Discussion Paper Series

RIEB

Kobe University

DP2017-09

Estimating Regional Returns to Education in India

Prabir BHATTACHARYA
Takahiro SATO

March 30, 2017

Research Institute for Economics and Business Administration
Kobe University
2-1 Rokkodai, Nada, Kobe 657-8501 JAPAN

Estimating Regional Returns to Education in India:[$]
A Fresh Look with Pseudo-Panel Data

Prabir Bhattacharya[*] Takahiro Sato[**]

March 30, 2017

Abstract

This study analyzes the effects of socio-economic factors on the real wage rates for male workers in India over the period 1983 to 2010. In particular, we examine the role of human capital by estimating the Mincerian wage equation. We construct a regional level pseudo panel data set for our analysis. Our findings show that while the return to primary education is remarkably high, the returns to other, higher, levels of education are equally remarkably low for all of India taken together, becoming progressively so as the level of education increases. These findings are in contradistinction to those of the other studies on returns to education in India, all of which, however, have relied on cross-sectional data for their analyses. We also find relatively little effects of caste, tribe and religion on real wage rates in India, suggesting that that these factors may not be as important as is sometimes believed.

JEL Classification: I24, I25, O15.
Key words: returns to education, India, regions, pseudo-panel data.

[$] We are grateful for the financial support of JSPS ``Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research: Global Initiatives" headed by Professor Tsukasa Mizushima.
[*] Visiting Foreign Scholar, RIEB, Kobe University, Rokkodai, Nada, Kobe, JAPAN & Associate Professor of Economics, School of Social Sciences, Heriot-Watt University, Edinburgh, Scotland, UNITED KINGDOM.
[**] Professor, RIEB, Kobe University, Rokkodai, Nada, Kobe, JAPAN.

1. Introduction

The purpose of this paper is to study the effects of socio-economic factors on real wage rates for male workers in India over the period 1983 to 2010. Caste, tribe, and religion are thought to play important roles in Indian society. It is, therefore, of some importance to examine the role of these factors in real wage rates determination in India. Do they indeed outweigh the impact of education, which, among other things, presumably enhances worker skills and productivity? And if education is important, what precise level of education is critical or most important? Previous studies on returns to education in India have often reached contradictory conclusions in this regard. Dutta (2006), for example, found a U-shaped pattern of return for the regular wage workers, with relatively lower returns for the primary level of education when compared to the secondary and graduate levels but higher than those at the middle level of education. Agarwal (2012), by contrast, found that returns to education in his sample increased with increases in the level of education (see also, among others, Duraiswamy, 2002, Mehta and Hasan, 2012, Azam, 2012, Chamarbaghwala, 2006, and Vatta et al., 2016). In this context, one may also note that Dreze and Sen (2002), in particular, have emphasized the vital importance of primary education for economic and human development.

Regarding the influence of caste and tribe status, Kijima (2006) found that the scheduled caste households had lower returns to education compared to the non-scheduled caste households. Madheswaran and Attewell (2007) too reached very similar conclusions. Ito (2009) found job discrimination against the members of the lower caste but not wage discrimination.

Most of the previous studies on returns to education in India have used only the cross-sectional information for the reference years chosen. However, the returns to education, as is well known, is likely to be closely correlated with unobservable factors such as individual abilities and/or motivations. There is an endogeneity problem here (Warunsiri and McNown, 2010) and to deal with it effectively, one needs, at the very least, a panel data set. For our analysis, we construct a pseudo panel dataset based on the unit level data available from India's National Sample Survey Organisation (NSSO). The data set and the methodology we use in our analysis are explained in due course.

The plan of the rest of the paper is as follows. Section 2 outlines the model we estimate. Section 3 describes the data and the variables used. Section 4 presents the results. Section 5 concludes.

