The Skills Space of Informal Workers: Evidence from Slums in Bangalore, India*

Nandana Sengupta[1], Sarthak Gaurav[2], and James Evans[3]

[1]Indian Institute of Technology Delhi
[2]Indian Institute of Technology Bombay
[3]University of Chicago, Santa Fe Institute

September 10, 2019

Abstract

We develop a framework for mapping and analyzing informal worker skills in the developing world. Using microdata from nearly 1500 workers residing in the slums of Bangalore in India, we also study skill determinants of wage level and wage regularity. Alongside econometric modelling, we employ techniques from machine learning to describe relationships between the large number of skills crowdsourced from respondents. We propose to retire the concept of "unskilled labor" by revealing that employment in the informal labor market relies not only on general and specialized task skills such as handling cash, cooking, and internet use but also a complex matrix of language, personal and social capacities, ranging from Kannada speaking to punctuality, dressing sense and a tolerance to work in unclean and unsafe environments. Policy implications of the study include insights about gender disparities in skill, the importance of English language and computer literacy skills and the central role of personal and social skills in the Indian informal labor market context.

KEYWORDS: India; Skills; Informal sector; Labor market; Crowdsourcing; Machine Learning.

*Financial Support from the National Science Foundation, University of Chicago Delhi Centre, and Azim Premji University. The authors are grateful to Indian Institute of Technology Bombay for administrative support. Amitabh Chaudhary at University of Chicago provided critical technical guidance and Vishal Vincent Joseph provided valuable research assistance. Comments from participants at a workshop at the University Chicago Center in Delhi were helpful.

I. Introduction

India is one of the youngest nations in the world with over half of the population below 25 years of age and an expected median age of 29 years by 2020 (Chandramouli & General, 2011). With one billion people in the 15-64 age group the Indian workforce is expected to be the largest in the world. In order to reap the demographic dividend that such statistics imply, however, there is an urgent need to overcome the twin challenges of available employment opportunities and labor force employability, such that skills supplied by job seekers better match those demanded by jobs. With approximately five million new entrants to the workforce every year and a pressing challenge of job creation, there is an unprecedented skilling gap that needs to be tackled in the context of a large and growing informality of the labour force (Himanshu, 2011; Mehrotra, Parida, Sinha, & Gandhi, 2014).

Over 90% of the work force of 453 million are employed as informal workers (Sengupta et al., 2009).[1] While the informal sector is dominant in both the rural and urban economy, the urban informal sector plays a crucial role in providing livelihood; particularly to migrant workers as well as low-income households residing in urban slums. Moreover, the structural transformation of the economy has been associated with a general movement of labour from the agricultural sector (Binswanger-Mkhize, 2013). Shrinking employments in the rural economy has led millions of workers with little formal educational to throng the cities. These migrants are more likely to be informally employed in urban centers. Entry and exit from informal work is easier compared to formal employment and the work is typically labour intensive, not requiring advanced education or elite service skills training. With little awareness about their rights as workers and low negotiating power, there is the potential for employers to exploit informal workers.

Informal employment is not restricted only to the "informal sector", which consists of enterprises characterized by heterogeneity in activities, low levels of organization, and the absence

[1] The Government of India "Report on Conditions of Work and Promotion of Livelihoods in the Unorganized Sector' (2008) estimates that informal (or unorganized) workers comprise about 92% of the total workforce in India. We use the same definition for informal or unorganized workers as this report: *"Unorganised workers consist of those working in the unorganised enterprises or households, excluding regular workers with social security benefits, and the workers in the formal sector without any employment/ social security benefits provided by the employers."*

of social security benefits for all employees. Even within the formal sector, a large proportion of workers are employed informally. These workers are casually employed without formal contracts. Workers are typically not provided social security such as disability insurance and pension. Our study focuses on informal workers who may be employed in either formal or informal sectors.

The study of informal workers has received a fair amount of academic interest from the sociological and ethnographic research community. For example, Ruthven (2008) investigates the moral bases for different types of informality observed within an artisanal industrial cluster in North India; Harriss-White (2009) analyzes the effect of globalization on the Indian informal economy from the lens of small commodity producers. Of particular relevance to our paper is the discussion of viewing the informal sector from a knowledge perspective by Basole (2014) who stresses on the need to investigate modes of knowledge production and innovation within the informal sector.

Quantitative and econometric scholarship on informal workers in India remains sparse and much of the work is based on aggregate statistics. Sengupta et al. (2009) focus on the broad findings of the National Commission for Enterprises in the Unorganized Sector which was set up by the Government of India in 2004 to understand the nature and magnitude of the Indian employment situation using nationally representative data from National Sample Surveys (NSS). A review of the empirical literature with special emphasis on whether informal work is a means of exploitation or a means of accumulation is provided in Maiti and Sen (2010). Marjit and Kar (2009) develop a estimate a theoretical model of the effect of regulatory changes on informal worker wages in India. Basole and Basu (2011) use aggregate-level data to trace the structural changes in the informal industrial sector. Mehrotra et al. (2014) analyze employment trends from 1993-2011 and estimate the need to create approximately 17 million jobs outside agriculture every year between 2012 and 2017. Justino (2007) uses data from a panel of 14 Indian states to estimate the importance of social security policies in developing countries.

A smaller set of papers study informal work at the individual level using quantitative methods. For example Bairagya (2012) and Shonchoy and Junankar (2014) use NSS and India Human Development Survey data respectively to estimate individual characteristics that determine informal employment. Goel and Deshpande (2016) investigate the relationship between caste identity and measures of self-worth among self-employed workers (who make up a large fraction of informal workers) using NSS data. Each of these studies uses nationally representative data but are restricted in the scope of their research questions by the structure of the original, fielded surveys. A few researchers have overcome this restriction by collecting primary data to answer specific research questions regarding informal work. For instance, Maiti (2008) collects primary data to study rural industrial and artisanal units of production in the Indian state of West Bengal. Banerjee (1983) uses primary data collected in Delhi and finds that rural to urban migration is likely to be led by opportunities both in the urban formal and informal sectors and that mobility from informal to formal sector is low. Gurtoo and Williams (2009) suggest that informal work is not always out of necessity or distress but often a voluntary entrepreneurial choice. [2]

In this paper we use primary data from slums in Bangalore to study the skills space of informal workers. We discover and map skills that workers perceive to be valued by the labor market, which range from language skills, to general and specific task skills to cultivated personal and social capacities. We attempt to examine if skills are clustered in their self-reportage and the degree to which they influence labour market outcomes, particularly the level and regularity of wages. To our knowledge this is among the first attempts to empirically analyze the effect of skills—including non-cognitive skills and personal qualities—within a developing economy labor market context. In particular, previous empirical research on skills in the Indian context is sparse. Helmers and Patnam (2011) study the determinants of cognitive and non-cognitive skills using data from two cohorts aged one to twelve from Andhra Pradesh. Krishnan and Krutikova (2013) evaluate the impact of a non-cognitive skill development program aimed at children and adolescents in urban Bombay. Pilz, Gengaiah, and Venkatram (2019) study skill development among street food vendors and highlight the importance of informal skilling and apprenticeship.

Despite sparse work on India, the study of skills and labour market outcomes has received a lot of attention globally; especially in the context of the US labor market (Heckman & Ru-

[2] The authors of the study note that their conclusions are not representative of any specific population due to their convenience sampling strategy.

binstein, 2001; Cunha & Heckman, 2007; Cunha, Heckman, & Schennach, 2010; Borghans, Duckworth, Heckman, & ter Weel, 2008). Heckman and Kautz (2012) emphasize the role of 'soft' social and personal capacities in predicting many dimensions of success. While desirable personal and social qualities like a grit, positive demeanor and social confidence are sometimes viewed as unalterable and unmeasurable traits, the demonstrated clinical success of approaches like cognitive-behavioral therapy (Hofmann & Smits, 2008; Butler, Chapman, Forman, & Beck, 2006) suggest that many of these seemingly non-cognitive skills can be cognitively cultivated, taught and exercised. These are often neglected as skills, likely because they require nontrivial assessment over a variety of contexts that map onto the diverse settings in which they were acquired and trained through a lifetime of family instruction, community and work experience. Weinberger (2014) finds complementarity between cognitive and social skills using data that links adolescent skill endowments to adult outcomes. Deming (2017) uses US National Longitudnal Survey of Youth (NLSY) and Occupational Information Network (O*NET) data to show that labor market returns to social skills are increasing with time. Recent work by Börner et al. (2018) uses machine learning techniques to analyze millions of publications, course syllabi, and job advertisements to shed light on skill discrepancies between research, education and jobs in the US, revealing a disconnect between the critical demand for social and communication skills, but a limited supply for the same in education and research. Each of these papers is aided by rich and detailed skills data and it is precisely the lack of related sources that challenges scholarship on skills in developing countries, particularly in geographies characterized by high informality.

Our study has a two-fold objective. First, we are interested in mapping the distribution of skills and cultivated personal capacities among informal workers in India. Second, we are interested in understanding to what extent diverse skills are associated with economic well-being as measured by higher and regular wages. We investigate these questions by collecting primary data from informal workers on the labor supply side from urban Bangalore in the southern state of Karnataka. Bangalore is synonymous with the Information Technology (IT) sector in India, considered to be a key driver of India's growth. Data is collected in two stages – in the first stage, workers provide information on their employment and training histories, their demographic characteristics, economic status, and a battery of general skills and talents. In this stage we also crowdsource diverse task-related skills, knowledge and personal or social capacities that respondents attribute to themselves and associate with value in the labor market. This crowdsourced data then becomes a critical input for the second stage of data collection where we conduct another survey that collects worker self-assessment regarding the most frequent crowdsourced skills from the first round along with a shorter survey on demographic characteristics, economic status and employment outcomes.

Given the growing policy emphasis on training the country's large and growing workforce through large scale public skilling programmes in both urban and rural areas, our study offers highly relevant insights for policy makers, training institutes and employers. Understanding the skill distribution across urban informal workers is also important in light of evidence regarding the low and falling labour force participation rate (LFPR), particularly for women (Bhalla & Kaur, 2011; Mehrotra & Parida, 2017). Insights from this paper may be especially useful in light of concerns about jobless growth (Bhalotra, 1998; Kannan & Raveendran, 2009) and the effects of automation on jobs (David, 2015; Arntz, Gregory, & Zierahn, 2016).

Our work makes a number of novel contributions to the literature. First, it adds to the quantitative literature on informal workers using primary data from a new geography. Second, our paper is among the first to study the role of diverse skills, including not only task-related abilities, but also cultivated personal and social talents, on employment outcomes from a hitherto unexplored population – informal workers in urban India. Third, along with standard econometric methods our study is designed to utilize machine learning techniques for manifold learning particularly well suited to the study of complex relationships between skills. In doing so our study marks an initial attempt to collect and analyze primary data in a fieldwork setting with a view towards using machine learning predictions to improve skill recommendation services. Finally, our unique data allows us to analyze the skills space in the Indian informal labor market from a multi-dimensional perspective. To our knowledge this is the first such attempt to apply a quantitative, bottom-up approach to understanding the distribution of skills in the Indian labor market.

The most striking results from our paper signal the significant gender disparity in skill distribution and wages. We also find that English language skills and computer literacy skills are consistently associated with better job outcomes. The importance of personal and social skills in the urban informal labor market is another important finding. We make the following policy recommendations based on our study: first, there is a case for large scale financial and computer literacy programs targeted specifically at women; second, we recommend prioritizing English and computer literacy skills in public schools; third, we recommend that personal and social skills training should be an inherent, if not central, part of any vocational training program; fourth: based on the large number of unique skills elicited from respondents we strongly caution policy makers against using the term 'unskilled' for informal workers, recommending instead the description of such workers as having 'transferable' or 'portable' skills, recognizing their demonstrated value. We argue that discovering the diversity of these so-called 'soft' skills, their complex relationships and complementarities, will facilitate training and staffing outcomes that improve human welfare.

The paper proceeds as follows: Section II provides background on the skilling question in the Indian context. Section III describes the sampling methodology, survey design and respondent characteristics. Section IV elaborates on our empirical methodology which employs both econometric and machine learning methods. The results of our analysis are presented in Section V. Section VI concludes.

II. Background

A. Skilling in India: Policy initiatives

Lack of marketable skills is likely to be an important underlying factor when it comes to an individual taking up a low paying and insecure job as an an informal worker and becoming blocked from wage advancement. The mode of skilling accessible to most workers is via informal and on-the-job training. Although these casually acquired skills enable workers to functionally perform specific tasks, lack of formal skilling hampers future employability and may adverse impact productivity. Vocational training has been advocated as the primary solution. The Indian government as well as certain non-governmental agencies have taken up the initiative and introduced farflung vocational training platforms, although a majority of the population has not yet gained access.

Over the past decade there have been considerable efforts at meeting the industrial demand for trained manpower as well as addressing the skilling gap in India. In 2007-08, along with significant credit extended to upgrade the infrastructure of Industrial Training Institutes (ITIs), the Skill Development Initiative Scheme was launched by the United Progressive Alliance government.[3] The National Skill Development Council (NSDC) was founded in 2008 to foster public private partnership in skill development.[4] This was followed by the National Skill Development Policy in 2009. In 2013, a competency based framework called the National Skills Qualification Framework was introduced to organize all qualifications according to a levels of knowledge, skills and aptitude[5]. Another scheme with a high budget outlay, the National Skill Certification and Monetary Reward Scheme, offers financial incentives for successful completion of approved training programs was launched in 2013. The National Democratic Alliance government that followed announced the National Policy for Skill Development and Entrepreneurship in 2015, and set up a Ministry of Skill Development & Entrepreneurship (MSD&E). The government also introduced the Pradhan Mantri Kaushal Vikash Yojana to impart skills training to 24 million youth for industry-relevant skill training.[6]

In the current institutional arrangement for skilling, three key institutions that drive skilling initiatives by the government are the National Skill Development Council (NSDC), National Skill Development Agency (NSDA), and the Director General of Training (DGT).[7] In addition, nearly 40 Sector Skill Councils (SSCs) have been set up. SSCs are industry-led and industry-

[3] One of the most important channels for vocational training are the Industrial Training Institutes (ITIs): around 12,000 ITIs, four fifths of which are private institutes. The ITIs impart training in 126 trades (73 Engineering, 48 Non- Engineering, 05 exclusively for visually impaired) over a duration ranging from one to two years. The ITIs are affiliated with the National Centre for Vocational Training (NCVT).

