Education-occupation (mis-)match and dispersion in returns to education: Evidence from India

Shweta Grover[1] and Ajay Sharma
Indian Institute of Management Indore

Abstract

Using a national level sample survey on labour market in India, we analyze the role of education-occupation (mis-)match (EOM) in explaining within-group dispersion in returns to education. Applying a double sample selection bias correction and Mincerian quantile wage regression estimation, we document that accounting for EOM reduces the dispersion in returns to education. Further, overeducated workers face a wage penalty of 7 percent and undereducated workers get a wage reward of 3 percent as compared to adequately matched workers. This study argues to focus on EOM to increase both pecuniary and social benefits of education in terms of productivity gains and wages as well as to reduce wage inequality.

Keywords: Education-occupation mismatch; Dispersion in returns to education; Wage dispersion; India; Quantile regression

JEL Code: I24, J24, J31

[1]***Correspondence Address***: Shweta Grover, Indian Institute of Management Indore, Rau-Pithampur Road, Prabandh Shikhar, Indore, (M.P) India-453556; e-mail: f15shwetag@iimidr.ac.in.

Ajay Sharma, Indian Institute of Managment Indore, Rau-Pithampur Road, Indore, (M.P.) India-453556; e-mail:ajays@iimidr.ac.in; ajaysharma87@gmail.com. Phone: +91-731-2439622

Education-occupation (mis-)match and dispersion in returns to education: Evidence from India

Abstract

Using a national level sample survey on labour market in India, we analyze the role of education-occupation (mis-)match (EOM) in explaining within group dispersion in returns to education. Applying a double sample selection bias correction and Mincerian quantile wage regression estimation, we document that accounting for EOM reduces the dispersion in returns to education. Further, overeducated workers face a wage penalty of 7 percent and undereducated workers get a wage reward of 3 percent as compared to adequately matched workers. This study argues to focus on EOM to increase both pecuniary and social benefits of education in terms of productivity gains and wages as well as to reduce wage inequality.

Keywords: Education-occupation mismatch; Dispersion in returns to education; Wage dispersion; India; Quantile regression

JEL Code: I24, J24, J31

Education-occupation (mis-)match and dispersion in returns to education: Evidence from India

1. Introduction

Understanding wage determination and resulting wage inequality/dispersion[2] remains one of the central themes of labour economics due to the interest of multiple agents *viz.*, workers, firms, and government. One unsolved puzzle in this literature relates to heterogeneity in the returns to same level of education, i.e., within-education group dispersion in returns. One explanation for this phenomenon can be drawn from human capital theory (HCT) (Becker 1964). The author asserts wages to be dependent on the combination of human capital characteristics and not only on education. This implies that workers with same education may draw different returns to education due to differences in other human capital aspects such as experience in the labour market, on the job training, and so on. Another elucidation can be inferred from job competition model (JCM) (Thurow 1975). JCM stress on job definition and its requirements as contributors to within-education group dispersion in returns. Further, apart from worker and firm characteristics, researchers have examined various other channels that can partially explain this heterogeneity, for example, unemployment duration (Eckstein and Van den Berg 2007), the nature of job-search process and consequent labour market frictions (Eckstein and Van den Berg 2007; Postel–Vinay and Robin 2002), social networks and job-referrals (Montgomery 1991), institutional structure and policy changes (such as active labour market policies, trade and immigration policy, etc.) (Kierzenkowski and Koske 2012), and labour market discrimination (Becker 2010). In similar spirit, this paper marginally contributes to this literature by providing an alternative explanation using the education-occupation (mis-)match (EOM) framework.

The term education-occupation mismatch refers to a difference between the attained years of education of an individual and required years of education for her occupation (Duncan and Hoffman 1981; Verdugo and Verdugo 1989). A person is said to be overeducated (undereducated) if her attained education is higher (lower) than the required education of occupation at which she works. In contrast, a person is categorized as adequately educated when attained and required education are aligned. A consistent finding in this literature is that

[2] Strictly speaking, the concepts of wage inequality and wage dispersion are slightly different (see Salverda and Checchi 2014 for conceptual discussion). However, in this paper, we will use these terms interchangeably.

overeducated workers endure significant wage penalties and undereducated workers receive considerable wage rewards as compared to their matched counterparts with the same level of education (Hartog 2000; Leuven and Oosterbeek 2011). This may signal that match status (adequately, over, or undereducated) in the labour market can affect returns to education and therefore, can explain a part of dispersion in returns to the same level of education (Martins and Pereira 2004). The main contribution of this paper lies in exploring whether and how much of the within-education group dispersion in returns can be attributed to EOM in the labour market. This is done by using the data from a nationwide sample survey on labour market particulars from India. We consider the context of a developing country because they exhibit a stronger relationship between education and wage dispersion than the developed countries (dos Reis and de Barros 1991).

The main contributions of this paper are as follows. First, this research enhances our understanding of (i) within-education group dispersion in returns and (ii) the impact of education-occupation (mis-)match on within-education group dispersion in returns. The issue of within-education group dispersion in returns has received its due share of attention. One important point to note here is that in the relevant literature, researchers have used dispersion in returns as an indicator of dispersion in wages but this may not always be true. Given this caveat, in the literature, the researchers document higher within-education wage dispersion for tertiary-educated workers (Azam 2012) and the positive relationship between education and within-education wage dispersion. This relation is termed as inequality increasing effect of education (Budría and Telhado-Pereira 2011) and is found for several countries (see Martins and Pereira 2004 for Europe and United States; Tansel and Bodur 2012 for Turkey). On the other side, research studies on the relationship between EOM and within-education group dispersion in returns are sparse. A notable study, Green and Zhu (2010), in case of Britain, focuses on estimating the relationship between changing levels of overeducation and dispersion in returns to education over time. The authors find no significant dispersion in returns to education for matched male workers whereas there was dispersion in returns for mismatched workers. Apart from this, the study by Budría and Moro-Egido (2008) finds that EOM only when combined with skill mismatch leads to within-education wage dispersion. Our contribution to this literature emerges from two fronts. On one side, we discuss dispersion in returns and not wage dispersion. The key reason being that dispersion in returns may not always show wage dispersion. Further, we use more nuanced methods to estimate dispersion in returns (such as gini coefficient, Lorenz curve, and coefficient of variation). On the other hand, we simultaneously deal with two crucial issues, *heterogeneity in the returns to education* (Henderson, Polacheck and Wang 2011) and *selection bias* (Heckman 1977). Ignoring these issues not only leads to biased but also inconsistent estimates. In the empirical literature on EOM, only one of the problems is considered at a time. To address the heterogeneity issue, we use quantile regression (QR) estimation that examines the effect of education and EOM across the wage distribution. To solve the second issue, i.e., selection bias, we employ double sample selection framework (Catsiapis and Robinson 1982) and consider two fundamental decisions – the decision to work (to be or not to be engaged in economic activity)[3] and the choice of economic activity status (waged/salary or self-employment).

Second, this study estimates returns to education occupation (mis-)match (EOM). Further, given that EOM status can affect returns to education, we provide corrected estimates for the same. If match status is ignored in the wage equation, there is a case for misspecification because of omitted variable and estimates of returns to education are likely to be downward bias. The reason being as we move up the ladder of education, overeducation is more likely to be observed and affects the returns to education negatively.

Third, this study also helps policymakers in developing countries in understanding some of the key issues such as, the role of human capital formation in the process of sustained economic growth and development. One major ongoing debate in developing countries is that whether available limited resources should be channelled for universal education or to be targeted for tertiary education for selected part of population. Recent studies have argued for the latter case (Castelló-Climent and Mukhopadhyay 2013). The missing piece in this debate is the situation of labour market and correct estimates of returns to education. Our study contributes to this debate by providing the argument that one need to understand the matching of workers' characteristics with the available jobs in the labour market and corresponding pecuniary returns to education. This is based on the argument that productivity of an individual is ascertained not only by the worker's ability but also by the adequacy of her match to the job (Devine and Kiefer 1991). Therefore, we aim to shed light on the efficiency of higher education from the perspective of differential returns for matched and mismatched workers.

The salient findings from our analysis are as follows. First, EOM affects the structure of dispersion in returns to education. We find that the within-education group dispersion in returns

[3] Economic activity is defined as an activity that results in the production of goods and services which, in turn, leads value addition to the national product (NSSO 2014).

decreases after the inclusion of EOM. Second, selection terms are statistically significant suggesting the evidence in favour of sample selection bias. Third, consistent with the earlier studies, we observe that, on an average, overeducated workers suffer a wage penalty of around 7 percent and undereducated workers receive a wage reward of approximately 3 percent as compared to their adequately educated counterparts. While the wage rewards to undereducated workers exhibit an increasing trend across the wage distribution up to 70th quantile, the wage penalty to overeducated workers decreases along the quantiles. It highlights that workers in the lower end of the wage distribution suffer higher wage losses if they are overeducated and earn lower wage rewards if they are undereducated. Lastly, returns to education increase with the level of education. Workers with graduation or above degree earn the highest returns. The key aspects which are generally missing in estimation of returns to education relates to information on education-occupation (mis-)match (EOM) and also correction for double sample selection bias. This paper considering these intricacies attempts to provide unbiased estimates of returns to education and thus aids the policymakers in decision for investments in education.

The rest of the paper is structured as follows. In Section 2, we discuss the measures to estimate EOM, followed by the description of data used in Section 3. Further, Section 4 provides the estimation methodology. Section 5 presents the results and discussion. The final section concludes.

2. Approaches to measures education-occupation (mis-)match

In the process of wage determination, the role of matching of worker and firm characteristics plays an important role. The relative importance of different agents in the matching process can lead to different approaches proposed in the literature.

When workers have more information to take well-informed decisions to choose their jobs and gauge their market wages, their self-assessment is considered appropriate. This approach is termed as *workers' self-assessment* (WA) in the EOM literature. The key assumptions in the case of WA are that workers are well educated, job-search mechanisms are well-defined, there is a prevalence of tight labour markets (number of jobs are more than number of workers), and workers are well placed to evaluate their jobs. In this method, individuals are asked about the educational requirements of their job and whether they are adequately educated for it. A fundamental criticism of WA approach is that workers may not be equipped with the expertise to give an unbiased opinion about their job requirements.

