logit specification. The results are reported in Table 4.

Table 4. Sensitivity analysis of estimated returns to education under alternative response link functions (%).

Model	Link Function	MAR	NMAR	NMAR-MAR
Full Parameter	Logit	10.74	9.31	-1.43
Full Parameter	Probit	10.75	9.17	-1.57
Full Parameter	Cloglog	10.76	9.26	-1.50
Semi-parameter	Logit	10.66	9.25	-1.41
Semi-parameter	Probit	10.68	9.12	-1.56
Semi-parameter	Cloglog	10.69	9.21	-1.49

Note: The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

The estimation results are generally consistent across the different link functions. In both the fully parametric and semiparametric models, the estimated return to education under NMAR is lower than the corresponding MAR estimate under the Logit, Probit, and complementary log-log (Cloglog) specifications, while the estimates vary only modestly across the alternative link functions. This suggests that the main conclusion that estimated returns to education are sensitive to assumptions about

the missingness mechanism does not depend on a particular functional form of the response probability and is therefore reasonably robust.

3.4. Robustness check with an extended set of control variables

To examine whether the baseline results are sensitive to the specification of control variables, this study further extends the set of controls included in the income equation. In addition to the baseline covariates, including years of schooling, work experience, and year effects, the extended specification further incorporates gender, marital status, hukou type, and regional fixed effects. The fully parametric model continues to include work experience and its squared term, whereas the semiparametric model continues to use a smooth function to capture the nonlinear effect of work experience. Because potential work experience in this study is constructed from age and years of schooling, age is not additionally included in the extended specification in order to avoid mechanical multicollinearity among age, years of schooling, and work experience. In addition, occupation and industry may lie along the pathway linking education to income, so including them as conventional control variables could alter the interpretation of the coefficient on years of schooling. They are therefore excluded from the main extended-control specification.

To prevent changes in sample composition caused by the additional control variables from confounding the comparison between the baseline and extended specifications, both models are re-estimated using the same common sample. The common sample contains 22,316 observations at the national level, including 14,155 urban observations and 8161 rural observations. Table 5 reports the estimated returns to education before and after extending the set of control variables.

Table 5. Robustness check of estimated returns to education with an extended set of control variables (%).

Model	Sample	Same Sample Baseline MAR	Same Sample Baseline NMAR	Extended Control MAR	Extended Control NMAR
Full Parameter	National	10.76***(0.34)	9.32***(0.34)	9.20***(0.41)	8.01***(0.41)
	Urban	10.84***(0.44)	10.25***(0.44)	10.14***(0.51)	9.62***(0.52)
	Rural	9.35***(0.63)	6.92***(0.61)	7.66***(0.68)	5.51***(0.66)
Semi-parameter	National	10.70***(0.34)	9.27***(0.34)	9.19***(0.41)	7.96***(0.41)
	Urban	10.77***(0.44)	10.22***(0.44)	10.11***(0.51)	9.59***(0.52)
	Rural	9.27***(0.64)	6.79***(0.62)	7.56***(0.68)	5.35***(0.67)

Note: Standard errors in parentheses are calculated using an individual-level PID cluster bootstrap ($B = 500$). *** indicates statistical significance at the 1% level. The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

Table 5 shows that, after additionally controlling for gender, marital status, hukou type, and regional fixed effects, the estimated returns to education decline under both the fully parametric and semiparametric models, but the core findings remain substantively unchanged. Under both model specifications, the estimated returns to education remain significantly positive, the urban estimates

continue to exceed the rural estimates, and the returns to education under the NMAR specification remain lower than the corresponding estimates under MAR. These results suggest that the main conclusions regarding the sensitivity of estimated returns to education to assumptions about the income missingness mechanism, as well as the urban–rural differences, are not driven by the relatively limited set of control variables in the baseline model and are reasonably robust.

3.5. Out-of-sample validation under artificially induced NMAR missingness

The preceding empirical results show that, under the model specifications adopted in this study, the estimated returns to education differ to some extent between the MAR and NMAR mechanisms. To further evaluate the recovery performance of the proposed NMAR estimation method in a setting with known nonrandom missingness, this study artificially induces NMAR missingness in the subsample with originally observed income and compares the estimates obtained under the MAR and NMAR approaches with those based on the complete data prior to artificial deletion.

To ensure consistency between the artificially induced missingness mechanism and the empirical specification used in the main analysis, the artificial response probability is likewise specified using a Logit-NMAR structure:

$$\begin{aligned}&\Pr\left(R_{it}=1 \mid Y_{it}^{(h)}, eduy_{it}, D_{2020,t}, D_{2022,t}, exper_{it}\right)\\&\quad=\frac{\exp\left(\eta_0+\alpha^{(h)}Y_{it}^{(h)}+\eta_1 eduy_{it}+\eta_2 D_{2020,t}+\eta_3 D_{2022,t}+\eta_4 exper_{it}\right)}{1+\exp\left(\eta_0+\alpha^{(h)}Y_{it}^{(h)}+\eta_1 eduy_{it}+\eta_2 D_{2020,t}+\eta_3 D_{2022,t}+\eta_4 exper_{it}\right)}.\end{aligned}$$

Here, $R_{it}=1$ indicates that the individual's income remains observed after the artificial missingness procedure, whereas $R_{it}=0$ indicates that the income is artificially masked. The slope parameters of the artificial missingness mechanism are first estimated using the original CFPS sample and the semiparametric NMAR model employed in the main analysis. The intercept is then recalibrated so that the average income missingness rate in the artificially generated sample matches the actual income missingness rate in the original sample. Thus, rather than arbitrarily imposing an alternative missingness mechanism, this experiment evaluates the recovery performance of the estimation method under the same Logit-NMAR structure as that used in the main analysis.