[4] NSDC was set up by the Ministry of Finance as a public private partnership company with the objective of supporting the skill development ecosystem and funding private sector vocational training institutions.

[5] NSQF requires five attributes for any level - process, professional knowledge, professional Skill, core Skill and responsibility. The levels are graded from one to ten, and are defined in terms of expected learning outcomes, regardless of how that learning was obtained —formally or informally.

[6] This flagship scheme of MSD&E has a target to train 14 million fresh trainees and assess and certify 10 million under recognition of prior training.

[7] NSDA is an autonomous body under MSD&E to coordinate and harmonize skill development activities across the country, including being the nodal agency for state skill development missions. DGT comprises a network of Industrial Training Institutes; Advanced Training institutes, Regional Vocational Training Institute and other institutes that offer training programs catered to students.

governed bodies for skill development in specific sectors. While most SSCs are clustered in the priority sector that includes auto, retail, IT and ITES, logistics, food processing and hospitality, 3 SSCs specifically cover informal work in plumbing, beauty & wellness, and domestic work. Large workforce sectors comprising telecom, agriculture, mining, and instrumentation comprise another cluster[8]. Finally, skilling initiatives such as Deen Dayal Upadhyay Grameen Kaushal Yojana implemented by the Ministry of Rural Development target the rural workforce. In total, 20 government ministries are involved in the space of skilling and vocational training, underscoring its urgency as a policy priority.

Even if access to training is obtained, however, the effectiveness and quality of the various skill development programs is unclear. Despite several institutional initiatives for imparting skills training, these initiatives have been plagued with a lack of skilled trainers, poor placement records, an inadequate industry interface, absence of appropriate in-plant apprenticeship programs, financing mechanisms and lack of standardization in the vocational training system, and overall poor coordination among multiple stakeholders (Prasad et al., 2017). Moreover, even if the state managed to effectively skill tens ofhttps://www.overleaf.com/6665539994hyhwngdgqwtk millions of job seekers, the arduous task of matching those skills to suitable jobs would remain a challenge.

In recent years, online portals such as ‘Babajobs’ have facilitated this skill matching at a primitive level for blue and grey collared workers. The pre-requisite skills for a particular job are posted on the website and those who satisfy requirements can apply. In recent years, technology platforms like the start-up ‘UrbanClap’ have jumped into the job-matching market with a focus on informal workers. ‘UrbanClap’ matches service providers like plumbers, carpenters and beauticians directly to consumers.

B. Informal Work in Bangalore

Bangalore (also called Bengaluru) is the capital city of the the state of Karnataka in southern India. It is a hub of IT and IT enabled services. The city houses start ups that became leading

[8] In the presence of a multi-agency approach to skilling, the institutional support for consolidation and coordination of skilling efforts has been facilitated under the National Skill Development Mission 2015. See (Prasad et al., 2017), pp.8-28 for details regarding the evolution of vocational education and training ecosystem in India.

global IT companies such as Infosys and Wipro; earning the city the moniker of the 'Silicon Valley of India'. The largest e-commerce company of Indian origin, Flipkart is also based in Bangalore. The urban agglomeration around Bangalore has a population of approximately nine million (2011 Census), making the city the third most populated in India, following Delhi and Mumbai. Apart from the software industry, the city offers considerable opportunities in the construction, real estate, hotels and hospitality, aerospace, biotechnology and agribusiness sectors.[9] Owing to its vibrant economy and dynamic workforce, the city has emerged as a popular destination for migrant workers across the skills spectrum.

In order to understand the composition of Bangalore's workforce, we analyzed unit level data from the 68^{th} Round of the nationally representative NSS Employment and Unemployment Survey, 2012-13. See Appendix A1 for the distribution of working age population (15-59 years of age) by activity, and activity based composition of formal and informal sector workers by Bangalore district.[10] More than half of the workers labour in the informal sector, with own account workers (self employed) being the most prevalent. Over a quarter of the informal workers are wage employees. Nearly 15 percent are involved in casual wage labour, whereas 13 percent are involved in unpaid family work. Among formal workers, regular salaried or regular wage workers are most prevalent.

In the context of growing Indian urbanization, informal settlements or slums are a stylized fact of the cities with more than a million inhabitants, housing more than a third of the country's slum dwellers (Roy et al., 2018). Poor and low income migrant workers as also native residents primarily employed in the informal sector are unable to afford formal housing and take up residence in the slums. Previous study on the socio-economic condition of Bangalore slums found that slum dwellers are mostly informal workers; with a little over 10 per cent being employed in formal sector work (Krishna, 2013; Krishna, Sriram, & Prakash, 2014). In this way, our study of urban informal skills is aptly set in the slums of Bangalore.

[9] See (Nair, 2005) for a historical perspective on the urbanization of the Bangalore metropolitan area and employment opportunities associated with the city's growth.

[10] While we focus on Usual Principal Activity Status that has a reference period of 365 days preceding the date of survey, the instrument also collected data on individuals' subsidiary activity status. Approximately 13 per cent of the sample is rural as there is both a Bangalore rural and urban district. Trends within the urban sample comprising 1296 individuals are qualitatively similar.

III. Data Collection

Data for our empirical analysis come from a two-stage survey of informal sector workers residing in Bangalore slums. In the first stage, we collect demographic characteristics and employment histories from workers along with their self-assessment on a set of basic skills and personal qualities. We also crowdsource skills, personal and social capacities and knowledge that respondents attribute to themselves and believe to be valued in the labor market. In the second stage of data collection we conduct surveys that contain a shorter segment on employment and demographic characteristics followed by self-assessment data on the most frequently occurring crowdsourced skills from the survey's first stage.

A. Sampling Strategy

For each of the two rounds we follow multistage sampling methods in order to sample 'declared' slums in the Bangalore urban district based on the 2011 Census of India. [11] Our sample is representative of workers living in Bangalore declared slums, who are typically informal workers. We do not claim that our sample is representative of the remaining types of slums in the city i.e., 'undeclared', 'non-notified' or 'de-notified' slums.

Using information on the total number of individuals in declared slums 306537 (N), we arrive at the required sample sizes (S) for our study using the following formula (Cochran, 1963):

$$S = \frac{n}{1+\frac{n}{N}}; \qquad \text{where} \qquad n = \frac{Z^2}{E^2}P(1-P). \tag{1}$$

($Z = 1.96$) is the Z-score for 5% level of significance. P is the estimated proportion of cases in the population of interest. In our study, we approximate P as the proportion of population in the city in the $15-59$ years age group ($P = 0.7$) based on the population pyramid of Bangalore detailed in the 2011 Census of India. [12] E is the error margin. In larger sample surveys comprising notified and non-notified slums in the city, a margin of 3% is commonly used (Roy et al., 2018). Our budget and resource constraints did not support this margin, but rather a

[11] Official agencies acknowledge only two broad types of slums: officially declared ("notified" or "recognized") slums. All others are described as low-income settlements.

[12] Note that variability in the working age population is captured in the product term of $P(1-P)$. Different estimates of the working age population in the city will result in different variability estimates of that population.

feasible error margin we could achieve for our surveys. For the first round, the margin of error chosen was $E = 3.5\%$ leading to a sample size of $S_1 = 657$. For the second round, we fixed the margin of error at $E = 3.25\%$ leading to a sample size of $S_2 = 762$.[13] Based on pilot surveys, a response rate of 93% was feasible. Therefore our sample sizes for final surveying in Round 1 and Round 2 were set at $S_1 = 706$ and $S_2 = 819$. Our final Round 1 and Round 2 datasets consist of 698 and 784 respondents respectively.

For each of the two rounds we employed a multistage sampling technique, a form of cluster sampling. In the first stage we randomly selected 5 town areas comprising multiple slums from a list of 15 town areas in the Bangalore urban district. The geographical locations of the town areas selected in both rounds is presented in Figure 1. In the second stage, we randomly selected two slums from the first stage units, resulting in a total of 10 slums to be considered for a single round of surveys. From the 10 slums sampled in the second stage, we employed a population proportional to size (PPS) sampling design for determining the number of units required to meet our required sample size[14]. In order to select households to be surveyed, we used systematic sampling based on random initialization within each slum. As indicated by the pre-survey demographic assessment of slums, each household was represented by approximately five members. Based on the roster of household members our trained investigators prepared, one working age respondent was selected randomly in each household as the main respondent in our surveys.

Figure 1 about here.

A pre-tested questionnaire was administered in CAPI (computer assisted personal interview) format on tablets for which the investigators had been well trained. Following respondents' informed consent to participate in the survey, questions were asked in the local languages of Kannada and Tamil by a team of trained surveyors. Wherever applicable, if the respondent was a migrant uncomfortable with the vernacular, the investigators conducted the survey in Hindi or English as per the respondent's preference. [15]

[13] We were able to marginally reduce the value of E from 3.5% to 3.25% due to the availability of additional funding for the project at the start of Round 2.

[14] Detailed tables of our sampling methodology are presented in the Appendix

[15] We did not use a sampling frame such as voter list as we expected to sample individuals in the 15-59 years age

Conducting surveys in slums posed several challenges. First, there were instances of discrepancy between the number of households reported by the government and the actual population. Our trained investigators conducted rapid appraisal of the history of the slum and vetted its status of notification. Second, several slums have undergone a transformation that no longer qualifies them to be a slum. In one instance, the slum selected for our sample had been redeveloped and de-notified. Following careful ethnographic observation, our investigators replaced the slum with another in the vicinity having a comparable number of households, similar housing stock and patterns of human behavior. Third, there were instances of initial resistance to the survey from those who lived in the slums. Our investigators deliberated with the community leaders regarding the objectives of the study, and surveys commenced following the approval of community leaders.

B. Survey Design

In Round 1 the main survey questions covered demographic characteristics (including age, gender, education, parents' education, social group, migration, family size, etc.), economic status (including household assets, savings and expenditure), employment history (including wages, experience, primary occupation etc) and employment seeking behaviour (including job search and training). Given the importance of basic skills in a changing economic environment (Murnane & Levy, 1996; Peter-Cookey M.A., 2017), respondents were also asked to assess themselves on 11 basic skills (ability to: read/write/speak Hindi and English; do simple math; use the internet; use a mobile phone; ride a motorbike; drive a car) and on 11 personal qualities (whether they consider themselves confident; ambitious; organized; sociable; talkative; make new friends easily; come up with new ideas; handle stress well; trust others easily; prefer routine work; or prefer high paying but intensive over low paying but relaxed jobs). Our choice of qualities was motivated by the literature on personal qualities and labour market outcomes Krueger and Schkade (2009); Lex Borghans and Weinberg (2009).

Respondents were scored at 1 if their self-assessment on basic skills was 'very comfortable', 0 if 'not at all comfortable' and 0.5 if 'somewhat comfortable'. Respondents were scored at 1

group. Voter IDs would only be available for those who 18 years or older. Furthermore, many migrants remain unlisted on the voter list of the constituency.

if a personality trait applied to them, 0 if not and 0.5 if they were unsure or indifferent. Finally, respondents were asked to come up with as many skills, qualities and knowledge that they attributed to themselves, which they believed had helped them in the job market.[16]

In Round 2, respondents were asked to assess themselves on the 100 most frequent skills elicited by Round 1 respondents. Along with these 100 skills, we added another six language skills: ability to read and write English, Hindi, and Kannada. The complete list of skills is presented in Table 1. Though Kannada is the lingua franca of the local population, Bangalore is home to several migrant workers in its IT sector whose first or second language is Hindi. For local workers in low wage employment, ability to speak English or Hindi may be associated with greater employment opportunity. Previously, young workers have been shown to receive a premium for spoken English in India (Azam, Chin, & Prakash, 2013a). Respondents were also asked a shorter set of questions focused on demography and employment including age, gender, education, parents' education, social group, migration, family size, employment status, experience, primary activity, wages and training.

C. Respondent Characteristics

Figure 2 provides a graphical description of respondent characteristics from Round 1 data with respect to a) distribution of male occupations, b) distribution of female occupations, c) mechanisms by which workers obtained their last job and d) percentage of workers who have a written contract for their current job. Summary statistics on skills, dis-aggregated by gender and on variables used in the empirical analysis are presented in Sections A. and B. respectively.

Figure 2 about here.

As at the national level, and in sync with our general description of the slums, both males and females are engaged in an extremely diverse set of occupations. Among jobs that could be labelled, construction sector jobs including painting jobs at construction sites comprise nearly a quarter of the jobs. Auto drivers, mechanics, drivers, delivery boys, retailers, caregivers,

[16] The question asked is reproduced here verbatim: *'What skills, knowledge or values do you have which you think help in getting or keeping a job? List as many as you can think of. (If no clear answers, surveyor should ask – if current, former or potential employers were considering both you and another person for the same job, why did they pick you?)'*

security staff and casual wage labourers are other occupation categories for males that emerged from our survey. Among females, housekeeping is the most commonly reported occupational category. Tailoring, floriculture, wedding work, retail, caregiving, and own enterprise are other categories that frequently emerged. Occupations reported by less than 2.5% of the sample (i.e. less than 10 men or 8 women) were collapsed into the 'Other' category – which makes up a large component of the distribution of occupations. Such heterogeneity in occupations is typical for informal workers who often switch frequently from one paid task to another.

Nearly two thirds of the jobs were obtained by referrals from a relative or a friend who knows the employer. This is an important way through which informal workers substitute formal mechanisms like job portals (e.g. Linkedin, Babajobs, UrbanClap). Direct inquiry to labor contractors at job locations also emerged as an important search mechanism. Dependence on web job portals is extremely low despite access to smart phones and moderate self reported ability to use the internet among men. This suggests that kinship networks and physical social networks dominate the job search space[17]. Finally, we note that only 10 respondents in our Round 1 sample reported having a written contract for their job, which constitutes 1.4% of the total sample and 1.7% of respondents in the sample who are currently employed. These distributions confirm that our sample consists almost exclusively of informal workers and corroborate our assumption that the data is representative of informal workers living in Bangalore's declared slums.

IV. Empirical Strategy

Our choice of empirical methods was guided by our two main interests to a) generate a map of skills among informal workers and b) investigate which skills are associated with better job outcomes. We utilize a mix of traditional econometric and machine learning techniques for our analysis. Machine learning techniques are particularly useful in applications involving units of interest with a large set of attributes: self-assessments on a large number of skills in the Round 2 data is a natural application area. Within machine learning methods, unsupervised

[17] There is also considerable variation in occupation by religion and caste but we have not reported them for brevity. These results are available to be shared upon request.

learning methods like k-means clustering and embedding diagrams shed light on the underlying structure within high dimensional skill spaces, whereas supervised learning methods like sparse regression (e.g., LASSO) can be used to select attributes that best predict relevant outcomes.

