Azim Premji University

CSE Working Paper

2019-03

How Comparable Are India's Labour Market Surveys?

A Comparison of NSS, Labour Bureau, and CMIE Estimates

Rosa Abraham and Anand Shrivastava

May 2019

Centre for Sustainable Employment

cse.azimpremjiuniversity.edu.in

How Comparable Are India's Labour Market Surveys? An Analysis of NSS, Labour Bureau, and CMIE Estimates *

Rosa Abraham† and Anand Shrivastava ‡

16th May 2019

Abstract

With the lack of official government data on unemployment and other labour market indicators, the most viable and recent source have been the regular household surveys conducted by the Centre for Monitoring the Indian Economy (CMIE). Given the differences in methods in data collection, it becomes exceedingly important to establish some comparability between the government and the CMIE datasets. This paper attempts to do that using two methods. First we fit a model of employment status on the CMIE data and see how well it predicts outcomes in the older Labour Bureau 2015-16 and NSS 2011-12 data. Then we compare state-level estimates of broad labour market indicators from CMIE 2016 and Labour Bureau 2015-16 datasets. The broad results are that despite differences in methodologies, the estimates for men are quite comparable between the surveys, while measures of women's participation in the labour force seem particularly sensitive to the way questions are asked in surveys.

JEL: J21, J64, C83,

Keywords: Employment data, household survey data, India, labour force participation rate, workforce participation rate

*We are very grateful to Amit Basole for many useful comments and suggestions and to Janaki Shibu for excellent research assistance.

†Azim Premji University, email: rosa.abraham@apu.edu.in

‡Azim Premji University, email: anand.shrivastava@apu.edu.in

1 Introduction

In recent years, unemployment has replaced poverty as the biggest economic issue that figures in political debates in India (Basole and Jayadev, 2019). An increasingly aspirational youth with college degrees have not been finding adequate number of desirable employment (Basole et al., 2018) evidenced in frequent news reports of millions applying for a few vacancies in government jobs.[1] The crisis seems to have been exacerbated by government policies such as demonetisation and the Goods and Services Tax (Shrivastava et al., 2019; Chodorow-Reich et al., 2018). Despite such a scenario, or probably because of it, the government has either delayed or refused to release data from two household surveys of employment conducted over the past few years. The Labour Bureau Employment Unemployment Survey (LB-EUS) conducted in 2016-17 and the Periodic Labour Force Survey (PLFS) conducted in 2017-18 have both been in the midst of some controversy. While the headline national estimates from these surveys got leaked (Jha, 2019b,a) , the detailed reports and the unit level data both remain outside the public domain.

While the headline numbers are consistent with the larger picture of an employment crisis, any detailed analysis of the situation or a detailed study of the impact of government measures is not possible without unit-level data. The only other source of national level household survey data on employment is the Consumer Pyramids Survey done by the Centre for Monitoring the India Economy (CMIE-CPHS), which started collecting employment data from 2016.Table 1 shows the comparison between the leaked estimates from LB-EUS 2016-17 and PLFS 2017-18 and CMIE-CPHS estimates for the same periods. While the larger trend of decreasing workforce participation rate and labour force participation rate is borne out by both the government data and CMIE data, there is considerable difference in the actual point estimates. Ex ante, one would expect some difference in estimates as the definition of employment, as well as the method of conducting the survey, are different in CMIE-CPHS as compared to the government surveys.

While some studies have used the CMIE-CPHS data (Chodorow-Reich et al., 2018), any analysis of employment trends prior to 2016 would require one to establish some comparability between CMIE-CPHS and the government surveys. In this paper we attempt to do this using two methods.

Since one of the most important uses of individual level employment data is to model the determinants of employment status, we see how similarly a given model

[1] See news reports at https://economictimes.indiatimes.com/news/politics-and-nation/over-93000-candidates-including-3700-phd-holders-apply-for-peon-job-in-up/articleshow/65604396.cms and http://www.bbc.com/capital/story/20180601-the-jobs-in-india-that-attract-millions-of-applicants.

	LFPR Overall	LFPR Female	LFPR Male	WPR Overall	WPR Female	WPR Male	UR Overall	UR Female	UR Male
2016-17 LB	52.8	26.9	76.8	50.7	25.3	74.3	3.9	6.1	3.3
2016 CMIE	46.8	15.6	74.8	43	12.1	70.7	8.2	22.4	5.5
2017-18 PLFS	49.8	23.3	75.8	46.8	-	-	6.1	-	-
2017 CMIE	43.9	11.9	72.6	41.9	10.6	70.1	4.4	10.9	3.5

Table 1: Estimates of labour force aggregate measures for India from CMIE and from leaked reports of government surveys

behaves in different datasets. We first fit a model of determinants of employment status on CMIE-CPHS 2016, and then use the model to generate predictions in LB-EUS 2015-16, which is the last government survey for which unit level data is available. We find that the model correctly predicts around 80% of the observations in LB-EUS which is the same rate of success that the model has in the original data. This shows that a model of individual employment status behaves very similarly in CMIE-CPHS as it does in government data. We further try to use this model to compare the different definitions of employment used in LB-EUS 2015-16 as well as the National Sample Survey Organisation's Employment Unemployment Survey 2011-12 (NSS-EUS 2011-12, henceforth), and find that the prediction success does not change much.

Then, we see how closely state level estimates of Labour Force Participation Rate (LFPR) and Workforce Participation Rate (WPR) [2] from CMIE-CPHS 2016 compare with those obtained from LB-EUS 2015-16. We find that while for men the estimates map quite well, which means that the bias as well as the variance are low, the same is not the case for women. This implies that the effect of the difference in the definition of employment and/or the method of surveying is largely observed in the responses for women. Hence, we recommend the usage of CMIE-CPHS data only for men wherever there is a need for comparing these estimates with that of government surveys hitherto.

The following section gives details of the surveys that we examine here and describes the difference in definitions of employment and survey methodology. Sections 3 and 4 respectively discuss the two methods we use to establish comparability. Section 5 concludes.

[2]WPR is defined as the proportion of working age population who are employed. LFPR is defined as the proportion of working age population whose employment status is 'Employed' or 'Unemployed', i.e. they are either working or looking for work. Working age is defined as 15 years or older.

2 Labour Market Surveys in India - an overview

We examine three surveys in this paper - the National Sample Statistics Organisation Employment Unemployment Survey (NSS-EUS) 2011-12, the Labour Bureau Employment Unemployment Survey (LB-EUS) 2015-16 and the Centre for Monitoring Indian Economy Consumer Pyramids Survey (CMIE-CPHS) 2016.