2. The Model

For simplicity *a la* Heckman, Lochner and Todd (2006), we assume that (i) the wage income (w) is determined by schooling years (s), (ii) an individual worker lives forever, and (iii) there is no educational cost during the schooling years. The worker is interested in maximizing the sum of the discounted stream of wage income:

$$\max_s \int_s^\infty w(s)e^{-\gamma t}dt = w(s)e^{-\gamma t}\frac{1}{\gamma}$$

where γ is the interest rate. The first order condition of this maximization problem is given by

$$\frac{w'(s)}{w(s)} = \gamma.$$

We obtain the following equation by integrating this condition with respect to schooling year (s).

$$\ln w = \gamma s + c$$

where c is a constant of integration. Thus the semi-log type equation is naturally generated by the dynamic optimization of a worker.

As is well known, the basic Mincerian wage equation is given by:

$$\ln w_i = \alpha + \sum \beta_k D_{ki} + \mathbf{X}'\boldsymbol{\delta} + e_i \qquad (1)$$

where ln w_i is the natural logarithm of wage for a given worker, D_{ki} is the dummy variable for ith level of education and **X'** is the vector of other variable that are expected to influence the wage of a worker such as age, experience, caste, religion, etc. e_{rt} is the random term. If we take the average annual returns (%) to a given level of education to be given by γ_k, then these can be estimated by using the following formula:

$$\gamma_k = \frac{(\beta_k - \beta_{k-1})}{(n_k - n_{k-1})} \qquad (2)$$

where, β_k is the coefficient of kth level of education and β_{k-1} is the coefficient of previous level of education and n_k is the number of years of schooling for the kth level and n_{k-1} is the number of years of schooling for the previous level. While different education level dummies are used as explanatory variables, other variables such as work experience, caste and tribe status and religious groupings can also be included as control variables.

One of the limitations of the above model, of course, is that of the possible correlation between unobservable factors and education which will lead to biased estimates. To address the endogeneity issue and to identify the determinants of wage rates which are specific to a particular geographical unit over time, we apply the pseudo panel data approach in which we aggregates the unit-level household data provided by the NSSO by regions that remain common across cross-sectional data sets in different years.

Once we take account of the regional wage, the equation (1) will become the model developed by Deaton (1985). We apply the pseudo panel for the unit r based on the regional classifications. The unit is denoted as r in the equation (3) below.

$$\ln \bar{w}_{rt} = \bar{\alpha} + \sum \beta_k \bar{D}_{rt} + \bar{\mathbf{X}}_{\mathbf{t}}' \boldsymbol{\delta} + \bar{u}_{rt} + \bar{\varepsilon}_{rt} \qquad (3)$$

where r denotes regional unit and t stands for survey years for six rounds of NSS, 1983, 1987-88, 1993-94, 1999-2000, 2004-05 and 2009-10 The upper bar means that the average of each variable is taken for each unit, k, for each round, t, $\bar{u}_{rt}$ is the unobservable individual effect specific to the cohort r (e.g. unobservable fixed effects which are not captured by explanatory variables), and $\bar{\varepsilon}_{rt}$ is an error term.

The following equation (4) can be estimated by the standard panel model, such as fixed effects or random effects model.

$$\ln w_{it} = \alpha + \sum \beta_k D_{kit} + \mathbf{X}_{\mathbf{t}}' \boldsymbol{\delta} + u_i + \varepsilon_{it} \qquad (4)$$

The issue is whether the equation (3) is a good approximation of the underlying household panel models for household in the equation (4) above. It is not straightforward to check this as we do not have 'real' panel data. However, as shown by Verbeek and Nijman (1992) and Verbeek

(1996), if the number of observations in cohort r tends to infinity, $\bar{u}_{rt} \rightarrow u_i$ and the estimator is consistent. In our case, r is reasonably large and thus the estimator is likely to be almost consistent.