A. Mapping the Skills Space

When mapping skills, we are interested in gender disparities among self-assessed skills, the co-occurrence of skills and the distributional hierarchy of skills.

A..1 Comparison of Means and Correlation Plots

Because gender disparities in the Indian labor market are well documented (?, ?; Filmer & King, 1999; Verick, 2018) we explore differences in self-assessed skills proficiency across male and female respondents. In order to do so we compute mean self assessment values for both groups and conduct standard comparison of means t-tests for each skill. We do this for self-assessments in both Round 1 and Round 2 data.

For Round 1 data we also present correlation plots for all basic skills and all personal qualities respectively. In both of plots we add a dummy variable for female respondents. The analysis is carried out in `R` using the package `corrplot` to generate correlation plots.

A..2 K-means clustering

We cluster the list of skills from Round 2 data using the standard k-means algorithm. As the name suggests, this technique partitions the data into k groups clustered around k central points. In our study, each of the 106 skills examined, s, has an associated vector of self-assessments denoted $x^{(s)}$. The algorithm runs iteratively, starting with an initialization of k random cluster centers, μ_k. The cluster center to which skill s is assigned is denoted $\mu^{(s)}$.

Step 1: Clustering The algorithm aims to find k clusters C_k with corresponding cluster centers μ_k such that the sum of the squares of the euclidean distance between each skill vector $x^{(s)}$ and the corresponding cluster centers $\mu^{(s)}$ is minimized. The associated optimization function is.

$$\min_{\mu_1,\cdots\mu_k} \frac{1}{106}\sum_{s=1}^{106} ||x^{(s)} - \mu^{(s)}||^2 \quad \text{where} \quad \mu^{(s)} \in (\mu_1, \cdots, \mu_k).$$

Step 2: Updating cluster centers. In the next step the cluster centers μ_k are updated by taking the mean of the vectors in each cluster C_k.

$$\mu_k = \frac{1}{|C_k|} \sum_{x^{(s)} \in C_k} x^{(s)}.$$

Steps 1 and 2 are repeated until convergence or a maximum number of iterations are reached. The primary input into the k-means clustering algorithm is the total number of clusters, k. This number itself can be estimated using the 'gap-statistic' as recommended in (Tibshirani, Walther, & Hastie, 2001). We implemented this in the algorithm in `R` using the command `kmean()` for clustering and the package `factoextra` for estimating optimal cluster number.

A..3 Hyperbolic embedding

Another popular class of unsupervised learning techniques involves the generation of low dimensional vectorized embeddings from discrete data within a structured system like persons within social networks, words within sentences, and in this paper, skills within people. Embedding techniques provide a map from discrete objects to numerical vectors that anchor a geometric space in which each object is uniquely placed to minimize distortion and optimally predict the original input data. Such maps can then be used to infer complex relationships between the modeled objects. One of the most popular low dimension linear embedding techniques for numerical data, Principal Components Analysis, is widely used by social scientists. Recent advances in computation have led to the ubiquitous use of neural network embedding techniques like `Word2vec`, `GLoVe`, and `BERT` for complex problems in natural language processing and information search, such as question-answering and analogy resolution (e.g., $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$). These approaches map words or phrases to a low-dimensional Euclidean space that reveals deep linkage between their underlying meanings. Related techniques generate embeddings in hyperbolic space, however, which captures not only similarity, but hierarchy among concepts (Chamberlain, Clough, & Deisenroth, 2017). Hyperbolic embeddings allow the precise representation of complex and intransitive associations between concepts or skills with far fewer dimensions than Euclidean space (Nickel & Kiela, 2017).

In this paper we generate hierarchical embedding figures for the 106 crowdsourced skills using hyperbolic embedding methods on the self-assessment data from Round 2[18].

We use the hyperbolic embedding algorithm developed by Nickel and Kiela (2017) in the open source package `PyTorch` [19]. The algorithm generates Poincaré disks, which represent tree-like hierarchical structures with high fidelity: points at the periphery represent the leaves and those at the center the trunk (see Nickel and Kiela (2017) and Börner et al. (2018) for more details). Each point on the disk is defined by a radius and an angle. A smaller radius indicates a more central position in the hierarchy. Skills lying close to the center are held in common by many workers in conjunction with diverse other skills. Smaller angular differences between any two skills indicate greater co-possession by the same and similar workers, suggesting skills that likely substitute or complement one another.

We use Round 2 self assessment data to generate a hyperbolic embedding of skills. The primary input for generating such embeddings is a list of object pairs. The algorithm estimates the underlying structure utilizing the co-occurence of skills. This is indicated by the frequency of each pair within the list. In this paper we created the input list as follows. For each individual we paired the skills that received the same self-assessment. If individual i's self-assessment was 'Very Comfortable' for both 'English Speaking' and 'Stitching', then those two skills form a pair. If many hold them in common, they will be close within the embedding, but if one is held in common with a much wider range of skills, it will be closer to the center. We consolidate all individual pair lists to create the final input list.

B. Determinants of Wages and Regular Work

The two job market outcomes with which we are interested are weekly wages and whether the respondent earns a regular monthly salary. The first of these is the usual metric for labor market success. Given the volatile nature of informal work, however, (respondents often earn wages on a daily or weekly basis) we are also interested in measuring job stability. We use receipt of regular monthly wages as a proxy for job stability.

[18] Principal Component embeddings were also generated, which represented comparable insights, and are available from the authors upon request.

[19] Available at https://github.com/facebookresearch/poincareembeddings

B..1 Econometric modelling

In order to analyze the association between basic skills and wages using Round 1 data we run different versions of the Mincerian wage equation ((Mincer, 1974)):

$$\log(wages) = \alpha_w + \beta_{edu}Edu + \beta_{exp}Exp + \beta_{exp2}Exp^2 + \sum_j \beta_{Xj}X_j + \sum_k \beta_{Sk}S_k + \beta_{imr}IMR + \varepsilon \quad (2)$$

where *wages* refers to weekly wages reported by respondents. Similarly for estimating the association between basic skills and regular monthly wages using Round 1 data we estimate the following outcome equation:

$$\text{logit}(Reg) = \alpha + \beta_{edu}Edu + \beta_{exp}Exp + \beta_{exp2}Exp^2 + \sum_j \beta_{Xj}X_j + \sum_k \beta_{Sk}S_k + \beta_{imr}IMR + \varepsilon \quad (3)$$

where *Reg* is an indicator variable for regular monthly wages. In the above regression equations *Edu* is number of years of schooling, *Exp* is number of years of experience [20] $X_j, (j = 1, \cdots J)$, is a set of controls including social identity (gender, religion, caste group) and training history. $S_k, (k = 1, \cdots K)$, is the respondent's self-assessment on basic skills and personal qualities. The term *IMR* is the Heckman correction term for addressing possible selection bias due to unobservable outcome data from respondents currently unemployed (Heckman, 1979). Specifically $IMR = \frac{\phi(\widehat{Emp})}{\Phi(\widehat{Emp})}$ is the inverse mills ratio estimated from the selection equation:

$$\text{logit}(Emp) = \alpha_{emp} + \sum_m \beta_{Zm}Z_m \quad (4)$$

where *Emp* is an indicator variable for respondent's currently employed and $Z_m, (m = 1, \cdots M)$, is a set of controls in the selection equation. We control for Mother's education, Father's education, dependent ratio in the household[21], asset index of the household, marital status, migrant status, whether schooling was in an English medium school and slum fixed effects. We also control for number of years experience, gender, caste and religious community. Finally we control for ownership of a smartphone, cycle and bike, as well possession of a PAN card

[20] This is calculated by subtracting the age at which the respondent started working from her current age.

[21] Calculated as the sum of dependents (members under 15 years of age and over 60 years of age) over the sum of wage earners in the household.

and a bank account, which indicate a basic level of financial inclusion. In some models we also include self-assessment of personal qualities in controls for the selection equation.

We estimate 4 versions of each model: 1) outcome equation and selection equation without controls for basic skills or personality traits; 2) controls for basic skills in the outcome equation; 3) controls for basic skills in the outcome equation and controls for personal qualities in the selection equation; 4) controls for basic skills and personal qualities in the outcome equation and controls for personal qualities in the selection equation.

B..2 LASSO regression

While the association between basic skills and job market outcomes is estimated using Mincerian wage equation specification for Round 1 data, the number of crowd-sourced skills for which respondents provide self-assessment in Round 2 is substantially larger at 106. In such cases, standard linear regression models often produce unstable estimates with low accuracy following from issues including multicollinearity. Sparse regression is one of the most popular techniques in the machine learning literature appropriate for such datasets, with the most popular being the LASSO technique (Tibshirani, 1996). LASSO simultaneously performs model selection and parameter estimation by adding a penalty term to the standard linear regression objective function. The additional term prevents over-fitting of the model by penalizing large values of the parameter estimates. Mathematically, given $i = [1, 2, \cdots N]$ observations of the dependent variable y_i and a set of covariates $x_{i1}, x_{i2} \cdots x_{iK}$ the LASSO predicts parameter values for $\beta_1, \beta_2 \cdots \beta_K$ by minimizing the following objective function:

$$Q(\beta_1, \beta_2, \cdots \beta_K) = \sum_{i=1}^{N} (y_i - \sum_{k=1}^{K} x_{ik}\beta_k)^2 + \lambda \sum_{k=1}^{K} |\beta_k| \tag{5}$$

where the first term is the standard linear regression loss function and the second term is the penalty for over-fitting. This form of the penalty term forces some of the parameter values β_k to zero, which makes it useful for the task of covariate selection given a large collection of possible covariates. λ can be interpreted as the weight given to the penalty term: when $\lambda = 0$ the estimated parameters are identical to linear regression, when $\lambda = \infty$ the estimated parameters will be equal to zero. In practice the value of λ is selected via a technique called cross-validation based on its ability to minimize the residual sum of squares plus lambda multiplied by the sum of absolute values of the coefficients.

While LASSO was originally developed within the linear regression setting, the technique has subsequently been extended to generalized linear models such as logistic regression. It is important to note that given the non-standard objective function, standard significance tests and confidence intervals do not hold for parameter estimates from LASSO.

We use LASSO techniques to analyze the association between skills and job market outcomes from the Round 2 data. As before, our outcomes of interest are 1) log of weekly wages, $\log(wages)$ and 2) whether the respondent received regular monthly wages, Emp. For both outcomes we consider 3 specifications. In the first baseline specification our covariates include all skills and demographic characteristics. In the second specification we add the heckman correction term, *IMR*, to the list of covariates where *IMR* is estimated as described in 4. In the final specification, the *IMR* term itself is obtained by a LASSO specification of the heckman equation that includes self-assessments of all skills. We implement LASSO regressions in `R` using the package `glmnet`[22].

V. Results

We present two sets of results in this paper. The first map the skills space of respondents in our sample. These provide insights into the distribution of skills within the population and particularly the disparity of skills across genders. We also present a mapping of how different skills are related to one another – both in terms of skill co-occurrence as well as skill hierarchy. The second set of results provide understanding regarding skill-based determinants of employment outcomes, in particular wages and regular monthly salaries. This set of results also points towards gender disparities as well as the relative benefits of skilling versus schooling in the informal labor market.

[22] We use the command `cv.glmnet` to estimate $\hat{\lambda}$ via crossvalidation and the command `glmnet` to fit the resulting LASSO model.

A. Mapping the Skills Space

A..1 Distribution of Basic Skills and Personality Traits

Figure 3 presents the mean self-reported assessments from respondents on the set of basic skills and personal qualities surveyed in Round 1 dis-aggregated by gender. The clear message from the basic skills figures (first column of Figure 3) is that across almost all basic skills, the mean self-reported assessments of males is significantly higher than females. This is also reflected in the negative correlations between dummies for female respondents and each basic skill. The disparity is most stark in the case of two-wheeler (less than 0.10 for females and almost 0.50 for males) and four-wheeler driving as well as internet usage (around 0.20 for females and 0.35 for males). The use of mobile phones has the highest mean assessment value of all basic skills – consistent with the recent proliferation of mobile phone ownership in India. In terms of language skills, when it comes to Hindi, the mean assessment for speaking the language (around 0.20 for females and 0.35 for males) is higher than reading and writing (around 0.10 for both females and males). English speaking scores, on the other hand, are *lower* than reading and writing. This implies that even though a number of individuals learn English as part of the school curriculum, they may not be confident speaking it in their daily lives. From the correlation plot we note that language skills are correlated with one another and with the use of mobile phones and internet, as expected.

Figure 3 about here.

While respondents assessed themselves low in basic skills questions, this was not the case when it came to questions about personality traits, as can be seen in the second column of Figure 3. Most respondents rated themselves very highly leading to mean values of 0.8 and above in most cases. Although females still tend to rate themselves lower than males on average, gender disparities are typically not significant in terms of self reported personal qualities. The two exceptions to this are the qualities of being 'easily trusting' and 'organized'. In both, the average female rates herself significantly lower than the average male. The only personal quality where the individuals rate themselves less than 0.8, on average, is the value of whether the respondent prefers a 'high effort, high pay job' (as opposed to a 'low effort, low pay job').

A..2 Distribution of Crowdsourced Skills

In Round 1 we also collected lists of skills and qualities respondents considered important for success. Respondents provided us with a rich taxonomy of skills, knowledge and qualities that affected employment outcomes according to them. Overall we recorded over 5000 entries, out of which over 1000 were unique. The large number of skills elicited is an important cultural finding in and of itself. Policy makers and economists tend to refer to informal workers in India as 'unskilled'. Our findings suggest that informal workers not only possess a large number of skills, however, but they are able to readily articulate these skills with, in many cases, a widely shared vocabulary. This is especially apparent when we consider some of the specific and nuanced responses elicited, which would have been impossible to have anticipated without discovery through an open-ended survey design. We report here some of these responses verbatim[23]. First, a set of personal qualities: 'tension-free', '(possessing a) happy face', 'taking initiative', '(making a good) self-introduction', 'ability to explain complex information clearly and simply', 'well behaved manner with higher persons'. Second, a set of skills that relate to unusual work environments: 'ability to work at heights', 'ability to work in unclean environments', 'ability to work independently at home'. Finally, our survey elicited highly specific skills: 'knowledge of removing skin from goats and beefs', 'demolition of houses', 'using granite polishing machine to polish granite and tiles', 'ability of knowing color coded wiring', 'shouting loudly to attract customers'.

