ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Wang, Le

Working Paper

Estimating returns to education when the IV sample is selective

IZA Discussion Papers, No. 7103

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Wang, Le (2012) : Estimating returns to education when the IV sample is selective, IZA Discussion Papers, No. 7103, Institute for the Study of Labor (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/69362

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

DISCUSSION PAPER SERIES

IZA DP No. 7103

Estimating Returns to Education when the IV Sample is Selective

Le Wang

December 2012

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Estimating Returns to Education when the IV Sample is Selective

Le Wang
University of New Hampshire,
Harvard University and IZA

Discussion Paper No. 7103
December 2012

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

Any opinions expressed here are those of the author(s) and not those of IZA. Research published in this series may include views on policy, but the institute itself takes no institutional policy positions. The IZA research network is committed to the IZA Guiding Principles of Research Integrity.

The Institute for the Study of Labor (IZA) in Bonn is a local and virtual international research center and a place of communication between science, politics and business. IZA is an independent nonprofit organization supported by Deutsche Post Foundation. The center is associated with the University of Bonn and offers a stimulating research environment through its international network, workshops and conferences, data service, project support, research visits and doctoral program. IZA engages in (i) original and internationally competitive research in all fields of labor economics, (ii) development of policy concepts, and (iii) dissemination of research results and concepts to the interested public.

IZA Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

IZA Discussion Paper No. 7103
December 2012

ABSTRACT

Estimating Returns to Education when the IV Sample is Selective*

The literature estimating returns to education has often utilized spousal education and parental education as instrument variables (IV). However, due to usual survey designs, both IVs are available only for the individuals whose spouse or parents are present in the same household. The IV estimates based on these selective sub-samples may be inconsistent, even when the IVs satisfy the standard assumptions. In this paper, we examine the empirical relevance of this issue in the Chinese context. To our surprise, unlike the selection issue in other situations, this kind of selection does not appear particularly worrisome, suggesting that the previous IV results are robust. In particular, using China Household Income Project 1995 and 2002, we find that correcting for this potential issue has only a modest impact on the magnitude of the standard IV estimates using parental education as an IV, but a negligible impact on those using spousal education. Using the specification tests proposed, we find that these impacts are generally not statistically significant. These results are further confirmed by our analysis using U.S. data. We believe that these results are of use to both policymakers and practitioners.

JEL Classification: J24, I21, C14, C31, P52

Keywords: returns to education, sample selection, instrument variable estimation, Chinese labor market

Corresponding author:

Le Wang
Department of Economics
University of New Hampshire
Durham, NH 03824
USA
E-mail: le.wang@unh.edu

* The author would like to thank Ian Walker, Yingying Dong, Daniel Millimet, and Chunbei Wang for their helpful comments. I also thank Yumin Li for putting together the U.S. data.

1. Introduction

Estimation of the returns to education is often complicated by potential endogeneity and measurement error problems. A typical solution is to employ an instrument variable (IV) approach. The literature has "a long tradition of using family background information" as an IV (see Card (1999) for a selective review of the past studies). Some recent studies (e.g. Chen and Hamori, 2009; Trostel et al., 2002; Arabsheibani and Mussurov, 2007) have utilized spousal education as an IV, while some have utilized parental education (e.g. Flabbi et al., 2008; Gao and Smyth, 2011; Heckman and Li, 2004; Li and Luo, 2004; Wang et al., 2009).[1] While the validity of these two IVs is controversial, the literature has utilized them and argued in different contexts that both variables satisfy the standard requirements for consistency of the IV estimates. For example, using the same data and specifications as in this paper, Wang (forthcoming) employs various statistical techniques to show that spousal education is strongly correlated with own education and at least *plausibly* exogenous in the Chinese context. Using a Bayesian approach and the German Socio-Economic Panel data, Hoogerheide et al. (2010) similarly find that moderate violation of the assumption of perfect instrument validity does not affect the results when parental education is used as an IV.

We want to emphasize that our purpose in this paper is not to argue the standard assumptions are met for two IVs. As noted in Trostel et al. (2002) in this journal, these IVs have been used in many applications, and their validity is "ultimately an empirical question". Instead, we would like to point out a novel problem that has been overlooked in the literature – even if these variables satisfy the standard IV requirements, IV estimations based on either spousal or parental education may still fail to deliver consistent estimates of the returns to education.

To see this, note that due to typical survey designs, information on spousal or parental

[1]Flabbi et al. (2008) use a control function approach to include parental education as an IV, which is actually equivalent to the two stage least squares approach in the linear context (see, e.g., Wooldridge, 2010, p.127).

characteristics is asked only when they are present in the same household. As a result, spousal education is only available for married individuals and parental education is only available for individuals whose parents are present in the same household. The IVs are missing for the rest of the population, which causes a selective estimation sample. There has been ample research work indicating that married workers systematically differ from non-married ones (e.g. Maasoumi et al., 2009) and that living arrangement is an endogenous choice (e.g. Lei et al., 2011). For example, married men may be more likely to possess desirable traits such as responsibility and maturity than single men, which is also a popular explanation for existence of marriage premium (see, e.g. Chiodo and Owyang (2002) for a review of competing theories). Similarly, Lei et al. (2011) (Table 3) show that individual characteristics such as age and education differ greatly across living arrangements. Altogether, this implies that the IVs are missing in a selective way and these samples are not necessarily representative of the population. Our data indeed show that this is true. In such cases, Mogstad and Wiswall (Forthcoming) (hereafter MW) show that IV estimations based on selective sub-samples cannot produce consistent estimates under general conditions.

The theoretical result in MW is not surprising. Past studies of selection issues encountered in other situations (such as female labor supply) have shown that the selection issue generally leads to inconsistent estimates, both theoretically and empirically. Estimation of returns to education is an important research area in which the results are useful not only because of their theoretical implications, but also owing to the fact that they provide a basis for sound policy advice about educational investment. It is thus of paramount importance to examine the empirical relevance of the selection issue discussed above and evaluate the magnitude of its impact on previous IV results.

In this paper, we adopt the robust approach developed in MW to investigate the impact of correcting for the selection bias on the IV estimates of returns to education (based on spousal and parental education) in urban China, while following the literature by assuming the standard IV assumptions are satisfied. To our surprise, unlike in other areas, the selection

issue does not appear to be particularly worrisome in this context. We further assess the robustness of this result using U.S. data and again find the existing IV results hold up, at least for the samples investigated. Confirming the prior results is equally important and useful because assessing the empirical impact of the selection issue is more than academic curiosity and has important policy implications. Moreover, our results indicate that the direction of the selection bias is *a priori* unknown and that if possible, the MW estimator, which is consistent regardless of existence of the selection issue, should always be used. However, the MW approach involving nonparametric estimation is often computationally costly. We thus also believe that such exercises and results in this paper are of use to practitioners – when the MW approach is computationally infeasible in practice (due to too many observations and control variables), our results do increase our confidence in the IV estimations using family backgrounds as IVs.

Even though focusing on estimating the returns to education in China and U.S. as specific examples, we believe that the issue discussed here is prevalent but overlooked in many empirical studies using the IV approach in the field of economics of education, even when the standard IV assumptions are met. For example, the gender of first two children is usually used as an IV for the number of children to identify the effect of family size on children's educational attainment (e.g. de Haan, 2010). However, this IV is only available for the households where parents choose to have two and more children (see MW for more examples). Although our illustration of a (first) application of the MW approach shows no significant impact of non-randomly missing IVs such as spousal and parental education, it is still important to re-evaluate the empirical relevance of other commonly used IVs in the existing literature that may potentially suffer from this issue. Moreover, the MW approach entails nonparametric estimation to construct the IV, but the authors discuss only the cases when the variables are either continuous or discrete, but not both. However, we often encounter *both* continuous *and* discrete control variables in practice. On the methodological front, we discuss how to operationalize the MW approach based on Generalized Kernel

Estimation in this type of situations. Furthermore, we propose two specification tests for the MW approach. We believe that these technical discussions are also of use for applied researchers who are interested in applying the method.

2. Empirical Methodology

To begin, we consider the following (augmented) Mincer equation:

$$ln(w_i) = \beta_1 S_i + X_i'\beta_2 + \epsilon_i \tag{1}$$

where $ln(w_i)$ is the log of earnings for an individual i; S_i is years of schooling. X_i includes all other covariates, $([1, E_i, E_i^2, Minority_i, Age_i, Province_i])$, where E_i is working experience, and E_i^2 working experience squared; $Minority_i$ is a dummy variable equal to one if an individual is a minority, zero otherwise; Age_i is a set of age group dummy variables, and $Province_i$ a set of provincial dummy variables; ϵ_i is the error term as usual, $\mathbb{E}[\epsilon_i] = 0$.[2] [3] β_1 measures the returns to education, the parameter of main interest in this paper.