As mentioned in the introduction, we have constructed a pseudo-panel data set to address the endogeneity problem that arises from the returns to education being correlated with unobservable factors such as individual abilities and/or motivation. However, there is the further problem in our case in that the high quality graduates may find it easier to move to high-waged regions and we need to address this endogeneity issue too. And this we do by taking lagged human capital as an explanatory variable. The model with the lagged human capital, therefore, is the model we consider to be more robust.

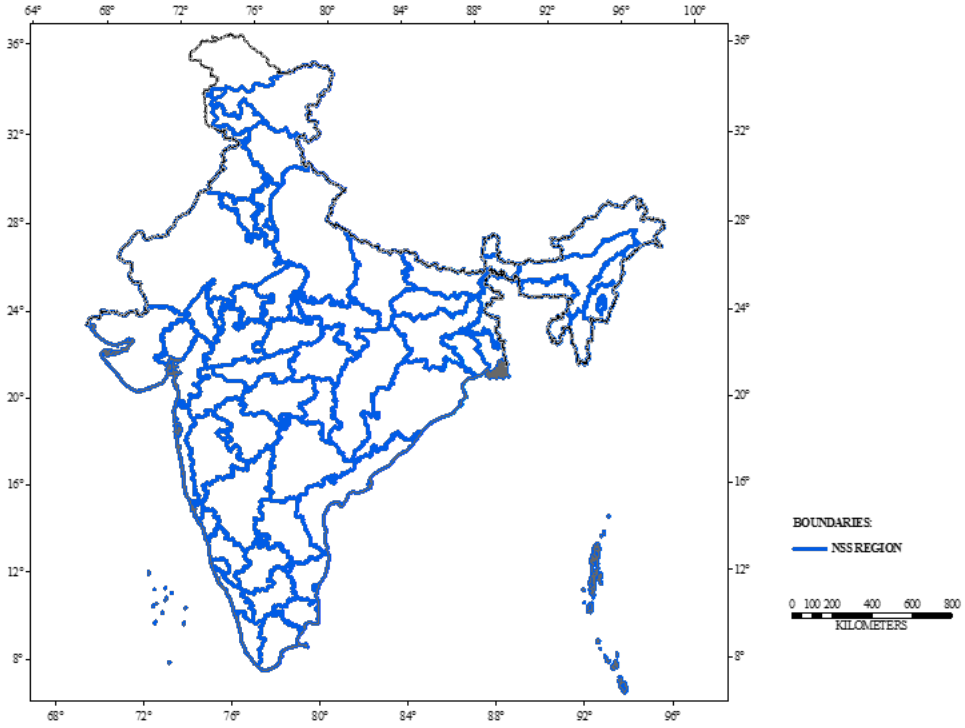
3. Data and Variables

The study uses the data on employment and unemployment in India from six rounds of the NSSO conducted during 1983 (38th round), 1987-1988 (43rd round), 1993-1994 (50th round), 1999-2000 (55th round), 2004-2005 (61st round), and 2009-2010 (66th round), respectively. Each round collected information about 120,000 households and more than half a million individuals, selected from rural and urban areas. The national level estimates of the labour and work force participation, industrial distribution of the workers and status of their employment and wages were prepared on the basis of the data collected from these surveys. The sample selection used two-stage stratified random sampling procedure where the first stage of sampling is the census villages and urban blocks and the second stage is the household in these villages and blocks. Apart from the information on employment and unemployment, the surveys also recorded information about household size, age and education of the household members, social group of the household, religion and land owned.

For our analysis, we have constructed a regional level pseudo panel data set based on the unit-level data of the NSSO. The NSSO defines a region as a ``grouping of contiguous districts having similar geographical features, rural population densities and crop-pattern. Generally, the regions were not found to be cutting across districts boundaries in any state except Gujarat" (NSSO, 2010, par. 2.1.4.). In order to make consistent time series over 1983 to 2010, we reclassified the NSS regions. The total number of NSS regions in the study is 65 as shown in Figure 1. We constructed regional units by aggregating individual variables into NSS regional-level variables with rural and urban areas separately. Thus, the total available number

of regions with rural and urban areas is 130. The approach adopted here has the merit of taking account of geographical diversity of India as well as of the improvements in the level of education over time.