We refer to all these critical inputs as skills going forward, categorizing them into 'language', 'task', 'specialized', 'personal' and 'social'. 'Language' skills like Hindi, English and Kannada speaking, writing and listening, as well as general 'task' skills including simple maths, measurement and traffic rules, are often the focus of general education and have a rich history of testing and evaluation. 'Specialized' task skills are linked with a specific occupation and are acquired through training and work experience. 'Personal' skills such as neatness, punctuality, honesty, stress management and dressing sense, as also 'social' skills like caring, friendliness and the ability to adjust one's attitude in a contentious interaction are associated with personal and social capacities trained over a lifetime of family instruction, community and work expe-

[23] The full list of skills is part of the supplementary material of the paper, available from the authors on request.

riences. These are often neglected as skills, and require nontrivial assessment over a variety of contexts that map onto the diverse settings in which they were exercised. This categorization is useful for exposition and respects the range of real capacities people bring to successful, sustainable employment. The most commonly occurring of these crowdsourced skills form the bulk of our second round of surveying, where respondents assessed themselves on each skill. Table 1 presents these crowdsourced skills, along with the mean self-assessment on each of these skills for female and male respondents in Round 2. The table also contains information on whether means across gender differ significantly. Skills within each category are presented in ascending order of mean self-assessment.

Table 1 about here.

As in Round 1, there is a large gender disparity in self-assessments in Round 2, particularly in the 'language', 'task' and 'specialized' skills. Part of these gender differentials can be explained by the gender-typing of occupations among informal workers – domestic help work is almost entirely handled by women, while transport and delivery services lie in the domain of men. As a result, skills associated with these occupations are also gendered. Consistent gender differences persist in favor of men with skills including 'driving' , 'riding a bike', 'traffic rules' and 'geographic knowledge', and in favor of women in skills such as 'cooking' , 'washing clothes', 'cleaning vessels' and 'housekeeping'. Nevertheless, there are also large gender gaps in basic task skills not associated with specific occupations, including 'using the internet', 'using mobile phones', 'bank work', 'general knowledge', and 'speaking in Hindi'. By contrast, most respondents rate themselves very highly on personal and social skills, for which evaluation is challenging and subtle: almost all respondents rate them selves highly in terms of their 'relationship with people' , 'work ethic', 'respectfulness' and 'honesty'.

A..3 Co-occurrence of Crowdsourced Skills

Our rich dataset of complete self-assessment data on over 100 skills by almost 800 respondents provides us with a unique opportunity to understand the distribution, co-occurrence and hierarchy of skills in the informal labour market using unsupervised learning methods. The first technique we use is k-means clustering (see Section A..2 for details on the algorithm). Table 2 presents the 7 optimal clusters obtained by the algorithm, in increasing order of number of skills per cluster. An informative way of analyzing the clusters is to see which skills co-exist within a cluster – and particularly if language, task, personal and social skills are clustered together with more specialized, occupation skills.

Table 2 about here.

Clustering also provides a good sanity check on the consistency and accuracy of respondent answers. Consider the skills collocated within cluster 2 – work in bathrooms, work in kitchens, cleaning house, washing clothes, ironing, cleaning vessels, cooking, housekeeping, and cutting meat are all associated with domestic work. The other skill in the cluster is ‘cleaning roads’, an activity for which local municipalities employ women. It is telling that no general language or task skills are associated with this cluster, confirming anecdotal evidence that domestic work is carried out by the most vulnerable (typically female) population. Similarly, Cluster 3 points towards respondents employed in the IT and IT enabled services sector as implied by the co-occurrence of skills in ‘internet usage’, ‘computer handling’ and ‘data entry’. Notably, this cluster also contains the important skills of ‘bank work’ as well as English language reading, writing and listening. Most interestingly, Kannada writing and reading are also part of the cluster because Kannada is the native language of the region, revealing that native Bangaloreans are more likely to be employed in the coveted IT sector.

The remaining clusters are not all as intuitive but surface interesting insights. Clusters 1 and 5 appear associated with sales work. They include skills linked with sellers and vendors such as ‘quality verification’, ‘inventory’, ‘packing’, ‘customer service’ and ‘sales’. A number of important basic skills also show up in these clusters such as ‘handling cash’, ‘simple maths’, ‘mobile phone usage’, ‘geographic knowledge’ and ‘tracking prices’. Critical social skills such as ‘management’, ‘supervising’, ‘teamwork’, and personal capacities like ’dressing sense’ and ’creativity’ also appear in these clusters. Cluster 7 brings most of the personal and social skills together, likely because respondents are less disciplined when assessing themselves as they link with valued identities and most respondents rate themselves very highly. Cluster 4 and 6 consist of language skills in Hindi, speaking skills in English, along with a number of specialized skills (from ‘driving’ and ‘working with tools’ to ‘stitching’ and ‘cement work’). These cluster likely corresponds to the population of respondents who can perform a large number of odd jobs and interact regularly with Bangalore’s large migrant non-Kannada speaking population of corporate and IT sector employees.

A..4 Hierarchy of Crowdsourced Skills

Hyperbolic embedding is an emerging technique in unsupervised learning that has gained popularity in part because it represents both the co-occurrence and hierarchy of concepts, in our case skills. In our application, skills that are more general, held in common across many workers and with a wider range of other skills, are placed closer to the center of the embedding (represented by lower radius values) and those that are less widely and diversely held are placed further from the center of the embedding (represented by higher radius values). What does a more or less general concept mean in the case of skills? Two types of skills will be found at the center of the embedding. First, those that are pathway skills – skills essential to acquiring other skills. Second, those that occur commonly – skills easily acquired or with an underlying sub-skill transferable from one specialized skill to another. From the analyst’s perspective, it is not possible to distinguish from the embedding whether a skill is more general because it is a gateway skill or an easily and widely acquired skill. With this caveat in mind, we analyze Figure 4.

Figure 4 about here.

Our first observation is that there is a large concentration of skills close to the center of the embedding – 27% of all skills are within the $r = 0.1$ disc and as many as 65% are within the $r = 0.2$ disc. If we consider our classification of basic, soft and specialized skills we note that, the concentration near the center is driven by soft and specialized skills. Soft skills are likely concentrated near the center because most respondents rated themselves highly on them. On the other hand we interpret the high concentration of specialized skills near the center to be their ease of acquisition due to an underlying set of transferable sub-skills.

Basic skills are more spread out in the embedding – implying that a number of skills which we may perceive as basic are rare within the population of our interest. Interestingly only 6 basic skills are contained in the $r = 0.1$ disc of the embedding: ‘English writing’, ‘English reading’, ‘English listening (understanding)’, ‘internet usage’, ‘following instructions’ and ‘general knowledge’. We know that the prevalence of English language skills and internet usage in our sample is comparatively low and so are not commonly occurring ones. Therefore we interpret these skills as being pathway skills which are useful for acquiring other skills. The $r = 0.1$ disc of the embedding contains 10 soft skills, of which 5 are social skills, namely: ‘customer service’, ‘communication’, ‘friendliness’, ‘relationship with people’ and ‘positive body language’. With the exception of ‘customer service’, all the other skills are extremely highly rated by respondents, therefore we refrain from interpreting these as gateway skills.

How do we think of the structure of the embedding generated by the crowdsourced skills list? It is true that informal workers are not ‘unskilled’, as is apparent from the large list of skills our respondents provide. However, our embedding results which show a high concentration near the center of the embedding implies that, at least the most commonly occurring specialized skills, are somewhat easy to acquire for our population of interest. We argue that there are likely some underlying sub-skills which are transferable across a number of specialized skills. For instance, we see that ‘ironing’ and ‘cement work’ are close together in the embedding, suggesting that both have some underlying characteristics in common. Similarly we see that ‘packing’ , ‘delivery’ and ‘handling heavy machines’ are close in the embedding – each of these require the ability to do physically demanding work with large packages making it easier to move between the specialized skills. Keeping these result in mind the authors therefore recommend the term ‘transferable’ or ‘portable’ skills to describe skills possessed by workers in the informal labor market.

B. Determinants of Wages and Regular Work

B..1 Round 1: Basic Skills and Personality Traits

Table 3 presents results of the outcome equation of the Mincerian wage estimation described in (2) from different specifications using the Round 1 survey data. Table 4 presents results of the estimation on whether the respondent receives regular monthly wages, corresponding to (3) using Round 1 survey data[24]. For ease of interpretation, the odds ratios of the coefficient estimates are presented in this table. Both tables include *IMR* terms from a first stage Heckman correction equation. The first column in both tables corresponds to the specification where neither the selection equation nor the outcome equation have any controls for basic skills and personality traits. Basic skills are included in the outcome equation specification for the remaining three columns. The selection equation in the second column does not include any personality traits. In the third column, personality traits are included in the selection equation specification. Finally in the fourth column personality traits are included in both selection and outcome equations.

Table 3 about here.

Estimates reported in the first column of Table 3 are along expected lines for the mincerian wage equation: log of weekly wages increases with number of years of education as well as experience. The effect of experience is slightly concave as evident from the small negative coefficient on the square of experience. Female workers start at a lower level of wages and are also subject to lower rates of wage increase with experience – this gender differential is along expected lines as well, and may be partly explained by gender specialization of occupations. Each of these estimates is significant at 95% level of confidence.

Interestingly the effect of education and gender are washed out once we control for basic skills as can be seen in the estimates reported in the last three columns. The strongest and most significant basic skill associated with higher log wages is ‘internet usage’. Next comes the effect of ‘English speaking’. ‘Hindi speaking’ also has a positive association with higher wages, even though the signal is weaker. In terms of personality traits, being ‘organized’, ‘sociable’ and ‘easily trusting’ are also associated with higher log wages. The R^2 values of the different specifications range from 28% to 36%. The adjusted R^2 value is highest for the fourth specification (with personality traits entering into both selection and outcome equations).

Table 4 about here.

[24] Selection equation results are available as supplementary material.

Estimates reported in the first column of Table 4 throw up some interesting insights: the odds of receiving regular monthly wages increase with the number of years of education and *decrease* with experience. The decreasing odds of receiving a regular salary with increasing experience seems counter-intuitive but may have a simple explanation. Experience here is most likely acting as a proxy for age, and this coefficient implies that younger workers are more likely to be working in jobs with regular wages. The interaction term for female and experience is positive implying that the odds of receiving a regular wage with increasing experience go up for female respondents. The square of experience is positively associated with regular wages, implying the increased odds of receiving regular wages after a number of years work experience. Hindu respondents have lower odds of receiving regular wages. Each of these estimates is significant at 95% levels of confidence. Unlike the previous set of results, the inclusion of skills and personality traits in the outcome equation does not render any of the original significant covariates insignificant. ‘English speaking’ has a very strong and positive associate with regular wages (odds of regular wages more than double with better English speaking abilities). ‘Hindi speaking’ on the other hand is associated with lower odds of regular wages. Being able to use the internet increases odds of a regular wage, however this signal is not very robust. Not surprisingly, a preference for ‘routine work’ is associated with a highly significant increase in odds of receiving a regular salary, as seen in the last column.

Overall the Round 1 results point towards gender disparities both in terms of levels and regularity of wages, albeit in opposite directions. They also point to the importance of education and experience (or age) in the determination of job market outcomes. The ability to speak in English and use the internet have positive effects on both levels and regularity of wages. In Round 1 the basic skills and personality traits questions are fewer in number and handpicked by the authors. Next we present results from Round 2 of the survey where we ask respondents to rate themselves on over 100 skills that were crowd-sourced from the Round 1 survey.

B..2 Round 2: Crowdsourced Skills

Table 5 presents skills selected by LASSO regression (along with estimated coefficient values) in a supervised learning problem using Round 2 data where the outcome of interest is log wages. Table 6 presents skills selected by LASSO regression (along with estimated odds ratios) where the outcome of interest is recipt of regular monthly wages. The first column in both tables presents results for a single stage model – ie *IMR* from a first stage Heckman selection model is not included in the LASSO estimation. The second and third columns present results where the *IMR* is part of the LASSO covariates. In the second column *IMR* is estimated using the standard Heckman regression (which does not include skills) whereas in the third column *IMR* itself is estimated using a LASSO regression which includes self-assessments on all the crowdsourced skills. We remind readers that LASSO regression results do not come with standard significance levels and confidence interval estimates. Rather LASSO regressions should be viewed as a model selection technique where the best predictors (along with associated coefficient estimates) of the outcome of interest are selected by the algorithm.

Table 5 about here.

In Table 5 we note that the different specifications lead to overlapping covariate selection indicating the robustness of our results. From the set of demographic characteristics, the only variable that is consistently selected is the dummy for female gender – indicating a strong negative association with log of weekly wages. Other demographic characteristics like number of years of education and experience do not get selected in any of the specifications. Out of the basic skills ‘English listening (understanding)’ and ‘computer handling’ are picked up as positive predictors of wages by all three specifications, whereas ‘internet usage’ and ‘bank work’ are more sensitive to model specification and also have smaller associated coefficients. Out of the soft skills, ‘supervising’ and ‘teamwork’ are selected as positive predictors of wages in all three specifications. Most of the predictive ability however lies with the specialized skills, which also provide insights into higher paying occupations. For instance ‘bike riding’, ‘knowledge of bike parts’, ‘knowledge of car parts’ and ‘geographic knowledge’ are picked up across all specifications, whereas ‘driving’ is picked up by two specifications. These skills are highly suggestive of the gains in wages associated with the recent growth in delivery services and ride-sharing services in Bangalore. Similarly ‘data entry’ is a specialized skill that, along with the basic skills of ‘computer handling’ and ‘internet usage’ correspond to the high-growth

IT sector in the city. Finally, 'washing clothes', a skill that two of the specifications pick up is the only skill which is *negatively* associated with wages.

Table 6 about here.