Note that following MW, β_1 is assumed to be constant and homogeneous across the population, and we abstract from heterogeneous returns to education in this paper. That is, we below consider the results using spousal and parental education as separate approaches to obtain β_1, and we discuss how selection bias affects the results in each case, but do not discuss why the results using different IVs may be different.

We can consistently estimate β_1 by Ordinary Least Squares (OLS) if $cov(\epsilon, S|X) = 0$.

[2]Such variables as tenure, occupation and sector are also available in the survey data, but these variables are potentially endogenous variables that themselves could be determined by schooling. That is, these variables could be the reasons why education affects individuals' earnings. As noted in the literature (e.g. Pearl, 2000; Frolich, 2004; Lee, 2005), controlling for these variables "would block the part of the causal effect that acts through these variables". Therefore, following the literature, we exclude various determinants of earnings such as tenure, occupation and sectors from the estimation. We condition on only exogenous variables here to simplify the interpretations of returns to education.

[3]It can be argued that the actual labor market experience itself could also be endogenous. There has been empirical evidence showing that "the treatment of labor market experience as exogenous does not introduce a significant bias in the IV estimates of the return to education" (see, e.g. Lemieux and Card, 2001). Moreover, as noted in Lemieux and Card (2001), there exists much evidence that earnings are better described by the current specification even though it may be endogenous.

However, if $cov(\epsilon, S|X) \neq 0$, the OLS estimates are generally inconsistent. A typical solution is to employ the IV estimation. The consistency of the IV estimates relies on the fact that the following conditions hold for the IV, Z:

(A1) Existence of First Stage Correlation: $\mathbb{E}[SZ|X] \neq 0$

(A2) Conditional Mean Independence: $\mathbb{E}[\epsilon|X, Z] = \mathbb{E}[\epsilon|X]$

Note that (A1) implies that conditioning on X, the endogenous variable is *sufficiently* correlated with the IV. When the IV is only weakly correlated with the endogenous variable, the IV estimates are biased toward the OLS estimates and inference is not reliable (Bound et al., 1995). (A2) implies that conditioning on X, the IV is exogenous (i.e. mean independent of the error term).

The standard IV estimation assumes the availability of an IV for the full sample. However, such an IV does not necessarily exist. Instead, the literature often relies on IVs such as spousal and parental education that are available only for certain subpopulations; and the estimations are performed on the sample for which the IV is available. Implicitly, these IV estimations are based on the following assumptions:

(A1') Existence of First Stage Correlation: $\mathbb{E}[SZ|X, D = 1] \neq 0$

(A2') Conditional Mean Independence: $\mathbb{E}[\epsilon|X, Z, D = 1] = \mathbb{E}[\epsilon|X, D = 1]$

where D is equal to one if the IV is not missing and zero otherwise. In other words, (A1) and (A2) are now stated based on the sub-samples where the IV is not missing. However, Mogstad and Wiswall (Forthcoming, Proposition 1) show that if the IV is not missing randomly and D is correlated with the error term ϵ, the IV estimates are generally inconsistent estimates of β_1.

In this paper, we focus on this particular issue while maintaining that the standard IV assumptions (A1') and (A2') are met. One solution is to construct a full-sample IV. An example is proposed by Angrist et al. (2010) as follows:

$$Z_{LP} = \begin{cases} Z - X'\lambda, & \text{if } D = 1 \\ 0, & \text{if } D = 0 \end{cases}$$

where $X'\lambda$ is the linear projection of Z on X using the sub-sample with non-missing IV. However, even though this full-sample IV approach *may* produce more efficient IV estimates asymptotically, Mogstad and Wiswall (Forthcoming, Lemma 1.) show that it is actually equivalent to the standard IV based on the non-missing IV sample, and therefore it is also *generally* inconsistent under (A1') and (A2').[4]

To solve this problem, MW propose a robust IV estimator based on the following full-sample IV

$$Z_{Robust} = \begin{cases} Z - \mathbb{E}[Z|X, D = 1], & \text{if } D = 1 \\ 0, & \text{if } D = 0 \end{cases}$$

Unlike the Angrist et al. (2010) approach, the MW approach is based on the true conditional expectation of Z on X, instead of the linear projection of Z on X. MW show that under (A1') and (A2'), this IV estimator produces consistent estimates of β_1.

Notice that if $\mathbb{E}[Z|X, D = 1]$ is linear in X (i.e. $\mathbb{E}[Z|X, D = 1] = X'\lambda$), the MW estimator nests both the standard IV and the Angrist et al. (2010) IV. The equivalence between these approaches in this special case implies that the standard IV and the Angrist et al. (2010) approaches are consistent when $\mathbb{E}[Z|X, D = 1]$ is linear in X; in other words, the linearity of $\mathbb{E}[Z|X, D = 1]$ is a *sufficient* condition for the consistency of both the standard IV and the Angrist et al. (2010) approach (see MW for other two sufficients conditions).

The underlying functional form could depart from linearity for many reasons. For exam-

[4]The equivalence between the standard IV and the Angrist et al approach is probably better understood when the sample analogs of both estimators are explicitly written out. The standard IV estimator is given by $[\sum_{i=1}^{N_1}(Z_i S_i)]^{-1}[\sum_{i=1}^{N_1}(Z_i ln(w_i))]$, and the Angrist et al IV estimator is given by $[\sum_{i=1}^{N}(Z_{LP,i} S_i)]^{-1}[\sum_{i=1}^{N}(Z_{LP,i} ln(w_i))]$ (see MW for detailed discussions). Recall that the only difference between Z_i and $Z_{LP,i}$ is that $Z_{LP,i}$ is zeros for those observations with missing Z_i. Adding more zeros to both denominator and numerator in the Angrist et al estimator would not change the value. It is essentially the standard IV estimator.

ple, minority students generally receive preferential treatments such as lower admission score for college entrance from the government, but the preferential treatments vary by the extent of the concentration of ethnic minority communities where they live (Wang, 2007). And there is indeed a very unequal distribution of minority communities across provinces. Moreover, the extent of the preferential treatments received by minority students also changes over time. Altogether, these facts suggest that there could be very complicated interactions between minority status, provincial dummies, and age cohort dummies (capturing time effects) in determining education. In addition, there were several nation-wide events such as Cultural Revolution in China that affected the educational attainment of then school-aged children and that the extent of the impacts of these events differs across provinces (see, e.g. Giles et al., 2004). This also suggests that there could be a potential interaction between geographic variable and age dummies, which is also ignored in a linear projection. Our nonparametric method allows for even more complicated forms of interactions between all control variables.

2.1. Practical Issues

Practical implementation of the robust IV estimator requires nonparametric estimation of $\mathbb{E}[Z|X, D = 1]$. In our case, X contains a continuous variable, (experience, E), two unordered categorical variables (minority status, $Minority$ and province, $Province$), and a ordered categorical variable, (age group, Age). Typical kernel methods, however, do not allow for smoothing categorical variables that are generally encountered in practice.[5] To overcome this issue, we adopt a variant of the local-linear least-squares (LLLS) estimator based on Generalized Kernel Estimation (see, e.g. Li and Racine, 2004; Racine and Li, 2004

[5]In their paper, MW employ a parametric regression with polynomial terms up to fourth order as an example to approximate the underlying conditional expectation. Given the specific Monte Carlo simulation design used by the authors (the conditional expectation is a quadratic function of X), this approximation works well by construction. However, this is still a parametric approach, instead of a nonparametric one. Despite computationally easier to implement, polynomial regressions using all the observations for estimations are still *global* functions and cannot capture local anomalies. It is thus generally not consistent and not a good approximation to the unknown function. Moreover, higher-order polynomial regressions are shown to have "undesirable nonlocal effects", which kernel estimators do not have (Magee, 1998).