2.1 NSS-EUS 2011-12

Since 1972-73, the National Sample Survey Organisation has conducted nationally representative household level Employment Unemployment surveys (EUS) to discern activity status of each member of the household, and their demographic features, including education, age, and social group. These surveys have occurred more or less regularly, once every five years. The 68^{th} Round (2011-12) NSS Employment Unemployment Survey (NSS-EUS) is the ninth and last such survey conducted. The survey spans around 170,000 households and 450,000 individuals.

The NSS-EUS schedule uses four different reference periods to arrive at four possible activity statuses - one year, one month, one week, and each day of the reference week. A person is identified as employed under Usual Principal Activity Status (UPS) if he/she spent a relatively long time either working or looking for work during the 365 days preceding the survey. If a person is not employed or looking for work for the majority of the year, but working for at least a month in the 365-day reference period (i.e. subsidiary status), then he/she is identified as employed as per Usual Principal and Subsidiary Activity Status (UPSS). Under Current Weekly Status (CWS), a person is identified as working if he/she worked for at least an hour during the seven days preceding the survey. A person's activity status on each day of the reference week determines the Current Daily Status (CDS), where he/she is considered as working a full day if engaged for four hours or more, or a half day if less than four hours. However, unlike the other definitions, the CDS definition demarcates a particular day as 'working' or 'not working', not an individual. Hence the CDS measures person days of employment rather than persons.

2.2 LB-EUS 2015-16

Responding to some discontentment with the large five-year gap in the NSS surveys, the Labour Bureau began conducting yearly Employment Unemployment Surveys. The first such survey was in 2009-10 and since then the LB-EUS began collecting yearly

surveys (more or less) regularly, until 2015-16.[3]. Although a survey was conducted in 2016-17, these results were never officially released (Abraham et al., 2019).

LB-EUS 2015-16 covers a nationally representative sample of about 160,000 households and 580,000 individuals. The LB-EUS uses a questionnaire very similar to that of the NSS EUS. However, the LB-EUS collects information on only two activity statuses - Usual Principal Activity Status and Usual Principal Subsidiary Status. The NSS-EUS and LB-EUS therefore, broadly identify a person as either (i) employed, or (ii) unemployed i.e. did not work but was seeking and/or available for work, or (iii) did not work and not looking for work (not in the labour force).

2.3 CMIE-CPHS

The Centre for Monitoring Indian Economy (CMIE), a private business information organisation, has been collecting data relating to employment and unemployment status since 2016. The Consumer Pyramid Survey as it is called (henceforth referred to as CMIE-CPHS) covers about 160,000 households and 522,000 individuals. The survey is conducted in three 'waves' with each wave spanning four months, beginning from January. Each individual is surveyed in every wave, so that for every year, the employment and unemployment status is available for three points in the year. Alongside employment status, the survey collects information on gender and age of the respondent.

While CMIE-CPHS is a household survey, the enumerator does not go to the household with a questionnaire to be filled point by point. Instead, the enumerator has a free-ranging discussion with the household head where all the questions from the survey are woven in. Additionally, the questions used to discern employment status differ between the government surveys (LB and NSS) and the CMIE survey.

The CMIE-CPHS categorises an individual into (i) employed, (ii) unemployed, willing and looking for a job, (iii) unemployed, willing but not looking for a job, (iv) unemployed, not willing and not looking for a job. However, the reference period used to discern activity status is different. The CMIE identifies an individual as employed if he/she "is engaged in any economic activity either on the day of the survey or on the day preceding the survey, or is generally regularly engaged in an economic activity". Individuals who were in some form of employment, but were not at work on that day of the survey due to various reasons such as illness, leave or holiday are still considered as employed when there is a reasonable surety of them going back to work.

Firstly, by identifying an individual's status as on the day of the survey, or on the

[3]For reports from all LB-EUS surveys see https://cse.azimpremjiuniversity.edu.in/resources/labour-bureau-employment-unemployment-survey-reports/

day preceding it, at first glance, the CMIE definition seems to be closest to the NSS CDS interpretation of employment. But as mentioned earlier, under CDS, the unit of observation is a day, rather than an individual. Therefore, CDS assigns an activity status to a day, whereas the CMIE-CPHS question assigns an activity status to an individual based on what they did on that day.

Secondly, by allowing for individuals who are 'generally regularly employed' to be also identified as employed, the CMIE interpretation of employment is similar to the NSS UPS/UPSS approach. Therefore, there is no one interpretation of employment that the NSS/LB uses that is perfectly equivalent to the CMIE-CPHS definition. The CMIE definition of employment is, in a sense, a combination of two or more definitions of employment as identified under NSS.

This change in the reference period when recording a person's employment status can influence labour market statistics. For instance, Heath et al. (2016) find that a shorter reference period results in a higher reporting of self-employed work, leading to greater incidence of work. Hence, we would expect there to be some difference in the employment statistics computed from CMIE-CPHS and LB/NSS surveys even if they were conducted in the same time period. While we do not have unit level data for overlapping time periods, we know from the national level aggregate numbers (from leaked reports) for years where both government surveys and CMIE-CPHS overlap, that the numbers are in fact different. Hence the differences in methodology do seem to lead to differences in estimates.

3 Comparing Surveys by fitting a model of employment status

The ideal way to check the comparability of two surveys would be if they had sampled the same individuals. Since this is obviously not the case, the next best thing to do is to look at individuals with similar demographic characteristics in the two surveys and see if the employment status recorded by the two surveys are similar, on average. The assumption here is that if the two surveys systematically differ in the classification of the employment status of some demographic subsection of society, then this would reflect as difference in the relationship between the employment status variable and the demographic variables across the two surveys.

Further, if there are multiple definitions of employment in the second survey, then we can say that the relationship between employment and demographic variables obtained in the first survey will match most with that definition that is most similar to the

definition of employment in the first survey.

We model the relationship between employment and demographic variables by constructing an econometric model at the individual level where the dependent variable is employment status and the independent variables are demographic characteristics of the individual. Given the difference in definitions used across surveys, we wanted to see whether a person identified as employed by the CMIE-CPHS definition would be similarly identified as per other definitions i.e. UPS/UPSS/CDS/CWS.