Keeping the limitations of WA in mind, the literature has proposed an alternative approach where job experts define the boundary of tasks to be performed at work and in turn, provide a mechanism for wage determination. The proponents of this method assume that it is not workers' expertise that is essential but instead the tasks that need to be performed in a job are decisive factor for stating job requirements and consequently level of wages. This approach is called *job analysis* (JA). Given that experts are involved in the process of job descriptions, it is devoid of any individual biases and provides an objective measure of EOM. However, this method cannot be applied if jobs are not perfectly defined, for example, in informal labour markets where workers are expected to perform assorted tasks and no unique classification of these jobs is possible.

Apart from these two, a third approach tackles the question of matching workers to jobs using the perspective of labour market and hence involves the interaction of demand and supply side factors. This method hinges on the assumption that observable characteristics of workers and their revealed preferences through the choice of occupation are essential to EOM. The literature argues that the most important characteristic of workers is education level reflecting their relative ability (Spence 1973). On the other hand, occupation of the worker reflects two aspects: (i) availability of job opportunities and (ii) the constrained optimal choice by the worker keeping in mind, compensating wage differentials, prevailing economic conditions and so on. This method combines these two aspects of a labour market and is called *realized matches* (RM). The prerequisite of using this measure is to have information on education of workers and broad set of occupational categories. Among the discussed methods, RM requires the least level of information and lacks any biases. In this method, worker’s attained education is compared against a statistical threshold (such as mean, mode etc.) for investigating whether she is adequately educated for a job. The statistical threshold acts as a proxy for required education due to unavailability of exogenous threshold. In the context of slack labour market (number of workers is more than number of jobs) and when average attainment of education among the job-seekers is relatively higher (Baker 2014; Hartog and Oosterbeek 1988), the statistical measure is likely to be upward bias. This has two implications. First, the upward bias reflects that minimum required education for a given occupation does not reflect current expectations of the occupation. With the increase in the average level of education of labour force, if occupational requirements do not get updated, we are likely to see upward bias in statistical threshold such as average education of workers in that occupation. Second, this also leads to underreporting of overeducated workers. These two implications, however, highlight the advantages of this method, i.e., using the full educational distribution of workers in an occupation provides a better picture of incidence of EOM. This aids in identifying not only the type of match status but also the extent of it. Besides, it presents the current scenario of labour market conditions irrespective of nature of workers and jobs. Thus, it is a better measure to capture the wage disparities than other available methods. Further, given the ease of measurement and being less data demanding, it is suitable for cross-country and cross-sectional analysis of EOM.

Given the pros and cons of all three approaches, in our context, we use RM method for the following reasons. First, our analysis deals with a developing country where informal sector is predominant (Ruppert Bulmer 2018), and thus jobs are either unstructured or lack specific requirements. Hence, WA and JA cannot serve as effective measures. Second, due to less frequent revision of job requirements under JA method as well as lack of availability of JA classification it remains inapplicable for developing countries. Third, in case of developing countries, WA surveys are scant and even when available cannot be used due to heterogeneity in the education distribution and micro-macro environment. The only approach remains applicable is RM. Another advantage in disguise is that many studies in the literature have used this approach in the context of developed and developing countries (e.g., Blázquez and Rendon 2014; Haddad and Habibi 2016; Herrera-Idárraga, López-Bazo and Motellón 2015; Sharma and Sharma 2017). Thus, provides us leverage to compare our findings with the literature.

3. Data and descriptive statistics

This section discusses the data set used and some descriptive statistics to set the context for the paper.

3.1. Source description

The data used in this study come from the employment and unemployment survey, 2011-12 (68th round) collected by National Sample Survey Office (NSSO) in every five years. It is being carried out in all the states and union territories of India, the exception being inaccessible places in parts of India and thus, is representative of the labour market conditions in India. The survey covers 101,724 households (59,700 in rural areas and 42,024 in urban areas) comprising a total of 456,999 persons (280,763 in rural areas and 176,236 in urban areas). NSSO gives a sampling weight to each observation which makes an estimated number of households and individuals equals to the number of households and population in India respectively. NSSO (2014) provides the details of sampling procedure for the survey.

Our analysis is restricted to the working-age group, i.e., 15-59 years, covering a sample of 288,782 individuals. The NSSO survey contains detailed information about demographic variables such as age, gender, place of residence, etc., together with the job-related characteristics such as type of enterprise, number of workers, industry, etc. The survey also collects data on daily wages (cash and in-kind) earned by regular wage/salaried employees and casual labour during the reference week.[4] This information is complemented by the details on the intensity of work (0.5 for half day and 1 for a full day) performed in a day for all the seven days during the reference period. Wages, in this study, are calculated by dividing the total wages (sum of cash and in-kind) received during the reference week by the total number of days a person reported to work in that week. The survey, however, does not collect income or wage information for self-employed workers.

In the next sub-section, we provide a detailed discussion on the measurement of EOM.

3.2. Education-occupation (mis-)match: definition and measurement

Based on the discussion in the previous section on measures to estimate EOM, we adopt Realized Matches (RM) method for our analysis.

Definition: RM categorizes a person to be matched or mismatched by comparing the attained education with the mean years of education of workers in a given occupation (Verdugo and Verdugo 1989). We use, following Verdugo and Verdugo (1989), one standard deviation limit to ascertain the boundaries of required years of education for an occupation. More precisely, suppose e_i represents attained years of education of an individual i and $e_m$and s_e are the mean and standard deviation of years of education respectively for her occupation. She will be considered:

Overeducated if: $e_i > e_m + s_e$;

Undereducated if: $e_i < e_m - s_e$; and

Adequately educated if: $e_m - s_e \leq e_i \leq e_m + s_e$

In particular, she is adequately educated if her years of education lie between mean plus and minus one standard deviation threshold and overeducated (undereducated) if above (below) mean plus (minus) one standard deviation.

[4] Regular wage/salaried employees are the individuals who work in others' farm or nonfarm enterprises and receive regular wage or salary for their contribution. On the other hand, casual labour receives wages according to the terms of the daily or periodic work contract (NSSO 2014).

Measurement: As mentioned above, we require two inputs to use RM, i.e., occupation categories and years of education of the workers. Occupation of an individual is collected as per National Classification of Occupation (NCO) 2004 three-digit codes using around 115 occupation titles in our survey data.[5] A full list of occupation titles is available in the report published by NSSO (2014). NCO integrates these occupations into nine categories.[6] Table 1 list these categories.

<TABLE 1 HERE>

It can be witnessed that occupation categories follow a hierarchical order with legislators, senior officials and managers (occupation 1) being at the top and elementary occupations (occupation 9) being at the bottom. The occupations have also been categorised on the basis of required education (NCO 2004).

The NCO code is available for employed workers only (145,706 in our sample), and therefore, our estimates of EOM are limited to them. In our survey, we get information on the general level of education which we converted into years of education following Duraisamy (2002) (Refer to Annexure 1). We follow the two-stage procedure to capture EOM. In the first stage, we estimate the required years of education. For that, we calculate the mean and standard deviation of years of education of individuals employed in a specific occupation category using sampling weights and then establishing the cut-offs of mean plus and minus one standard deviation. In the second, i.e., identification stage, we categorize individuals into adequately, under, and overeducated workers using the required years of education as a threshold (Verdugo and Verdugo 1989). The next sub-section discusses the measures to capture the within-education group dispersion in returns.

3.3. Within-education group dispersion in returns: definition and measurement

The literature provides various methods to capture the dispersion such as Lorenz curve, range, Kuznets ratio, Gini coefficient, and coefficient of variation (Ray 1998).

So far, in the education literature, range and Kuznets ratio has been widely accepted as measure of within-group dispersion (Buchinsky 1998; Martins and Pereira 2004; Tansel and Bodur 2012). However, these methods have two problems. First, policymakers and researchers are

[5] The number of observations in different occupation titles ranges from 14 to 36,041.

[6] NCO 2004 also identifies the tenth category, workers not classified by occupations which include, new workers seeking employment, workers reporting occupations unidentifiable or inadequately described, and workers not reporting any occupations. We exclude this category from our analysis. In the process, we lose 377 observations.

often interested in examining the whole distribution of returns to summarize the dispersion. Range by using only maximum and minimum returns to education ignores the intricacies across the wage distribution. Second, the methods do not satisfy the Dalton principle of inequality. The principle states that any regressive transfer should increase the prevailing inequality in the economy. However, by considering only the sub-set of data, particularly extremes, range and Kuznets ratio overlook this aspect. Therefore, these methods can be misleading. Lorenz Curve and coefficient of variation considers the whole data to produce the dispersion measures. Also, Lorenz curve and coefficient of variation satisfy all the four principles viz., anonymity, population, relative income, and Dalton principle of inequality measurement (Ray 1998). Therefore, in this paper, within-education group dispersion in returns is captured using Lorenz curve and coefficient of variation along with range and Kuznets ratio. Ratio and Kuznets ratio are calculated to remain consistent with the literature and provide comparative results.

Definition and measurement:

a) Range – Range is calculated using the difference between the returns to education at different segments of wage distribution.

b) Kuznets Ratio – Kuznets ratio is depicted using the ratio of returns at different segments of wage distribution.

c) Lorenz Curve - Lorenz curve can be generated for the set of n ordered numbers, $y_1 \leq y_2 \leq y_3 \leq \cdots \leq y_n$. In this case, Lorenz curve at points $\frac{i}{n}$ $(i = 0, \dots., n)$ is defined by $L(0) = 0$ and $L\left(\frac{i}{n}\right) = S_i/S_n$, where $S_i = y_1 + y_2 + \cdots + y_i$. Therefore, we can define Lorenz curve $L(p)$ for all percentiles (p). In simple words, it expresses the fraction of total possessed by smallest p^{th} fraction (Gastwirth 1972).

d) Coefficient of variation –Coefficient of variation measures the relative variability. It is the ratio of standard deviation of the series to its mean.