In Table 6 again we note that the different specifications lead to overlapping covariate selection. Here, we note that 'experience' (a proxy for age) leads to lower odds of having a regular monthly wage and the dummy variable for the female gender leads to higher odds of having a regular monthly wage. Both of these are consistent with our findings from Round 1. However, we note that number of years of education does not make an appearance in any of the model specifications in Round 2. Here too, in terms of basic skills, 'English listening (understanding)' and 'computer handling' are selected across specifications and are associated with higher odds of a regular wage. Interestingly there are a number of soft skills that are associated with much higher odds of a regular wage. In particular 'relationship with owners' is associated with about 4 times higher odds of regular wage. 'Punctuality' is another soft skill which more than doubles the odds of a regular wage. 'Following instructions' and 'dressing sense' are associated with about 50% higher odds of regular wages. These skills are selected across all 3 specifications. A number of specialized skills are also selected across specifications, indicating specific occupations that are more stable. Skills related to domestic work such as 'cleaning vessels', 'work in bathrooms' and 'cleaning houses' are associated with somewhat higher odds of a regular wage. Similarly 'cleaning roads' is associated with regular monthly wages, as is to expected since the city municipal corporation hires women for the job of picking up trash and cleaning roads in the city. On the other hand skills related to the construction sector such as 'cement work', 'painting walls', 'applying putty' and 'construction work' are associated with lower odds of a regular wage.

Overall the empirical results from Round 2 reiterate some results from Round 1. Women are more likely to receive regular monthly wages, however women's labor is associated with lower wages. Younger workers are more likely to be employed in jobs which provide regular monthly wages. English language skills and computer literacy skills consistently come up as having a positive association with both level and regularity of wages. The inclusion of crowdsourced skills leads to a number of novel insights. In particular 'relationship with owners' is a strong predictor of regular wages – suggesting the importance of network connections particularly with individuals heading the social hierarchy in cities. Other soft skills like 'punctuality' and 'dressing sense' also show up as important predictors of regular wages. Finally, the set of specialized skills shed light on occupation specific differences in levels and regularity of wages. Skills that service the IT sector, particularly data entry jobs, are associated with both higher and more regular wages. Skills that service the delivery and transport sectors, related to bikes and cars, are associated with higher wages. Domestic workers are more likely and construction workers are less likely to be regular wage earners.

VI. Summary

This paper uses primary data from Bangalore slums to study the labor supply side in India's informal labor market. The primary data is collected in two separate rounds. In Round 1 we collect information from 698 respondents on their demographic characteristics, employment history, self-assessments on a set of pre-specified questions and skills that respondents associate with better job market outcomes. In Round 2 we collect information from 784 new respondents on a parsimonious set of questions related to demographic characteristics and employment history, as well as detailed self-assessment data on the most commonly reported skills from Round 1. Our rich primary data allows us to analyze skill disparity, co-occurrence, and hierarchy as well as determinants of informal labor market outcomes from a quantitative lens. We employ both econometric and machine learning techniques for an improved understanding and mapping of skills held by informal workers in one of the fastest growing emerging economies.

Our study makes a number of novel contributions to the literature. First, despite the fact that the vast majority of Indian workers are informally employed, quantitative scholarship on informal labor markets in India is scare. While some authors utilize macroeconomic data to empirically study informal labor at the national level, detailed microeconomic analysis of informal labor is limited to a handful of papers which collect primary household level data. Our paper adds to this literature using primary data collected from the slums in Bangalore. Second, our paper also adds to the growing literature on the role of “soft” and general task skills in determining employment outcomes. The study of skills has been gaining prominence globally as a reaction to the growing ‘gig’ economy fueled by the proliferation of online platforms for job-matching and skill acquisition. Our paper contributes to this literature by placing the spotlight on a hitherto unexplored population – informal workers in urban India. Third, our study utilizes standard econometric analysis, crowdsourcing as well as machine learning methods. This mixed method approach allows us to utilize the structure of standard social science surveys as well as the flexibility of a bottom-up approach towards analyzing the importance of skills in an urban informal setting. Finally, whereas the majority of applied machine learning research uses secondary data sources or primary data collected over devices and online, this is among the first attempts to collect and analyze primary data in a fieldwork setting with a view towards machine learning applications.

Using data from Round 1 we estimate large and significant disparities between male and female workers across most basic skills. We do not observe significant gender differences in self reporting of personality traits. Mincerian wage estimation results using a heckman correction show significant gender effects on both level and regularity of wages – while female workers earn less than male workers, they are more likely to be in stable jobs and receive regular monthly wages. We also find that younger and more educated workers are more likely to receive regular wages. The ability to speak in English and use the internet have positive effects on both levels and regularity of wages. In fact the effect of education on wage levels stops being significant on controlling for basic skills.

Crowdsourcing in Round 1 is a critical component of our study because it provides us with a rich source of data grounded in everyday realities. This approach allows for the discovery of skills which the respondents know to be important, yet are not readily recognized by quantitative researchers. Out of the most commonly occurring crowdsourced skills, many turn out to be cultivated personal and social capacities including honesty, punctuality and interpersonal confidence - signalling the importance attached by respondents on these critical qualities. Another set of skills are specialized, providing us with a detailed menu of occupation-specific task skills. While the mean self assessments from female respondents is significantly lower than those from males for most skills, the crowdsourced list contains a small set of skills where mean self assessments of female respondents is significantly higher than males. In this way, a bottom-up approach can shed light on skills possessed by sub-populations otherwise under-represented. Many non-standard crowdsourced skills turned out to be important predictors of job market outcomes, such as ‘geographic knowledge’ and ‘relationship with owners’ that positively predict weekly wages and regular monthly wages, respectively.

Data from Round 2 provides insight on the clustering and hierarchy of skills, which augment our understanding of the skills space beyond traditional econometric techniques. For example, we find that no general task, personal or social skills are part of a cluster clearly representing domestic work (an occupation that is almost exclusively female). On the other hand, we discovery that the cluster representing IT sector work is also associated with native residents of Karnataka. The embedding results suggest that the most commonly occurring skills associated with informal work may be relatively inexpensive to acquire and substitute into or out of. Supervised learning using LASSO helps us acquire a more dis-aggregated view of the skills associated with desirable job outcomes.

A number of policy implications emerge from our study. The most striking results signal the significant gender disparity in both skill distribution and wages. The mean self-assessment by females on all but a few skills is significantly lower than males. The few skills that females rate themselves higher on are related to either domestic and care work or tailoring. Our clustering results show that domestic work skills are isolated from any general task, personal or social skills. The more concerning finding is the large and significant differences in general task and language skills that are increasingly essential for self-sufficiency. For instance: English and Hindi language skills and the use of mobile phones, banking services and the internet. A simple policy prerogative this recommends is the development of large scale financial and computer literacy programs targeted specifically at women.

We find a robust association between English language skills and computer/internet literacy skills with improved job outcomes. In fact, we find that the association of education with higher wages washes out when we control for English and internet usage skills. This result makes the case for our second policy implication: prioritizing English and computer literacy skills in public schools. Notably, the there is an emphasis on computer training, personal and social skills in India's skilling programmes such as the Deen Dayal Upadhyaya Grameen Kaushalya Yojana (DDU-GKY), which is a large scale employment scheme for rural youth.

Our study also reveals the importance of personal and social skills in the urban informal labor market – something that has not previously been studied in a quantitative context within India. The importance of non-task, personal and social capacities became apparent from the large proportion of crowdsourced responses that contained such skills. A number of these skills are associated with regular monthly wages – for instance: 'relationship with owners', 'punctuality' and 'dressing sense'. While the National Skills Development Corporation of India (NSDC) has formulated a set of National Occupation Standards, these are largely silent on personal and social skills, focusing instead on an exhaustive list of key task skills required for various job functions. As a third policy implication, our results strongly indicate that "soft" personal and social skills training should become an inherent part of any such occupation standards. This is particularly relevant for migrant workers who are mostly employed in the construction and service sector of a metropolis such as Bangalore. For the same reasons these skills have been difficult to measure, they may also be difficult to train outside the family, community and work experiences through which they have traditionally been acquired, but their economic importance warrants substantial research on these topics.

Finally, our study emphasizes the need for a new vocabulary when addressing skills for the informal labor market in India. Based on the large number of unique skills elicited from respondents, researchers and policy makers are cautioned against using the term 'unskilled' to describe informal workers. Informal workers are not 'unskilled' because our technology for skills measurement and representation has been limited. Informal workers are not 'unskilled' simply because our technology for skills measurement and representation has been limited. The term 'low skill' is also misleading as many skills elicited are highly specialized. Nevertheless, we find that most skills reside at comparable levels in the skills hierarchy, which signals high transferability and the potentially to acquire even specialized skills. Our findings contextualize the relevance of 'tacit knowledge' (Polyani (1983)), which may be embodied in apprenticeship as well as on the job training experience of informal workers, and which may not be separated from the skills that workers have reasons to value. And yet the very concept of 'tacit' skills highlights the lack of research and educational investments made in naming, training and valuing these critical skills. It is in this context that we propose a more nuanced notion of 'transferable' or 'portable' skills to describe skills possessed by informal workers. We also emphasize the diversity of general and specialized task skills, language, social and personal talents, and the importance for discovering them and their complex complementarities. There is not a single mass of 'soft' skills, but a high-dimensional continuous space of task, personal and social capabilities that, if better measured and strategically taught, could enable more rational training and staffing procedures and a punctuated burst in human welfare.

The study also has limitations. The chief caveat is that data on skills is self-reported. The collection of household level demographic and economic data via sample surveys is standard practice, however self-reporting on skills is likely to be noisy and self-aggrandizing. We expect and find that respondents tend to rate themselves positively in personal and social skills in particular. Fortunately we find large variation in assessments for a majority of the remaining skills implying substantial signal in the data. Gender differences in self-assessments may also be partly attributed to generally lower levels of confidence among women. We note that in the majority of personal and social skill assessments, however, mean female assessments were not significantly different from mean male assessments, somewhat allaying our concerns. Finally, our sampling frame is based on the list of declared slums in Bangalore area and is representative of workers residing there. While residents of declared slums are typically informally employed, they do not represent the entire population of informal workers in the city. Our study should not therefore be viewed as representative of all informal workers.

Future research directions include studying the demand side of informal labor from a skills perspective and understanding if skills mapping can offer insights into efficient matching of jobs by incorporating demand side preferences. Randomized control trials to measure the impact of specific skill interventions suggested by our results are another natural next step. Given the high dimensionality of the skills space, an important methodological extension is the use of adaptive surveys for skill mapping, which may be less sensitive to attrition. Further, an ideal application of research in this domain is the prediction of personalized skill acquisition paths instead of one-size-fits-all training modules currently being used for skills training in the Indian labor market.

References

Abraham, R. (2017). Informality in the indian labour market: An analysis of forms and determinants. *The Indian Journal of Labour Economics*, *60*(2), 191–215.

Arntz, M., Gregory, T., & Zierahn, U. (2016). The risk of automation for jobs in oecd countries.

Azam, M., Chin, A., & Prakash, N. (2013a). The returns to english-language skills in india. *Economic Development and Cultural Change*, *61*(2), 335–367.

Azam, M., Chin, A., & Prakash, N. (2013b). The returns to english-language skills in india. *Economic Development and Cultural Change*, *61*(2), 335–367.

Bairagya, I. (2012). Employment in india's informal sector: size, patterns, growth and determinants. *Journal of the Asia Pacific Economy*, *17*(4), 593–615.

Bairagya, I. (2018). Why is unemployment higher among the educated? *Economic & Political Weekly*, *53*(7), 43.

Bairagya, I., et al. (2015). *Socio-economic determinants of educated unemployment in india*. Institute for Social and Economic Change.

Banerjee, B. (1983). The role of the informal sector in the migration process: A test of probabilistic migration models and labour market segmentation for india. *Oxford Economic Papers*, *35*(3), 399–422.

Basole, A. (2014). The informal sector from a knowledge perspective. *Yojana*, *58*, 8–13.

Basole, A. (2016). Informality and flexible specialization: Apprenticeships and knowledge spillovers in an indian silk weaving cluster. *Development and Change*, *47*(1), 157–187.

Basole, A., & Basu, D. (2011). Relations of production and modes of surplus extraction in india: Part ii-'informal'industry. *Economic and Political Weekly*, 63–79.

Bhalla, S., & Kaur, R. (2011). Labour force participation of women in india: some facts, some queries.

Bhalotra, S. R. (1998). The puzzle of jobless growth in indian manufacturing. *Oxford bulletin of economics and statistics*, *60*(1), 5–32.

Binswanger-Mkhize, H. (2013). The stunted structural transformation of the indian economy:agriculture, manufacturing and the rural non-farm sector. *Economic & Political Weekly*, *48*, 5–13.

Borghans, L., Duckworth, A. L., Heckman, J. J., & ter Weel, B. (2008). Estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, *43*(4), 972–1059.

Börner, K., Scrivner, O., Gallant, M., Ma, S., Liu, X., Chewning, K., … Evans, J. A. (2018). Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy. *Proceedings of the National Academy of Sciences*, *115*(50), 12630–12637.

Bourse, M., Harzallah, M., Leclère, M., & Trichet, F. (2002). Commoncv: modeling the competencies underlying a curriculum vitae. In *Proceedings of the 14th international conference on software engineering and knowledge engineering (seke'2002)* (pp. 65–73).

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: a review of meta-analyses. *Clinical psychology review*, *26*(1), 17–31.

Chamberlain, B. P., Clough, J., & Deisenroth, M. P. (2017). Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*.

Chandramouli, C., & General, R. (2011). Census of india 2011. *Provisional Population Totals. New Delhi: Government of India*.

Cochran, W. G. (1963). *Sampling techniques.2nd ed.* John Wiley & Sons.

Cunha, F., & Heckman, J. (2007). The technology of skill formation. *American Economic Review*, *97*(2), 31–47.

Cunha, F., Heckman, J. J., & Schennach, S. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883–931.

David, H. (2015). Why are there still so many jobs? the history and future of workplace automation. *Journal of economic perspectives*, *29*(3), 3–30.

Deaton, A., & Grosh, M. (n.d.). *Designing household survey questionnaires for developing countries: Lessons from ten years of lsms experience.*

Deming, D. J. (2017). The growing importance of social skills in the labor market. *The Quarterly Journal of Economics*, *132*(4), 1593–1640.

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.

Filmer, D., & King, E. (1999). *Gender disparity in south asia: comparisons between and within countries*. The World Bank.

for Enterprises in the Unorganised Sector, I. N. C., & Academic Foundation (New Delhi, I. (2008). *Report on conditions of work and promotion of livelihoods in the unorganised sector*. Academic Foundation.