3.1 A model of Employment Status

Constructing an econometric model of demographic determinants of employment status is also instructive because one of the most important uses of individual-level employment data is to study the determinants of labour market outcomes (Kingdon and Unni, 2001; Klasen and Pieters, 2012; Srivastava and Srivastava, 2010). Establishing a comparability of such a model across datasets would enable researchers to map the changes in the effects of determinants like age and gender across time.

The approach we take is to first estimate a multinomial logit model of activity status on CMIE-CPHS and then use the model to predict outcomes in LB-EUS 2015-16 and see what percentage of observations the model succeeds in predicting correctly. Since the survey periods are not overlapping, an additional source of variation could come from changes in labour demand. Hence, an underlying assumption in the analysis that follows is that labour demand did not change significantly between 2015 and 2016.

The model that we estimate is given below. Here the dependent variable takes one of three of values representing an individual's activity status - Employed, Unemployed, and Out of the Labour Force (OOLF). We classify the fourth category in CMIE - unemployed and willing but not looking for work - as out of the labour force.

$$\frac{P(y_{ij}=k)}{P(y_{ij}=0)} = exp(\alpha+\beta_{1k}age_{ij}+\beta_{2k}age_{ij}^2+\beta_{\mathbf{3k}}education\ status_{ij}+\gamma_{jk}+\varepsilon_{ijk}),\ \ k\in\{1,2\}$$

Here, the subscript i indicates an individual and j indicates a state. y_{ij} is the employment status of an individual that can be 0, 1 or 2 indicating Out of labour force (OOLF), Unemployed and Employed respectively.[4] The variable age_{ij} is a continuous variable indicating age in years, and $education\ status_{ij}$ is a categorical variable representing levels of education, going from illiterate, primary, middle, secondary, higher secondary, to graduate and above. γ_{jk} indicates state fixed effects.

[4] We classify the fourth category in CMIE - unemployed and willing but not looking for work - as out of the labour force for comparability with LB-EUS.

The model is estimated separately for men and women to allow for different coefficients. The CMIE-CPHS survey is a panel survey, with each individual being interviewed thrice in a year. Therefore, for any one individual, there are three possible employment statuses associated with him/her for the year 2016. To make this data comparable with the other surveys which have a single activity status associated with an individual, a pseudo-cross section was constructed out of the panel data by simulating a sampling scheme where all the sample households would be randomly allocated to one of the waves. This was done by randomly choosing and retaining only one of the three possible employment statuses for an individual in a year. We later show that using a pooled cross section instead of a pseudo-cross section does not change the result.

Then, the model estimated on CMIE-CPHS 2016 is used to predict the employment status in LB-EUS 2015-16. Being a multinomial model, the predictions for every individual will be the probabilities of that individual being in each of the three employment statuses. We choose that status that has the highest predicted probability as the final predicted employment status. We then compare these predictions with the actual observed employment status. We use the Usual Principal Status definition of employment for the baseline exercise, and compare it to different definitions later.

An observation is identified as 'Matched' if the predicted employment status is the same as the actual employment status. If the predicted and actual status are not the same then it may be classified in one of four categories. If the predicted outcome is Employed but the actual status is Unemployed, then the observation is classified as 'Employment Overpredicted', and vice versa for 'Employment Underpredicted'. An observation is classified as 'LFP overpredicted' if the prediction is the individual is in the labour force, i.e. they are either Employed or Unemployed, but the actual employment status is Out Of the Labour Force, and vice versa for 'LFP Underpredicted'. Table A1 in the Appendix describes the possible outcomes for every combination of predicted and actual economic status.

3.2 Results

The actual estimations results from the model are given in the appendix (Table A2), but we are more interested in the result of the matching exercise. Overall, the model estimated on CMIE-CPHS is able to predict the activity status of approximately 80% of individuals correctly in the LB-EUS data.

Now the question is 'how good is 80%?' To answer this we use the same model to predict outcomes in the CMIE-CPHS data itself. We also do the opposite exercise, i.e. estimating the model on LB-EUS and then using it to predict outcomes in LB-EUS

and in CMIE-CPHS. The results are in Table 2. We can see that the rate of correct prediction is approximately 80% in all four cases. This implies that the model is as good at predicting employment statuses within the data it is estimated on, as it is in predicting employment status on another dataset.

Model run on Prediction matched on	CMIE LB	CMIE CMIE	LB LB	LB CMIE
Matched	80.5	79.6	81.8	78.5
Employment overpredicted	1.0	1.4	1.0	1.6
Employment underpredicted	0.1	0	0	0
LFP overpredicted	4.3	4.5	5.2	6.9
LFP underpredicted	14.2	14.4	11.9	13.0
Total	100	100	100	100

Table 2: Comparing results of different models: overall

This provides a very strong argument that the definition of employment in CMIE-CPHS classifies around 80% of the population in the same way that LB-EUS does.[5] As for the rest of the 20%, the model fails to predict their actual status, quite likely because of important factors that are not measured.[6]

It is instructive to see who the unmatched are and what is the nature of the mismatch. Most of the mismatch comes from LFP underpredictions, i.e. individuals who are in reality in the labour force are identified by the CMIE model as being out of the labour force. Of the matched observations, 54% were men. Considering that there was a fairly equal distribution of men and women in the population as a whole, this suggests an under-representation of women in the Matched sample. We estimated the models separately for men and women to investigate this further.[7].

In the women-only model, the CMIE is able to predict correctly for only 76% of the sample (Table 4). This holds true for all four estimation-prediction combinations (Table2). This could mean that some important factors that determine women's employment status are not captured in the surveys and are omitted from the model. This fits well with the literature on women's labour force participation which says that sur-

[5]It could be the case that there is a section of population that the model predicts correctly in the CMIE-CPHS data, but it gets replaced by another section in LB-EUS. This means that although the two surveys classified them differently, they both got predicted correctly within each dataset. However, going by the consistency of the results across slices along various variables, this seems unlikely.

[6]Two important factors that are likely to influence labour force participation, but that are not included because of their unavailability in LB-EUS are rural/urban location and number of children. But including these factors in a similar analysis with NSS-EUS 2011 does not increase the prediction rate.

[7]Slicing the data along other lines, including estimating separately for different states, did not yield interesting results

veys do not adequately capture women's work (Hussmanns et al., 1990) and that a significant reason for low labour force participation of women are often factors on labour demand side, including the nature of work available rather than the woman's or her household's characteristics (Fletcher et al., 2017; Verick, 2014). Non-inclusion of the labour demand-side factors in the model, and higher sensitivity of women towards these factors could explain the lower prediction rate of the model for women.