Given that the same individual may appear in multiple CFPS survey waves, this study partitions the sample into training and test sets at the individual PID level, using an 80%–20% split, in order to avoid information leakage arising from observations of the same individual appearing in both sets. The experiment is independently repeated 100 times. In each replication, the estimated return to education based on the complete data prior to artificial deletion is treated as the Oracle benchmark, after which the model is re-estimated under the MAR and NMAR specifications, respectively. For both the fully parametric and semiparametric models, the outcome equation, response mechanism, and MCEM implementation are kept consistent with those used in the baseline analysis. The Monte Carlo sample size is set to $M=500$, the convergence threshold for parameter changes is 0.05, and the maximum number of iterations is 10. The evaluation metrics include the absolute error of the estimated return to education relative to the Oracle benchmark, as well as the root mean square error (RMSE) and mean absolute error (MAE) calculated on the artificially masked income observations.

Table 6 shows that, under artificially induced NMAR missingness, the two model classes yield highly consistent results. Compared with the MAR specification, the NMAR-adjusted estimates of

the return to education are closer to the Oracle benchmark obtained prior to artificial deletion, and their absolute errors, RMSEs, and MAEs are all lower, indicating that the proposed method has good recovery performance under the specified NMAR setting. It should be emphasized, however, that this validation exercise is intended only to assess the performance of the proposed estimation method under a known NMAR mechanism and should not be interpreted as evidence that the true income missingness mechanism in the CFPS data is necessarily NMAR.

Table 6. Recovery performance of the MAR and NMAR models under artificially induced NMAR missingness.

Model	Oracle Education Return Rate (%)	MAR (%)	NMAR (%)	MAR Absolute Error	NMAR Absolute Error	MAR RMSE	NMAR RMSE	MAR MAE	NMAR MAE
Full Parameter	10.90	11.83	10.50	0.94	0.39	1.371	1.181	1.276	1.061
Semi-parameter	10.86	11.80	10.45	0.94	0.41	1.369	1.177	1.274	1.058

Note: The absolute error in the return to education is defined as the absolute difference between the corresponding estimate and the Oracle estimate obtained from the complete data prior to artificial deletion, and is reported in percentage points. The RMSE and MAE are both calculated using the artificially masked income observations. The results reported in the table are averages over 100 repeated experiments. The return to education is reported as the estimated coefficient on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

4. Urban–rural differences in returns to education in China

4.1. Estimated returns to education in urban and rural areas

Building on the preceding estimates of overall returns to education, this study further examines urban–rural differences in returns to education. Table 7 reports the estimated returns to education in rural and urban areas under different model specifications, missingness mechanisms, and sample classification schemes. Overall, regardless of whether the fully parametric or semiparametric model is used, and whether the original or adjusted sample classification is adopted, the estimated returns to education are consistently higher in urban areas than in rural areas, indicating that the urban–rural disparity in returns to education is highly robust across alternative specifications.

As shown in Table 7, in the original sample, the estimated returns to education for rural and urban workers under the fully parametric model are 9.19% and 10.90%, respectively, under the MAR mechanism, and 6.78% and 10.32%, respectively, under the NMAR mechanism. The semiparametric model exhibits the same overall pattern. When the missingness mechanism is changed from MAR to NMAR, the decline in the estimated returns to education is substantially greater for the rural sample than for the urban sample, indicating that the rural estimates are more sensitive to the specification of the income missingness mechanism. Such differences may be related to characteristics of rural workers, including their sources of income, forms of employment, and the completeness of income reporting, although this study does not attempt to establish these mechanisms causally. Accordingly, the results here mainly indicate that alternative missingness assumptions lead to larger differences in the estimates for the rural sample, rather than implying that the true income missingness mechanism in rural areas must necessarily be NMAR.

Table 7. Comparison of urban and rural returns to education under different models, missingness mechanisms, and sample classifications (%).

Model	Sample Classification	Rural (MAR)	Rural (NMAR)	Urban (MAR)	Urban (NMAR)
Fully Parametric	Original Sample	9.19***(0.63)	6.78***(0.60)	10.90***(0.45)	10.32***(0.45)
	Adjusted Sample	9.18***(0.48)	7.38***(0.46)	13.67***(0.70)	13.53***(0.71)
Semiparametric	Original Sample	9.13***(0.64)	6.64***(0.60)	10.81***(0.46)	10.25***(0.45)
	Adjusted Sample	9.12***(0.49)	8.23***(0.46)	13.61***(0.71)	13.51***(0.71)

Note: Standard errors in parentheses are calculated using an individual-level PID cluster bootstrap ($B = 500$). *** indicates statistical significance at the 1% level. The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

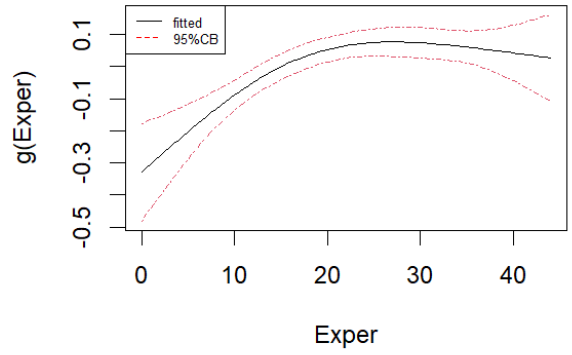
The higher estimated return to education in urban areas indicates that the association between years of schooling and labor income is stronger in the urban sample. One possible explanation is that urban and rural workers differ in terms of job types, occupational positions, and industry distribution, and these structural differences may be related to urban–rural heterogeneity in educational returns. Accordingly, this study interprets the urban–rural difference in returns to education as reflecting differences in the strength of the education–income association across distinct labor market environments, rather than as direct evidence of a causal effect of education itself.