In the next sub-section, we provide estimates of EOM and wage dispersion.

3.4. Descriptive statistics

This sub-section provides the descriptive statistics on EOM and prevalence of within-education group dispersion in the Indian labour market.

3.4.1. Education-occupation (mis-)match in India

In 2011-2012, the total stock of mismatched workers was around 141 million with an almost equal proportion of undereducated (71 million) and overeducated (70 million) workers. Table 2 depicts the match status of workers at an overall level and by gender.

<TABLE 2 HERE>

Table 2 highlights that the incidence of undereducation is marginally higher than that of overeducation in India. Further, overeducation is a more common phenomenon among male rather than female workers. This is in contrast with earlier findings (Groot and Van Den Brink 2000). A plausible explanation can be that women being secondary earners choose not to work when they are not able to find the adequate job for their level of education. Hence, they are not a part of employed population. For example, in India, around 73 percent of graduate and above females in the working-age group are either unemployed or out of labour force. Further, among graduates and above, the proportion of females is only 39 percent. Thus, the other explanation can be attributed to the lower level of education among all females than males.

Besides, we find a higher level of overeducation among the younger cohorts as compared to the older groups (Table 3). Sicherman (1991) contemplated that people at the start of their career may choose jobs that require lower education than their attained level as it helps them to gain necessary training and experience for subsequent career mobility. Therefore, young overeducated workers should not be of much policy concern (Robst 2007). Further, increasing rates of undereducation with age level can indicate the substitutability amongst various human capital variables. In particular, higher experience can aid workers to command jobs for which they are undereducated.

<TABLE 3 HERE>

3.4.2. Education-occupation (mis-)match and wage dispersion

Average daily wages for our sample is rupees (Rs.) 238 with workers residing in urban areas faring better (Rs. 384) than that of rural areas (Rs. 172). The gender wage gap is also evident with male and female workers earning an average daily wage of Rs. 264 and Rs.174 respectively. Table 4 shows average daily wages across different education groups by match status. We find that undereducated workers earn the highest wages followed by adequately educated and overeducated workers (the exception being graduates and above category where adequately educated workers earn the highest wages). The phenomenon represents a situation where workers with same level of education earn different wages due to their employment in different occupations. Undereducated workers are employed in occupation that requires more education as compared to their respective attained education and thus have a pay premium attached to it. In other words, they can be regarded as excellent performers (Verdugo and Verdugo 1989) or are lucky to get occupation premium. On the other hand, overeducated workers are generally employed in the jobs which have lower mean education and consequently reflect low-paying jobs. Also, increase in education does not necessarily leads to increase in productivity. Therefore, lower wages earned by overeducated workers highlight that they either underutilize the skills or they possess lower skills (Mateos-Romero and del Mar Salinas-Jiménez 2015). In a nutshell, there are significant wage differences among adequately, over, and undereducated workers at all levels of education groups. Hence, indicate the presence of wage differential among workers similar in terms of education but different in match status.

<TABLE 4 HERE>

Table 5 exhibits wages at 90th, 50th, and 10th quantiles by education. The positive relationship between education and wages exists at all the quantiles. However, a closer look indicates that as we move up the ladder of education, the difference and ratio of wages at different quantiles increases and hence confirms the "inequality increasing effect of education" found by previous studies (Martins and Pereira 2004). One possible reason could be that with increase in education, spectrum of job opportunities available widens. Hence, leads to varied wage profiles. However, note that these figures do not take account of heterogeneity in job and other demographic characteristics. Therefore, presents the partial picture.

<TABLE 5 HERE>

4. Estimation strategy

This section has two sub-parts. The first part explains outcome and explanatory variables. The next part discusses the empirical model, first unveiling the econometric model followed by various empirical specifications to be used.

4.1. Variables used

In our case, the outcome variable is the logarithm of daily wages calculated as the sum of cash and in-kind emoluments. Since individuals may have a varied number of working days which can influence wages, all the comparisons in this paper are made using daily wages. Further, other than the key interest variable, i.e., dummy for highest level of education (no formal schooling, primary, middle, secondary, higher secondary, and graduates and above), we include three kinds of covariates in our analysis.

First comprises human capital variables, namely, age, and age squared. The age of a worker is used as a proxy for her experience in the labour market. Also, the quadratic term for age allows for possible diminishing returns to experience. Most of the past studies use age minus years of schooling minus five as potential labour market experience. However, due to significant proportion of workers with no formal schooling, the measure was rendered unsuitable for India (Goel 2017).

Second, we include indicators of job characteristics, namely, occupation category (legislators, senior officials and managers, professionals, associate professionals, clerks, service workers and market sales workers, skilled agricultural and fishery workers, craft and related trades workers, plant and machine operators and assemblers, and elementary occupations), industry type (agriculture, manufacturing, construction, and services), the location of workplace (rural, urban, and no fixed location), enterprise type (proprietary, partnership, government, public/private, and other), firm size (less than 10 workers, 10-20, 20 and above, and unknown), and type of work contract (unwritten and written contract).

Lastly, the analysis captures set of personal and household characteristics, namely, gender (male and female), marital status (unmarried, married, and others), the interaction of gender and marital status, social group (scheduled tribe, scheduled caste, other backward class and others), religion (Hindu, Muslim, Christian, and others), sector (rural and urban), and state of residence. Introducing dummy for state helps to control for state-level heterogeneity. These are the standard variables that have been used by the previous studies as well (Agrawal 2012; Duraisamy 2002). Table 6 provides the summary statistics for all the variables.

<TABLE 6 HERE>

4.2. Empirical model

Given that our purpose is to ascertain whether EOM can explain the within-education group dispersion in returns, we employ the following methodology. In the first step, we measure the within-education group dispersion in returns taking account of human capital variables, job characteristics, and other personal and regional characteristics. In the later stage, we take account of EOM to explore how matching status of workers affects within-education dispersion in returns.

Given that we need estimates at different percentile to ascertain the covariates of dispersion in returns, the usual approach is to estimate a quantile regression (QR) model which is an extension of Mincerian (Mincer 1974)[7] wage equation in the context of distributional analysis. QR method was first introduced by Koenker and Basset (1978). The following equation represents QR:

$$y_i = X_i\beta_\theta + \ e_{\theta i} \ with \ Quant_\theta(y_i|X_i) = X_i\beta_\theta \quad (1)$$

where β and X_i are the vectors of parameters and explanatory variables respectively. $Quant_\theta(y_i|X_i)$ represents the θ^{th} conditional quantile of y given X.[8] QR method involves minimizing the weighted absolute value of the residuals. This can be done using linear programming methods where standard errors are obtained using bootstrap methods.

Using the above framework, Mincerian quantile wage equation to estimate the within-education group dispersion in returns is:

$$logw_i = \beta_0 + \ \beta_{1\theta i,k}S_{i,k} + \xi_\theta X_i + \ e_{\theta i} \quad (2)$$

where the subscript θ denotes the estimate at the θ^{th} conditional quantile. The dependent variable is the logarithm of daily wages. S_{ik} represent the dummy variables identifying the highest level of education (k) attained by individual i. The vector of other covariates is denoted by X_i and e is the error term.

One of the critical issues in Mincerian wage equation is that wages are not observed for all individuals. Not taking the note of the fact that wage distribution is truncated can lead to the problem of sample selection bias (Heckman 1977) which, in turn, results in biased and inconsistent estimates in the standard OLS as well as QR framework. Another issue which is specific to survey data in developing countries is the unavailability of wage/income information for the self-employed workers. This may be due to two reasons. First, income from self-employment represents a partial picture of economic prosperity for the self-employed. There are several other non-pecuniary benefits of self-employment such as higher autonomy,

[7] See, Lochner and Todd (2003) for detailed discussion on Mincerian wage equation.

[8] The θ^{th} regression quantile, $0 < \theta < 1$, is estimated by solving the following minimization problem: $\min_\beta \frac{1}{n}\left[\sum_{i:y_i \geq X_i\beta} \theta \ |y_i - X_i\beta| + \sum_{i:y_i < X_i\beta}(1-\theta) \ |y_i - X_i\beta|\right]$ (Buchinsky 1998). If we define 'check function,' $\rho_\theta(\varepsilon)$ as: $\rho_\theta(\varepsilon) = \begin{cases}(\theta - 1)\varepsilon, & \varepsilon < 0 \\ \theta\varepsilon, & \varepsilon \geq 0\end{cases}$; the minimization problem can be rewritten as $\min_\beta \frac{1}{n}\sum_{i=1}^{n} \rho_\theta(\varepsilon_{\theta i})$ (see Martins and Pereira 2004).

better socio-economic status, etc. (Le 1999). Thus, collecting information on income earned by self-employed is a crude measure of reflecting gains from self-employment. Second, self-reported income suffers from substantial underreporting (Baumeister 1982). Hence, instead of providing an incomplete picture or incurring measurement error or both, researchers/data collectors often find it easier not to collect this information. However, the choice of self-employment versus wage employed is not random and thus disregarding the possible sample selection into self-employment may again lead to sample selection bias (Dolton and Makepeace 1990).

Standard models in the literature have considered only the employment selection and thus ignored this aspect. A few notable exceptions are Agrawal and Agrawal (2018), Dolton and Makepeace (1990), among others. Hence, to take this into account, we use the double sample selection framework and consider two decisions – the decision to work (to be or not to be engaged in economic activity) and the choice of economic activity status (waged/salary or self-employment). The two-stage estimation procedure suggested by Heckman (1977) is the widely accepted method to correct for sample selection bias. In the first stage, i.e., participation equation, the probability that a person participate in the labour market is estimated to obtain the selection term captured by inverse mills ratio[9]. In the later stage, wage equation is estimated using selection term as one of the covariates. This model works on the standard assumptions of OLS. However, in the case of QR, the bias term is of unknown form, and hence, the premise of normality of errors is usually rejected. Thus, renders the unsuitability of Heckman method for the analysis (Buchinsky 2002). Therefore, as suggested by Buchinsky (1998, 2002) we employ the semi non-parametric correction method.[10] In the first step, we estimate the participation equations using semi-non-parametric estimation:

$$Y_{1i}^{*} = z_{1i}\xi + u_{1i} \qquad (3a)$$

$$Y_{2i}^{*} = z_{2i}\xi + u_{2i} \qquad (3b)$$

where * indicates the unobserved variable. The dependent variable in equation (3a) is a binary variable which takes a value of 1 when a person is employed and 0 when a person is not employed, and dependent variable in equation (3b), is binary variable which takes a value of 1 when an individual is employed and his wage is observed (wage/salary employee) and 0

[9] Inverse Mills ratio is defined as the ratio of the probability density function to the cumulative distribution function.