Additionally, while for men the LFP under or over-prediction are roughly equal suggesting errors being equally likely to occur both ways, in women almost all the mismatch comes from LFP undeprediction. This shows that the model consistently predicts that women are out of the labour force when they are actually in it. Hence, we can conclude that it is likely that the effects of differences in the definition of employment between CMIE-CPHS and LB-EUS are likely to show up in differences in classification of women who are in the labour force. Hence one would expect that the workforce participation rate and unemployment rate for women could be very different between the two surveys, which is what we find in the next section.

Model run on	CMIE	CMIE	LB	LB
Prediction matched on	LB	CMIE	LB	CMIE
Matched	76.4	79.4	78.1	77.4
Employment overpredicted	0	0.1	0.1	0.7
Employment underpredicted	0.1	0	0	0
LFP overpredicted	0.3	0.3	3.3	5.6
LFP underpredicted	23.3	20.2	18.5	16.3
Total	100	100	100	100

Table 3: Comparing results of different models: female

	All	Male	Female
Matched	80.5	84.3	76.4
Employment overpredicted	1	1.9	0
Employment underpredicted	0.1	0.1	0
LFP overpredicted	4.3	7.9	0.3
LFP underpredicted	14.2	5.8	23.3
Total	100	100	100

Table 4: Results of predicting LB outcomes using CMIE model

3.3 Comparison of different definitions of employment

The LB-EUS has information on two definitions of employment - Usual Principal Activity Status (UPS) and Usual Principal and Subsidiary Activity Status (UPSS). We have used only the UPS definition in the analysis above. If using the UPSS definition, increases (or decreases) the prediction success rate, this could imply that the UPSS definition is closer to the CMIE-CPHS defintion of employment. While doing the exercise, we expanded it to include the NSS-EUS 2011-12, which also collected Current Weekly Status (CWS) and Current Daily Status (CDS) definitions of employment. Even though the survey was carried out five years before the CMIE-CPHS, a relative comparison of the prediction success of different definitions could still be instructive.

Since the CDS is a measure of person day, we modified it so as to approximate the CMIE definition. We interpret two versions of the CDS employment status that are definitionally closest to the CMIE measure. CDS-1 identifies an individual as employed if he/she was reported as working on the day of the survey, or the day prior to the survey, or for the majority of the year (i.e. by UPS status). CDS-2 identifies a person as employed if he/she worked for at least half the week.

The multinomial logit model is estimated on the CMIE-CPHS data as earlier, and the predictions are created for the NSS-EUS 2011-12 data. Predictions for the LB-EUS dataset were already available from the previous exercise. Then in both datasets, predictions are matched to the actual employment status according to the various defintions. The results are in Table 5.

Employment definition used	UPA	UPA	UPSS	UPSS	CWS	CDS 1	CDS2
Model run on	CMIE	CMIE	CMIE	CMIE	CMIE	CMIE	CMIE
Prediction matched on	LB	NSS	LB	NSS	NSS	NSS	NSS
Matched	80.5	81.7	78.9	77.8	79.1	79.5	79.6
Employment overpredicted	1.0	0.6	0.7	0.6	1.0	1.0	1.4
Employment underpredicted	0.1	0.0	0.1	0.0	0.0	0.0	0.0
LFP overpredicted	4.3	2.9	4.2	2.7	3.1	3.0	3.2
LFP underpredicted	14.2	14.8	16.2	18.9	16.8	16.5	15.8
Total	100	100	100	100	100	100	100

Table 5: Comparing matching results from different definitions of employment across LB-EUS 2015-16 and NSS-EUS 2011-12

We find that there is no significant difference in the prediction success rate between the different definitions. Hence, the effects of the differences, if any, would only effect the classification of those whose employment status could not be predicted by the model.

3.4 Robsutness check: CMIE-CPHS as a pooled cross section

The CMIE-CPHS survey is a panel survey, which we convert to a pseudo-cross section for the analysis presented above. As a robustness check, we treated the CMIE-CPHS survey as a pooled cross section, with every observation representing a separate individual. The results did not vary significantly when estimated on a pooled-cross section.

So the results are robust to all definitions of employment status, and is not affected by the differences in sampling methodology across surveys.

4 Comparing State-Level Estimates

Next we conduct an exercise comparing the aggregate estimates of employment status from the CMIE-CPHS survey with comparable government data. For most policy and discussion purposes, these aggregate estimates are the ones that are used. Hence, the most important question is of the kind : 'If the CMIE-CPHS estimate of the unemployment rate is 6%, how much is it according to the NSS/LB definition?' While we do not attempt to provide a direct answer to this question, we try to provide a framework of thinking about the differences in the aggregate estimates in the two surveys.

There are a couple of ways one could think about how the differences in the definition and survey method could translate into differences in aggregate measures. One is to assume that there are certain sections of the population that are going to be classified differently in the two surveys. And, states that have higher proportion of these populations are going to show more divergence between estimates from the two surveys.

This follows from the approach we took in the previous section, and allows us to make some predictions based on our results there. We had found that for women the differences in classification are likely to be for those who are in the workforce, but for men there was no such pattern. Hence, we can predict that states who have a higher labour force participation for women are going to see more mismatch between the workforce participation rates in the two surveys, while this will not be the case for men. We find exactly this result in the data (Table 6).

While this is instructive, it does not bring us any closer to answering the question posed at the beginning of this section. To do that we adopt a different model of thinking about how the differences in methods and definitions aggregate up to state level estimates. We can model the classification as a stochastic process. Imagine the ideal case where the sample of the two surveys is exactly the same - by doing this we abstract from any sampling errors that may contribute to differences in estimates.[8] Now, every

[8]As we are looking at state level aggregates, the sample sizes are still large enough to make this

	(1) diff_wpr_abs_male	(2) diff_wpr_abs_female	(3) diff_wpr_abs_female
lb_lfpr_ps_male	0.0921		
	(0.0875)		
lb_lfpr_ps_female		0.507***	0.284***
		(0.0953)	(0.0830)
_cons	-3.953	-2.167	1.974
	(6.467)	(2.434)	(1.950)
N	24	24	23
R^2	0.048	0.563	0.359

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: *The last column drops the state Chattisgarh, which is an outlier with a difference in estimates of 40 percentage points*