The adjusted-sample results show that the urban–rural gap in returns to education widens further after reclassifying employed individuals who currently reside in urban areas but hold rural hukou into the rural sample. Under the fully parametric model, the estimated returns to education for the rural and urban subsamples are 9.18% and 13.67%, respectively, under MAR, and 7.38% and 13.53%, respectively, under NMAR. The semiparametric model yields a consistent pattern. One possible explanation is that the sample in this study is restricted to employed individuals. Although rural migrants are classified as rural in terms of household registration or place of origin, their income returns are already influenced by urban employment opportunities, industrial structure, and job characteristics. Reassigning these individuals to the rural sample increases heterogeneity within the rural subsample, while the urban subsample becomes more concentrated among workers with higher returns to education and more stable employment quality. As a result, the estimated urban–rural gap in returns to education becomes more pronounced.

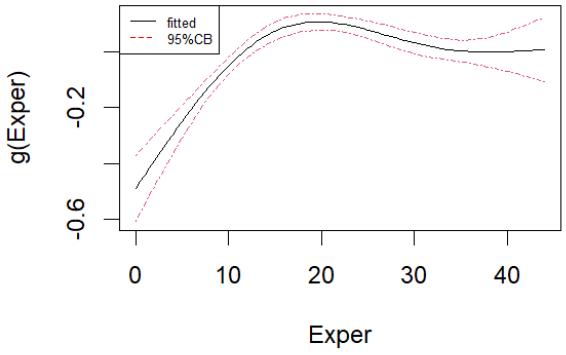
Supplementary analyses of the three population groups in Appendix B further show that the estimated return to education is generally higher among urban residents than among rural residents and rural migrants. Rural migrants were originally included in the urban sample; after they are reclassified into the rural sample, the adjusted urban sample retains a larger proportion of urban residents with relatively higher educational returns, thereby raising the estimated urban return to education and widening the urban–rural gap. In addition, when moving from the MAR specification to the NMAR specification, the decline in the estimated return to education is most pronounced among rural residents, indicating that the impact of the income missingness mechanism differs across groups with different migration statuses.

4.2. Nonparametric estimation of the work experience effect

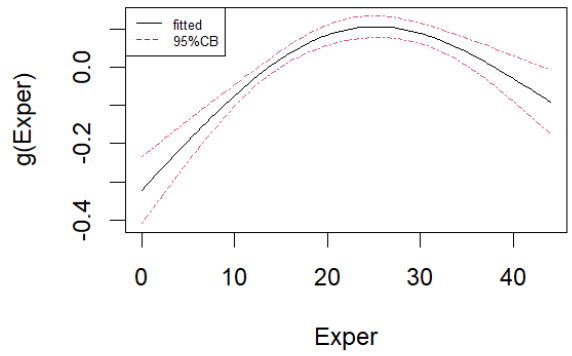
Building on the estimates of urban–rural returns to education presented in the previous section, this study further examines the nonlinear characteristics of the work experience effect in the semiparametric model. Under the NMAR mechanism, the smooth functions of work experience are plotted separately for the rural and urban subsamples under both the original and adjusted sample classifications, as shown in Figure 1. In the figure, the black solid line represents the estimated smooth function of work experience, while the red dashed lines represent the 95% confidence bands.



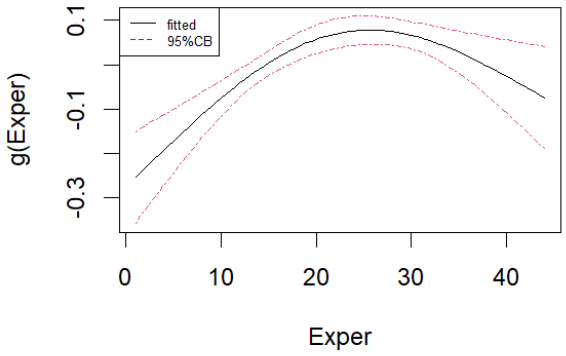
(a) Original rural sample



(b) Original urban sample



(c) Adjusted rural sample



(d) Adjusted urban sample

Figure 1. Smooth function estimates of the work experience effect on income under the NMAR mechanism.

Overall, the smooth functions of work experience exhibit clear nonlinear patterns across all four samples, indicating that the effect of work experience on labor income is not simply linear. As work experience increases, its effect on income generally rises at first, reaches a relatively high level at the medium or medium-to-high experience stage, and then tends to level off or decline, displaying an

inverted U-shaped pattern. This suggests that, in the sample, the positive association between labor income and work experience gradually weakens at higher levels of experience.

The curves above indicate that the relationship between work experience and income exhibits a clear stage-dependent pattern. The positive association is relatively strong at low to moderate levels of experience, but gradually weakens at higher levels of experience. Differences in the shapes of the urban and rural curves further suggest heterogeneity in the work experience–income relationship across groups. Because this study does not directly identify specific mechanisms such as job promotion, industry transitions, or skill accumulation, interpretations of these mechanisms are presented only as possible explanations.

These results indicate that it is necessary to flexibly control for the work experience effect when estimating urban–rural differences in returns to education. Compared with the fully parametric model, the semiparametric model can better capture the experience–income relationship and reduce potential bias arising from functional-form misspecification. Combined with the results in Table 7, the broadly similar estimates obtained from the two model specifications suggest that the main conclusions regarding urban–rural differences in returns to education are reasonably robust. At the same time, the smooth-function estimates reveal differences in the pathways through which work experience affects income in rural and urban samples, indicating that urban–rural differences in returns to education are related not only to years of schooling, but also to employment experience, stages of career development, and labor market conditions.

4.3. Comparison of the sources of urban–rural differences in returns to education

To compare changes in the urban–rural gap in returns to education across different specifications, this study compares the urban–rural gap in returns to education along three dimensions: The missingness mechanism, model specification, and sample classification. Tables 8–10 show that the numerical changes in the urban–rural gap induced by alternative missingness assumptions and sample classifications are generally larger than those resulting from changes in the functional form of the model.

Table 8. Urban–rural gaps in returns to education under different missingness mechanisms (percentage points).