Goel, D., & Deshpande, A. (2016). Identity, perceptions and institutions: Caste differences in earnings from self-employment in india.

Günther, I., & Launov, A. (2012). Informal employment in developing countries: Opportunity or last resort? *Journal of development economics*, *97*(1), 88–98.

Gurtoo, A., & Williams, C. C. (2009). Entrepreneurship and the informal sector: some lessons from india. *The International Journal of Entrepreneurship and Innovation*, *10*(1), 55–62.

Harriss-White, B. (2009). Globalization, the financial crisis and petty production in india's socially regulated informal economy. *Globalization and Labour in China and India*, *12*, 131–150.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the econometric society*, 153–161.

Heckman, J. J., & Kautz, T. (2012). Hard evidence on soft skills. *Labour economics*, *19*(4), 451–464.

Heckman, J. J., & Rubinstein, Y. (2001). The importance of noncognitive skills: Lessons from the ged testing program. *American Economic Review*, *91*(2), 145–149.

Helmers, C., & Patnam, M. (2011). The formation and evolution of childhood skill acquisition: Evidence from india. *Journal of Development Economics*, *95*(2), 252–266.

Himanshu. (2011). Employment trends in india: A re-examination. *Economic and Political Weekly*, 43–59.

Hofmann, S. G., & Smits, J. A. (2008). Cognitive-behavioral therapy for adult anxiety disorders: a meta-analysis of randomized placebo-controlled trials. *The Journal of clinical psychiatry*, *69*(4), 621.

Justino, P. (2007). Social security in developing countries: Myth or necessity? evidence from india. *Journal of International Development: The Journal of the Development Studies Association*, *19*(3), 367–382.

Kamath, A. (2014). *Industrial innovation, networks, and economic development: Informal information sharing in low-technology clusters in india*. Routledge.

Kannan, K., & Raveendran, G. (2009). Growth sans employment: A quarter century of jobless growth in india's organised manufacturing. *Economic and Political weekly*, 80–91.

Krishna, A. (2013). Stuck in place: Investigating social mobility in 14 bangalore slums. *Journal of Development Studies*, *49*, 1010-1028.

Krishna, A., Sriram, M., & Prakash, P. (2014). Slum types and adaptation strategies: identifying policy-relevant differences in bangalore. *Environment and Urbanization*, *49*, 568-585.

Krishnan, P., & Krutikova, S. (2013). Non-cognitive skill formation in poor neighbourhoods of urban india. *Labour Economics*, *24*, 68–85.

Krueger, A. B., & Schkade, D. (2009). Sorting in the labor market: Do gregarious workers flock to interactive jobs? *Journal of Human Resources*, *43*, 859–883.

Kuhn, P., & Weinberger, C. (2005). Leadership skills and wages. *Journal of Labor Economics*, *23*(3), 395–436.

La Porta, R., & Shleifer, A. (2014). Informality and development. *Journal of Economic Perspectives*, *28*(3), 109–26.

Lex Borghans, B. t. W., & Weinberg, B. A. (2009). Interpersonal styles and labor market outcomes. *Journal of Human Resources*, *43*, 815–858.

Maiti, D. (2008). The organisational morphology of rural industries and its dynamics in liberalised india: a study of west bengal. *Cambridge Journal of Economics*, *32*(4), 577–591.

Maiti, D., & Sen, K. (2010). The informal sector in india: A means of exploitation or accumulation? *Journal of South Asian Development*, *5*(1), 1–13.

Marjit, S., & Kar, S. (2004). Pro-market reform and informal wage: Theory and the contemporary indian perspective. *India Macroeconomics Annual*, *5*, 130–156.

Marjit, S., & Kar, S. (2009). A contemporary perspective on the informal labour market:

theory, policy and the indian experience. *Economic and Political weekly*, 60–71.

Mehrotra, S., Kalaiyarasan, A., Kumra, N., & Raman, K. R. (2015). Vocational training in india and the duality principle: A case for evidence-based reform. *Prospects*, *45*(2), 259–273.

Mehrotra, S., Parida, J., Sinha, S., & Gandhi, A. (2014). Explaining employment trends in the indian economy: 1993-94 to 2011-12. *Economic and Political Weekly*, *49*(32), 49–57.

Mehrotra, S., & Parida, J. K. (2017). Why is the labour force participation of women declining in india? *World Development*, *98*, 360–380.

Mincer, J. (1974). Schooling, experience, and earnings. human behavior & social institutions no. 2.

Murnane, R. J., & Levy, F. (1996). *Teaching the new basic skills: Principles for educating children to thrive in a changing economy*. Free Press.

Nair, J. (2005). *The promise of the metropolis: Bangalore's twentieth century*. Oxford University Press.

Nickel, M., & Kiela, D. (2017). Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems* (pp. 6338–6347).

Peter-Cookey M.A., K., & Janyam. (2017). Reaping just what is sown: Low-skills and low-productivity of informal economy workers and the skill acquisition process in developing countries. *International Journal of Educational Development*, *56*(56), 11–27.

Pilz, M., Gengaiah, U., & Venkatram, R. (2019). Skills development in the informal sector in india: The case of street food vendors. *International Review of Education*, *61*, 10.1007/s11159-015-9485-x.

Polyani, M. (1983). *The tacit dimension*. University of Chicago.

Prasad, S., et al. (2017). Report of the committee for rationalization & optimization of the functioning of the sector skill councils. *New Delhi: MSDE*.

Roy, D., Palavalli, B., Menon, N., Pfeffer, K., Lees, M., & Sloot, P. M. (2018). Survey-based socio-economic data from slums in bangalore, india. *Scientific Data*, *5*, 10.1038/sdata.2017.200.

Ruthven, O. (2008). Metals & morals in moradabad. *Perspective on ethics in the*.

Sahoo, B. K., & Neog, B. J. (2017). Heterogeneity and participation in informal employment among non-cultivator workers in india. *International Review of Applied Economics*, *31*(4), 437–467.

Sengupta, A. K., Srivastava, R., Kannan, K., Malhotra, V., Yugandhar, B., & Papola, T. (2009). The challenge of employment in india: an informal economy perspective. *Report of the National Commission for Enterprises in the Unorganised Sector, Government of India, Volumes*, *1*.

Shonchoy, A. S., & Junankar, P. R. (2014). The informal labour market in india: transitory or permanent employment for migrants? In *Development economics* (pp. 173–202). Springer.

Thomas, J. J. (2014). The demographic challenge and employment growth in india. *Economic and Political Weekly*, *49*(6), 15–17.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*(2), 411–423.

Verick, S. (2018). Female labor force participation and development. *IZA World of Labor*.

Weinberger, C. J. (2014). The increasing complementarity between cognitive and social skills. *Review of Economics and Statistics*, *96*(4), 849–861.

Figures and Tables

Figure 1: Bangalore has 15 town areas which served as the first stage sampling frame in the multistage sampling strategy utilized for the analysis. The purple (starred) and green (drop shaped) pins in the map (generated using Google My Maps application) indicate the 5 town areas randomly selected in Round 1 and Round 2 respectively. In the second stage 2 slums each were selected randomly from each of the 5 town areas, leading to 10 (non-overlapping) slums being surveyed in each round.

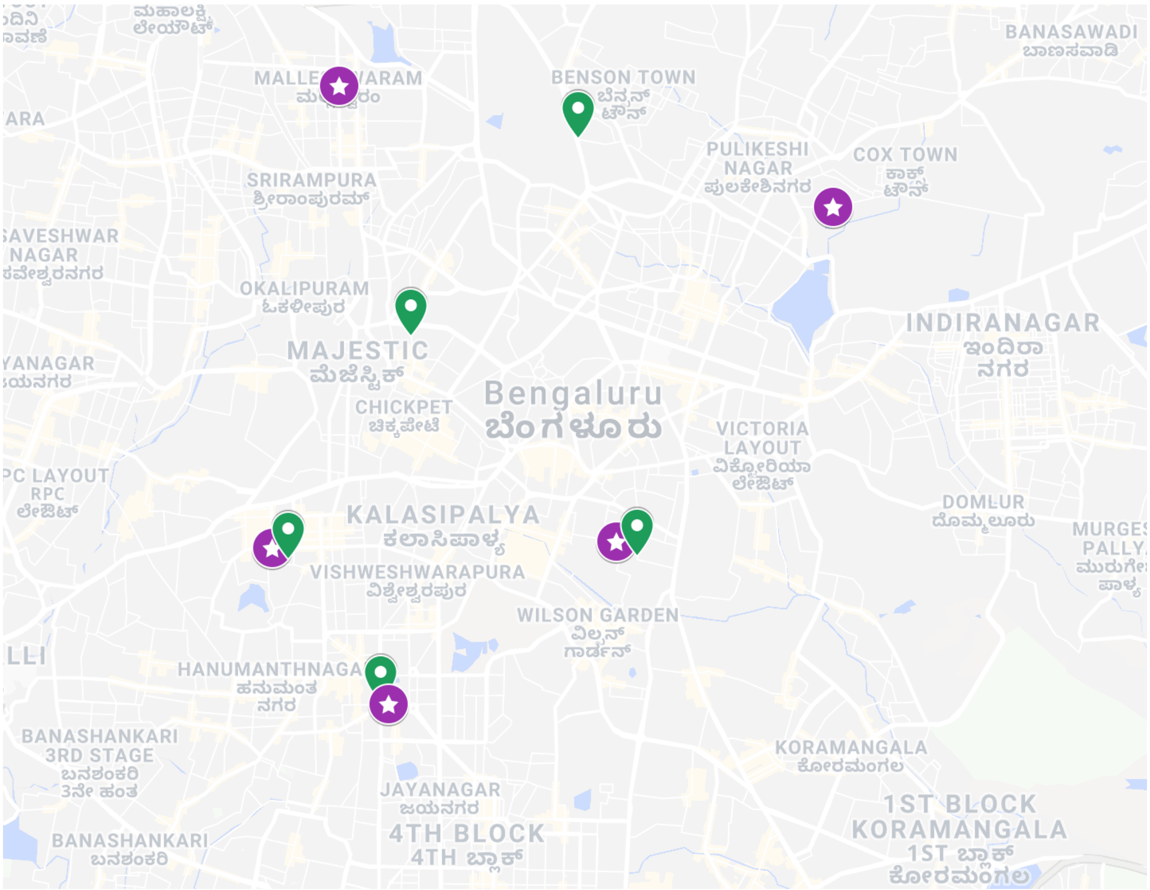


Figure 2: Respondent characteristics from Round 1 of the survey provides us with confirmation that the vast majority of respondents are informal workers. Male and female occupation distributions are presented in (a) and (b) respectively. Methods for job search are presented in (c). The pie chart in (d) is the answer to whether respondents who are currently employed reported having signed a written contract for their jobs.