Figure 1: Map of NSS region



In India, the constituent states can be divided into two categories, viz., the main states and special category states. The special category states consist of smaller states and union territories and they generally receive more financial transfers (in relation to their revenues) from the central government. We present results for all states combined as well as for the main states separately.

In calculating the returns to education by estimating Mincerian wage equations, the times taken to complete the primary, middle, secondary and graduate levels of education are assumed to be 5, 8, 12 and 15 years, respectively. Accordingly, the time interval for each educational level dummy category is taken as 5, 3, 4 and 3 years, respectively. As the state-wise consumer price indices are not easily available in India, real wage (w) is computed by using the implicit deflator of the net state domestic product (NSDP). D_{krt} is the kth education level in region r at time t (with below primary being the reference). The X vector consists of scheduled castes, scheduled tribes, Muslims, work experience in years, industry share in state s at time t (with agricultural

sector being the reference). Descriptions, means and standard deviations of the main variables used in our analysis are presented in Table 1.

Table 1: Descriptive statistics

	All States, N=770				Main States, N=564			
	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max
ln*w*: ln (real wage rate)	4.824	0.780	−0.516	6.794	4.751	0.739	−0.516	6.234
belowp: Below primary	0.200	0.147	0.000	0.859	0.223	0.149	0.006	0.859
prim: Primary	0.121	0.065	0.000	0.427	0.124	0.063	0.000	0.397
middle: Middle	0.173	0.074	0.000	0.484	0.167	0.068	0.000	0.388
secon: Secondary	0.296	0.095	0.025	1.000	0.280	0.086	0.025	0.620
grad: Graduate	0.209	0.119	0.000	0.630	0.205	0.115	0.000	0.598
st: Scheduled Tribe	0.127	0.223	0.000	1.000	0.066	0.093	0.000	0.621
sc: Scheduled Caste	0.134	0.097	0.000	0.525	0.153	0.092	0.000	0.525
muslim: Muslim	0.097	0.145	0.000	0.982	0.087	0.073	0.000	0.603
expyear: Working experience in years	27.85	2.55	20.33	35.15	27.54	2.31	20.33	34.61

4. Results

We present two sets of results: one without and the other with the lagged human capital. For reasons already stated, we regard the model with the lagged human capital as the preferred one. However, for purposes of comparison, we also present the results of the model without the lagged human capital (which we call the basic model).[1]

The basic model (Table 2) shows the returns to all levels of education to be positive and statistically significant across all specifications. However, the results with the lagged human capital (Table 3) show that while the returns to primary and secondary levels of education are positive and significant, the returns to other, higher levels of education are statically insignificant across all regressions. Figure 2 presents estimated returns by levels of education from the basic model for both the main and all states combined. Figure 3 does the same for the model with lagged human capital.

Both Figure 2 and Figure 3 clearly bring out the very high return to primary school education. This is in sharp contrast to most other studies which find the return to primary level of education to be very low.[2] In our case the return to primary education is seen to be as high as 30 per cent. The return to graduate level education, by contrast, is seen to be insignificant in the model with the lagged human capital (the robust model). These results would seem to suggest that the motivation hypothesis may have some strength in that people with high motivation may succeed without necessarily having to acquire higher education.[3] There is, of course, the other (in some ways complementary) hypothesis that the quality of most higher levels of education in India may be so low that the value added of these levels of education is not particularly significant.

Of the social and religion group variables, the Muslim variable is statistically insignificant across all specifications in both the basic model and regressions with the lagged human capital. The ST variable is statistically insignificant across all specifications in regressions with lagged human capital. The ST variable is also statistically insignificant for all states in the basic regression (though it is positive and statistically significant in three of the other four specifications in the basic regression). The SC variable is statistically insignificant across all

[1] We also exclude Manipur with the least share of primary education in all states, Bihar with the least share of primary education in main states, and Delhi (which is not usually included as a main state) from the regression samples for robustness checks.