Model	Sample Classification	MAR Urban–Rural Gap	NMAR Urban–Rural Gap	Change
Fully Parametric	Original Sample	1.71	3.54	+1.83
	Adjusted Sample	4.49	6.15	+1.66
Semiparametric	Original Sample	1.68	3.61	+1.93
	Adjusted Sample	4.49	5.28	+0.79

Note: Change = NMAR urban–rural gap − MAR urban–rural gap.

From the perspective of the missingness mechanism, the urban–rural gap in returns to education is consistently larger under the NMAR specification than under MAR. In the original sample, the urban–rural gap increases from 1.71 to 3.54 percentage points under the fully parametric model and from 1.68 to 3.61 percentage points under the semiparametric model. In the adjusted sample, the corresponding gaps increase from 4.49 to 6.15 percentage points and from 4.49 to 5.28 percentage points, respectively.

Under the NMAR specification adopted in this study, the estimated return to education in the rural sample declines more substantially relative to the MAR specification, thereby widening the urban–rural gap. This result indicates that the urban–rural difference in returns to education is clearly sensitive to assumptions regarding the income missingness mechanism.

Table 9. Urban–rural gaps in returns to education under different model specifications (percentage points).

Sample Classification	Missingness Mechanism	Fully Parametric Urban–Rural Gap	Semiparametric Urban–Rural Gap	Change
Original Sample	MAR	1.71	1.68	-0.03
	NMAR	3.54	3.61	+0.07
Adjusted Sample	MAR	4.49	4.49	+0.00
	NMAR	6.15	5.28	-0.87

Note: Change = semiparametric urban–rural gap − fully parametric urban–rural gap.

From the perspective of model specification, the urban–rural gaps in returns to education estimated by the fully parametric and semiparametric models are generally similar. In the original sample, the semiparametric estimates differ from the fully parametric estimates by only 0.03 and 0.07 percentage points. In the adjusted sample, the two models yield identical results under the MAR mechanism, whereas under the NMAR mechanism, the urban–rural gap estimated by the semiparametric model is 0.87 percentage points lower than that estimated by the fully parametric model. This indicates that, after accounting for nonrandom missingness and sample reclassification, the functional form of the model still has some influence on the estimation results. One possible explanation is that the semiparametric model captures the relationship between work experience and income more flexibly, thereby reducing the restrictions imposed by prespecifying the functional form of work experience. Overall, however, the use of a smooth function does not alter the basic conclusions that an urban–rural gap in returns to education exists and that this gap is larger under the NMAR mechanism, indicating a certain degree of robustness in the main findings of this study.

Table 10. Urban–rural gaps in returns to education under different sample classifications (percentage points).

Model	Missingness Mechanism	Original-Sample Urban–Rural Gap	Adjusted-Sample Urban–Rural Gap	Change
Fully Parametric	MAR	1.71	4.49	+2.78
	NMAR	3.54	6.15	+2.61
Semiparametric	MAR	1.68	4.49	+2.81
	NMAR	3.61	5.28	+1.67

Note: Change = adjusted-sample urban–rural gap − original-sample urban–rural gap.

From the perspective of sample classification, the urban–rural gap in returns to education is

substantially larger in the adjusted sample than in the original sample. Under the fully parametric model, the urban–rural gap increases from 1.71 to 4.49 percentage points under the MAR specification and from 3.54 to 6.15 percentage points under the NMAR specification. The semiparametric model exhibits the same pattern. These results indicate that the classification of migrant workers can significantly affect the estimated urban–rural gap in returns to education. Particularly in a sample restricted to employed individuals, reclassifying rural migrants changes the internal composition of both the rural and urban samples, thereby altering the estimated urban–rural gap in returns to education across different classification schemes.

4.4. Regional analysis of urban–rural differences in returns to education

The preceding analysis shows that urban–rural differences in returns to education persist at the aggregate level, but these differences may not be evenly distributed across regions. Given the substantial regional disparities in economic development, industrial structure, and labor market conditions in China, this study further divides the sample into the eastern, central, western, and northeastern regions to examine urban–rural differences in returns to education within each region.

To examine the robustness of the above results across different model specifications, Table 12 further reports the regional urban–rural estimates of returns to education obtained from the semiparametric model.

Tables 11 and 12 show that the urban–rural gaps in returns to education are generally larger in the central, western, and northeastern regions than in the eastern region, indicating a certain degree of regional heterogeneity in urban–rural differences in educational returns. Such heterogeneity may be related to differences across regions in industrial structure, the distribution of employment opportunities, the level of urbanization, and urban–rural labor market conditions. Urban and rural labor markets are relatively more integrated in the eastern region, whereas more pronounced urban–rural differences in job structure and regional development conditions may exist in the central, western, and northeastern regions. It should be noted that this study does not causally identify specific mechanisms such as industrial structure, employment opportunities, or labor market segmentation; the above discussion is therefore intended only to provide possible explanations for the observed regional heterogeneity.

Overall, the estimated returns to education are consistently higher in urban areas than in rural areas across different model specifications and sample classifications. The urban–rural gap generally widens when moving from the MAR specification to the parametric NMAR specification, and it increases further when the sample classification is changed from the original scheme to the adjusted scheme. These findings indicate that urban–rural differences in returns to education are relatively sensitive to both assumptions about the income missingness mechanism and sample-classification rules. Within the range of specifications considered in this study, the numerical differences induced by changes in the missingness mechanism and sample classification are generally larger than those between the fully parametric and semiparametric models.

Table 11. Estimates of urban–rural returns to education by region under the fully parametric model (%).