for more detailed discussions). The nonparametric regression model is given by:

$$Z_i = \mathbb{E}[Z|X, D = 1] + u_i = m(X_i) + u_i$$

where $m(\cdot)$ is the unknown conditional expectation, a smooth regression function with X_i, and u_i is the error term as usual. To ease the discussion, we arrange the variables with respect to their types: the first q_c are continuous variables, X^c, the next q_u are unordered, X^u, and the last q_o are ordered, X^o. Note that q^c, q^u, q^o represent the dimension of each type of variable, respectively. In our case, $q_c = 1, q_u = 2, q_o = 1$. Notice that taking a first-order expansion of (2.1) with respect to X yields

$$\begin{aligned} Z_i &= m(X_i) + u_i \\ &\approx m(X) + (X_i^c - X^c)m'(X) + u_i \end{aligned} \tag{2}$$

Treating $m(X)$ and $m'(X)$ as parameters to be estimated, we have the following model:

$$\begin{aligned} Z_i &= m(X) + (X_i^c - X^c)m'(X) + u_i \\ &= \delta_1 + (X_i^c - X^c)\delta_2 + u_i \end{aligned}$$

The LLLS estimator, ($\delta = [\delta_1, \delta_2]'$), minimizes the following objective function:

$$\min_{\delta_1, \delta_2} \sum (Z_i - \delta_1 - (X_i^c - X^c)\delta_2)^2 K(\frac{X_i - X}{h})$$

where $K(\cdot)$ is a generalized kernel function and h a bandwidth vector. Notice that the minimization problem is similar to generalized least squares problem but with the weight $K(\frac{X_i - X}{h})$. We have a closed-form solution for the LLLS estimator:

$$\delta = [\sum K(\frac{X_i - X}{h})\overline{X_i}'\overline{X_i}]^{-1}[\sum K(\frac{X_i - X}{h})\overline{X_i}Z_i]$$

where $\overline{X_i} = (1, [(X_i^c - X^c)]')$.

The first practical issue of implementation of LLLS is concerned with the choice of the kernel function, $K(\cdot)$. Unlike the conventional nonparametric estimation, for which the kernel function is designed for continuous variables only (thus usually a popular density function), the Generalized Kernel Estimation permits both continuous and discrete variables. In particular, the generalized kernel is the product of different kernel functions specifically designed for each type as follows (recall that $X_i = [X_i^{c'}, X_i^{u'}, X_i^{o'}] = [(X_{1i}^c, \ldots, X_{q^c i}^c)', (X_{1i}^u, \ldots, X_{q^u i}^u)', (X_{i1}^o, \ldots, X_{q^o i}^o)']$.):

$$K(\frac{X_i - X}{h}) = \Pi_{s=1}^{q_c} k^c(X_{si}^c, X_s^c, h_s^c) \Pi_{s=1}^{q_u} k^u(X_{si}^u, X_s^u, h_s^u) \Pi_{s=1}^{q_o} k^o(X_{si}^o, X_s^o, h_s^o)$$

where the kernel function for continuous variables is given by

$$k^c(X_{si}^c, X_s^c, h_s^c) = \frac{1}{\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{X_i^c - X^c}{h^c})^2\}$$

The kernel function for unordered discrete variables (Aitchison and Aitken, 1976) is given by

$$k^u(X_{si}^u, X_u^c, h_s^u) = \begin{cases} 1 - h_s^u & \text{if } X_{si}^u = X_s^u \\ \frac{h_s^u}{d_s - 1} & \text{otherwise} \end{cases}$$

where d_s is the number of unique values a variable can take. The kernel function for ordered discrete variables is given by:

$$k^o(X_{si}^o, X_s^o, h_s^o) = \begin{pmatrix} d_s \\ j \end{pmatrix} (h_s^o)^j (1 - h_s^o)^{d_s - j} \quad \text{when } |X_{si}^o - X_s^o| = j$$

Notice that the rate of convergence of the estimator depends solely on the number of continuous variables, and the number of discrete variables does not add to the "curse of dimensionality" problem (Henderson, 2009). It is widely believed in the literature that the choice of kernel functions matters little in the nonparametric estimation (see, e.g. textbook

discussions in Hardle (1990) and Li and Racine (2007)).

However, selection of bandwidths is often considered to be the most salient factor in the nonparametric estimations. The second practical issue is thus concerned with selection of an optimal bandwidth vector. To see the issue, consider (2.1). Given the choice of a kernel function, the value of h determines the size of the neighborhood around a point X, and the observations within this neighborhood are given more weights in estimations. A very small bandwidth means a very small neighborhood and very few points will be given weights in estimations, resulting in estimates with smaller bias yet less precision. On the other hand, a large bandwidth means a large neighborhood and more points will be utilized in estimations, resulting in estimates with larger bias yet more precision. The key issue is to balance the trade-off between bias and precision. To avoid any arbitrariness in our selection, we opt for a popular choice of optimal bandwidth selection method – least square cross validation (LSCV). Stone (1984) shows that this method is asymptotically optimal "in the sense of minimizing the estimation integrated square error" (Li and Racine, 2007, p.18).

Another useful feature of the LSCV procedure, among others, is its ability to detect whether a continuous variable enters the function linearly in the LLLS case (Hall et al., 2007). In (2.1), a very large bandwidth ($h \to \infty$) (thus $K(\cdot) \to K(0)$, a constant) implies each observation is given an equal weight in estimation, which makes the original minimization problem essentially an OLS problem over the whole support. In this case, the true functional form *is* linear. Recall that the linearity of $\mathbb{E}[Z|X, D = 1]$ is a sufficient condition for the consistency of the standard IV estimator. The LSCV bandwidth is thus particularly informative of the source of the (in)consistency of the conventional IV estimator by identifying the linearity of the conditional expectation function in the continuous variable(s).

2.2. Specification Tests

We have thus far discussed the practical details of the implementation of the MW approach. It would also be useful to provide a formal assessment of whether the MW approach is preferred to the standard IV or Angrist et al. approach and is needed in practice. Here,

we propose two statistical tests for this purpose: the first test is an indirect test, which tests whether the parametric functional form used in the Angrist et al approach is misspecified; and the second test is a direct comparison of these two estimators, which determines whether or not the differences are indeed statistically significant. An ideal test should be able to detect the significant difference between these two approaches and thus be informative of the severity of the selection issue.

2.2.1. Test of Functional Form

As discussed above, the difference between the Angrist et al. approach and the MW approach lies in the assumed functional form of $\mathbb{E}[Z|X, D = 1]$. If the parametric form is misspecified (and thus the Angrist et al. estimator is inconsistent), then the MW approach (relying on nonparametric estimation and thus free of misspecification of functional form) is preferred. To test the correctness of the parametric form, we employ the specification test recently developed in Hsiao et al. (2007). The null hypothesis is that a parametric model is correctly specified ($H_0 : \Pr[\mathbb{E}[Z|X, D = 1] = X'\lambda] = 1$), and the alternative is that the model is misspecified ($H_A : \Pr[\mathbb{E}[Z|X, D = 1] = X'\lambda] < 1$). Notice that the alternative hypothesis includes all other alternatives than the specification in the null hypothesis. Therefore, unlike most popular parametric tests that lack power in certain directions, this test is a consistent test. Moreover, this test, again, permits both continuous and discrete data, which is usually not allowed in the existing kernel-based tests. The test statistic is given by:

$$J_N = N(h_1^c, \ldots, h_{q_c}^c)^{\frac{1}{2}} \widehat{I_N} / \sigma_\alpha \sim \mathcal{N}(0, 1)$$

where $\widehat{I_N} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \widehat{u}_i \widehat{u}_j K(\cdot)$; $\sigma_\alpha = \frac{2(h_1^c, \ldots, h_{q_c}^c)}{N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \widehat{u}_i^2 \widehat{u}_j^2 K(\cdot)^2$. Following the notations above, $K(\cdot)$ is the product kernel, $h_1^c, \ldots, h_{q_c}^c$ are the optimal bandwidths for the continuous variables, q_c is the number of continuous variables. As noticed in Hsiao et al. (2007), the asymptotic approximation usually performs poorly in finite samples, therefore we conduct inference based on a bootstrap procedure to improve the finite properties of

the test statistic. This test is useful since rejection of the null hypothesis implies that there would be a difference between the Angrist et al. approach and MW approach.