[10] This is done by using snp command in STATA v15. See De Luca (2008) for its detailed application in STATA.

when a person is employed but wage is not observed (self-employed). Here, z_i, i =1,2 captures the observed variables and u_i, i =1,2 corresponds to the error term that captures the impact of unobserved variables. It is a prerequisite to identify at least one variable that does not affect the wages but influences the probability of participation, called as exclusion variable(s) to be used in this method. This implies that z_{1i} and z_{2i} should include at least one exclusion variable that influences the decision to work and choice of economic activity status respectively but not the wages. For equation (3a) we use land holding by the household,[11] the number of dependent members (aged below 6 years or above 60 years) in a household, household type, and household size as exclusion variables. This is in line with the literature that advocates the use of family characteristic variables as appropriate exclusion variable for the choice of work (Buchinsky 2002; Agrawal 2012). Further, Dolton and Makepeace (1990) argue that the choice of self-employment versus wage/salary employment depends on personal attributes, knowledge, and ability. Thus, we use vocational and technical education, household head's education and occupation, and primary occupation of a household as the exclusion variables for equation (3b). Hence, participation equations are estimated using the human capital variables, other personal and regional characteristics, and respective exclusion variables. We do not include job characteristics while estimating participation equations since these are only observed after a person is employed and thus cannot be considered as determinants of choice of work or economic activity status.

In the second step, QR is estimated using inverse Mills ratios and its square as one of the independent variables.[12] Hence, equation (2) can be modified as follows:

$$\begin{aligned} logw_i = \beta_0 + \ \beta_{1\theta i,k} S_{i,k} + \ \xi_\theta X_i + \delta_{\theta 1} Sel_{wc} + \delta_{\theta 2} Sel_{wc}^2 + \ \delta_{\theta 3} Sel_{we} \\ + \delta_{\theta 4} Sel_{we}^2 \ + \ e_{\theta i} \qquad (4) \end{aligned}$$

where Sel_{wc} and Sel_{we} refer to inverse mills ratio estimated respectively from participation equation 1, i.e., the decision to work and participation equation 2, i.e., the choice of economic activity status. Other variables are interpreted as before. We estimate this model for every decile from 10th to 90th percentile.

In the next sub-section, we provide the detailed description of empirical specifications used in the econometric analysis.

[11] Land holding is estimated by considering the maximum of land owned and land possessed.

[12] We use both mills ratio and its square term in the wage equation following Buchinsky (1998, 2001).

4.3. Empirical specifications

Given the near absence of a theoretical framework for examining the relationship between EOM and wages, we adopt from theories of wage determination, i.e., human capital theory (HCT) (Becker 1964), job competition model (JCM) (Thurow 1975), and assignment model (Sattinger 1993) in this context. HCT affirms that individuals earn wages based on their respective productivity level which is determined by the attained education, experience in the labour market, etc. combined to be called human capital. It contends that workers with similar human capital earn equal wages irrespective of their job characteristics. Contrary to HCT, JCM contends wages to be a function of job characteristics only. Hence, it argues that the job definition and its requirements are the main contributors to the dispersion in returns among similar workers. Assignment model lies between HCT and JCM and argues wages to be dependent on both human capital and job characteristics.

Keeping them in mind, we use the following empirical specifications to answer the question in hand, i.e., what is the impact of education-occupation (mis-)match (EOM) on within-education group dispersion in returns?

The first empirical specification (specf. 1) includes apart from education, other human capital variables, personal and regional characteristics and self-selection terms.

$$logw_i = \beta_0 + \beta_{1\theta i,k}S_{i,k} + \xi_\theta X_i + \delta_{\theta 1}Sel_{wc} + \delta_{\theta 2}Sel_{wc}^2 + \delta_{\theta 3}Sel_{we} + \delta_{\theta 4}Sel_{we}^2 + e_{\theta i} \quad (5)$$

Variables are interpreted as before. However, in this equation, to capture the sole effect of human capital characteristics on wages, we do not include job characteristics in X_i.

The second empirical specification (specf. 2) is consistent with assignment model and includes both human capital and job characteristics.[13]

$$logw_i = \beta_0 + \beta_{1\theta i,k}S_{i,k} + \xi_\theta X_i + \delta_{\theta 1}Sel_{wc} + \delta_{\theta 2}Sel_{wc}^2 + \delta_{\theta 3}Sel_{we} + \delta_{\theta 4}Sel_{we}^2 + e_{\theta i} \quad (6)$$

Variables are interpreted as before. Moreover, in this equation X_i includes all the three kinds of covariates, i.e., human capital, job, and other personal and regional characteristics. We hypothesize that equation (6) should exhibit lower dispersion in returns than previous

[13] To be consistent with job competition model, we also estimated the solitary impact of job characteristics on wages. For the sake of brevity, the JCM estimation and corresponding results are not provided, but are available from the authors on request.

specification. This is because workers with same education may witness dispersion in returns due to differences in their job characteristics. Hence, by taking account of both human capital and job characteristics, the above specification expands the similarity quotient of workers. Thus, it should reduce, if not alleviate the dispersion in returns among workers with similar education.

Subsequently, in third specification (specf. 3), we include EOM variables to measure the impact of EOM on dispersion in returns to education. This is done using Verdugo and Verdugo (1989)'s procedure.[14]

$$logw_i = \beta_0 + \beta_{1\theta i,k} S_{i,k} + \xi_\theta X_i + \beta_{3\theta i} D_i^o + \beta_{4\theta i} D_i^u + \delta_{\theta 1} Sel_{wc} + \delta_{\theta 2} Sel_{wc}^2 + \delta_{\theta 3} Sel_{we} + \delta_{\theta 4} Sel_{we}^2 + e_{\theta i} \quad (7)$$

where $D_i^o$and D_i^u are the dummies for overeducated and undereducated respectively. Other variables are interpreted as before.

For EOM to be a significant determinant of wages, spec. (3) should explain variation in wages better that specf. (1) and (2). This can be tested using likelihood ratio test (LRT). LRT is a statistical tool that enables to test two statistical models where one is a special case of other. In our case, equation (6) represents a nested model of equation (7) and hence, we can employ LRT to determine the apt model in our context.

Also, we expect this specification to explain the within-education group dispersion in returns further. There are two possible reasons for the same. First, by categorizing the workers by match status, we further accentuate the criterion of similarity among workers. Workers have not only same education and other characteristics but also the match status. Second, EOM acts as a signal in the labour market and therefore, leads workers to command different wages as per their match status. Hence, taking account of differential match status, the within-education group dispersion should be further explained.

The next section contains the detailed results.

[14] Duncan and Hoffman (1981) provide another approach to estimate the impact of EOM on wages. The authors used continuous variable for education, i.e., years of education and divided this into two parts: required years of education and over/under years of education. However, the problem is by including continuous years of education, it does not allow us to consider level of education and hence fails to solve the purpose of this paper.

5. Results and discussion

This section has three sub-parts. The first part discusses the ordinary least squares (OLS) results. The second part presents the quantile regression (QR) estimates. Finally, the last part answers the critical question of our paper by measuring the differences in the within-education group dispersion in returns across empirical specifications.

5.1 Results from ordinary least square (OLS)

Table 7 presents OLS estimates of the wage equation for different specifications as mentioned in the previous sub-section (Section 4.3).[15] In specf. (1), we find positive association between level of education and wages (in line with Barrett, Callan and Nolan 1999). In specf. (2), even though the association between wages and level of education remains unchanged, the magnitude of coefficients reduces. A possible reason can be that level of education and job characteristics are correlated and taking only one set of variables can lead to omitted variable bias and thus upward or downward bias in estimates of coefficients. The results highlight that other things being constant, workers with different combinations of education and occupation, earn differential returns in the labour market. This brings us to the importance of interplay of attained education and occupational requirements, i.e., *education-occupation (mis-)match*.

Coming to specf. (3) which accounts for EOM, we find that, on an average, overeducated workers suffer a wage penalty of around 7 percent and undereducated workers receive a wage reward of about 3 percent as compared to their adequately educated counterparts. The earlier studies have also found similar results for other countries (Hartog 2000; Leuven and Oosterbeek 2011). The finding could be due to following reasons: (i) overeducated workers are generally employed in the jobs which have lower mean education and thus reflect low-paying jobs and (ii) undereducated workers work in the occupations that are more challenging in comparison to their education, and thus they command higher wages as a premium to perform the job tasks.

Further, we find that there is a considerable increase in returns to education as compared to specf (2). EOM being correlated to wages and also to education leads to omitted variable bias if ignored. As shown in descriptive statistics (Section 3.3), the overeducated workers earn lower wages followed by adequately educated and then undereducated workers for a given level of education. Also, the penalty of being overeducated is higher than reward associated

[15] For the sake of brevity, the estimates of only concerned variables are given. The estimates for the full set of variables are available upon request.

with being undereducated. Therefore, the presence of overeducated workers brings down the average returns for a given level of education.

<TABLE 7 HERE>

As mentioned earlier (Section 4.3), we use likelihood ratio test to conclude the most suitable model for explaining wages. Likelihood ratio test (value = 59.29) rejects the null hypothesis that model without EOM (specf.2) is a better model to explain wages and hence concludes that specf. (3) provides the better estimates of returns to education. We also found the evidence of sample selection bias in our data. The selection term for the choice of economic activity status is statistically significant for all the specifications. However, the selection term for the decision to work was statistically insignificant for specf. (2) and (3).