Region	Category	Original Sample MAR	Original Sample NMAR	Adjusted Sample MAR	Adjusted Sample NMAR
Eastern Region	Rural	10.57***(1.16)	9.99***(1.20)	10.13***(0.84)	9.71***(0.84)
	Urban	11.94***(0.77)	11.56***(0.76)	14.55***(1.28)	14.49***(1.26)
Central Region	Rural	8.60***(1.22)	6.75***(1.20)	8.40***(0.97)	7.34***(0.96)
	Urban	10.78***(0.87)	10.13***(0.86)	14.33***(1.26)	14.25***(1.25)
Western Region	Rural	8.79***(1.06)	4.81***(0.95)	8.65***(0.80)	6.28***(0.75)
	Urban	10.70***(0.87)	9.80***(0.85)	13.26***(1.47)	13.16***(1.54)
Northeastern Region	Rural	6.16***(1.93)	5.35***(1.95)	5.99***(1.63)	5.45***(1.61)
	Urban	9.43***(1.17)	9.38***(1.18)	12.20***(1.64)	12.18***(1.66)

Note: Standard errors in parentheses are calculated using an individual-level PID cluster bootstrap ($B = 500$). *** indicates statistical significance at the 1% level. The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

Table 12. Estimates of urban–rural returns to education by region under the semiparametric model (%).

Region	Category	Original Sample MAR	Original Sample NMAR	Adjusted Sample MAR	Adjusted Sample NMAR
Eastern Region	Rural	10.56***(1.17)	9.93***(1.20)	10.06***(0.84)	9.72***(0.84)
	Urban	11.88***(0.77)	11.50***(0.76)	14.26***(1.31)	14.19***(1.28)
Central Region	Rural	8.43***(1.22)	6.48***(1.19)	8.28***(0.98)	7.66***(0.96)
	Urban	10.75***(0.88)	10.09***(0.86)	14.27***(1.27)	14.20***(1.25)
Western Region	Rural	8.72***(1.08)	4.61***(0.98)	8.54***(0.81)	6.90***(0.77)
	Urban	10.51***(0.84)	9.62***(0.85)	13.39***(1.48)	13.23***(1.54)
Northeastern Region	Rural	6.02***(1.96)	5.26***(1.94)	5.97***(1.65)	5.66***(1.63)
	Urban	9.45***(1.17)	9.41***(1.18)	12.31***(1.63)	12.29***(1.65)

Note: Standard errors in parentheses are calculated using an individual-level PID cluster bootstrap ($B = 500$). *** indicates statistical significance at the 1% level. The returns to education are reported as the estimated coefficients on years of schooling multiplied by 100, with the standard errors in parentheses reported on the same scale.

5. Conclusions

Using three waves of panel data from the China Family Panel Studies (CFPS) for 2018, 2020, and 2022, this study re-estimates urban and rural returns to education under both MAR and NMAR using

fully parametric and semiparametric models. It further examines how the reclassification of migrant samples affects the measurement of urban–rural differences in returns to education. The main findings are as follows.

First, the estimated returns to education are clearly sensitive to the assumed income missingness mechanism. In both the national sample and the urban–rural subsamples, the estimated returns to education under the parametric NMAR specification are generally lower than the corresponding estimates under MAR. For example, in the national sample, the estimated return to education under the fully parametric model declines from 10.74% under MAR to 9.31% under NMAR, while under the semiparametric model, it declines from 10.66% to 9.25%. These results indicate that different assumptions about the income missingness mechanism can lead to appreciable differences in the estimated returns to education. Therefore, when survey data are used to measure returns to education, it is important to examine the sensitivity of the results to alternative missingness assumptions rather than relying on a single missingness mechanism.

Second, the estimated returns to education are generally higher in urban areas than in rural areas, and the rural sample appears to be more sensitive to the assumed missingness mechanism. Regardless of whether the fully parametric or semiparametric model is used, and regardless of whether the original or adjusted sample classification is adopted, the estimated return to education is consistently higher in urban areas than in rural areas. At the same time, when moving from the MAR specification to the parametric NMAR specification, the change in the estimated rural return to education is generally larger than that in the urban estimate. This suggests that the sensitivity of estimated returns to education to assumptions about the income missingness mechanism differs between urban and rural populations. Accordingly, comparisons of urban and rural returns to education should also take into account group differences in the structure of income missingness.

Third, the fully parametric and semiparametric models yield highly consistent substantive conclusions. The semiparametric model provides a more flexible representation of the relationship between work experience and income through a smooth function, yet its estimates are broadly similar to those from the fully parametric model. Under both model classes, the estimated return to education is higher in urban areas than in rural areas, and the estimates under the parametric NMAR specification are generally lower than their corresponding MAR estimates. This suggests that the main conclusions regarding urban–rural differences in returns to education and their sensitivity to the assumed missingness mechanism do not depend strongly on a particular functional-form specification for work experience.

Fourth, the way migrant populations are classified between urban and rural samples can substantially affect the measured urban–rural difference in returns to education. Under the employed-sample definition and the reclassification scheme adopted in this study, reassigning rural migrants to the rural sample further widens the estimated urban–rural gap in returns to education. Supplementary analyses based on three population groups show clear differences across urban residents, rural residents, and rural migrants in both the estimated returns to education and their sensitivity to the assumed missingness mechanism. In particular, the estimated return to education among rural migrants is lower than that among urban residents. Reassigning rural migrants from the original urban sample to the rural sample leaves the adjusted urban sample with a larger proportion of urban residents who exhibit higher educational returns, thereby widening the estimated urban–rural gap. This finding indicates that the classification of migrant populations can affect estimates of returns to education

and their urban–rural differences by altering the internal composition of the urban and rural samples. The resulting estimates should therefore be interpreted in light of the specific sample scope and classification rules employed.