2.2.2. Hausman Type of Test

The test of functional form, though useful, does not necessarily provide any information regarding the magnitude of the difference and whether such difference is statistically significant. Thus, we provide an alternative test to further examine this issue. The test is based on the principle that was first developed by Hausman (1978). The idea is as follows. Under the null hypothesis (a parametric model is correctly specified), both the Angrist et al. estimator ($\widehat{\beta_1^{angrist}}$) and MW estimator ($\widehat{\beta_1^{MW}}$) are consistent, which implies the difference $\widehat{q} = \widehat{\beta_1^{angrist}} - \widehat{\beta_1^{MW}}$ tends to zero. Under the alternative hypothesis (a parametric model is misspecified), only $\widehat{\beta_1^{MW}}$ is consistent, and $\widehat{q}$ differs from zero. The test is to detect whether $\widehat{q}$ departs from zero. Notice that under the null hypothesis, both estimators are normally distributed with variances $V^{angrist}$ and V^{MW}, respectively. It follows that $\widehat{q}$ is also normally distributed with variance V^q, and with this result, we could easily construct the Hausman type of test. Moreover, if $\widehat{\beta_1^{angrist}}$ is asymptotically efficient, asymptotic variance is further simplified $V^q = V^{MW} - V^{angrist}$. However, as noted in Wooldridge (2010), the assumption of asymptotic efficiency is only needed for simplification of the asymptotic variance, and "it is essentially irrelevant" when applying the Hausman's (1978) principle. To relax the assumption of asymptotic efficiency and allow for more complicated variance structure, we also conduct inference based on a bootstrap procedure. We believe that the Hausman test is of more empirical use, since we just need to calculate the parametric estimates, which are easy to obtain, and then contrast them with the MW estimates.

3. Data

The data are obtained from the the urban sample of the China Household Income Project 1995 and 2002. The data have been used widely in the literature; we thus provide limited discussions here (see e.g. Gustafsson et al., 2008). Interested readers are referred to Wang

(forthcoming) for detailed explanations of sample restrictions and choices of variables, which we follow here. The outcome variable is annual wages, measured by the total individual salary (equal to the sum of regular salary, bonuses and subsidies, allowance for temporarily laid off workers, and other income from the work unit) plus income from private enterprise.[6] The independent variable is years of schooling. Also included in estimations are working experience, minority status, a set of age group dummy variables and provincial dummy variables. We restrict the sample to individuals aged 16 to 65. The legal minimum working age is set at 16 in China, and the announced maximum retirement age is 65. Summary statistics of the data used in the analysis are reported in Table (1).

To gain a sense of potential severity of the selection issue, we also report in Table (2) the percent of sample that have spouses and the percent that have parents living at home. We first notice that utilization of the IVs reduces the sample size. However, the size of reduction and the pattern of changes over time are different across IVs. In the case of spousal IV, we notice that whereas the percent of married women slightly decreased over time, the percent of married men actually increased. Regardless, marriage is still prevalent in urban China – at least 80 percent of the samples for both men and women are married. By contrast, the number of individuals living with their parents is relatively small, comprising roughly 10 percent of the samples. The number of observations is 795 for men and 614 for women in 1995, and it decreases further to 615 for men and to 449 for women in 2002. This pattern is consistent with the increasing trend of Chinese elderly living alone.

The above results indicate that use of either spousal education or parental education as an

[6]In the Chinese context, it is difficult to separate different income variables; the distinction between different types of income from the work unit (*Danwei*) is often blurry in different units. Separating them may create additional measurement errors. Moreover, private enterprise income is rare especially in the early stages of economic reforms in China, accounting for only a small fraction of individual incomes (Shu and Bian, 2003). Also, employees in private business and individual entrepreneurs (*getihu*) are substantially under-sampled (Shu and Bian, 2002). Moreover, as mentioned above, nonparametric estimations are computationally expensive. Given the focus of this paper on assessing the robustness of the previous results, we follow the literature by using the total annual income as our dependent variable. However, this issue is worth further explorations to assess the returns to education for different income measures, and we leave it for future research.

IV precludes certain individuals, resulting selective samples. And the selection issue is likely more severe with using parental education as an IV due to a drastic reduction in the sample sizes. The next question is whether the selection is random. To answer this question, we examine the summary statistics of observable characteristics used in our analysis by marital status and by living arrangements. These results are reported in Tables (3) and (4). As we can clearly see, married workers are different from their single counterparts. For example, in 1995, single men have less working experiences (-14.129) and are much younger than married men. These differences are statistically significant at at least $p \leq 0.001$. We observe a similar pattern across living arrangement. For example, workers with parents living at home are more likely to be younger and have more education and experiences, consistent with the results reported in Lei et al. (2011) using CHARLS data. These differences confirm that the selection is not random. More worrisome, the selection may be partly due to nonrandomness along unobservable dimensions such as ability, thereby biasing the estimations that cannot be eliminated simply by including observable characteristics in estimations. Thus, we turn to the MW approach that could help to address this issue.

4. Results

Recall that the difference between the Angrist et al approach and the MW approach arises from the non-linearity of the underlying conditional expectation, $\mathbb{E}[Z|X, D = 1]$. In the LLLS case, the size of the bandwidth for a continuous variable provides useful information about the linearity – a large bandwidth (tending to infinity) determines that the variable enters linearly in $\mathbb{E}[Z|X, D = 1]$. In practice, a rule of thumb is that when using a Gaussian kernel function (as in our analysis), if the bandwidth on a continuous variable is larger than two standard deviations of the variable, the conditional expectation is linear in this variable (Hall et al., 2007; Henderson et al., Forthcoming). The LSCV results are presented in Table (5). Indeed, we find that none of the bandwidths on the continuous variable, experience (E) exceeds two standard deviations of it (reported in parentheses). This result implies that the

conditional expectation $\mathbb{E}[Z|X, D = 1]$ is at least not linear in experience and the nonlinear form could be potentially highly complicated, which *may* render the standard IV estimates inconsistent; recall linearity is a sufficient but not necessary condition.

Thus, we turn to a more formal assessment of reasonableness of the linearity assumption and conduct the Hsiao et al. (2007) test for functional form. The results are reported in Table (6). When using spousal education as an IV, we easily reject the null of a correctly specified linear model at $p \leq 0.0001$ level, consistent with the interpretation of the bandwidths discussed above. This result implies that the nonparametric specification is strongly preferable and failure to correctly specify the functional form could lead to inconsistent IV estimates. On the other hand, when using parental education as an IV, we *fail* to reject the parametric functional form at the conventional level, except for female in 2002. This result suggests that the selection issue may not be too severe and the traditional IV and Angrist et al. approaches may be preferable on efficiency grounds. However, these results do not directly speak to the size of the actual impact of selection on estimates and whether the impact is indeed statistically significant.

We therefore turn to the actual estimates obtained from different approaches. The results are presented in Table (7). Column (1) reports the OLS estimates of the returns to an additional year of schooling. The estimates for men are .0358 in 1995 and 0.0662 in 2002, while those for women are 0.0562 in 1995 and 0.081 in 2002. Even though the OLS estimates imply positive returns to education, these figures are generally smaller relative to the international standard (roughly 10 percent reported in Psacharopoulos and Patrinos, 2004). These OLS results are similar to those reported in the literature (e.g. Li, 2003). Correcting for the potential endogeneity and measurement error problem, the IV estimates in Columns (2) and (3) imply larger returns to education compared to the OLS estimates, regardless of which IV is used. For example, when using parental education as IV, the implied returns to an additional year of schooling are about 11 percent for men and 14 percent for women in 2002. This result is consistent with the literature that generally finds the OLS approach

underestimates the returns to education (Card, 1999).

Columns (4) and (5) report the IV estimates using the Angrist et al. (2010) approach. As expected, the estimates are identical to the standard IV estimates. However, pooling the samples together to construct a full-sample IV does not improve the precision of the estimates. The standard errors of the Angrist et al IV estimates are even slightly larger than those of the standard IV estimates. This difference between two standard errors is not statistically significant. Lack of efficiency gains may be explained by the fact that pooling together the whole sample simply adds more zeros to the IV, which does not necessarily increase useful variation for identification of the parameter. However, we should see more definitive efficiency gains for other variables in this case. To verify this conjecture, we report the standard errors of all control variables for both approaches in Table (A1). We find that differences in standard errors of the conventional and Angrist et al. IV estimates of control variables are indeed positive, suggesting that pooling the whole sample (the Angrist et al. approach) produces smaller standard errors and thus the efficiency gains!