Given the restrictive results towards mean outcomes, OLS does not unveil significant differences in the returns across wage distribution. Therefore, the recent literature makes use of quantile regression to uncover this dimension (e.g., Agrawal 2012; Azam 2012). We present the results for conditional quantile regression in the next sub-section.

5.2. Results from quantile regression (QR)

The results for conditional quantile regression are provided in Table 8. We restrict our estimation to specification 2 and 3 only. Also, to ease the comparison of estimates at different quantiles with the average returns, OLS estimates are reproduced in column 1. The key results are also presented in Figure 1.

<TABLE 8 HERE>

The following results are noticeable. First, in consistency with OLS estimates, there is a positive relationship between level of education and associated returns, irrespective of the wage quantiles. Thus, suggest that there are increasing additional returns to increase in education. Further, for any education level there is increase in returns as we move from lower to higher quantiles of wages. A plausible reason often quoted in the literature is that it can be indicative of increase in ability and quality of education of workers as we move to higher wage quantiles (Herrera-Idárraga, López-Bazo and Motellón 2015).

<FIGURE 1 HERE>

Second, the reward associated with undereducation increases up to 70th quantile and then decreases (however, the difference between the returns at 70th and 80th quantile is statistically insignificant) and the penalty associated with overeducation decreases over the quantiles up to 80th quantile. This finding can be looked from the perspective of ability segments (McGuinness and Bennett 2007). The authors categorize and associate top end of wage distribution with the high-ability segments and vice-versa. Thus, workers tend to compensate their overeducation and complement their undereducation with high ability. The opposite holds true for workers in low-ability segments. The finding indicates that workers at the lower end of the wage distribution suffer double penalty. On one side, they earn lower returns to education and on the other side, they receive higher penalty for being overeducated and lower rewards for being undereducated.

The key takeaways from these results are that, there is heterogeneity in returns to same level of education, as well as in coefficients of over and under education indicators. In the next sub-section, we do a more detailed analysis of this heterogeneity and its explanation. Also, we will turn to answer the primary question of this paper, i.e., whether EOM can explain within-education group dispersion in returns?

5.3. Education-occupation mismatch and within-education group dispersion in returns

The differences in the returns to education across the wage distribution indicate the presence of dispersion in the returns to same level of education. This sub-section explores that whether EOM affects this dispersion.

Figure 2 presents the Lorenz curve for returns to education as per specf. (2) and (3). The line at the 45° angle indicates perfect equality in returns to education, while the other lines show the actual distribution of returns to education as per both the specifications. The further away from the diagonal, the more unequal the returns to education or, in other words, there is higher dispersion in returns to education. We find that except for primary education, in all the other education groups, specf. (3) exhibits lower dispersion in returns than specf. (2).

<FIGURE 2 HERE>

Also, using the coefficient of variation as a measure of inequality, we find that there is a significant decline in coefficient of variation for middle, secondary, and higher secondary after we take account of EOM (Table 9). For the other two education groups, i.e., primary and graduate and above, there is a marginal increase in coefficient of variation as we move from specf. (2) to (3).

<TABLE 9 HERE>

Table 10 presents the difference between returns at 90th and 10th quantile (Range: θ90 – θ10) and ratio of returns at 90th and 10th quantile (Kuznets Ratio: θ90/θ10) for different specifications. Note that it is not the absolute but excess dispersion in returns within the group compared to the reference group. We further divide θ90/θ10 (θ90 – θ10) into θ90/θ50 (θ90 – θ50) and θ50/θ10 (θ50 – θ10) percentile to analyse the top-end and low-end disparities respectively. We observe that relatively large disparities in returns take place mainly in the lower half of the distribution for all levels of education. This is indicated by higher θ50 – θ10 spread and θ50/θ10 ratio as compared to θ90 – θ50 spread and θ90/θ50 ratio respectively. The flat returns at top end of the wage distribution are an obvious reason for this finding. Further, as per specf. (3), we find insignificant within-education dispersion in returns for primary, higher secondary, and graduate and above. For rest, *viz.*, middle and secondary, the dispersion is lower as compared to specf. 2. For example, after considering EOM, θ90 – θ10 spread is 4.9 percent and 4.7 percent in middle and secondary level respectively, and the corresponding spreads rockets to 5.7 percent and 6.1 percent, if we do not take account of prevailing EOM in the labour market. Similar results are witnessed for wage ratios.

<TABLE 10 HERE>

Contradictory to earlier studies that have found inequality increasing effect of education (e.g., Azam 2012; Martins and Pereira 2004), we do not see a positive relationship between education and within-education dispersion in returns across all the measures of inequality. Omitted variables such as, information on occupation, and EOM in the previous studies may be responsible for the deviation in the results. On the basis of the observation of higher wage inequality amongst higher educated workers, Azam (2012) warned that wage inequality in urban India will increase with the increase in the proportion of people getting higher education. Our study, on the other hand, claims that differences in the match type can be responsible for the within-education group wage dispersion in returns and hence EOM should be paid a due share of attention by the policymakers.

To summarize, as hypothesized, EOM does explain some part of the within-education group dispersion in returns. The finding can be owed to following reasons. First, wages are not decided individually by education or occupation but by combination of these two. Depending on the extent and magnitude of EOM for a particular education group, the degree of within-group dispersion in returns may vary. Hence, workers with same level of education may receive higher or lower returns to education than average depending on their alignment with the respective occupation which leads to within-education group dispersion in returns. Therefore, taking account of EOM succinct match type in a particular education group and consequently leads to better explanation of within-group dispersion. Second, returns to education comprises of returns to two segments: required education and over/under education. While the returns to former are always positive, the latter may lead to differential returns and thus can lead workers with same education to command different returns. Thus, education in itself do not lead to within-group dispersion in returns but the differences in the match status among workers with same education result in the varied return profiles. This indicates that EOM is an important factor that aids in explaining the within-education group dispersion. EOM does not only lead to varied returns between education groups but also dispersion in returns for a particular education group. The results are crucial since they highlight that EOM needs to be corrected if policymakers intend to reduce within-education group dispersion in returns, as well as need to estimate correct returns to education level as inputs in policies.

6. Conclusion

The paper empirically examines the relevance of education-occupation (mis-)match in explaining heterogeneity in returns to same level of education. Through EOM, we capture the interplay between workers and job characteristics in understanding within-group dispersion in returns to education.

Our main findings can be summarized as follows. First, we observe that, on an average, overeducated workers suffer a wage penalty of around 7 percent and undereducated workers receive a wage reward of approximately 3 percent as compared to their adequately educated counterparts. Second, we also provide estimates for returns to education while taking account of double sample selection bias. This is a novel contribution. Last and the most important finding is that the inclusion of match status affects within-education group dispersion in returns. The finding highlights that ignoring the EOM and thus, adopting a restrictive view of similarity across workers leads to overestimation of the within-education group dispersion in returns. Also, contrary to previous studies, we do not find inequality increasing effect of education once we integrate the various labour market theories. This indicates that capturing the impact of isolated labour market aspects presents the partial picture of dispersion in returns across similarly educated workers.

The study highlights that policymakers interested in reducing wage inequality should pay attention to mismatch between workers and firm characteristics. This is crucial to harness the wage benefits of education. This is depicted by the higher returns to education after we take account of EOM. Hence, the varied match type will bring the average returns to a particular education group. From the workers' perspective, there should be labour market institutions facilitating the proper matching of skills and education to the requirement of occupation. Such institutions should not just focus on providing employment but a decent and well-matched. A failure to do so can lead to hampering of future prospects in terms of lower wages, limited upward career mobility choices, and lower satisfaction. Besides, when firms hire mismatched workers especially overeducated workers, given the wage penalty and job dissatisfaction due to wage dispersion (Fleming and Kler 2008), the attrition is likely to be higher. This leads to increase in the cost in terms of training, hiring, etc. Providing excessive education without considering the demand side can lead to potential productivity losses especially based on pecuniary returns and barring the social returns to education. This is a severe issue especially for developing countries where resources are even scarce as compared to the developed countries.

These results can also be used for policymaking in developing countries. One such issue is the long-standing debate of choice between universal primary education versus selected population with tertiary education in accelerating and sustaining the economic growth when resources are limited (Castelló-Climent and Mukhopadhyay 2013). There is supporting evidence towards, the former highlighting that education and returns to education holds a negative relationship with each other (Psacharopoulos and Patrinos 2004). However, the recent evidence suggests the opposite phenomenon to be true, i.e., returns to education increases with the level of education (Agrawal 2012; Barro and Lee 2013). Hence, supports the view that higher education is the key to growth. Our results further substantiate and complement these findings. We argue that providing higher education and facilitating the adequate match between workers' education and required education by the occupation will lead to higher returns and therefore accentuates the pecuniary benefits of education. However, focusing only on one aspect, i.e., providing higher education or creating education-intensive jobs may not only lead to loss of productivity but also lower wages. In a nutshell, we contend that providing higher education may not convert into higher growth if the available job opportunities are inadequate and inapt. Therefore, in developing countries, where resources are scarce, the adequate balance between providing higher education and creating education-intensive jobs should be maintained.

Although the study provides a rich description of within-education group dispersion in returns, the limitations are inevitable. The unavailability of information such as, quality of schooling, field of training, cognitive and non-cognitive skills etc. limits our scope to differentiate workers within the same education group. Another dimension which can further enrich our study is to include spatial aspects of labour market. The reason being that the probability of employment, and consequent match status is affected by the prevailing conditions in local labour market. When the person chooses her job or occupation, the physical limits of labour market plays a crucial role.

To capture the dynamic aspect of dispersion in returns in labour market, the natural extension of this study would be to conduct the inter-temporal analysis. This would help us in understand changing dynamics of educational inequality and its long- term relationship with wage inequality using the current EOM framework.

References

Agrawal, T. (2012). Returns to education in India: Some recent evidence. *Journal of Quantitative Economics*, 10, 131–151.

Agrawal, T., & Agrawal, A. (2018). Who Gains More from Education? A Comparative Analysis of Business, Farm and Wage Workers in India. *The Journal of Development Studies*, 1-18.

Azam, M. (2012). Changes in wage structure in urban India, 1983–2004: A quantile regression decomposition. *World Development*, 40(6), 1135-1150.

