The Self-Selection of Workers to the Formal and Informal in Transition Economies: Evidence from Tajikistan

Peter HUBER – Austrian Institute for Economic Research, Austria and Mendel University Brno - Faculty of Economics and Business Administration (Peter.Huber@wifo.ac.at)

Ulugbek RAHIMOV – Faculty of Economics, Westminster International University in Tashkent (urahimov@wiut.uz)

Abstract
This paper analyzes the selection of workers to informal and formal sector employment in Tajikistan. It estimates multinomial selection bias correction model to assess the impact of observable and unobservable characteristics on the self-selection of workers into the formal and informal sector and on sector specific wages using individual level data from the Tajikistan Standards of Living Survey of 2007. The results suggest that selection of workers on both observable and unobservable characteristics to these sectors is broadly consistent with self-selection on comparative advantages, that the self-selection of workers on unobservable characteristics is the main reason for higher wages in the informal than the formal sector in Tajikistan and that relative wages in the two sectors have a rather strong impact on the decision of workers to work in the formal or informal sector.

1. Introduction

Labor markets in developing countries often consist of two sectors: a formal sector offering high earnings and secure employment and an informal sector with low earnings and insecure employment. Two competing hypotheses aim to explain employment in these sectors. The segmentation hypothesis assumes that individuals working in the informal sector would prefer to work in the formal sector, but cannot find employment there. It thus predicts that informal sector employment is mostly involuntary, that individuals are poor as a result of informal sector employment and that informal sector employees could increase their wages by moving to the formal sector. The comparative advantage hypothesis, by contrast, claims that workers sort into the sector where they can earn the highest income and thus could not increase wages by moving to the other sector. As a consequence, this hypothesis suggests that workers in the informal sector are poor because they are poorly endowed with characteristics necessary for employment in the formal sector (see Maloney, 2004, Loayza and Rigolini, 2006, Amaral and Quintin, 2006, Bennett and Estrin, 2010 and Bennett, 2011 for theoretical work and World Bank, 2007 and Battini et al., 2010 for surveys).

Quite a few contributions have tested for the relevance of the comparative advantage hypothesis by analyzing formal and informal sector wage gaps. For instance, Magnac (1991) tests for this by estimating a generalized Roy model. In complementary papers, Gindling (1991) uses regressions with sample selection and Pratap and Quintin (2006) use a semi-parametric approach to test for the labor market segmentation hypothesis. The results of these papers, mostly suggest that individuals

self-select into different sectors according to comparative advantages. Other authors (e.g. Maloney, 1999, Bosch and Maloney, 2005 Lehmann and Pignatti, 2007 and Pages and Stampini, 2009) use data on worker flows between formal and informal sector employment to differentiate between the segmentation and the comparative advantage hypothesis. These studies lead to less univocal results. Some authors argue that high bi-directional flows between these sectors suggest no segmentation (Bosch and Maloney, 2005), but others find evidence for segmentation on account of predominantly unidirectional flows from the formal to the informal sector (Pages and Stampini, 2007).

In addition, recent studies often find that formal-informal sector wage differentials are lower among high income groups than among low income groups (e.g. Bargain and Kwenda, 2015, Nguyen et al., 2013, Tansel and Kan, 2012). This is interpreted as evidence that informal labor markets consist of an upper and a lower tier, with the upper tier representing the competitive part of the informal sector and the lower tier of persons rationed out of the formal sector (Cunningham and Maloney, 2001, Günther and Launov, 2012). This suggests that segmentation and sorting on comparative advantages may not be mutually exclusive, but may apply to different sub-segments of the labor market.

Only a few contributions in this literature so far focus on transition countries. Furthermore, these studies - as pointed out in a recent survey by Lehmann (2015) – remain inconclusive with respect to the relevance of the comparative advantage or segmentation hypotheses. In particular Lehmann (2015) suggests that findings depend on the definition of informality used, the treatment of self-employed in the formal-informal sector definition and on whether wage or worker flow data are used to differentiate between the hypotheses. Thus, an early study by Lehmann and Pignatti (2007) using panel data for the Ukraine differentiates between voluntary informal, involuntary informal and formal salaried work as well as between formal and informal sector self-employment. This study finds evidence of a wage premium for voluntary informal sector employees as well as formal and informal self-employed relative to formal sector employees. Evidence on worker flows, however, suggests that workers queue for formal sector employment. Pages and Stampini (2009), by contrast, based on panel data for Albania, Ukraine and Georgia find no evidence of a wage premium for informal salaried workers but some evidence for a preference for salaried formal sector employment over salaried informal sector employment from worker flow data. The same, however, does not apply to formal and informal sector self-employment.