Both the bandwidths results and the results of functional forms indicate the potential severity of the selection problem among studies using spousal and parental education as IVs, especially in more recent years. To address this problem, the results using the robust IV are reported in Columns (6) and (7). We observe that the estimates using both IVs change, but differently. In particular, when using spousal education as an IV, correcting for the non-random sample problem generally produces smaller IV estimates, except in the case for men in 2002. The magnitude of the difference between the robust and standard IV estimates is relatively small, ranging from -.23 (=(0.0704-0.0727)*100) and .06 (=(0.0888-0.0882)*100) percentage points. By contrast, when using parental education as IV, correcting for the non-random sample problem has a larger impact, producing larger IV estimates. Compared to the case of spousal education, the magnitude of the difference between the robust and standard IV estimates, when using parental education, is much larger. For example, the difference is 1.26 (=(0.12-0.1074)*100) percentage points for men and 1.21(=(0.1504-0.1383)*100) per-

centage points for women in 2002. These differences imply that the magnitude of the bias due to the non-randomly missing IV is sizable – about 10.5 percent for men and 8 percent for women. Another interesting finding is that there does not appear to be a discernible pattern of the direction of the selection bias. This is primarily because the selection mechanism (and the nonparametric conditional expectation) is unknown and may differ substantially across IVs and samples (also evident in the results of functional form tests). Given the fact that the MW estimator is always consistent, we recommend that it should always be used in practice, if possible.

In sum, we find that the IV estimates using spousal education suffer much less from the selection bias in terms of the magnitude than do those using parental education, especially in recent years. This result may be explained by the fact that marriage is still prevalent and divorce rate is low in China, compared to other developed countries. For example, the divorce rates, according to China Internet Information Center[7], are 0.18 percent in both 1995 and 2002. Despite these results, the Hausman test results, however, suggest that these differences are not statistically significant at the conventional levels, except for the sample of female workers in 1995. The results remain unchanged when we take into account potential unknown variance structure using the bootstrap procedure. Our results suggest that the selection issue exists but is not as severe as in other contexts. A note of caution is in order concerning this test. As noted in Hausman (1978), power considerations of this type of test are important in practice, and a test may fail to reject the null hypothesis may be simply due to lack of power (which in turn depends such factors as sample size). Thus, further explorative analysis is warranted of finite sample properties of this test.

4.1. Further Results Excluding Age Dummies

We include age dummies in our specification to control for several nation-wide events in the past. Here, we also examine what impact the correlation between age and experience

[7] `http://www.china.org.cn/english/en-sz2005/sh/biao/23-42.htm`

may have on our results.

To capture the correlation between age dummies and education and experience, we report the R-squared values from regressions of schooling or experience on age group dummies (Table B1). As we can see, age and experience is correlated, but not perfectly. Specifically, R^2 ranges from .6 to .8. Given the focus in this paper is on estimations of the returns to education, the correlation between age group and experience should matter little for the results. What really matters is the correlation between schooling and age groups. And indeed, we find statistically significant relationship, but R^2 is smaller as opposed to the case of experience. To assess the impact of the correlation between age and other variables, we exclude age group dummies and repeat all our analyses. The results are reported in Tables (B2) and (B3) in the Appendix. We find that the results remain mostly unchanged.

5. International Evidence

In the Chinese context, there is only modest evidence to suggest that selection may lead to inconsistent estimates, especially when utilizing parental education as an IV. However, this result is not necessarily going to hold true universally. Given the potential importance of this topic, we feel it imperative that we should also assess the impact of selection on the results reported in the previous important studies using data from other countries. We do this by first replicating the results in Trostel et al. (2002) and then repeat our analysis using the same data. The data are from International Social Survey Programme data (U.S. sample), 1985 – 1995. To focus on the selection issue, we follow their sample selection criteria and model specifications. In particular, we utilize in estimations log of hourly wages as dependent variable, years of schooling as measure of education, and such control variables as year fixed effects, marital status, union status, age and age squared (when appropriate). The sample consists of only employed workers aged 21-59.

The results are reported in Tables (C1)-(C2). Examining the results for specification tests of functional forms, we notice that for both the sample of men using spousal education

and the sample of women using parental education as IVs, the null hypothesis of correctly specified parametric form is rejected at least at $p \leq 0.10$ levels. But we cannot reject the null hypothesis for the sample of men using parental education and the sample of women using spousal education as IVs. These results suggest that selection issue could potentially bias the IV estimates for certain samples and that there may be room for improvement by addressing the selection issue.

Turning to actual estimates, we are able to successfully replicate the standard IV results in Trostel et al. (2002) (Tables 5 and 6, first row in their paper). The estimates in U.S. are larger than the estimates in China in 1995, but close to those in China in 2002. The drastic increase in the returns to education over time in China, catching up with U.S., seems to suggest that the ongoing economic transition during this period increase the value of education, which could be due to either increased demand for highly-educated workers or more appropriately valuing human capital at its market rate (Wang, forthcoming). When addressing the nonrandom selection issue, we find the results change only slightly, and the selection issue does not appear to bias the results much, even though the tests results of functional form suggest such possibility. Confirming the casual observation, the Hausman test results suggest that we cannot reject the null hypothesis of the difference between the two methods being zero at the conventional levels.

6. Conclusions

This paper first estimates the returns to education in China, adjusting for the selection bias due to non-randomly missing IV. As noted in Fleisher et al. (forthcoming), due to the difficulty in finding a valid instrument variable, very few studies estimating returns to education have attempted to address the endogeneity and measurement error problems in the Chinese context. A good IV candidate is much needed for future studies. Our results suggest that the non-random selection does not bias the IV results much (at least the effect is not statistically significant). This result is important since it suggests family backgrounds

such as spousal and parental educations could be used as IVs for own education provided that the standard IV assumptions are met. For example, *together* with the evidence in Wang (forthcoming) (verifying that the standard IV assumptions are indeed met for spousal education), our results thus show some supporting evidence for the use of spousal education as an IV in the Chinese context. We also further assess the robustness of our results using the U.S. samples. We similarly find evidence that the selection issue exists but does not bias the results significantly. While these results cannot necessarily be generalized to other countries, they do increase our confidence in the IV results using family backgrounds as IVs, especially when implementation of the MW method is not computationally viable.

References

Aitchison, J. and C.G.G. Aitken. 1976. "Multivariate Binary Discrimination by Kernel Method." *Biometrika* 63:413–420.

Angrist, J., V. Lavy, and A. Schlosser. 2010. "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *Journal of Labor Economics* 28:773–824.

Arabsheibani, G.R. and A. Mussurov. 2007. "Returns to Schooling in Kazakhstan." *Economics of Transition* 15:341–364.

Bound, J., D. A. Jaeger, and R. M. Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association,* 90:443–450.

Card, D. 1999. *Handbook of Labor Economics*, volume 3, chapter The causal effect of education on earnings. Amsterdam: North Holland.

Chen, G. and Shigeyuki Hamori. 2009. "Economic returns to schooling in urban China: OLS and the instrumental variables approach." *China Econoimc Review* 20:143–152.

Chiodo, A. and M. Owyang. 2002. "For Love or Money: Why Married Men Make More." *The Regional Economist* April:10–11.

de Haan, M. 2010. "Birth Order, Family Size and Educational Attainment." *Economics of Education Review* 29:576–588.

Flabbi, L., S. Paternostro, and E.R. Tiongson. 2008. "Returns to Education in the Economics Transition: A Systematic Assessment Using Comparable DAta." *Economics of Education Review* 27:724–740.

Fleisher, B.M., Y. Hu, and H. Li. forthcoming. "Economic Transition, Higher Education and Worker Productivity in China." *Journal of Development Economics* .

Frolich, M. 2004. "Programme Evaluation with Multiple Treatments." *Journal of Economic Surveys* 18:181–224.

Gao, W. and R. Smyth. 2011. "Economic Returns to Speaking 'Standard Mandarin' Among Migrants in China's Urban Labour Market." *Economics of Education Review* 30:342–352.

Giles, J., A. Park, and J. Zhang. 2004. "The Great Proletarian Cultural Revolution, Discruption to Education, and Returns to Schooling in Urban China." *Working Paper, Department of Economics, University of Michigan* .

Gustafsson, B.A., L. Shi, and T. Sicular. 2008. *Inequality and Public Policy in China*. Cambridge, UK: Cambridge University Press.

Hall, P., Q. Li, and J.S. Racine. 2007. "Nonparametric Estimation of Regression Functions in the Presence of Irrelevant Regressors." *Review of Economics and Statistics* 89:784–89.

Hardle, W. 1990. *Applied Nonparametric Regression*. Cambridge: Cambridge University Press.

Hausman, J.A. 1978. "specification Tests in Econometrics." *Econometrica* 46:1251–1271.