[2] See, for example, Kamal et al. (2016) which shows that the return to primary education of male workers varies from -2 per cent to +6% during the period 1983 to 2010.

[3] This finding is consistent with Warunsiri and Mcnown (2010) which estimate the return to education in Thailand by employing a pseudo-panel approach.

specifications in the basic regression model. It is also statistically insignificant for all states in the regressions with lagged human capital. Taken together, these results would seem to suggest that the caste, tribe, and religion based factors may not be as important in real wage rates determination in India as is sometimes believed.

As expected, work experience in years has positive and statistically significant effect across all regressions.

Table 2: Basic regression results

	(1)	(2)	(3)	(4)	(5)
	All states	Main states	All states without Manipur	Main states without Delhi	Main states without Bihar
prim	1.531	1.989	1.807	1.95	1.775
	(2.95)**	(3.99)**	(4.23)**	(3.83)**	(3.84)**
middle	1.745	2.128	2.128	2.115	1.881
	(2.72)*	(3.33)**	(3.57)**	(3.22)**	(3.05)**
secon	2.207	2.593	2.14	2.595	2.363
	(5.70)**	(5.96)**	(5.15)**	(5.86)**	(6.51)**
grad	2.768	3.272	2.814	3.285	3.204
	(6.67)**	(7.59)**	(6.34)**	(7.59)**	(7.42)**
st	0.034	0.834	0.473	0.722	0.722
	(0.07)	(2.99)**	(1.85)	(2.83)*	(2.82)*
sc	0.024	0.026	0.026	0.082	−0.037
	(0.09)	(0.09)	(0.10)	(0.27)	(0.14)
muslim	0.317	0.42	0.216	0.499	0.56
	(1.67)	(1.24)	(1.18)	(1.53)	(1.81)
expyear	0.036	0.053	0.038	0.053	0.052
	(4.29)**	(5.35)**	(4.21)**	(5.36)**	(5.15)**
Constant	2.506	0.666	2.546	1.207	1.814
	(4.85)**	(0.53)	(4.65)**	(1.37)	(2.79)*
Location Fixed Effect	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes
State-level Industrial Structure	Yes	Yes	Yes	Yes	Yes
Observations	770	564	747	552	540
R-squared	0.72	0.78	0.75	0.78	0.80

Cluster-robust t statistics in parentheses(cluster=state)
* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 2: Estimated returns to education (the basic model)