Table 6: Regression results checking the correlation between difference in WPR estimates and LFPR for men and women

individual i is classified according to the LB definition as either Employed or Not Employed. The proportion of individuals classified as Employed gives us the LB estimate of the WPR (Workforce Participation Rate). For an individual classified as Employed in the LB data, let the probability of him/her being classified as Employed in the CMIE-CPHS data be p_1. Similarly, For an individual classified as Not Employed in the LB data, let the probability of him/her being classified as Employed in the CMIE-CPHS data be p_2. If we define binary variables x_i and y_i that take the value 1 if individual i is classified as Employed in the LB and CMIE data respectively, then we can write

$$Pr(y_i = 1) = p_1^{x_i} p_2^{1-x_i}, \quad x_i \in \{0,1\}$$

Hence if the sample size is n out of which there are k individuals classified as Employed in the LB data, then the probability distribution of the number of Employed in the CMIE-CPHS data is a Poisson Binomial distribution, with its expected value given by $E[k'] = kp_1 + (n-k)p_2$. This would be the expected number of people classified as Employed in the CMIE-CPHS data. Hence, the expected value of the WPR in the

abstraction.

CMIE-CPHS data for a state j is related to the WPR in the LB data through the equation

$$E[WPR_j^{CMIE}] = E\left[\frac{k'_j}{n_j}\right] = p_{2j} + (p_{1j} - p_{2j})WPR_j^{LB}$$

This can be rewritten as follows.

$$WPR_j^{CMIE} = p_{2j} + (p_{1j} - p_{2j})WPR_j^{LB} + \varepsilon_j \tag{1}$$

In this general form the probabilities are allowed to be different for each state, aligning to the idea of states having different proportions of populations that are prone to being differently classified. But this cannot be estimated from the data we have. In order to estimate this and get some useful interpretation out of the estimation, we make two assumptions

- We assume that the value of p_1 is the same for all individuals. $p_{1j} = p \; \forall \; j$.
- We assume $p_2 = 0$. We know that CMIE, on average underpredicts both LFPR and WPR. Hence, we make this assumption to easily interpret the regression results.[9]

Hence, the regression model becomes.

$$WPR_j^{CMIE} = pWPR_j^{LB} + \varepsilon_j \tag{2}$$

We estimate this using aggregate estimates for WPR and LFPR for 24 states obtained from the LB-EUS 2015-16 data and the pseudo-cross section constructed from the CMIE 2016 data.[10]

4.1 Unemployment rate

The question we started this section with was about unemployment rate. In an attempt to answer the question, let us consider the LFPR counterpart of equation 2.

$$LFPR_j^{CMIE} = p'LFPR_j^{LB} + \nu_j \tag{3}$$

Now, we can get an expression for the unemployment rate (UR) as estimated from CMIE data in terms of that estimated from LB data.

[9] We also present estimates with p_2 not set to zero in the appendix.

[10] Note that here take the Usual Principal Activity Status (UPS) numbers for the Labour Bureau survey, and we define LFPR for the CMIE-CPHS data to not include include people who are unemployed and willing to work but not actively looking for work.

$$UR_j^{CMIE} = 1 - \frac{WPR_j^{CMIE}}{LFPR_j^{CMIE}} = 1 - \frac{pWPR_j^{LB} + \varepsilon_j}{p'LFPR_j^{LB} + \nu_j} \tag{4}$$

From 4 we can see that there is no way of deriving a linear relationship between UR_j^{CMIE} and $UR_j^{LB} = 1 - \frac{WPR_j^{LB}}{LFPR_j^{LB}}$. Hence, running a linear regression between state-level unemployment rate estimates from CMIE-CPHS and LB-EUS would not make any sense as the coefficient would not have any interpretation under this model. Hence, we restrict ourselves to LFPR and WPR.

4.2 Regression results

Figure 1 shows the comparison with the LB numbers on the horizontal axis and the CMIE numbers on the vertical axis. The regression results are presented in Table 7. We can see that our estimates for p are 0.931 and 0.901 for LFPR and WPR respectively. This can be interpreted as an under-reporting of LFPR and WPR in the CMIE data compared to the LB-EUS data, of around 7% and 10% respectively.

	(1)	(2)
	LFPR CMIE	WPR CMIE
LFPR LB	0.931***	
	(0.0233)	
WPR LB		0.902***
		(0.0216)
Observations	24	24
R^2	0.986	0.987

Standard errors in parentheses

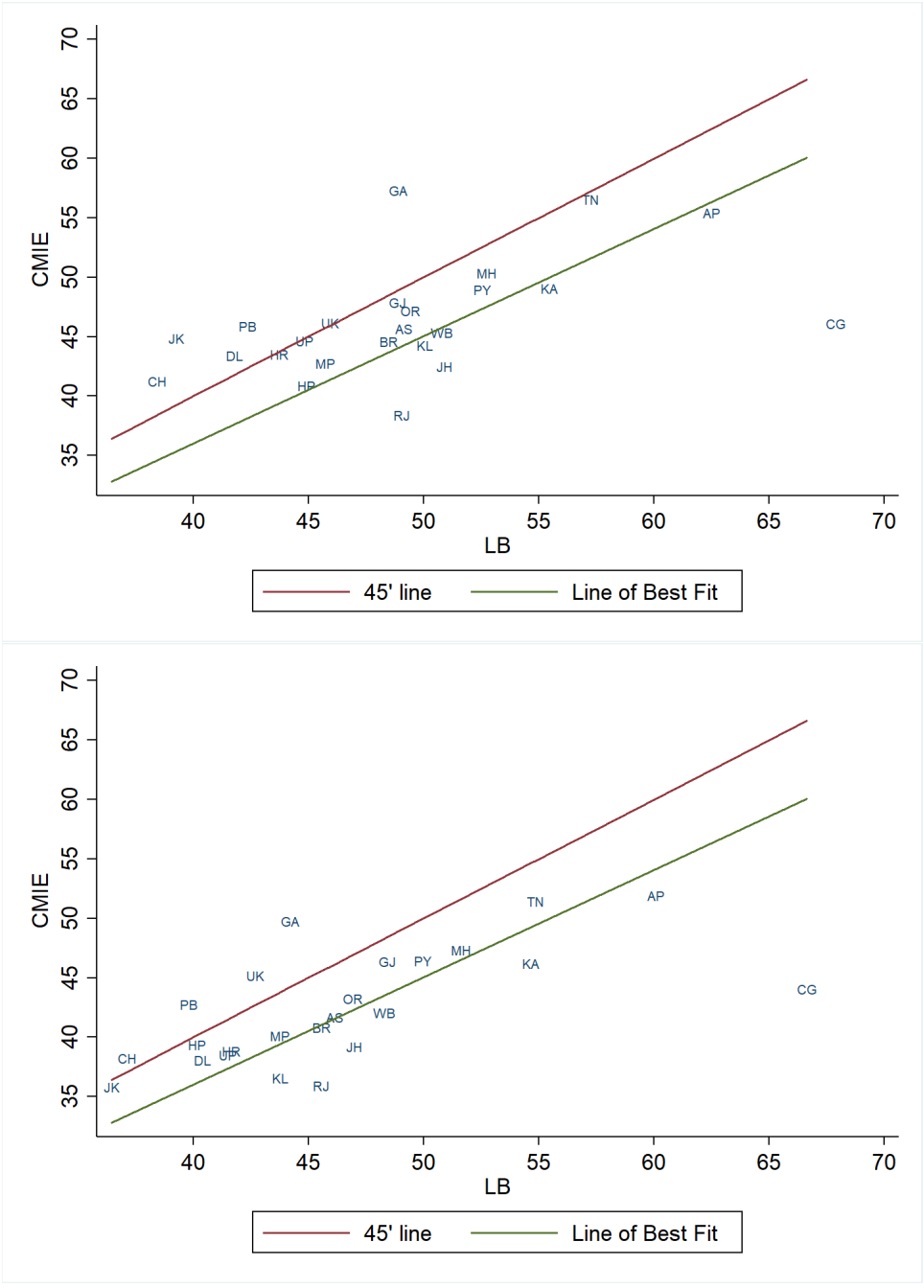
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 7: Regression results for overall state level LFPR and WPR