Heckman, J.J. and X. Li. 2004. "Selection bias, comparative advantage and heterogeneous returns to education: evidence from China in 2000." *Pacific Economic Review* 9:155–171.

Henderson, D.J. 2009. "A Non-parametric Examination of Capital-kill Complementarity." *Oxford Bulletin of Economics and Statistics* 71:519–538.

Henderson, D.J., C. Papageorigiou, and C.F. Parmeter. Forthcoming. "Growth Empirics without Parameters." *Economic Journal* .

Hoogerheide, L., J.H. Block, and R. Thurik. 2010. "Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis." *Tinbergen Institute Discussion Paper* 0753.

Hsiao, C., Q. Li, and J. Racine. 2007. "A Consistent Model Specification Test with Mixed Categorical and Continuous Data." *Journal of Econometrics* 140:802–826.

Lee, M.J. 2005. *Micro-Econometrics for Policy, Program, and Treatment Effects*. Oxford University Press.

Lei, X., J. Strauss, M. Tian, and Y. Zhao. 2011. "Living Arrangements of the Elderly in China Evidence from CHARLS." *Rand Working Paper Series* WR-866.

Lemieux, T. and D. Card. 2001. "Education, Earnings, and the 'Canadian G.I. Bill'." *Canadian Journal of Economics* 34:313–344.

Li, H. 2003. "Economic Transition and Returns to Education in China." *Economics of Education Review* 22:317–328.

Li, H. and Y. Luo. 2004. "Reporting errors, ability heterogeneity, and returns to schooling in China." *Pacific Economic Review* 9:191–207.

Li, Q. and J. Racine. 2004. "Cross-Validated Local Linear Nonparametric Regression." *Statistica Sinica* 14:485–512.

Li, Q. and J.S. Racine. 2007. *Nonparametric Econometrics: Theory and Practice*. New Jersey: Princeton University Press.

Maasoumi, E., D.L. Millimet, and D. Sarkar. 2009. "Who Benefits from Marriage." *Oxford Bulletin of Economics and Statistics* 71:1–33.

Magee, L. 1998. "Nonlocal Behavior in Polynomial Regressions." *The American Statistician* 52:20–22.

Mogstad, M. and M. Wiswall. Forthcoming. "Instrumental Variables Estimation with Partially Missing Instruments." *Economics Letters* .

Pearl, J. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press, 1st edition.

Psacharopoulos, G. and H.A. Patrinos. 2004. "Returns to Investiment in Education: A Further Update." *Education Economics* 12:111–134.

Racine, J. and Q. Li. 2004. "Nonparametric Estimation of Regression Functions with Both Categorical and Continuous Data." *Journal of Econometrics* 119:99–130.

Stone, C.J. 1984. "An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates." *Annals of Statistics* 12:1285–1297.

Trostel, P., I. Walker, and P. Woolley. 2002. "Estimates of the economic return to schooling for 28 countries." *Labour Economics* 9:1–16.

Wang, L. forthcoming. "How Does Education Affect the Earnings Distribution in Urban China?" *Oxford Bulletin of Economics and Statistics* .

Wang, T. 2007. "Preferential policies for ethnic minority students in China's college/university admission." *Asian Ethnicity* 8:149 – 163.

Wang, X., B.M. Fleisher, H. Li, and S. Li. 2009. "Access to Higher Education and Inequality: The Chinese Experiment." *Unpublished Manuscrpit* .

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT University Press, 2e edition.

Table 1: Summary Statistics

Variables	1995				2002			
	Male		Female		Male		Female	
	Mean (S.D.)	N	Mean (S.D.)	N	Mean (S.D.)	N	Mean (S.D.)	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Dependent Variable								
Log Annual Wages	8.584	6274	8.365	5642	9.298	5618	9.067	4510
	(0.586)		(0.687)		(0.643)		(0.679)	
Independent Variable								
Years of Schooling	11.097	6274	10.421	5642	11.471	5618	11.347	4510
	(2.990)		(2.846)		(3.064)		(2.898)	
Spouses Years of Schooling	10.042	5125	11.089	4597	10.6	4702	11.534	3610
	(2.970)		(3.003)		(3.034)		(3.105)	
Parent's Years of Schooling	10.535	795	10.422	614	10.024	615	10.241	449
	(3.710)		(3.824)		(3.068)		(3.223)	
Minority (Yes = 1)	0.042	6274	0.044	5642	0.04	5618	0.043	4510
	(0.201)		(0.206)		(0.197)		(0.202)	
Experience	20.987	6274	17.867	5642	21.691	5618	18.229	4510
	(10.293)		(8.744)		(9.875)		(9.080)	
Experience Squared	546.384	6274	395.673	5642	567.98	5618	414.714	4510
	(444.77)		(318.404)		(415.831)		(332.947)	
Age 16 - 25	0.101	6274	0.117	5642	0.057	5618	0.082	4510
	(0.301)		(0.321)		(0.232)		(0.274)	
Age 26 - 35	0.231	6274	0.283	5642	0.201	5618	0.261	4510
	(0.421)		(0.450)		(0.401)		(0.439)	
Age 36 - 45	0.366	6274	0.428	5642	0.349	5618	0.414	4510
	(0.482)		(0.495)		(0.477)		(0.493)	
Age 46 - 55	0.227	6274	0.16	5642	0.336	5618	0.229	4510
	(0.419)		(0.367)		(0.473)		(0.420)	
Age 56 - 65	0.075	6274	0.012	5642	0.056	5618	0.015	4510
	(0.263)		(0.108)		(0.231)		(0.121)	

[1] Notes: Standard deviations in parentheses. Samples are as follows: 1995 and 2002 China Household Income Project (CHIP). Provincial dummies are also included.

Table 2: Summary Statistics (Number of Observations by Marital Status and Living Arrangement)

Gender	Year	Full Sample (1)	Married (2)	Living with Parents (3)
Panel A: Male	1995			
	No. of Individuals	6274	5125	795
	Percentage		0.817	0.127
	2002			
	No. of Individuals	5618	4702	615
	Percentage		0.837	0.109
Panel B. Female	1995			
	No. of Individuals	5642	4597	614
	Percentage		0.815	0.109
	2002			
	No. of Individuals	4510	3610	449
	Percentage		0.800	0.100

[1] Notes: Samples are as follows: 1995 and 2002 China Household Income Project (CHIP).

Table 3: Summary Statistics by Marital Status

Variables	1995						2002					
	Male			Female			Male			Female		
	Married	Single	Diff	Married	Single	Diff	Married	Single	Diff	Married	Single	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent Variable												
Log Annual Wages	8.666	8.219	-0.447	8.432	8.073	-0.359	9.362	8.967	-0.395	9.112	8.882	-0.230
	(0.518)	(0.717)	0.000	(0.633)	(0.826)	0.000	(0.599)	(0.753)	0.000	(0.643)	(0.780)	0.000
Independent Variable												
Years of Schooling	11.083	11.159	0.076	10.292	10.988	0.696	11.322	12.239	0.918	11.163	12.084	0.921
	(3.045)	(2.731)	0.403	(2.849)	(2.765)	0.000	(3.093)	(2.785)	0.000	(2.886)	(2.827)	0.000
Minority (Yes = 1)	0.042	0.044	0.003	0.044	0.047	0.003	0.038	0.052	0.014	0.040	0.054	0.015
	(0.200)	(0.206)	0.694	(0.205)	(0.212)	0.683	(0.191)	(0.223)	0.069	(0.195)	(0.227)	0.072
Experience	23.575	9.446	-14.129	19.815	9.297	-10.518	24.025	9.709	-14.316	20.207	10.293	-9.914
	(8.671)	(8.951)	0.000	(7.522)	(8.592)	0.000	(8.405)	(7.994)	0.000	(7.947)	(9.011)	0.000
Experience Squared	630.932	169.268	-461.664	449.208	160.171	-289.037	647.831	158.091	-489.740	471.468	187.067	-284.4
	(423.742)	(324.437)	0.000	(303.203)	(273.755)	0.000	(393.781)	(252.717)	0.000	(318.927)	(287.839)	0.000
Age 16 - 25	0.006	0.525	0.519	0.018	0.551	0.533	0.001	0.342	0.340	0.005	0.389	0.384
	(0.075)	(0.500)	0.000	(0.134)	(0.498)	0.000	(0.039)	(0.475)	0.000	(0.072)	(0.488)	0.000
Age 26 - 35	0.211	0.319	0.107	0.286	0.270	-0.016	0.147	0.481	0.335	0.236	0.360	0.124
	(0.408)	(0.466)	0.000	(0.452)	(0.444)	0.295	(0.354)	(0.500)	0.000	(0.425)	(0.480)	0.000
Age 36 - 45	0.428	0.089	-0.340	0.503	0.100	-0.403	0.393	0.124	-0.268	0.478	0.154	-0.324
	(0.495)	(0.285)	0.000	(0.500)	(0.300)	0.000	(0.488)	(0.330)	0.000	(0.500)	(0.362)	0.000
Age 46 - 55	0.267	0.049	-0.218	0.183	0.058	-0.125	0.394	0.043	-0.351	0.267	0.076	-0.191
	(0.442)	(0.215)	0.000	(0.387)	(0.235)	0.000	(0.489)	(0.202)	0.000	(0.442)	(0.264)	0.000
Age 56 - 65	0.087	0.019	-0.068	0.010	0.021	0.011	0.066	0.010	-0.056	0.013	0.021	0.008
	(0.282)	(0.137)	0.000	(0.097)	(0.144)	0.014	(0.247)	(0.099)	0.000	(0.115)	(0.144)	0.130