Overall, the findings show that urban–rural differences in returns to education are influenced not only by the functional-form specification of the outcome equation and the sample-classification scheme, but also by assumptions regarding the income missingness mechanism. In particular, under the NMAR specification adopted in this study, the estimated rural return to education changes more substantially, which in turn leads to a wider urban–rural gap relative to that under MAR. The MAR and NMAR results are therefore more appropriately interpreted as a sensitivity comparison under alternative assumptions about the missingness mechanism, rather than as evidence that one particular mechanism must represent the true data-generating process. This interpretation is consistent with the theoretical setup of the NMAR model in this study, whose identification relies on assumptions concerning the joint distribution, the response mechanism, and the direction of the nonignorability effect.

Based on the above findings, this study derives the following policy implications:

First, further improve the conditions under which rural workers can realize returns to human capital. The results show that the estimated return to education is generally lower in the rural sample than in the urban sample, indicating that the association between years of schooling and labor income is relatively weaker in rural areas. Accordingly, efforts should continue to improve the allocation of educational resources in rural areas, strengthen vocational education and skills training, and expand rural workers' access to high-quality employment opportunities.

Second, further promote the integration of urban and rural labor markets. This study finds that the estimated return to education is consistently higher in urban areas than in rural areas, suggesting that differences in labor market conditions may be associated with disparities in educational returns. In light of existing urban–rural employment conditions, further efforts are needed to improve a unified urban–rural employment service system, reduce institutional barriers related to hukou status, social security, and access to public services, and provide rural workers with more equal opportunities for employment and mobility.

Third, greater attention should be paid to income missingness mechanisms and their potential selectivity in household and income surveys. The results show that estimates of returns to education and urban–rural differences vary substantially under different assumptions about the income missingness mechanism, with the rural sample being more sensitive to such assumptions. Future survey design and data dissemination should therefore further improve the formulation and administration of income-related questions, reduce selective nonresponse among particular income groups, and strengthen sensitivity analyses under alternative missingness assumptions in related empirical studies, thereby improving the reliability and transparency of analyses based on micro-level survey data.

Fourth, greater caution is needed in defining the urban–rural classification of migrant populations. The results show that, under the employed-sample definition and the reclassification scheme adopted in this study, reassigning rural migrants to the rural sample does not narrow the estimated urban–rural gap in returns to education; instead, the gap widens further. This indicates that the effect of migrant reclassification depends on the specific classification scheme adopted. Accordingly, policy evaluation and academic research should avoid relying solely on static urban–rural labels when assessing differences in educational returns and should instead incorporate population mobility,

employment structure, and sample composition into a more dynamic analytical framework.

Appendix

Appendix A. Stability check of the Monte Carlo sample size

To assess the numerical stability of the Monte Carlo sample size $M = 500$ used in the baseline estimation, this study considers $M = 50$, 100, 200, 500, and 1000 under the NMAR specification for the national sample, with each setting independently re-estimated 20 times. Table A1 reports the mean estimated returns to education across replications and the corresponding Monte Carlo standard deviations (MC SD).

Table A1. Stability of NMAR estimates of returns to education under different Monte Carlo sample sizes.

M	Fully parametric estimate (%)	Fully parametric MC SD	Semiparametric estimate (%)	Semiparametric MC SD
50	9.3416	0.0163	9.2813	0.0170
100	9.3253	0.0112	9.2646	0.0112
200	9.3200	0.0073	9.2590	0.0073
500	9.3142	0.0040	9.2530	0.0042
1000	9.3121	0.0027	9.2506	0.0028

Note: Each value of M is independently estimated over 20 replications. MC SD denotes the standard deviation of the estimated returns to education across repeated Monte Carlo estimations with a fixed M, which measures numerical variation induced by finite Monte Carlo sampling rather than the statistical standard error of the parameter estimates. The reported returns to education are obtained by multiplying the estimated coefficients on years of schooling by 100.

As M increases, the Monte Carlo standard deviations of both the fully parametric and semiparametric models continuously decrease, and the estimated returns to education gradually become more stable. When $M = 500$, the estimates from both models are already very close to those obtained with $M = 1000$, with relatively small Monte Carlo variation. Therefore, the baseline estimation in this study adopts $M = 500$ to achieve a balance between computational cost and numerical stability.

Appendix B. Reclassification of migrant populations and supplementary analyses of three population groups

To further illustrate the specific procedure of migrant reclassification and its impact on sample composition, this appendix first reports the urban and rural sample sizes before and after reclassification for each survey year. It then divides the sample into three groups—urban residents, rural residents, and rural migrants—and compares their income missingness patterns and estimated returns to education. Specifically, individuals who currently reside in urban areas and hold urban hukou are classified as

urban residents; individuals who currently reside in rural areas are classified as rural residents; and individuals who currently reside in urban areas but hold rural hukou are classified as rural migrants. The three-group classification is used only to provide supplementary interpretation of the compositional changes induced by sample reclassification and does not alter the baseline urban–rural classification or estimation framework adopted in the main text.

Table B1 shows that 2760, 2249, and 2247 individual-year observations were reclassified in 2018, 2020, and 2022, respectively, with a total of 7256 observations across the three waves. After reclassification, the rural sample increases from 8429 to 15,685 observations, while the urban sample decreases from 14,523 to 7267 observations, indicating that the reclassification of migrant populations has a substantial impact on the composition of urban and rural samples.

Table B1. Reclassification of migrant populations and changes in urban–rural sample sizes.

Year	Original Rural	Original Urban	Reclassified to Rural	Adjusted Rural	Adjusted Urban	Total Sample
2018	3328	5564	2760	6088	2804	8892
2020	2715	4516	2249	4964	2267	7231
2022	2386	4443	2247	4633	2196	6829
Total	8429	14,523	7256	15,685	7267	22,952

Note: The number of reclassified observations refers to the number of individual-year observations of rural migrants that are reassigned from the original urban sample to the rural sample according to the adjusted sample classification scheme adopted in this study.