Here we have assumed that the p is constant for all individuals. But from our previous exercise we know that it is likely to be different for men and women. Hence, we estimate the models separately for men and women. The results are presented in Figures 2 and 3 and Tables 8 and 9.

We find that while the underreporting is estiated as significantly higher for women, for men it is close to zero. For men, we cannot statistically reject the hypothesis that *p=1*. The most likely explanation is that for men the differences in classification may either be very small or be in different directions that get averaged, while for women

Figure 1: Comparing CMIE-CPHS and LB-EUS estimates: Overall



Note: The graph on the top is for LFPR and the one on the bottom is for WPR

they are systematically in one direction, which shows up in the graph and regression.

Now we look at this separately for men and women, knowing that capturing women?s work has been difficult and we have also found evidence of this in our earlier analysis of determinants of labour supply. Figures 3 and 2 show the comparison for men and women respectively. As we can see the estimated LFPR and WPR for men from the two surveys are very close with a difference of less than 5%. For women, the LFPR estimates are close but WPR estimates show a huge bias where CMIE underpredicts WPR by 40%.

	(1) LFPR CMIE (Female)	(2) WPR CMIE (Female)
LFPR LB (Female)	0.654***	
	(0.0644)	
WPR LB (Female)		0.560***
		(0.0570)
Observations	24	24
R^2	0.818	0.807

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 8: Regression results for state level female LFPR and WPR

	(1) LFPR CMIE (Male)	(2) WPR CMIE (Male)
LFPR LB (Male)	0.999***	
	(0.0100)	
WPR LB (Male)		0.988***
		(0.00968)
Observations	24	24
R^2	0.998	0.998

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9: Regression results for state level male LFPR and WPR

And additional point to note is that for men not only is the estimate close to 1 but the standard errors are also very small. This shows that not only is p close to 1 on average, it is likely to be close to 1 for most states too. This implies that whatever section of the male working age population gets differently classified in the two surveys, the

misclassification averages out in most states resulting in close matching between the male LFPR and WPR numbers obtained from the two surveys.

This gender difference is also seen if we compare the CMIE estimates with the leaked estimates from two unreleased government surveys - the Labour Bureau survey of 2016-17 and the NSSO Periodic Labour Force Survey of 2017-18. The comparison is presented in Table 1.

The low WPR for women in CMIE may be due to multiple reasons. Some modes of doing the survey - which in the case of CMIE is an extended conversation with the household head - may lend themselves to more bias (Bardasi et al., 2011). Also the difference in the question about unemployment could mean that some people who would be classified as employed according to the principal status question would be classified as unemployed in the CMIE definition. This is likely to be people who are in irregular employment, and women are more likely to be in this situation.

Hence, the practical conclusion we can draw from this exercise is that while comparing LFPR and WPR numbers obtained from CMIE data to past government data, it would be better to only use the estimates for men as the divergence in the estimates is quite significant for women. We can also conclude that whichever way we thing about the process that generates the difference in estimates, the trends are going to be in the same direction i.e. if the the LFPR/WPR decreases according to CMIE, it will also decrease according to the government definition. This is because regardless of whether the difference is caused by sections of the population being classified differently, or by a stochastic process that generates the difference probabilistically, the parameters of these processes are not going to change with time, at least not over the period of a few years.