[1] Notes: Standard deviations in parentheses. Samples are as follows: 1995 and 2002 China Household Income Project (CHIP).

[2] Diff means difference in means between two groups. Column (3) = (2) - (1); Column (6) = (5) - (4); Column (9) = (8) - (7); Column (12) = (11) - (10).

Table 4: Summary Statistics by Living Arrangement

Variables	1995						2002					
	Male			Female			Male			Female		
	With Parents	Not with Parents	Diff	With Parents	Not with Parents	Diff	With Parents	Not with Parents	Diff	With Parents	Not with Parents	Diff
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Dependent Variable												
Log Annual Wages	8.147	8.647	0.501	8.003	8.409	0.406	8.963	9.339	0.376	8.883	9.087	0.203
	(0.720)	(0.535)	0.000	(0.784)	(0.661)	0.000	(0.760)	(0.615)	0.000	(0.770)	(0.665)	0.000
Independent Variable												
Years of Schooling	11.511	11.037	-0.474	11.539	10.284	-1.255	12.499	11.345	-1.154	12.989	11.166	-1.823
	(2.539)	(3.045)	0.000	(2.484)	(2.858)	0.000	(2.683)	(3.084)	0.000	(2.335)	(2.897)	0.000
Minority (Yes = 1)	0.042	0.042	0.001	0.044	0.045	0.001	0.060	0.038	-0.022	0.056	0.041	-0.015
	(0.200)	(0.201)	0.912	(0.205)	(0.206)	0.948	(0.238)	(0.191)	0.026	(0.230)	(0.199)	0.197
Experience	5.931	23.172	17.241	4.915	19.448	14.533	6.446	23.565	17.119	4.873	19.705	14.832
	(4.536)	(8.981)	0.000	(4.120)	(7.794)	0.000	(4.833)	(8.634)	0.000	(4.136)	(8.232)	0.000
Experience Squared	55.722	617.579	561.857	41.107	438.972	397.864	64.865	629.826	564.961	40.815	456.054	415.239
	(119.840)	(429.464)	0.000	(81.759)	(309.384)	0.000	(101.974)	(397.432)	0.000	(99.386)	(323.810)	0.000
Age 16 - 25	0.677	0.017	-0.660	0.805	0.033	-0.772	0.463	0.007	-0.456	0.675	0.016	-0.659
	(0.468)	(0.130)	0.000	(0.397)	(0.179)	0.000	(0.499)	(0.083)	0.000	(0.469)	(0.126)	0.000
Age 26 - 35	0.306	0.220	-0.086	0.178	0.296	0.118	0.509	0.164	-0.345	0.316	0.255	-0.061
	(0.461)	(0.414)	0.000	(0.382)	(0.456)	0.000	(0.500)	(0.370)	0.000	(0.466)	(0.436)	0.008
Age 36 - 45	0.016	0.417	0.401	0.016	0.479	0.462	0.028	0.388	0.361	0.009	0.459	0.450
	(0.127)	(0.493)	0.000	(0.127)	(0.500)	0.000	(0.164)	(0.487)	0.000	(0.094)	(0.498)	0.000
Age 46 - 55	0.001	0.260	0.259	0.000	0.180	0.180	0.000	0.378	0.378	0.000	0.254	0.254
	(0.035)	(0.439)	0.000	(0.000)	(0.384)	0.000	(0.000)	(0.485)	0.000	(0.000)	(0.435)	0.000
Age 56 - 65	0.000	0.086	0.086	0.002	0.013	0.011	0.000	0.063	0.063	0.000	0.016	0.016
	(0.000)	(0.280)	0.000	(0.040)	(0.113)	0.000	(0.000)	(0.244)	0.000	(0.000)	(0.127)	0.000

[1] Notes: Standard deviations in parentheses. Samples are as follows: 1995 and 2002 China Household Income Project (CHIP).

[2] Diff means difference in means between two groups. Column (3) = (2) - (1); Column (6) = (5) - (4); Column (9) = (8) - (7); Column (12) = (11) - (10).

Table 5: Least Square Cross Validation Bandwidths

		1995		2002	
		Spousal Education (1)	Parental Education (2)	Spousal Education (3)	Parental Education (4)
Male	Age Group	0.120	1.000	0.087	1.000
	Province	0.160	0.351	0.351	0.769
	Minority	0.249	0.173	0.500	0.041
	Experience	5.998	8.525	5.654	3.241
		(20.586)	(20.586)	(19.750)	(19.750)
Female	Age Group	0.134	1.000	0.041	0.206
	Province	0.647	0.511	0.459	0.698
	Minority	0.500	0.500	0.500	0.500
	Experience	1.619	1.203	3.593	2.711
		(17.488)	(17.488)	(18.160)	(18.160)

[1] Notes: Reported in parenthesis are two standard deviations of experience for different samples.

Table 6: Conditional Moment Test of Parametric Specification: J_n Test

Gender	Year	Spousal Education (1)	Parental Education (2)
Panel A: Male	1995	5.0969	-2.0343
	p-value	$[p = 0.001]$	$[p = 0.414]$
	2002	7.2852	-1.896
	p-value	$[p = 0.001]$	$[p = 0.965]$
Panel B. Female	1995	4.0549	-2.3479
	p-value	$[p = 0.001]$	$[p = 0.569]$
	2002	5.1034	-0.6305
	p-value	$[p = 0.001]$	$[p = 0.048]$

[1] Notes: Samples are as follows: 1995 and 2002 China Household Income Project (CHIP).

Table 7: Baseline Results

Gender	Year	Full Sample OLS	Non-missing Sample, IV		Full Sample, IV				Hausman Tests	
					Linear Projection		Nonparametric Projection			
			Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) (6) - (4)	(9) = (7) - (5)
Panel A: Male	1995	0.0358***	0.0444***	0.0868***	0.0444***	0.0868***	0.0435***	0.0894***	-0.0009	0.0026
		(0.0022)	(0.0046)	(0.0254)	(0.0046)	(0.0256)	(0.0047)	(0.0261)	(0.0010)	(0.0051)
	No. of Obs.	6274	5125	795	6274	6274	6274	6274	[0.0009]	[0.0071]
	2002	0.0662***	0.0882***	0.1074**	0.0882***	0.1074**	0.0888***	0.1200**	0.0006	0.0126
		(0.0026)	(0.0053)	(0.0353)	(0.0053)	(0.0359)	(0.0055)	(0.0389)	(0.0015)	(0.0150)
	No. of Obs.	5618	4702	615	5618	5618	5618	5618	[0.0011]	[0.0131]
Panel B. Female	1995	0.0562***	0.0727***	0.0854**	0.0727***	0.0854**	0.0704***	0.0859**	-0.0023	0.0005
		(0.0028)	(0.0061)	(0.0286)	(0.0061)	(0.03)	(0.0062)	(0.0303)	(0.0011)**	(0.0043)
	No. of Obs.	5642	4597	614	5642	5642	5642	5642	[0.0013]*	[0.0132]
	2002	0.0810***	0.1183***	0.1383***	0.1183***	0.1383***	0.1172***	0.1504***	-0.0011	0.0121
		(0.0032)	(0.0067)	(0.0377)	(0.0068)	(0.0384)	(0.0071)	(0.0396)	(0.0020)	(0.0097)
	No. of Obs.	4510	3610	449	4510	4510	4510	4510	[0.0017]	[0.0173]

[1] Notes: Robust standard errors in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.01$

[2] Hausman Test is based on the difference between IV estimates based on nonparametric projection and those based on linear projection. Reported in the parentheses and brackets are asymptotic standard errors (see text for details) and bootstrapped standard errors, respectively.