Table B2 shows that, in the pooled sample, urban residents, rural residents, and rural migrants account for 31.66%, 36.72%, and 31.61% of the sample, respectively, with corresponding income missingness rates of 3.39%, 7.40%, and 4.18%. Based on these figures, the income missingness rates in the original rural and urban samples are calculated as 7.40% and 3.78%, respectively, while those in the adjusted rural and urban samples are 5.91% and 3.39%, respectively. These results indicate that the reclassification of migrant populations not only changes the sizes of the urban and rural samples but also alters the structure of income missingness under different urban–rural classification schemes.

Table B3 shows that the results for the three population groups are highly consistent across the two model specifications. Urban residents exhibit the highest returns to education, with MAR estimates of 13.67% and 13.61% under the fully parametric and semiparametric models, respectively, and NMAR estimates of 13.49% and 13.44%, respectively. The estimated returns to education among rural residents and rural migrants are generally lower. When moving from the MAR specification to the NMAR specification, the decline in the estimated return to education is most pronounced among rural residents, followed by rural migrants, whereas the estimates for urban residents change only slightly. This indicates that groups with different migration statuses differ in their sensitivity to assumptions regarding the income missingness mechanism.

Moreover, the estimated return to education among rural migrants is lower than that among urban residents. Therefore, when rural migrants are reassigned from the original urban sample to the rural sample, the adjusted urban sample retains a larger proportion of urban residents with higher educational returns. This provides further evidence from the perspective of sample composition for the finding in

the main text that the urban–rural gap in returns to education increases under the adjusted classification scheme.

Table B2. Sample composition and income missingness among the three population groups.

Year	Group	Sample Size	Income Observations	Missing Income	Missing Rate (%)	Sample Proportion (%)
2018	Urban Residents	2804	2590	214	7.63	31.53
2018	Rural Resident Population	3328	2779	549	16.50	37.43
2018	Rural Floating Population	2760	2515	245	8.88	31.04
2020	Urban Residents	2267	2247	20	0.88	31.35
2020	Rural Resident Population	2715	2660	55	2.03	37.55
2020	Rural Floating Population	2249	2210	39	1.73	31.10
2022	Urban Residents	2196	2184	12	0.55	32.16
2022	Rural Resident Population	2386	2366	20	0.84	34.94
2022	Rural Floating Population	2247	2228	19	0.85	32.90
Total	Urban Residents	7267	7021	246	3.39	31.66
Total	Rural Resident Population	8429	7805	624	7.40	36.72
Total	Rural Floating Population	7256	6953	303	4.18	31.61

Note: Urban residents, rural residents, and rural migrants are classified according to the sample reclassification scheme adopted in this study. "Total" refers to the pooled individual-year observations from the three survey waves of 2018, 2020, and 2022.

Table B3. Supplementary estimates of returns to education for the three population groups (%).

Group	Full Parameter Model			Semiparametric Model		
	MAR	NMAR	NMAR–MAR	MAR	NMAR	NMAR–MAR
Urban Residents	13.67	13.49	-0.18	13.61	13.44	-0.17
Rural Resident Population	9.19	6.79	-2.40	9.13	6.67	-2.46
Rural Floating Population	8.71	7.80	-0.92	8.58	7.71	-0.87

Note: The returns to education are reported in percentage terms, obtained by multiplying the estimated coefficient on years of schooling by 100. NMAR−MAR denotes the difference between the NMAR estimate and the MAR estimate under the same model specification. This table provides supplementary comparisons across the three population groups, and PID-cluster bootstrap standard errors are not reported again.

Appendix C. Local numerical identification diagnostics of the response-mechanism estimating equations

Table C1 reports the local numerical identification diagnostics of the response-mechanism estimating equations evaluated in the neighborhood of the baseline fully parametric NMAR estimates.

Table C1. Local numerical identification diagnostics of the response-mechanism estimating equations.

Diagnostic Indicator	Value
Standardized Jacobian Condition Number	11671.9
Min/Max Absolute Eigenvalue Ratio	8.6×10^{-5}
Nuisance-adjusted information ratio	0.000699

Note: The above metrics are calculated based on the Jacobian matrix of the response-mechanism estimating equations evaluated in the neighborhood of the baseline fully parametric NMAR estimates. They are used only to assess the local numerical dependence among parameters and should not be interpreted as a formal likelihood ratio test or as a criterion for determining sufficient identification.

Author contributions

Chao Dai: Formal analysis, investigation, validation, software, visualization, writing–original draft; Tingfang Luo: Resources, data curation; Jierui Du: Conceptualization, methodology; Hanji He: Supervision, writing–review and editing. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported by the Guangxi Natural Science Foundation (2025GXNSFBA069438), Guangxi Natural Science Foundation (2026GXNSFBA00640158), and China Association for Business Statistics 2025 annual planning key projects (2025STZD15).

Conflict of interest

The authors declare that there is no conflict of interest in the publication of this paper.

References

1. T. Sicular, X. M. Yue, B. Gustafsson, S. Li, The urban–rural income gap and inequality in China, *Rev. Income Wealth*, **53** (2007), 93–126. https://doi.org/10.1111/j.1475-4991.2007.00219.x
2. J. A. Mincer, *Schooling, experience, and earnings*, New York: National Bureau of Economic Research, 1974.
3. G. Psacharopoulos, H. A. Patrinos, Returns to investment in education: A decennial review of the global literature, *Educ. Econ.*, **26** (2018), 445–458. https://doi.org/10.1080/09645292.2018.1484426
4. J. S. Zhang, Y. H. Zhao, A. Park, X. Q. Song, Economic returns to schooling in urban China, 1988 to 2001, *J. Comp. Econ.*, **33** (2005), 730–752. https://doi.org/10.1016/j.jce.2005.05.008
5. X. Zhou, Increasing returns to education, changing labor force structure, and the rise of earnings inequality in urban China, 1996–2010, *Soc. Forces*, **93** (2014), 429–455. https://doi.org/10.1093/sf/sou073
6. X. X. Ma, I. Iwasaki, Return to schooling in China: A large meta-analysis, *Educ. Econ.*, **29** (2021), 379–410. https://doi.org/10.1080/09645292.2021.1900791
7. X. Gao, M. S. Li, Differences between returns to education in urban and rural China and its evolution from 1989 to 2019, *PLoS One*, **17** (2022), e0274506. https://doi.org/10.1371/journal.pone.0274506
8. B. Huang, M. Tani, Y. Wei, Y. Zhu, Returns to education in China: Evidence from the great higher education expansion, *China Econ. Rev.*, **74** (2022), 101804. https://doi.org/10.1016/j.chieco.2022.101804
9. D. X. Hu, H. C. Li, H. X. Song, J. M. Liang, Trends and urban-rural disparities in the returns to higher education in China amidst scale expansion: A correlational analysis, *Front. Educ.*, **11** (2026), 1752579. https://doi.org/10.3389/feduc.2026.1752579