5 Conclusion

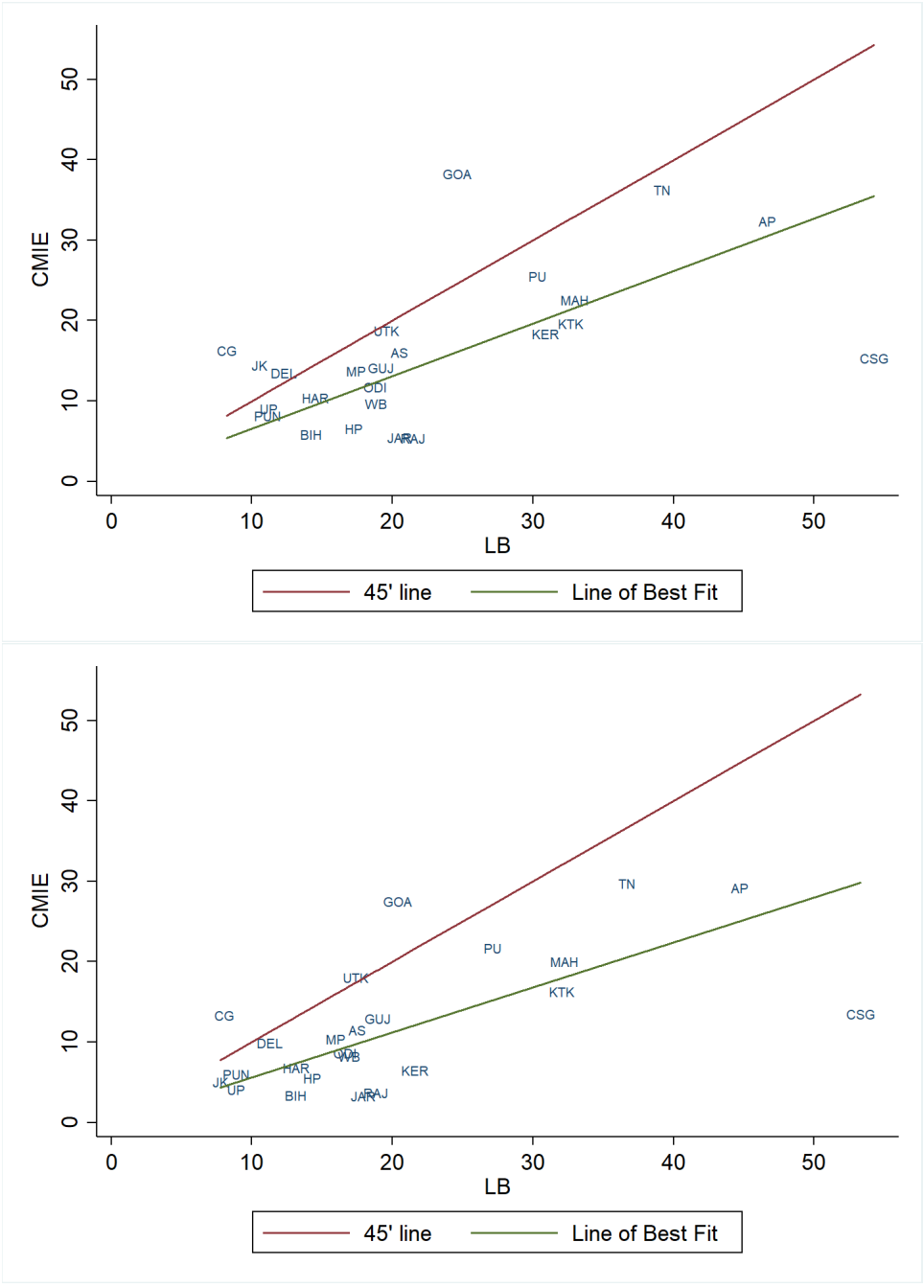
Amidst the lack of government survey data to generate labour market estimates, the CMIE-CPHS has emerged as the only source of household level nationally representative employment data. In this paper, we take into account the variations in definitions of economic activity status and differing reference periods between the government surveys and CMIE-CPHS and see to what extent the estimates from these surveys are comparable.

We first run an econometric model of employment status on a pseudo-cross section constructed from CMIE-CPHS 2016 data and find that 80% of the observations in the government surveys matched with the predicted status obtained from the model. Next, we develop a stochastic model of employment status classification and use it to compare

state level aggregate estimates from the CMIE-CPHS 2016 and LB-EUS 2015-16.

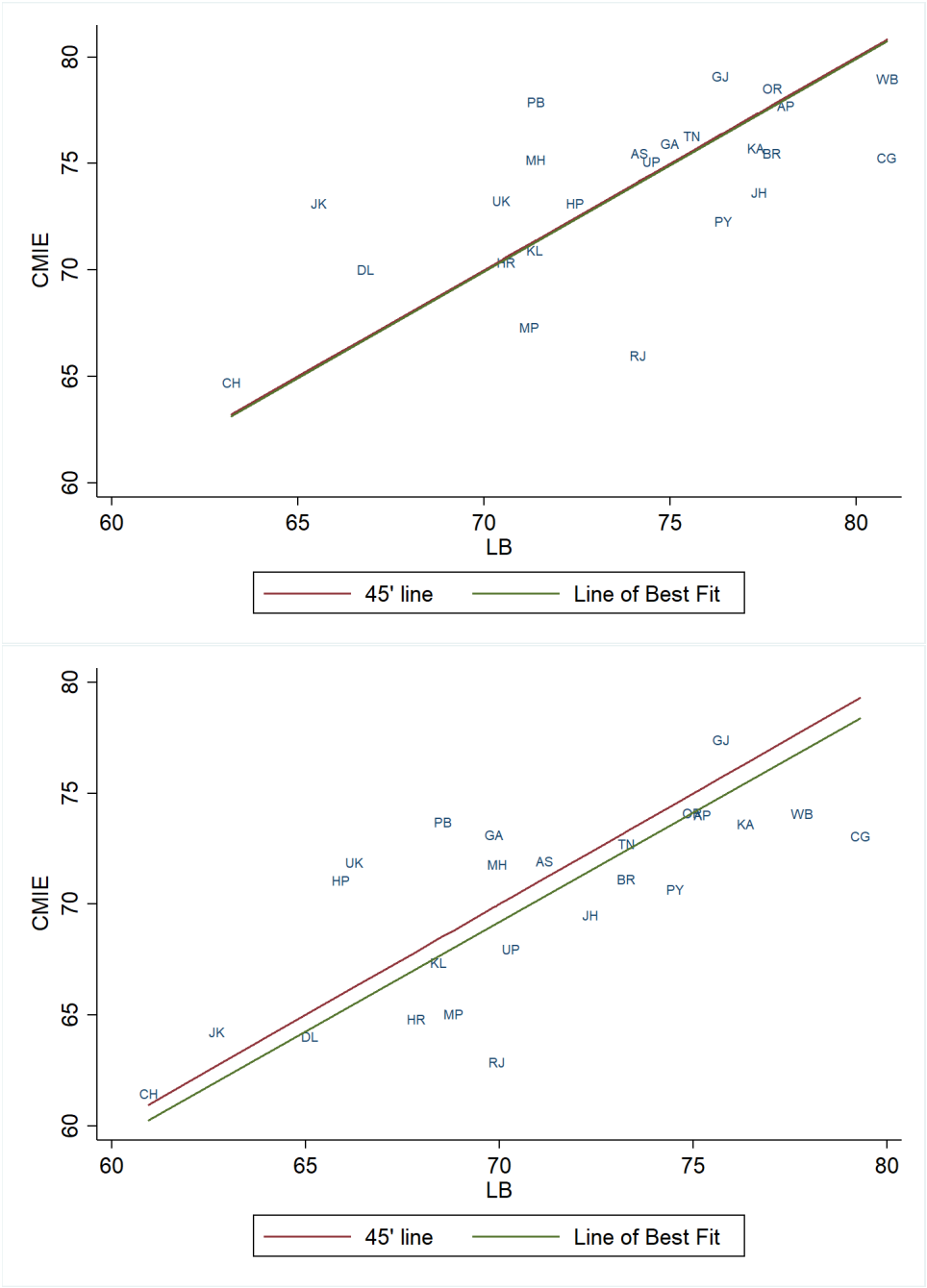
Taken together, both the econometric analysis and analysis of state-level variations indicate that measures of women's participation in the labour force seem particularly sensitive to the way questions are asked in surveys, and predictions of women's LFPR based on standard labour supply variables are much less reliable than those for men. When using data for men, the level of comparability is quite high and aggregate estimates like LFPR and WPR are found to match very closely.