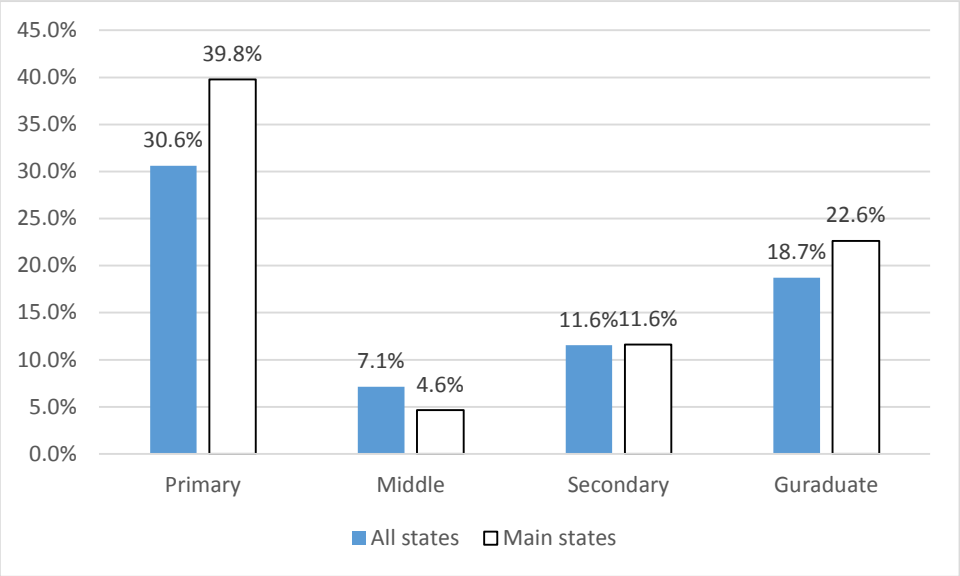


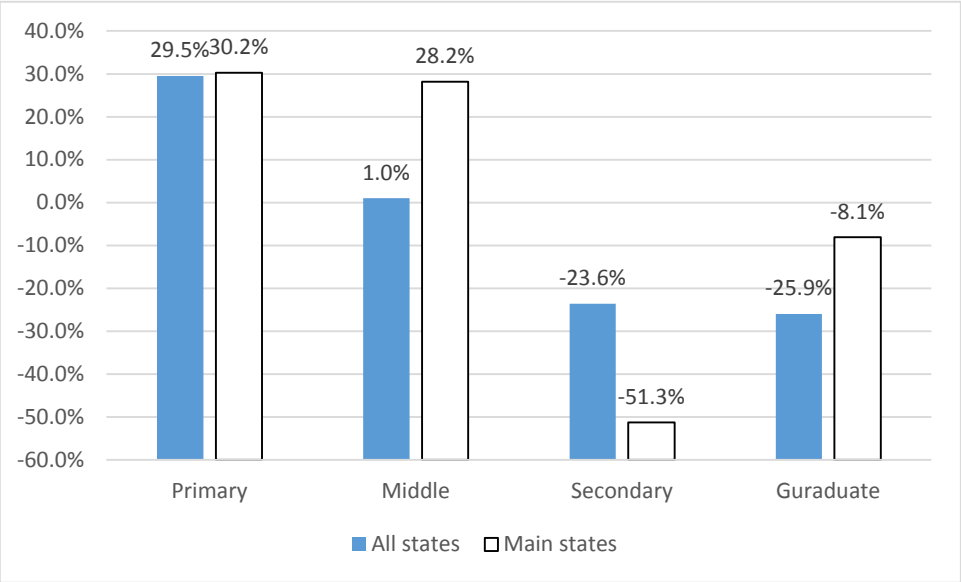
Table 3: Regression results with lagged human capital

	(1) All states	(2) Main states	(3) All states without Manipur	(4) Main states without Delhi	(5) Main states without Bihar
prim(lagged)	1.258	1.487	1.327	1.48	1.372
	(2.85)**	(3.75)**	(2.85)**	(3.56)**	(3.83)**
middle(lagged)	1.424	2.403	1.661	2.449	2.28
	(2.63)*	(4.37)**	(2.99)**	(4.29)**	(4.15)**
secon(lagged)	0.581	0.337	0.424	0.288	0.531
	(1.18)	(0.59)	(0.90)	(0.51)	(0.91)
grad(lagged)	−0.256	0.079	−0.169	0.003	0.039
	(0.72)	(0.19)	(0.47)	(0.01)	(0.09)
st	−0.562	−0.451	−0.454	−0.674	−0.579
	(1.34)	(0.87)	(1.02)	(1.48)	(1.16)
sc	−0.672	−1.137	−0.687	−1.035	−1.105
	(2.03)	(2.89)**	(2.14)*	(2.80)*	(2.59)*
muslim	0.236	−0.169	0.203	−0.119	0.158
	(1.09)	(0.38)	(0.87)	(0.26)	(0.44)
expyear	0.042	0.058	0.045	0.06	0.056
	(3.34)**	(4.36)**	(3.61)**	(4.50)**	(4.10)**
Constant	2.867	1.807	2.904	2.371	2.108
	(3.68)**	(1.15)	(3.70)**	(1.81)	(2.56)*
Location Fixed Effect	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes
State-level Industrial Structure	Yes	Yes	Yes	Yes	Yes
Observations	638	470	619	460	450
R-squared	0.70	0.74	0.72	0.74	0.75