10. Z. Q. Liu, Institution and inequality: The hukou system in China, *J. Comp. Econ.*, **33** (2005), 133–157. https://doi.org/10.1016/j.jce.2004.11.001

11. C. R. Bollinger, B. T. Hirsch, C. M. Hokayem, J. P. Ziliak, Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch, *J. Polit. Econ.*, **127** (2019), 2143–2185. https://doi.org/10.1086/701807

12. D. B. Rubin, Inference and missing data, *Biometrika*, **63** (1976), 581–592. https://doi.org/10.1093/biomet/63.3.581

13. L. Q. Xia, X. L. Wang, P. X. Zhao, Y. Q. Song, Empirical likelihood for varying coefficient partially nonlinear model with missing responses, *AIMS Math.*, **6** (2021), 7125–7152. https://doi.org/10.3934/math.2021418

14. Z. Q. Liang, Y. Q. Zhou, Model averaging based on weighted generalized method of moments with missing responses, *AIMS Math.*, **8** (2023), 21683–21699. https://doi.org/10.3934/math.20231106

15. H. J. He, M. N. Li, G. M. Deng, Group feature screening for ultrahigh-dimensional data missing at random, *AIMS Math.*, **9** (2024), 4032–4056. https://doi.org/10.3934/math.2024197

16. J. K. Kim, C. L. Yu, A semiparametric estimation of mean functionals with nonignorable missing data, *J. Am. Stat. Assoc.*, **106** (2011), 157–165. https://doi.org/10.1198/jasa.2011.tm10104

17. J. W. Zhao, J. Shao, Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data, *J. Am. Stat. Assoc.*, **110** (2015), 1577–1590. https://doi.org/10.1080/01621459.2014.983234

18. J. Shao, L. Wang, Semiparametric inverse propensity weighting for nonignorable missing data, *Biometrika*, **103** (2016), 175–187. https://doi.org/10.1093/biomet/asv071

19. W. Miao, P. Ding, Z. Geng, Identifiability of normal and normal mixture models with nonignorable missing data, *J. Am. Stat. Assoc.*, **111** (2016), 1673–1683. https://doi.org/10.1080/01621459.2015.1105808

20. J. R. Du, X. Cui, Semiparametric estimation in generalized additive partial linear models with nonignorable nonresponse data, *Stat. Papers*, **65** (2024), 3235–3259. https://doi.org/10.1007/s00362-023-01522-0

21. Y. T. Xiao, Y. F. Shi, Robust estimation for varying-coefficient partially nonlinear model with nonignorable missing response, *AIMS Math.*, **8** (2023), 29849–29871. https://doi.org/10.3934/math.20231526

22. M. L. Cheng, Y. K. Liu, J. Qin, A novel semiparametric approach to nonignorable missing data by catching covariate marginal information, *Scand. J. Stat.*, **52** (2025), 691–709. https://doi.org/10.1111/sjos.12767

23. S. Shetty, Y. Y. Ma, J. W. Zhao, Robust estimation under a semiparametric propensity model for nonignorable missing data, *Electron. J. Stat.*, **19** (2025), 956–981. https://doi.org/10.1214/25-EJS2355

24. Y. H. Zhao, Leaving the countryside: Rural-to-urban migration decisions in China, *Am. Econ. Rev.*, **89** (1999), 281–286. https://doi.org/10.1257/aer.89.2.281

25. X. Meng, J. S. Zhang, The two-tier labor market in urban China: Occupational segregation and wage differentials between urban residents and rural migrants in Shanghai, *J. Comp. Econ.*, **29** (2001), 485–504. https://doi.org/10.1006/jcec.2001.1730

26. Y. Yao, G. S. Chen, R. Salim, X. J. Yu, Schooling returns for migrant workers in China: Estimations from the perspective of the institutional environment in a rural setting, *China Econ. Rev.*, **51** (2018), 240–256. https://doi.org/10.1016/j.chieco.2017.09.008

27. X. Gao, M. S. Li, The urban–rural difference in returns to education in China from the perspective of household registration conversion, *Asia Pac. Educ. Rev.*, **27** (2026), 969–981. https://doi.org/10.1007/s12564-025-10058-x

28. X. Y. Wu, M. Amorim, Return migration and educational investments in children in China: Hukou differences and the role of parental coresidence, *Demogr. Res.*, **54** (2026), 591–644. https://doi.org/10.4054/DemRes.2026.54.20

29. J. R. Du, Y. Li, X. Cui, Identification and estimation of generalized additive partial linear models with nonignorable missing response, *Commun. Math. Stat.*, **12** (2024), 113–156. https://doi.org/10.1007/s40304-022-00284-9

30. E. Flachaire, N. Lustig, A. Vigorito, Underreporting of top incomes and inequality: A comparison of correction methods using simulations and linked survey and tax data, *Rev. Income Wealth*, **69** (2023), 1033–1059. https://doi.org/10.1111/roiw.12618



AIMS Press

© 2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)