Baker, D. (2014). *The schooled society: The educational transformation of global culture*. Stanford University Press.

Barrett, A., Callan, T., & Nolan, B. (1999). Rising wage inequality, returns to education and labour market institutions: evidence from Ireland. *British Journal of Industrial Relations*, 37(1), 77-100.

Barro, R. J., & Lee, J. W. (2013). A new data set of educational attainment in the world, 1950–2010. *Journal of Development Economics*, 104, 184-198.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin*, 91(1), 3.

Becker, G. S. (1964). *Human capital theory*. Columbia, New York, 1964.

Becker, G. S. (2010). *The economics of discrimination*. University of Chicago press.

Blázquez, M., & Rendon, S. (2014). Over-education in multilingual economies: Evidence from Catalonia. *International Migration*, 52(6), 149-164.

Buchinsky, M. (1998). The dynamics of changes in the female wage distribution in the USA: a quantile regression approach. *Journal of Applied Econometrics,* 1-30.

Buchinsky, M. (2002). Quantile regression with sample selection: Estimating women's return to education in the US. In Economic Applications of Quantile Regression (pp. 87-113). Physica, Heidelberg.

Budría, S., & Moro-Egido, A. I. (2008). Education, educational mismatch, and wage inequality: Evidence for Spain. *Economics of Education Review*, 27(3), 332-341.

Budría, S., &Telhado-Pereira, P. (2011). Educational qualifications and wage inequality: evidence for Europe. *Revista de Economía Aplicada*, 19(56), 5.

Castelló-Climent, A., & Mukhopadhyay, A. (2013). Mass education or a minority well educated elite in the process of growth: the case of India. *Journal of Development Economics*, 105, 303-320.

Catsiapis, G., & Robinson, C. (1982). Sample selection bias with multiple selection rules: An application to student aid grants. *Journal of Econometrics*, 18(3), 351-368.

De Luca, G. (2008). SNP and SML estimation of univariate and bivariate binary-choice models. Stata Journal, 8(2), 190.

Devine, T. J., Devine, T. J., & Kiefer, N. M. (1991). *Empirical labor economics: the search approach*. Oxford University Press on Demand.

Dolton, P. J., & Makepeace, G. H. (1990). Self employment among graduates. *Bulletin of Economic Research*, 42(1), 35-54.

dos Reis, J. G. A., & de Barros, R. P. (1991). Wage inequality and the distribution of education: A study of the evolution of regional differences in inequality in metropolitan Brazil. *Journal of Development Economics*, 36(1), 117-143.

Duncan, G. J., & Hoffman, S. D. (1981). The incidence and wage effects of overeducation. *Economics of Education Review*, 1(1), 75-86.

Duraisamy, P. (2002). Changes in returns to education in India, 1983–94: by gender, age-cohort and location. *Economics of Education Review*, 21(6), 609-622.

Eckstein, Z., & Van den Berg, G. J. (2007). Empirical labor search: A survey. *Journal of Econometrics*, 136(2), 531-564.

Fleming, C. M., &Kler, P. (2008). I'm too clever for this job: a bivariate probit analysis on overeducation and job satisfaction in Australia. *Applied Economics*, 40(9), 1123-1138.

Gastwirth, J. L. (1972). The estimation of the Lorenz curve and Gini index. *The Review of Economics and Statistics*, 306-316.

Goel, M. (2017). Inequality between and within skill groups: The curious case of India. *World Development*, 93, 153-176.

Green, F., & Zhu, Y. (2010). Overqualification, job dissatisfaction, and increasing dispersion in the returns to graduate education. *Oxford Economic Papers*, 62(4), 740-763.

Groot, W., & Van Den Brink, H. M. (2000). Overeducation in the labor market: a meta-analysis. *Economics of Education Review*, 19(2), 149-158.

Haddad, G. K., & Habibi, N. (2017). Vertical skill mismatch and wage consequences in low-skilled jobs: Evidence from Iran. *International Labour Review*, *156*(1), 45-72.

Hartog, J. (2000). Over-education and earnings: where are we, where should we go?. *Economics of Education Review*, 19(2), 131-147.

Hartog, J., &Oosterbeek, H. (1988). Education, allocation and earnings in the Netherlands: 0verschooling?. *Economics of Education Review*, 7(2), 185-194.

Heckman, J. J. (1977). Sample selection bias as a specification error (with an application to the estimation of labor supply functions).

Heckman, J. J., Lochner, L. J., & Todd, P. E. (2003). Fifty years of Mincer earnings regressions (No. w9732). National Bureau of Economic Research.

Henderson, D. J., Polachek, S. W., & Wang, L. (2011). Heterogeneity in schooling rates of return. *Economics of Education Review*, 30(6), 1202-1214.

Herrera-Idárraga, P., López-Bazo, E., &Motellón, E. (2015). Double penalty in returns to education: informality and educational mismatch in the Colombian labour market. *The Journal of Development Studies*, 51(12), 1683-1701.

Kierzenkowski, R., &Koske, I. (2012). Less income inequality and more growth–are they compatible? Part 8. The drivers of labour income inequality–a literature review.

Koenker, R., & Bassett Jr, G. (1978). Regression quantiles. *Econometrica: Journal of the Econometric Society*, 33-50.

Le, A. T. (1999). Empirical studies of self-employment. *Journal of Economic Surveys*, 13(4), 381-416.

Leuven, E., & Oosterbeek, H. (2011). *Overeducation and mismatch in the labor market*. In Handbook of the Economics of Education (Vol. 4, pp. 283-326). Elsevier.

Martins, P. S., & Pereira, P. T. (2004). Does education reduce wage inequality? Quantile regression evidence from 16 countries. *Labour Economics*, 11(3), 355-371.

Mateos-Romero, L., & del Mar Salinas-Jiménez, M. (2015). Skills heterogeneity among graduate workers: real and apparent overeducation in the Spanish labor market. *Social Indicators Research*, 1-18.

McGuinness, S., & Bennett, J. (2007). Overeducation in the graduate labour market: A quantile regression approach. *Economics of Education Review*, 26(5), 521-531.

Mincer, J. (1974). *Schooling, Experience, and Earnings*. Human Behavior & Social Institutions No. 2.

Montgomery, J. D. (1991). Equilibrium wage dispersion and interindustry wage differentials. *The Quarterly Journal of Economics*, 106(1), 163-179.

National Classification of Occupations (NCO) (2014). "Introduction of National Classification of Occupation, 2004". http://www.dget.nic.in/upload/uploadfiles/files/publication/1%20preface.pdf

National Sample Survey Office (NSSO) (2014). "Employment and Unemployment Situation in India". Report No. 554 (68/10/1). New Delhi: Government of India.

Postel–Vinay, F., & Robin, J. M. (2002). Equilibrium wage dispersion with worker and employer heterogeneity. *Econometrica*, 70(6), 2295-2350.

Psacharopoulos, G., & Patrinos, H. A. (2004). Returns to investment in education: a further update. *Education Economics*, 12(2), 111-134.

Ray, D. (1998). Development economics. Princeton University Press.

Robst, J. (2007). Education and job match: The relatedness of college major and work. *Economics of Education Review*, 26(4), 397-407.

Ruppert Bulmer, E. (2018). Defining informality vs mitigating its negative effects. *IZA World of Labor*:442. doi: 10.15185/izawol.442

Salverda, W., & Checchi, D. (2015). Labor market institutions and the dispersion of wage earnings. In Handbook of income distribution (Vol. 2, pp. 1535-1727). Elsevier.

Sattinger, M. (1993). Assignment models of the distribution of earnings. *Journal of Economic Literature*, 31(2), 831-880.

Sharma, S., & Sharma, P. (2017). Educational mismatch and its impact on earnings: evidence from Indian labour market. *International Journal of Social Economics*, 44(12), 1778-1795.

Sicherman, N. (1991). " Overeducation" in the Labor Market. *Journal of Labor Economics*, 9(2), 101-122.

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, 87(3), 355-374.

Tansel, A., &Bodur, F. B. (2012). Wage inequality and returns to education in Turkey: A quantile regression analysis. *Review of Development Economics*, 16(1), 107-121.

Thurow, L. C. (1975). Generating inequality. Basic books.

Verdugo, R. R., &Verdugo, N. T. (1989). The impact of surplus schooling on earnings: Some additional findings. *Journal of Human Resources*, 629-643.

Tables

Table 1: Division and respective occupation title

Division	Title	Skill level	Years of formal education
1	Legislators, Senior Officials and Managers	Skill level not defined*	NA
2	Professionals	4	15 and above
3	Associate Professionals	3	14-15
4	Clerks	2	11-13
5	Service Workers and Shop & Market Sales Workers	2	11-13
6	Skilled Agricultural and Fishery Workers	2	11-13
7	Craft and Related Trades Workers	2	11-13
8	Plant and Machine Operators and Assemblers	2	11-13
9	Elementary Occupations	1	0-10

Source: National Classification of Occupation (2004).

*The concept of skill level has not been applied in the case of legislators, senior officials & managers as skills for executing task and duties of these occupations vary to such an extent that it would be impossible to link them with any of the four broad skill levels.

Table 2: Education and occupation (mis-)match by gender and location (in percentage)

Match Type	Overall	Gender	
		Male	Female
Undereducated	20.15	17.79	28.33
Adequately educated	59.99	59.96	60.07
Overeducated	19.87	22.24	11.6

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.

Note: Sampling weights have been used.

Table 3: Education and occupation (mis-)match by age cohorts (in percentage)

Match Type	15-24 years	25-34 years	35-44 years	45-59 years
Undereducated	12.41	15.23	21.55	28.94
Adequately Educated	61.77	60.05	59.6	59.19
Overeducated	25.82	24.72	18.85	11.88

Source: Authors’ calculation based on NSSO employment and unemployment survey, 2011-12.
Note: Sampling weights have been used.