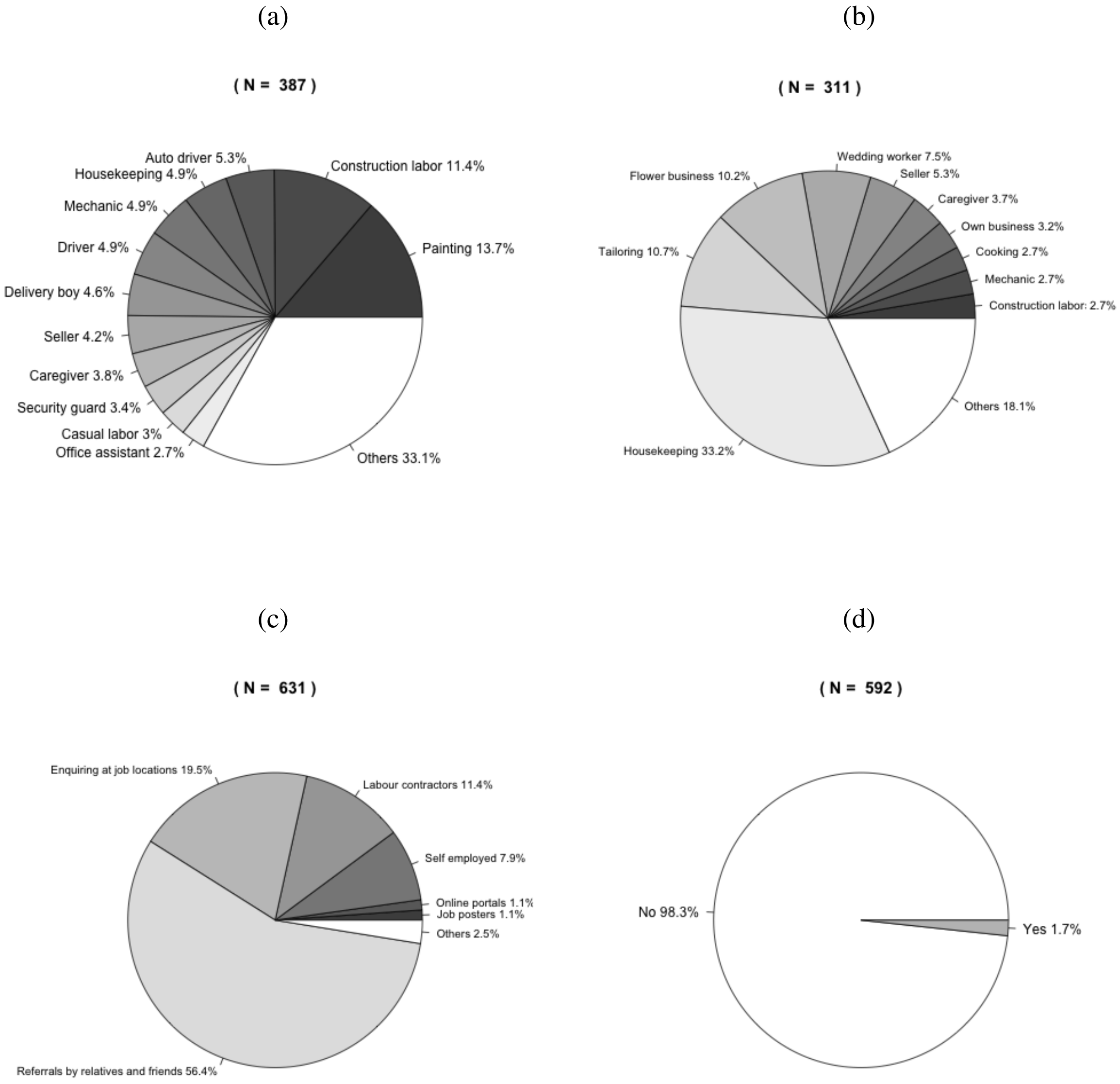
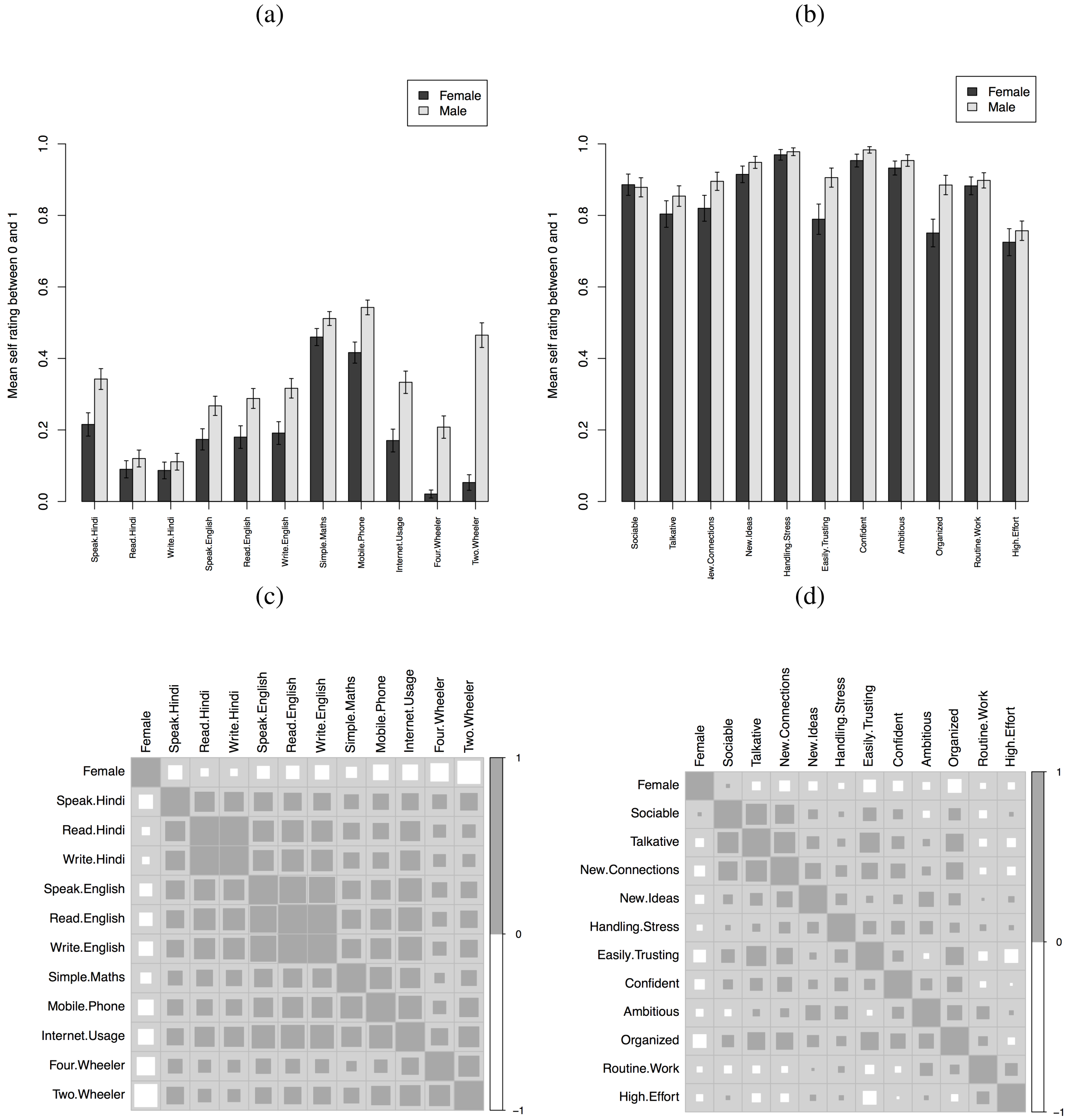


Figure 3: Mean self-reported assessments of respondents on the set of basic skills and personality traits from Round 1 dis-aggregated by gender (along with 95 % error bars) are presented in (a) and (b) respectively. While there is clear gender disparity in self-assessment of basic skills with female mean assessments being significantly lower than males, this was not the case when it came to questions about personality traits, where most respondents rated themselves very highly irrespective of gender. Correlation plots between the different basic and the different personality traits, (along with a dummy variable for females) are presented in (c) and (d).



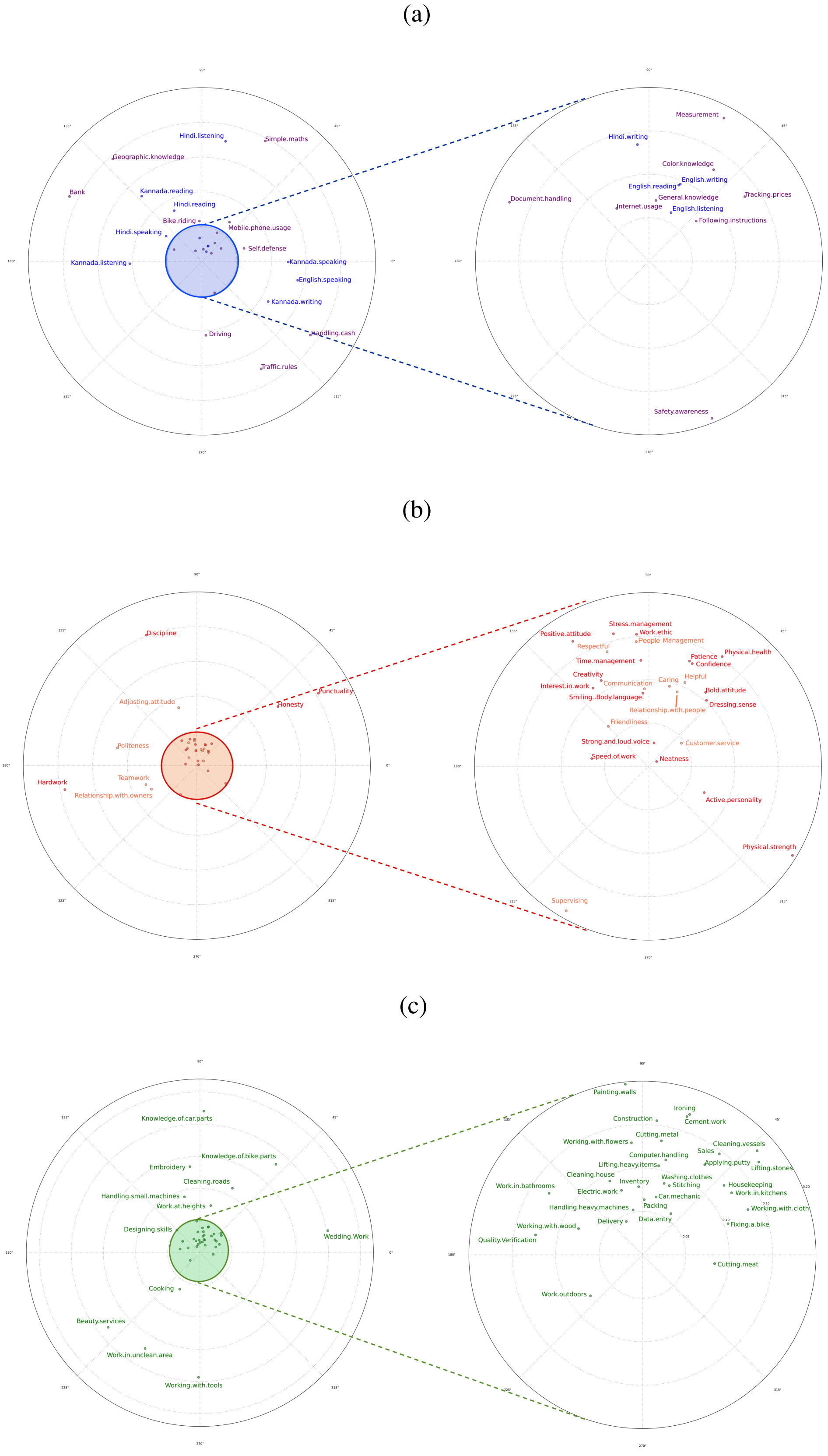
(a)

(b)

(c)

(d)

Figure 4: The hyperbolic embedding of skills presented here provide both a hierarchy of skills along with their co-occurence. Skills which are more general are placed closer to the center of the embedding. Our embedding, disaggregated by (a) 'basic', (b) 'soft' and (c) 'specialized' skills shows a high concentration near the center of the embedding implying that there are likely some underlying sub-skills which are transferable across a number of specialized skills in particular. Basic skills are more spread out in the embedding – implying that a number of skills which we may perceive as basic are rare within the population of our interest


(a)

(b)

(c)

Table 1: Mean self-assessments on skills in Round 2, disaggregated by gender are presented. Most respondents rate themselves very highly on 'soft' skills. Large gender disparity are observed in 'basic' and 'specialized' skills. The handful of skills where females' mean self-assessment is significantly higher than males are highlighted in bold.

Skill Category	Skill Description	$\overline{F}$	$\overline{M}$	$\overline{F}$ - $\overline{M}$
Basic	Hindi writing	0.083	0.077	0.006
	Hindi reading	0.080	0.080	0.000
	English speaking	0.120	0.169	-0.049 ***
	Driving	0.014	0.401	-0.387 ***
	English reading	0.222	0.341	-0.119 ***
	English writing	0.241	0.350	-0.109 ***
	English listening	0.244	0.350	-0.106 ***
	Internet usage	0.186	0.445	-0.259 ***
	Kannada writing	0.297	0.417	-0.120 ***
	Bank	0.256	0.456	-0.200 ***
	Kannada reading	0.325	0.463	-0.138 ***
	Hindi speaking	0.295	0.511	-0.216 ***
	Hindi listening	0.345	0.563	-0.218 ***
	Bike riding	0.079	0.779	-0.700 ***
	General knowledge	0.347	0.565	-0.218 ***
	Measurement	0.440	0.602	-0.162 ***
	Simple maths	0.567	0.667	-0.100 ***
	Geographic knowledge	0.444	0.779	-0.335 ***
	Tracking prices	0.670	0.771	-0.101 ***
	Handling cash	0.678	0.821	-0.143 ***
	Traffic rules	0.586	0.917	-0.331 ***
	Mobile phone usage	0.659	0.867	-0.208 ***
	Color knowledge	0.751	0.815	-0.064 ***
	Following instructions	0.799	0.800	-0.001
	Document handling	0.827	0.907	-0.080 ***
	Self defense	0.867	0.942	-0.075 ***
	Kannada speaking	0.897	0.942	-0.045 ***
	Kannada listening	0.923	0.966	-0.043 ***
	Safety awareness	0.954	0.967	-0.013
Soft	Management	0.289	0.351	-0.062 **
	Customer service	0.448	0.515	-0.067 **
	Creativity	0.476	0.601	-0.125 ***
	Supervising	0.549	0.731	-0.182 ***
	Teamwork	0.695	0.837	-0.142 ***
	Dressing sense	0.756	0.815	-0.059 **
	Confidence	0.857	0.838	0.019
	Adjusting attitude	0.861	0.861	0.000
	Active personality	0.861	0.899	-0.038 *
	Strong and loud.voice	0.891	0.904	-0.013
	Punctuality	0.884	0.930	-0.046 ***
	Neatness	0.901	0.923	-0.022
	Time management	0.912	0.936	-0.024
	Politeness	0.923	0.958	-0.035 **
	Physical health	0.943	0.955	-0.012
	Smiling Body language.	0.947	0.955	-0.008
	Physical strength	0.953	0.966	-0.013
	Stress management	0.948	0.972	-0.024 **
	Relationship with owners	0.960	0.964	-0.004
	Hardwork	0.953	0.978	-0.025 ***
	Helpful	0.973	0.969	0.004

··· continued

Soft	Patience	0.964	0.976	-0.012
	Speed of work	0.963	0.977	-0.014
	Positive attitude	0.970	0.975	-0.005
	Discipline	0.976	0.979	-0.003
	Communication	0.981	0.984	-0.003
	Friendliness	0.984	0.990	-0.006
	Bold attitude	0.987	0.987	0.000
	Caring	0.997	0.981	**0.016** ***
	Interest in work	0.989	0.992	-0.003
	Relationship with people	0.994	0.989	0.005
	Work ethic	0.990	0.993	-0.003
	Respectful	0.996	0.997	-0.001
	Honesty	0.997	0.999	-0.002
Specialized	Lifting.stones	0.007	0.023	-0.016 **
	Car mechanic	0.020	0.049	-0.029 ***
	Handling heavy machines	0.014	0.056	-0.042 ***
	Embroidery	0.099	0.021	**0.078** ***
	Applying putty	0.013	0.089	-0.076 ***
	Work at eights	0.013	0.091	-0.078 ***
	Lifting heavy items	0.017	0.093	-0.076 ***
	Fixing a bike	0.007	0.112	-0.105 ***
	Designing skills	0.063	0.080	-0.017
	Painting walls	0.030	0.128	-0.098 ***
	Working with wood	0.013	0.143	-0.130 ***
	Wedding work	0.138	0.077	**0.061** ***
	Electric work	0.007	0.182	-0.175 ***
	Knowledge of car parts	0.011	0.194	-0.183 ***
	Construction	0.029	0.203	-0.174 ***
	Stitching	0.268	0.037	**0.231** ***
	Cutting metal	0.059	0.213	-0.154 ***
	Working with flowers	0.261	0.083	**0.178** ***
	Computer handling	0.136	0.208	-0.072 ***
	Data entry	0.145	0.244	-0.099 ***
	Delivery	0.072	0.329	-0.257 ***
	Working with cloth	0.304	0.151	**0.153** ***
	Handling small machines	0.242	0.303	-0.061 **
	Knowledge of bike parts	0.034	0.483	-0.449 ***
	Sales	0.288	0.314	-0.026
	Cement work	0.181	0.416	-0.235 ***
	Work in unclean area	0.309	0.437	-0.128 ***
	Beauty services	0.309	0.554	-0.245 ***
	Cleaning roads	0.560	0.377	**0.183** ***
	Work in bathrooms	0.712	0.365	**0.347** ***
	Working with tools	0.285	0.722	-0.437 ***
	Packing	0.562	0.580	-0.018
	Cutting meat	0.688	0.482	**0.206** ***
	Cooking	0.956	0.370	**0.586** ***
	Ironing	0.751	0.610	**0.141** ***
	Inventory	0.645	0.718	-0.073 ***
	Work in kitchens	0.973	0.492	**0.481** ***
	Housekeeping	0.958	0.522	**0.436** ***
	Cleaning vessels	0.960	0.533	**0.427** ***
	Washing clothes	0.990	0.530	**0.460** ***
	Quality Verification	0.692	0.849	-0.157 ***
	Cleaning house	0.981	0.697	**0.284** ***
	Work outdoors	0.885	0.970	-0.085 ***
Note:		*p$<$0.1; **p$<$0.05; ***p$<$0.01		

Table 2: The 7 optimal clusters obtained by the k-means algorithm, in increasing order of number of skills per cluster. Cluster 2 provides us a good sanity check – the skills contained in it are all associated with domestic work and no basic skills are associated with this cluster – confirming anecdotal evidence that such work is carried out by the most vulnerable (typically female) population. Cluster 3 points towards respondents employed in the IT sector. This cluster also contains the important skills of 'bank work' as well as English reading, writing and listening. Native Bangaloreans are more likely to be employed in this sector implied by the presence of Kannada reading and writing skills in this cluster.