Cluster-robust t statistics in parentheses(cluster=state)

* significant at 10%; ** significant at 5%; *** significant at 1%

Figure 3: Estimated returns to education with lagged human capital



5 Conclusions

Our study has underlined the importance of primary education. This adds one more reason for making the provision of universal primary education a top priority goal for the policy makers. We have, however, found the returns to higher levels of education to be remarkably low, in some cases almost insignificant. Clearly, this needs an explanation. Could it be that the jobs that are being created mostly do not require the skills that are thought to be enhanced by higher levels of education? Or, equally, could it be that the quality of education itself at these higher levels leave much to be desired? Both are plausible and we plan to explore these issues in future research. So far as the effects of caste and religion are concerned, we found these to be statistically insignificant in most cases, suggesting that these factors may not be as important in wage determination as is sometimes believed.

References:

Aggarwal, Tushar (2012) Returns to education in India: Some recent evidence, *Journal of Quantitative Economics*, 10 (2), pp.131-151.

Azam, Mehtabul (2012) Changes in wage structure in Urban India, 1983-2004: A quantile

regression decomposition, *World Development*, 40, pp. 1135-50.

Chamarbagwala, Rubiana (2006) Economic liberalization and wage inequality in India, *World Development*, 34, pp. 1997-2015.

Deaton, A. (1985) Panel data from the time series of cross-sections, *Journal of Econometrics*, 30, pp.109-126.

Dreze, J. and A. Sen (2002) *India: Development and Participation*, Oxford University Press.

Duraiswamy, P. (2002) Changes in returns to education in India, 1983-94: by gender, age-cohort and location, *Economics of Education Review*, 21, pp. 609-22.

Dutta, Puja Vaseudeva (2006) Returns to education: New evidence for India, 1983-99, *Education Economics*, 14, pp. 431-51.

James J. Heckman, Lance J. Lochner, and Petra E. Todd (2006) Fifty Years of Mincer Earnings Regressions, in Hanushek, E. and F. Welch (eds.) *Handbook of Education Economics*, Vol. 1. , Elsevier.

Ito, Takahiro (2009) Caste discrimination and transaction costs in the labor market: Evidence from rural North India, *Journal of Development Economics*, 88, pp. 292-300.

Kamal Vatta, Takahiro Sato and Garima Taneja (2016) Indian Labour Markets and Returns to Education, *Millennial Asia*, 7 (2), pp.1-24.

Kijima, Yoko (2006) Caste and tribal inequality: Evidence from India, 1983-1999, *Economic Development and Cultural Change*, 54, pp. 369-404.

Kumar, Utsav and Prachi Mishra (2008) Trade liberalization and wage inequality: Evidence from India, *Review of Development Economics*, 12, pp. 291-311.

Madheswaran, S and Paul Attewel (2007) Caste discrimination in the Indian urban labour market: Evidence from the National Sample Survey, *Economic and Political Weekly*, 41, pp. 4146-53.

Mehta, Ashish and Rana Hasan (2012) The effects of trade and services liberalization on wage inequality in India, *International Review of Economics and Finance*, 23, pp. 75-90.

National Sample Survey Organisation (NSSO) (2010) *Concepts and Definitions Used in NSS*, NSSO.

Verbeek, M. (1996) Pseudo Panel Data, in Matyas, L. and Sevestre, P. (eds.), *The Econometrics of Panel Data*, second edition, Kluwer Academic, pp. 280-92.

Verbeek, M. and Nijman T. E. (1992) Can cohort data be treated as genuine panel data? *Empirical Economics*, 17, pp. 9-23.

Warunsiri, S. and R. Mcnown (2010) The return to education in Thailand: A pseudo-panel Approach, *World Development*, 38 (11), pp. 1616-1625.