Table 4: Average daily wages (in Indian rupees) across match status by education

Education level	Undereducated	Adequately Educated	Overeducated	Total
No Formal Schooling	153	118	-	133
	(86.11)	(16.81)		(27.43)
Primary or Below	222	151	-	155
	(8.64)	(36.89)		(23.87)
Middle	296	203	149	184
	(1.92)	(21.92)	(20.1)	(17.53)
Secondary	471	280	176	244
	(2.01)	(9.37)	(33.21)	(12.63)
Higher Secondary	678	410	236	351
	(1.12)	(7.01)	(21.35)	(8.67)
Graduate and Above	589	750	550	680
	(0.19)	(7.99)	(25.34)	(9.87)
Total	198	254	261	247

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.

Note: Sampling weights have been used.

Numbers in parenthesis indicate proportion of workers in that category.

Table 5: Daily wages (in Indian rupees) across education

	Quantile			Ratio		
	10th	50th	90th	50/10	90/50	90/10
Education:						
No Formal Schooling	64	117	200	1.82	1.71	3.11
Primary or Below	71	130	250	1.82	1.92	3.5
Middle	78	150	333	1.91	2.22	4.25
Secondary	97	171	500	1.77	2.92	5.16
Higher Secondary	100	233	743	2.33	3.18	7.43
Graduate and Above	140	536	1300	3.83	2.43	9.29

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.

Note: Sampling weights have been used.

Table 6: Summary statistics by match status

	Overall	Overeducated	Adequately	Undereducated
Number of Observations	70,588	13,399	44,630	10,681
Education				
No Formal Schooling	0.19	-	0.12	0.69
Primary	0.2	-	0.29	0.11
Middle	0.17	0.28	0.16	0.07
Secondary	0.13	0.24	0.11	0.09
Higher Secondary	0.12	0.19	0.12	0.04
Graduate and Above	0.19	0.29	0.21	0.01
Age	36.05	33.19	36.36	38.72
Occupation				
Legislators, Senior Officials and Managers	0.02	0	0.03	0.02
Professionals	0.08	0.04	0.09	0.07
Associate Professionals	0.11	0.1	0.12	0.09
Clerks	0.06	0.04	0.07	0.05
Service Workers and Shop & Market Sales Workers	0.1	0.08	0.11	0.1
Skilled Agricultural and Fishery Workers	0.03	0.03	0.03	0.05
Craft and Related Trades Workers	0.17	0.16	0.16	0.22
Plant and Machine Operators and Assemblers	0.08	0.05	0.09	0.09
Elementary Occupations	0.33	0.5	0.29	0.31
Industry				
Agriculture	0.17	0.21	0.19	0.07
Manufacturing	0.15	0.15	0.14	0.16
Construction	0.2	0.27	0.14	0.37
Services	0.48	0.37	0.53	0.4
Work Location				
Rural	0.44	0.44	0.42	0.51
Urban	0.53	0.54	0.56	0.45
No Fixed Location	0.03	0.03	0.03	0.04
Enterprise Type				
Proprietary	0.44	0.43	0.41	0.55
Partnership	0.02	0.02	0.02	0.02
Government	0.33	0.3	0.37	0.24
Public/Private	0.1	0.13	0.1	0.07
Other	0.11	0.11	0.1	0.12
Firm Size				
Less than 20	0.53	0.51	0.51	0.6
10-20	0.13	0.12	0.13	0.12
20 & Above	0.26	0.28	0.27	0.19
Unknown	0.09	0.09	0.08	0.09
Type of Contract				
Unwritten	0.7	0.72	0.67	0.81
Written	0.3	0.28	0.33	0.19
Gender				
Male	0.78	0.85	0.78	0.79
Female	0.22	0.15	0.22	0.21
Marital Status				
Unmarried	0.20	0.27	0.2	0.12
Married	0.75	0.71	0.76	0.8
Others	0.05	0.02	0.05	0.08
Social Group				
Scheduled Tribe	0.14	0.11	0.15	0.15
Scheduled Caste	0.20	0.2	0.19	0.24
Other Backward Class	0.37	0.39	0.37	0.37

Table 6: Summary statistics by match status

	Overall	Overeducated	Adequately	Undereducated
Others	0.29	0.3	0.3	0.24
Religion				
Hindu	0.77	0.8	0.76	0.75
Muslim	0.12	0.1	0.12	0.17
Christian	0.07	0.07	0.08	0.05
Others	0.04	0.04	0.04	0.04
Sector				
Rural	0.55	0.56	0.53	0.58
Urban	0.45	0.44	0.47	0.42

Source: Authors’ calculation based on NSSO employment and unemployment survey, 2011-12.

Note: Sampling weights have not been used.

Table 7: Returns to education –Ordinary Least Square (OLS) estimates

Explanatory Variables	Outcome Variable: Logarithm of Daily Wages		
	(1)	(3)	(4)
	Specf. 1	Specf. 2	Specf. 3
Education (Base Cat.: No Formal Schooling)			
Primary	0.129***	0.0743***	0.104***
	(0.0095)	(0.0098)	(0.0144)
Middle	0.230***	0.117***	0.171***
	(0.0110)	(0.0109)	(0.0170)
Secondary	0.427***	0.194***	0.261***
	(0.0128)	(0.0125)	(0.0188)
Higher Secondary	0.760***	0.342***	0.426***
	(0.0141)	(0.0142)	(0.0220)
Graduate & Above	1.186***	0.532***	0.636***
	(0.0139)	(0.0162)	(0.0258)
Match Status (Base Cat.: Adequately Educated)			
Undereducated			0.0346**
			(0.0135)
Overeducated			-0.0708***
			(0.0101)
Selection Term 1	-0.0391*	-0.0339	-0.0337
	(0.0218)	(0.0229)	(0.0229)
Selection Term 2	-0.0405***	-0.107***	-0.105***
	(0.00942)	(0.0126)	(0.0126)
F-Value	526***	504.01***	492.87***
R-Squared	0.3907	0.4822	0.483
Likelihood Ratio Test			59.29***
Number of Observations	69,381	55,873	55,873

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.
Note: (i) *** signals significant at 1% level
(ii) robust standard errors are given in parenthesis

Table 8: Returns to education- OLS and QR

Explanatory Variable	Outcome Variable: Logarithm of Daily Wages									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	OLS	θ = 10	θ = 20	θ = 30	θ = 40	θ = 50	θ = 60	θ = 70	θ = 80	θ = 90
Specification 2										
Education (Base Cat.: No Formal Schooling)										
Primary	0.0743***	0.0508***	0.0577***	0.0673***	0.0741***	0.0675***	0.0691***	0.0769***	0.0755***	0.0714***
	-0.0098	(0.0146)	(0.0108)	(0.00942)	(0.00806)	(0.0086)	(0.00691)	(0.00754)	(0.00830)	(0.0120)
Middle	0.117***	0.0670***	0.0828***	0.0904***	0.115***	0.111***	0.114***	0.125***	0.125***	0.124***
	(0.0109)	(0.0160)	(0.0115)	(0.00956)	(0.00935)	(0.00948)	(0.00928)	(0.00899)	(0.00946)	(0.0116)
Secondary	0.194***	0.149***	0.147***	0.156***	0.176***	0.178***	0.190***	0.197***	0.212***	0.210***
	(0.0125)	(0.0175)	(0.0121)	(0.0127)	(0.0112)	(0.0121)	(0.0103)	(0.0109)	(0.0114)	(0.0133)
Higher Secondary	0.342***	0.276***	0.281***	0.298***	0.321***	0.321***	0.333***	0.330***	0.335***	0.326***
	(0.0142)	(0.0198)	(0.0148)	(0.0143)	(0.0134)	(0.0136)	(0.0133)	(0.0118)	(0.0123)	(0.0158)
Graduate & Above	0.532***	0.510***	0.516***	0.530***	0.548***	0.534***	0.530***	0.519***	0.514***	0.510***
	(0.0162)	(0.0181)	(0.0177)	(0.0179)	(0.0172)	(0.0162)	(0.0152)	(0.0142)	(0.0139)	(0.0170)
Specification 3										
Education (Base Cat.: No Formal Schooling)										
Primary	0.104***	0.0743***	0.0827***	0.0937***	0.105***	0.104***	0.110***	0.122***	0.115***	0.110***
	(0.0144)	(0.0203)	(0.0156)	(0.0137)	(0.0133)	(0.0117)	(0.0120)	(0.0102)	(0.0132)	(0.0154)
Middle	0.171***	0.129***	0.143***	0.153***	0.176***	0.171***	0.172***	0.187***	0.183***	0.178***
	(0.0170)	(0.0212)	(0.0184)	(0.0168)	(0.0163)	(0.0147)	(0.0138)	(0.0127)	(0.0147)	(0.0178)
Secondary	0.261***	0.226***	0.234***	0.240***	0.255***	0.251***	0.256***	0.265***	0.274***	0.273***
	(0.0188)	(0.0252)	(0.0189)	(0.0179)	(0.0169)	(0.0160)	(0.0149)	(0.0138)	(0.0178)	(0.0208)
Higher Secondary	0.426***	0.377***	0.395***	0.398***	0.418***	0.406***	0.416***	0.416***	0.409***	0.400***
	(0.0220)	(0.0252)	(0.0257)	(0.0212)	(0.0197)	(0.0181)	(0.0190)	(0.0169)	(0.0202)	(0.0222)
Graduate & Above	0.636***	0.629***	0.652***	0.649***	0.665***	0.638***	0.628***	0.617***	0.604***	0.605***
	(0.0258)	(0.0320)	(0.0278)	(0.0252)	(0.0253)	(0.0212)	(0.0226)	(0.0207)	(0.0232)	(0.0272)
Match Status (Base Cat.: Adequately Educated)										
Undereducated	0.0346**	0.0209	0.0271*	0.0317**	0.0356***	0.0415***	0.0485***	0.0574***	0.0519***	0.0516***
	(0.0135)	(0.0187)	(0.0155)	(0.0127)	(0.0116)	(0.0120)	(0.0118)	(0.0102)	(0.0107)	(0.0149)
Overeducated	-0.0708***	-0.110***	-0.110***	-0.0983***	-0.0834***	-0.0572***	-0.0504***	- 0.0354***	- 0.0315***	-0.0344***
	(0.0101)	(0.0145)	(0.0119)	(0.0110)	(0.0101)	(0.00761)	(0.00880)	(0.00776)	(0.00882)	(0.0107)

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.