Table A1: Differences in Standard Errors of Conventional and Linear Projection IV Estimates of Control Variables

	Male Sample				Female Sample			
	Spousal Education		Parental Education		Spousal Education		Parental Education	
	1995	2002	1995	2002	1995	2002	1995	2002
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Age 26 - 35	0.0420	0.0171	0.1729	0.0293	0.0350	0.0552	0.0865	0.0330
Age 36 - 45	0.0354	0.1105	0.1677	0.0983	0.0314	0.1401	0.0838	0.1492
Age 46 - 55	0.0335	0.0328	0.1659		0.0305	-0.0768	0.0825	
Age 56 - 65	0.0294		0.1597		0.0601	0.4169	0.0739	
Province Dummy 2	0.0002	0.0748	0.0002	0.0865	0.0008	0.1097	0.0024	0.0984
Province Dummy 3	0.0016	0.0581	0.0021	0.0484	0.0015	0.0865	0.0029	0.0619
Province Dummy 4	0.0009	0.0591	0.0008	0.0704	0.0014	0.0680	0.0016	0.0616
Province Dummy 5	0.0014	0.0646	0.0007	0.0846	-0.0015	0.0748	-0.0010	0.1096
Province Dummy 6	0.0001	0.0704	0.0015	0.0756	0.0026	0.0682	0.0036	0.0585
Province Dummy 7	0.0004	0.0784	0.0007	0.0812	0.0017	0.0827	0.0020	0.0778
Province Dummy 8	0.0017	0.0657	0.0015	0.0756	0.0021	0.0778	0.0027	0.0496
Province Dummy 9	0.0005	0.0743	0.0004	0.0931	0.0007	0.0989	0.0044	0.1311
Province Dummy 10	0.0011	0.0568	0.0008	0.0704	-0.0006	0.1200	0.0035	0.0193
Province Dummy 11	0.0032	0.0713	0.0001	0.0783	0.0028	0.0834	0.0015	0.1688
Province Dummy 12			0.0002	0.0973			0.0035	0.0889
Minority	0.0028	0.0651	0.0047	0.0450	-0.0162	0.1017	0.0080	0.0917
Experience	0.0004	0.0076	0.0004	0.0153	0.0008	0.0167	0.0005	0.0154
Experience Squared	0.0000	0.0002	0.0000	0.0007	0.0000	0.0010	0.0000	0.0006
Constant	0.0309	0.0180	0.1472	0.0307	0.0230	0.0360	0.0587	0.0051

Table B1: Relationship between Age Group Dummies and Other Variables

Gender	Year	Statistics	Dependent Variable	
			Schooling (1)	Experience (2)
Panel A: Male	1995	R Squared	0.025	0.819
		F Test	40.94	7099.741
		p-value	0.000	0.000
	2002	R Squared	0.058	0.734
		F Test	86.723	3877.822
		p-value	0.000	0.000
Panel B. Female	1995	R Squared	0.037	0.685
		F Test	54.268	3063.478
		p-value	0.000	0.000
	2002	R Squared	0.062	0.612
		F Test	74.54	1772.949
		p-value	0.000	0.000

[1] Notes: Reported are the R-squared values from regressions of schooling (column (1)) or experience (column (2)) on age group dummies. The F-test statistics and its corresponding p values are performed for joint significance of age group dummies in these regressions.

Table B2: Conditional Moment Test of Parametric Specification (Excluding Age Group Dummies): J_n Test

Gender	Year	Spousal Education (1)	Parental Education (2)
Panel A: Male	1995	5.0969	-2.0343
	p-value	$[p = 0.001]$	$[p = 0.414]$
	2002	7.2852	-1.896
	p-value	$[p = 0.001]$	$[p = 0.965]$
Panel B. Female	1995	4.0549	-2.3479
	p-value	$[p = 0.001]$	$[p = 0.569]$
	2002	5.1034	-0.6305
	p-value	$[p = 0.001]$	$[p = 0.048]$

[1] Notes: Samples are as follows: 1995 and 2002 China Household Income Project (CHIP).
[2] In all estimations, age group dummies are excluded

Table B3: Results (Excluding Age Group Dummies)

Gender	Year	Full Sample OLS	Non-missing Sample, IV		Full Sample, IV				Hausman Tests	
					Linear Projection		Nonparametric Projection			
			Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) (6) - (4)	(9) = (7) - (5)
Panel A: Male	1995	0.0362***	0.0442***	0.0881***	0.0442***	0.0881***	0.0447***	0.0879***	0.0005	-0.0002
		0.0021	0.0046	0.0257	0.0046	0.0259	0.0046	0.0265	(0.0009)	(0.0056)
	No. of Obs.	6274	5125	795	6274	6274	6274	6274	[0.0007]	[0.0065]
	2002	0.0671***	0.0884***	0.1079**	0.0884***	0.1079**	0.0901***	0.1223**	0.0017	0.0144
		0.0027	0.0053	0.0339	0.0054	0.0351	0.0055	0.0387	(0.0010)	(0.0163)
	No. of Obs.	5618	4702	615	5618	5618	5618	5618	[0.001]	[0.0128]
Panel B. Female	1995	0.0574***	0.0695***	0.0899**	0.0695***	0.0899**	0.0699***	0.0870**	0.0004	-0.0029
		0.0028	0.0061	0.0292	0.0061	0.0301	0.0062	0.0299	(0.0011)	.
	No. of Obs.	5642	4597	614	5642	5642	5642	5642	[0.0012]	[0.0105]
	2002	0.0795***	0.1166***	0.1365***	0.1166***	0.1365***	0.1172***	0.1508***	0.0006	0.0143
		0.0032	0.0063	0.0353	0.0064	0.0366	0.0065	0.0384	(0.0011)	(0.0116)
	No. of Obs.	4510	3610	449	4510	4510	4510	4510	[0.0011]	[0.0181]

[1] Notes: Robust standard errors in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.01$

[2] Hausman Test is based on the difference between IV estimates based on nonparametric projection and those based on linear projection. Reported in the parentheses and brackets are asymptotic standard errors (see text for details) and bootstrapped standard errors, respectively.

Table C1: Conditional Moment Test of Parametric Specification (U.S. Sample): J_n Test

Gender	Year	Spousal Education (1)	Parental Education (2)
Panel A: Male	Test Statistic	0.3962	-1.3434
	p-value	$[p = 0.015]$	$[p = 0.7469]$
Panel B. Female	Test Statistic	-1.2718	0.4909
	p-value	$[p = 0.411]$	$[p = 0.0576]$

[1] Notes: Samples are as follows: International Social Survey Programme (1985 - 1993).

Table C2: Results (US Sample)

Gender	Full Sample OLS	Non-missing Sample, IV		Full Sample, IV				Hausman Tests	
				Linear Projection		Nonparametric Projection			
		Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education	Spousal Education	Parental Education
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8) = (6) - (4)	(9) = (7) - (5)
Panel A: Male	0.0740***	0.0870***	0.1095***	0.0870***	0.1095***	0.0883***	0.1085***	0.0013	-0.0010
	(0.0053)	(0.0095)	(0.0190)	(0.0095)	(0.0193)	(0.0096)	(0.0191)	(0.0014)	.
No. of Obs.	2076	1310	800	2076	2076	2076	2076	[0.0017]	[0.0041]
Panel B. Female	0.0952***	0.1029***	0.1050***	0.1029***	0.1050***	0.1029***	0.1051***	0.0000	0.0001
	(0.0066)	(0.0165)	(0.0259)	(0.0166)	(0.0261)	(0.0167)	(0.0269)	(0.0018)	(0.0065)
No. of Obs.	2126	1080	827	2126	2126	2126	2126	[0.002]	[0.0051]

[1] Notes: Robust standard errors in brackets. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.01$. Data source: International Social Survey Programme (1985 - 1993). Included in estimations are age, age squared, marital status, union status, and year fixed effects.

[2] Hausman Test is based on the difference between IV estimates based on nonparametric projection and those based on linear projection. Reported in the parentheses and brackets are asymptotic standard errors (see text for details) and bootstrapped standard errors, respectively.