Figure 2: Comparing CMIE-CPHS and LB-EUS estimates: Female



Note: The graph on the top is for LFPR and the one on the bottom is for WPR

Figure 3: Comparing CMIE-CPHS and LB-EUS estimates: Male



Note: The graph on the top is for LFPR and the one on the bottom is for WPR

References

Abraham, R., J. Shibu, and R. Narayanan (2019). Where has all the jobs data gone? *Business Standard*.

Bardasi, E., K. Beegle, A. Dillon, and P. Serneels (2011). Do labor statistics depend on how and to whom the questions are asked? results from a survey experiment in tanzania. *The World Bank Economic Review 25*(3), 418–447.

Basole, A. and A. Jayadev (2019). The employment question in india: Politics, economics, and the way forward. *The India Forum*.

Basole, A., A. Jayadev, A. Shrivastava, and R. Abraham (2018). State of working india, 2018. , Centre for Sustainable Employment, Azim Premji University.

Chodorow-Reich, G., G. Gopinath, P. Mishra, and A. Narayanan (2018). Cash and the economy: Evidence from india's demonetization. Working paper, National Bureau of Economic Research.

Fletcher, E., R. Pande, and C. M. T. Moore (2017). Women and work in india: Descriptive evidence and a review of potential policies.

Heath, R., G. Mansuri, D. Sharma, B. Rijkers, and W. Seitz (2016). Measuring employment in developing countries. evidence from a survey experiment. Working paper, Working paper.

Hussmanns, R., F. Mehran, and V. Varmā (1990). *Surveys of economically active population, employment, unemployment, and underemployment: an ILO manual on concepts and methods*. International Labour Organization.

Jha, S. (2019a). Unemployment rate at four-decade high: Nsso survey compared past figures. *Business Standard*.

Jha, S. (2019b). Unemployment rose to a 4-year high during demonetisation: Govt survey. *Business Standard*.

Kingdon, G. G. and J. Unni (2001). Education and women's labour market outcomes in india. *Education Economics 9*(2), 173–195.

Klasen, S. and J. Pieters (2012). Push or pull? drivers of female labor force participation during india's economic boom.

Shrivastava, A., R. Abraham, and A. Basole (2019). What do household surveys reveal about employment in india since 2016? In *State of Working India, 2019*. Centre for Sustainable Employment, Azim Premji University.

Srivastava, N. and R. Srivastava (2010). Women, work, and employment outcomes in rural india. *Economic and political weekly*, 49–63.

Verick, S. (2014). Women's labour force participation in india: Why is it so low. *International Labor Organization*.

Appendix

Actual status	Predicted status	Category
OOLF	OOLF	Matched
OOLF	Unemployed	LFP overpredicted
OOLF	Employed	LFP overpredicted
Unemployed	OOLF	LFP underpredicted
Unemployed	Unemployed	Matched
Unemployed	Employed	Employment overpredicted
Employed	OOLF	LFP underpredicted
Employed	Unemployed	Employment underpredicted
Employed	Employed	Matched

Table A1: Categorising Actual and Predicted Status combinations

	(1) Employment status	(2) Employment status
For outcome 1		
Age	0.0712***	-0.0495***
	(0.00994)	(0.00702)
Age squared	-0.00155***	0.000138
	(0.000142)	(0.0000946)
Education categories ('Below primary' is the base category)		
Primary	0.131	0.165**
	(0.149)	(0.0814)
Middle	0.209	0.248***
	(0.133)	(0.0842)
Secondary	0.662***	0.635***
	(0.128)	(0.0740)
Higher secondary	1.154***	1.045***
	(0.126)	(0.0740)
Graduate	1.971***	1.741***
	(0.125)	(0.0707)
Post-graduate	1.761***	1.834***
	(0.170)	(0.105)
For outcome 2		
Age	0.527***	0.188***
	(0.00411)	(0.00476)
Age sqyared	-0.00605***	-0.00231***
	(0.0000505)	(0.0000581)
Education categories ('Below primary' is the base category)		
Primary	0.169***	-0.406***
	(0.0449)	(0.0287)
Middle	-0.0291	-0.647***
	(0.0441)	(0.0388)
Secondary	-0.357***	-0.992***
	(0.0407)	(0.0372)
Higher secondary	-0.660***	-0.871***
	(0.0418)	(0.0470)
Graduate	-0.639***	-0.188***
	(0.0417)	(0.0389)
Post-graduate	-0.613***	0.648***
	(0.0623)	(0.0618)
N	226168	205167

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A2: Estimates of multinomial logit model of employment status on CMIE-CPHS

	(1) LFPR CMIE	(2) WPR CMIE
LFPR LB	0.356**	
	(0.127)	
WPR LB		0.429***
		(0.108)
Constant	28.81***	22.49***
	(6.312)	(5.103)
Observations	24	24
R^2	0.263	0.416

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: Regression results for state level LFPR and WPR

	(1) LFPR CMIE (Female)	(2) WPR CMIE (Female)
LFPR LB (Female)	0.450***	
	(0.138)	
WPR LB (Female)		0.446***
		(0.117)
Constant	5.814	3.062
	(3.522)	(2.746)
Observations	24	24
R^2	0.326	0.399

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Regression results for state level female LFPR and WPR

	(1) LFPR CMIE (Male)	(2) WPR CMIE (Male)
LFPR LB (Male)	0.562***	
	(0.143)	
WPR LB (Male)		0.660***
		(0.136)
Constant	32.36***	23.31**
	(10.59)	(9.622)
Observations	24	24
R^2	0.411	0.519

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: Regression results for state level male LFPR and WPR