Note: (i) *** signals significant at 1% level, ** signals significant at 5% level, and * signals significant at 10% level and (ii) robust standard errors are given in parenthesis

Table 9: Coefficient of variation –Returns to education

Education (Base Cat.: No Formal Schooling)	Specf 2	Specf 3
Primary	0.12	0.14
Middle	0.19	0.11
Secondary	0.13	0.06
Higher Secondary	0.07	0.03
Graduate and Above	0.02	0.03

Source: Authors’ calculation based on NSSO employment and unemployment survey, 2011-12.

Table 10: Range and Kuznets ratio of returns to education

Education (Base Cat.: No Formal Schooling)	Specification 2			Specification 3		
	θ90/θ10	θ90/θ50	θ50/θ10	θ90/θ10	θ90/θ50	θ50/θ10
	(θ90 - θ10)	(θ90 - θ50)	(θ50 - θ10)	(θ90 - θ10)	(θ90 - θ50)	(θ50 - θ10)
Primary	1.4055	1.0578	1.3287	1.4805	1.0577	1.3997
	(0.0206)	(0.0039)	(0.0167)	(0.0357)	(0.006)	(0.0297)
Middle	1.8507	1.1171	1.6567	1.3798	1.0409	1.3256
	(0.057)***	(0.013)	(0.044)***	(0.049)**	(0.007)	(0.042)*
Secondary	1.4094	1.1798	1.1946	1.208	1.0876	1.1106
	(0.061)***	(0.032)**	(0.029)*	(0.047)*	(0.022)	(0.025)
Higher Secondary	1.1812	1.0156	1.163	1.061	0.9852	1.0769
	(0.05)**	(0.005)	(0.045)***	(0.023)	(-0.006)	(0.029)
Graduate & Above	1	0.9551	1.0471	0.9618	0.9483	1.0143
	(0)	(-0.024)	(0.024)	(-0.024)	(-0.033)	(0.009)

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.

Note: (i) *** signals significant at 1% level, ** signals significant at 5% level, and * signals significant at 10% level

(ii) Differences between the returns at conditional quantiles are in parenthesis.

Figures

Figure 1: Returns to education – Quantile regression

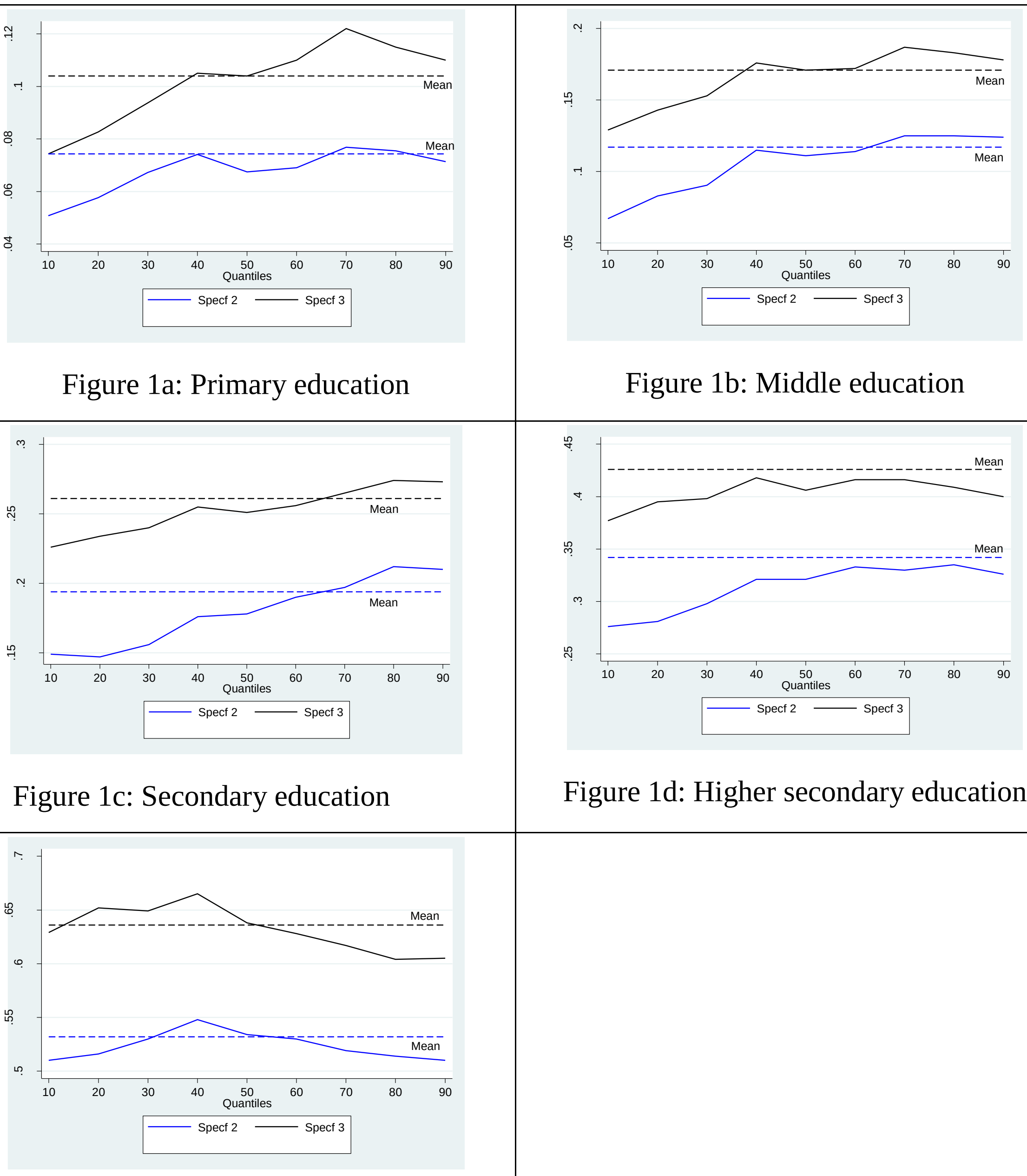

Figure 1a: Primary education

Figure 1b: Middle education

Figure 1c: Secondary education

Figure 1d: Higher secondary education

Figure 1e: Graduates and above education

Source: Authors' calculation based on NSSO employment and unemployment survey, 2011-12.
Note: Reference category is 'no formal schooling'.

Figure 2: Lorenz Curves – Returns to education

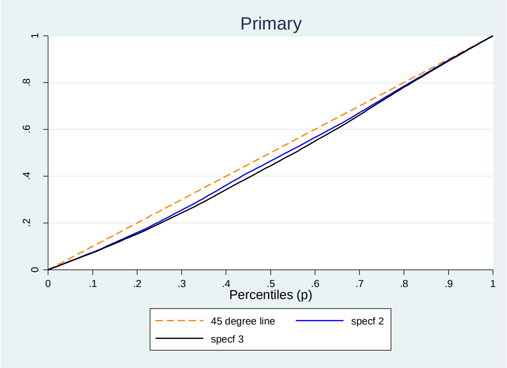

Figure 3a: Primary education

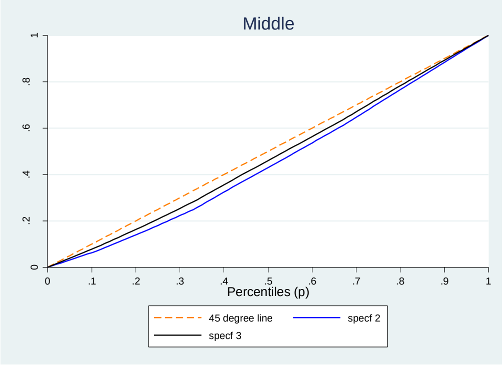

Figure 3b: Middle education

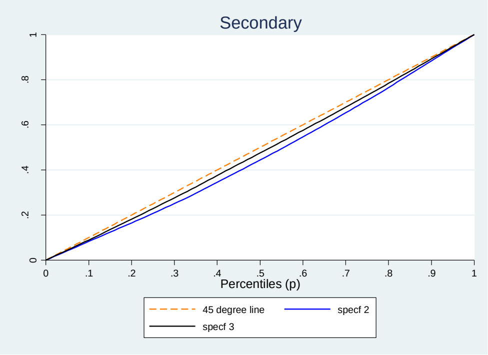

Figure 3c: Secondary education

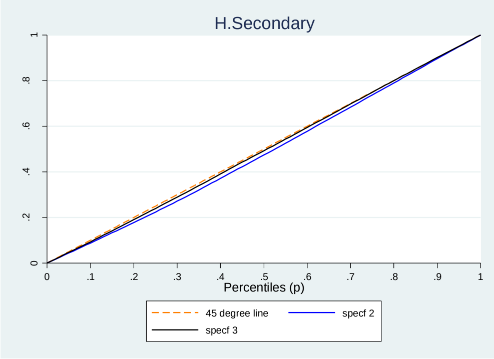

Figure 3d: Higher secondary education

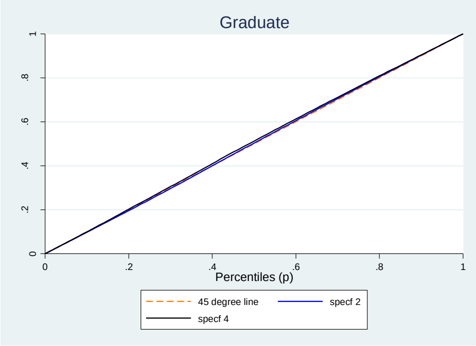

Figure 3e: Graduates and above education

Source: Authors’ calculation based on NSSO employment and unemployment survey, 2011-12.
Note: Reference category is ‘no formal schooling’.

Annexures

Table D: Level of education with corresponding years of education

Education Category	Years of Schooling
No Formal Schooling	0
Below Primary	3
Primary	5
Middle	8
Secondary	10
Higher Secondary	12
Graduate	15
Post Graduate and Above	17