Cluster 1	Packing, Management, Customer service, Sales, Creativity, Cleaning roads
Cluster 2	Cutting meat, Work in bathrooms, Work in kitchens, Cleaning house, Washing clothes, Ironing, Cleaning vessels, Cooking, Housekeeping
Cluster 3	Bank, Internet usage, English listening, English reading, English writing, Kannada reading, Kannada writing, Computer handling, Data entry
Cluster 4	Measurement, General knowledge, Hindi listening, Hindi speaking, Beauty services, Working with tools, Work in unclean area, Bike riding, Knowledge of bike parts, Driving
Cluster 5	Handling cash, Simple maths, Mobile phone usage, Geographic knowledge, Tracking prices, Traffic rules, Color knowledge, Quality Verification, Inventory, Supervising, Dressing sense, Teamwork
Cluster 6	English speaking, Hindi reading, Hindi writing, Designing skills, Stitching, Embroidery, Working with flowers, Working with cloth, Working with wood, Handling small machines, Handling heavy machines, Cutting metal, Lifting heavy items, Lifting stones, Cement work, Construction, Work at heights, Wedding Work, Fixing a bike, Knowledge of car parts, Car mechanic, Delivery, Electric work, Applying putty, Painting walls
Cluster 7	Document handling, Safety awareness, Kannada listening, Kannada speaking, Work outdoors, Following instructions, Self defense, Caring, Smiling Body language , Communication, Confidence, Friendliness, Helpful, Neatness, Patience, Physical health, Physical strength, Hardwork, Positive attitude, Punctuality, Politeness, Honesty, Respectful, Discipline, Adjusting attitude, Strong and loud voice, Bold attitude, Active personality, Relationship with people, Relationship with owners, Speed of work, Interest in work, Stress management, Time management, Work ethic

Table 3: Results of the outcome equation of the Mincerian wage equation with Heckman correction corresponding to (2) from different specifications using the Round 1 survey data.

	Dependent variable: log of weekly wages			
	(1)	(2)	(3)	(4)
Years of education	0.023*** (0.005)	0.007 (0.006)	0.008 (0.006)	0.005 (0.006)
Years of experience	0.014* (0.007)	0.022*** (0.007)	0.020** (0.006)	0.021** (0.006)
(Years of experience)^2	−0.0004** (0.0001)	−0.0005*** (0.0001)	−0.0005** (0.0001)	−0.0005*** (0.0001)
Whether female	−0.357*** (0.077)	−0.150$^{\cdot}$ (0.083)	−0.112 (0.081)	−0.132 (0.082)
(Years of experience)(Whether female)	−0.009* (0.004)	−0.013*** (0.004)	−0.014*** (0.004)	−0.012** (0.004)
Whether SC	0.007 (0.045)	0.002 (0.044)	0.008 (0.043)	−0.039 (0.044)
Whether ST	0.140 (0.125)	0.137 (0.122)	0.180 (0.121)	0.107 (0.121)
Whether Hindu	−0.069 (0.051)	−0.034 (0.051)	−0.029 (0.051)	0.005 (0.051)
Received some training	−0.031 (0.076)	−0.064 (0.075)	−0.015 (0.071)	−0.005 (0.071)
Length of training	0.0001 (0.0002)	0.00001 (0.0002)	0.00001 (0.0002)	0.00001 (0.0002)
Hindi speaking		0.065$^{\cdot}$ (0.038)	0.072$^{\cdot}$ (0.038)	0.079* (0.039)
English speaking		0.124* (0.058)	0.119* (0.058)	0.098$^{\cdot}$ (0.058)
Simple maths		−0.099$^{\cdot}$ (0.056)	−0.103$^{\cdot}$ (0.056)	−0.087 (0.057)
Internet usage		0.174*** (0.049)	0.177*** (0.049)	0.143** (0.050)
Two-Wheeler driving		0.055 (0.039)	0.052 (0.039)	0.068$^{\cdot}$ (0.039)
Sociable				0.094* (0.044)
Easily trusting				0.075* (0.037)
Organized				0.080* (0.041)
Inverse Mills Ratio	0.345 (0.276)	0.259 (0.270)	−0.211 (0.181)	0.061 (0.205)
Constant	7.747*** (0.101)	7.584*** (0.115)	7.595*** (0.115)	7.055*** (0.260)
Outcome Equation: Basic Skills Controls	No	Yes	Yes	Yes
Outcome Equation: Personality Trait Controls	No	No	No	Yes
Outcome Equation: Other Demographic Controls	Yes	Yes	Yes	Yes
Selection Equation: Personality Trait Controls	No	No	Yes	Yes
Observations	634	634	634	634
R^2	0.275	0.336	0.336	0.363
Adjusted R^2	0.262	0.314	0.314	0.332

Note: $^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

Table 4: Results of the estimation on whether the respondent receives regular monthly wages with Heckman correction corresponding to (3) using Round 1 data. For ease of interpretation, the odds ratios of the coefficient estimates are presented in this table.

	Dependent variable: whether regular monthly wage			
	(1)	(2)	(3)	(4)
Years of education	1.110*** (0.029)	1.065* (0.033)	1.067* (0.033)	1.071* (0.034)
Years of experience	0.849*** (0.027)	0.869*** (0.029)	0.866*** (0.028)	0.861*** (0.029)
(Years of experience)^2	1.003*** (0.001)	1.002*** (0.001)	1.003*** (0.001)	1.003*** (0.001)
Whether female	0.929 (0.336)	1.044 (0.441)	1.105 (0.460)	1.122 (0.485)
(Years of experience)(Whether female)	1.077*** (0.020)	1.073*** (0.021)	1.071*** (0.021)	1.076*** (0.022)
SC	1.402 (0.294)	1.258 (0.273)	1.288 (0.278)	1.213 (0.273)
ST	2.860$^{\cdot}$ (1.806)	2.452 (1.571)	2.640 (1.673)	2.237 (1.419)
Hindu	0.579* (0.139)	0.577* (0.146)	0.569* (0.145)	0.593* (0.155)
Received some training	0.801 (0.312)	0.843 (0.341)	0.944 (0.361)	0.926 (0.364)
Length of training	0.999 (0.001)	0.998 (0.001)	0.998 (0.001)	0.998 (0.001)
Hindi speaking		0.700$^{\cdot}$ (0.134)	0.682* (0.133)	0.647* (0.132)
English speaking		2.315** (0.671)	2.367** (0.686)	2.149* (0.645)
Internet usage		1.501 (0.372)	1.505$^{\cdot}$ (0.373)	1.662* (0.426)
Prefers routine work				2.530*** (0.647)
Inverse Mills Ratio	6.866 (9.500)	8.215 (11.791)	3.413 (3.039)	3.552 (3.769)
Constant	1.662 (0.790)	1.421 (0.818)	1.534 (0.881)	0.423 (0.556)
Outcome Equation: Basic Skills Controls	No	Yes	Yes	Yes
Outcome Equation: Personality Trait Controls	No	No	No	Yes
Outcome Equation: Other Demographic Controls	Yes	Yes	Yes	Yes
Selection Equation: Personality Trait Controls	No	No	Yes	Yes
Observations	630	630	630	630
Log Likelihood	−368.862	−358.147	−358.296	−348.095
Note:		$^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001		

Table 5: Variables selected by LASSO regression (along with estimated coefficient values) in a supervised learning problem using Round 2 data where the outcome of interest is log wages. The first column in corresponds to a single stage model. The second column includes results from a model with *IMR* estimated using the standard Heckman regression (which does not include skills). In the third column *IMR* itself is estimated using a LASSO regression including all skills.

	Variables Selected:	Dependent variable: log of weekly wages		
		(1)	(2)	(3)
(A)	Whether female	-0.3645	-0.3551	-0.3649
	English listening	0.0246	0.0094	0.0323
	Knowledge of car parts	0.0306	0.0107	0.0358
	Knowledge of bike parts	0.0449	0.0389	0.0500
	Teamwork	0.0391	0.0095	0.0521
	Data entry	0.0688	0.0653	0.0698
	Bike riding	0.0742	0.0792	0.0699
	Computer handling	0.0679	0.0443	0.0787
	Supervising	0.1072	0.0899	0.1133
	Geographic knowledge	0.1959	0.2053	0.1923
(B)	Washing clothes	-0.0092		-0.0213
	Driving	0.0131		0.0213
	Bank			0.0004
	Internet usage		0.0006	
	Observations	727	727	727
	Selection Equation	No	Yes,	Yes, LASSO

Table 6: Variables selected by LASSO regression (along with estimated odds ratios) where the outcome of interest is receipt of regular monthly wages. The first column in corresponds to a single stage model. The second column includes results from a model with *IMR* estimated using the standard Heckman regression (which does not include skills). In the third column *IMR* itself is estimated using a LASSO regression including all skills.

	Variables Selected:	Dependent variable:		
		whether regular monthly wage		
		(1)	(2)	(3)
(A)	Cement work	0.5202	0.5281	0.4973
	Applying putty	0.8685	0.9242	0.8034
	Painting walls	0.8761	0.9243	0.8124
	Construction	0.9563	0.9893	0.9536
	Years of experience	0.9843	0.9851	0.9856
	Cleaning house	1.0380	1.0132	1.0452
	Cleaning vessels	1.1874	1.1841	1.1851
	Work in bathrooms	1.2386	1.2452	1.2300
	English listening	1.1498	1.1100	1.2396
	Cleaning roads	1.4495	1.4194	1.4436
	Computer handling	1.3751	1.3257	1.4490
	Dressing sense	1.4049	1.3797	1.5188
	Following instructions	1.5856	1.5598	1.6061
	Data entry	1.8546	1.8651	1.8781
	Whether female	1.9521	1.8914	2.1602
	Punctuality	2.9875	2.8140	2.5881
	Relationship with owners	4.3082	3.8343	4.8146
(B)	Lifting heavy items	1.1053		1.2335
	Inverse Mill's Ratio			0.1538
	Caring			0.8276
	Confidence			1.0193
	Time management			1.1332
	Observations	727	727	727
	Selection Equation	No	Yes,	Yes, LASSO

Appendix

Table A1: Distribution of the working age population by activity

	Sector of Activity		
Activity	All (%)	Formal Sector (%)	Informal Sector (%)
Own Account Worker	13.80	2.87	43.65
Self employed (Employer)	1.00	0.86	3.05
Unpaid Family Worker	4.76	3.15	12.94
Regular Salaried or Wage Employee	27.06	86.82	25.63
Casual Wage Labour (Public Work)	0.00	0.00	0.00
Casual Wage Labour (Other Work)	6.56	6.30	14.72
Actively Seeking Work	1.47		
Attended Educational Institution	13.06		
Attended Domestic Duties	26.93		
Other Domestic Duties	3.55		
Rentiers, Pensioners etc.	0.54		
Unable to Work due to Disability	1.07		
Others (Beggars, Prostitution etc.)	0.20		
Total	100	100	100
Observations	1493	349	394

Estimates are based on unit level data from NSS 68th round survey on Employment and Unemployment in India. Activities are based on Usual Principal Activity Status (UPAS).

Table A2: First stage sampling framework of town area in urban Bangalore district

Sl.No	Town Areas (Ist Stage)	Population
1	Gandhinagar	7214
2	Chikkapete	7193
3	Binnypete	9105
4	Chamrajpete	26829
5	Shanthinagar	7095
6	Basavangudi	5985
7	Yelahanka	19596
8	Jayamahal	13111
9	Malleswaram	25833
10	Bharathinagar	11448
11	Shivajinagar	458
12	Jayanagar	42723
13	Rajajinagar	13677
14	Varthur	26646
15	Uttarhali	89624
	Total	306537

Table A3: Round 1 sampling frame

Sl.No	Selected Town Areas (IInd Stage)	Population	Selected Slums (IIIrd Stage)	Population
1	Chamrajpete	26829	Appajiyappa Garden	600
			Rajagopal Garden Phase 1	1100
2	Shanthinagar	7095	MayaBazar	2804
			Thimmaraiyappa Garden	317
3	Basavangudi	5985	Kempegowda Nagar	791
			Sanyasikunte	1080
4	Malleswaram	25833	Valluvarpuram	1837
			Sunnadugudu	865
5	Bharathinagar	11448	Kadirappa 9th Cross	150
			G.Muniyappa Garden	1450
	Total	77190	Total	10994
	Proportion of Ist Stage Population	25.18 %	Proportion of IInd Stage Population	14.24 %

Table A4: Round 2 sampling frame

Sl.No	Selected Town Areas (IInd Stage)	Population	Selected Slums (IIIrd Stage)	Population
1	Gandhinagar	7214	Hanumanthappa	924
			Vivekananda	685
2	Chamrajpete	26829	Fireworks colony	200
			Gurappa garden	960
3	Shanthinagar	7095	Jyothinivas	300
			Shaktivelu	900
4	Basavangudi	5985	Bhovi colony	355
			Madival Machaiah	369
5	Jayamahal	13111	A.A.N.Block	400
			Gangenahalli	300
	Total	60234	Total	5393
	Proportion of Ist Stage Population	19.65 %	Proportion of IInd Stage Population	8.95 %