More recently Nguyen et al. (2013) – adding to these findings – suggest that the formal-informal sector earnings gap depends on workers' job status in Vietnam, such that informal jobs are at least as rewarding as formal sector jobs in many cases, and Lehmann et al. (2012) show that in Russia displaced workers have a significantly higher probability to end up in involuntary informal employment after one year, while those who quit their job more often move to voluntary informal sector employment, while Lehmann and Muravyev (2014) find that the size of the informal sector in both transition and Latin American countries is related to national labor market institutions and tax wedges. Finally, Staneva and Arabsheibani (2014), in the contribution that is most closely related to the current one, decompose differences in earnings between formal and informal sector employees in Tajikistan in the year 2007. Using quantile regression and matching techniques they find higher informal sector than formal sector

wages in all parts of the wage distribution in Tajikistan. Furthermore, performing Oaxaca-Blinder decompositions of the wage differentials these authors also find that 60-70% of the observed wage gaps can be ascribed to differences in the distribution of the observed characteristics between formal and informal sector workers.

This paper extends on these findings in two respects. First, in contrast to Staneva and Arabsheibani (2014), who focus on the formal and informal sector wage differential, the current paper focuses on the self-selection of workers into the formal and informal sector. In particular, it pays particular attention to the impact of self-selection on observable and unobservable characteristics on the wages in the informal and formal sector. Second, the current paper - next to the choice between formal and informal sector employment modeled by Staneva and Arabsheibani (2014) - also considers the choice between inactivity and either formal or informal sector employment.

To achieve its objectives the paper estimates a version of a Heckman type selection model based on a multinomial choice between inactivity, formal sector and informal sector employment developed by Dubin and McFadden (1994) and theoretically founded in the comparative advantage hypothesis. The results of this estimation are then used to address three sets of questions related to formal and informal sector employment and remuneration in Tajikistan. First, the paper assesses whether the patterns of self-selection on observable characteristics of workers and the returns to observable and unobservable characteristics suggest that workers are employed in the sector, which offers the highest returns for their observed and unobserved characteristics. Second, it decomposes wages in the two sectors by (selectivity corrected) Oaxaca-Blinder decompositions. This provides insights as to what share of the informal-formal sector wage differential is due to a) differences in observable characteristics, b) differences in the returns to observable characteristics and c) the self-selection on unobservable characteristics of workers in these sectors. Third, it estimates structural responses of labor supply to changes in relative wages to assess whether policies impacting on relative remuneration could have an impact on the self-selection of workers to the formal and informal sector. The next section describes the theoretical basis of the empirical model used and discusses the estimation strategy. Section 3 then introduces the data, while section 4 presents the estimation results. Section 5 finally concludes and interprets results in terms of the segmentation and comparative advantage hypothesis.

2. Model and estimation

As a theoretical framework for our analysis we consider a standard choice model based on the comparative advantage hypothesis. In this, potentially risk averse workers (indexed by j) choose to be in one of three states. These are non-employment, informal and formal sector employment (indexed by k $\in \{n, i, f\}$). Each of these states is associated with earnings w_{jk} such that:

$$w_{jk} = X_j \beta_k + u_{jk}$$

where X_j is a vector of productive characteristics (such as education and age) determining wages, β_k is a vector of parameters measuring the sector specific returns to these characteristics and u_{jk} is a random component (with variance σ_u). This

measures any productive attributes (such as intelligence, entrepreneurial talent or other individual traits relevant to remuneration in the sector) that are known to the individual but are unobserved by the researcher.

In addition, working in a particular sector is associated with non-pecuniary benefits (b_{jk}). These are also related to the observable characteristics of workers (Z_j) as well as to characteristics which are unobserved in the data but known to the individuals (such as their risk preferences or their preferences with respect to the working conditions in these sectors or for the higher employment stability in the formal sector) and thus captured in an error term ε_{jk}, (with variance σ_ε) such that:

$$b_{jk} = Z_j \gamma_k + \varepsilon_{jk}$$

In this setup income maximizing workers will aim to select to the sector, in which they can expect to obtain the highest total payoff. The payoff in sector k is given by $\pi_{jk}^* = b_{jk} + w_{jk} = \Gamma_j \delta_k + \eta_{jk}$ with $\eta_{jk} = u_{jk} + \varepsilon_{jk}$, π_{jk}^* the (unobserved) expected payoff of working in sector k and $\Gamma_j = (X_j \ Z_j)$ and $\delta'_k = (\beta_k \ \gamma_k)$. Thus denoting s as the choice of sector of employment and encoding this as 0 for non-employment, 1 for informal sector employment and 2 for formal sector employment a rational worker will choose the sector of employment according to:

$$s = \begin{cases} 0 & if & \pi_{jn}^* > \max[\pi_{jf}^*, \pi_{ji}^*] \\ 1 & if & \pi_{ji}^* > \max[\pi_{jn}^*, \pi_{jf}^*] \\ 2 & if & \pi_{jf}^* > \max[\pi_{jn}^*, \pi_{ji}^*] \end{cases} \qquad (1)$$

Under the assumption that the η_{jk} are independently and identically extreme value distributed, an appropriate econometric model to estimate the parameters governing the choice of the sector of employment represented in equation (1) is a multinomial logit model (McFadden, 1973). This estimates the probability that a worker works in one of the sectors relative to that of being in an arbitrarily chosen reference sector (which is non-employment) depending on a set of exogenous characteristics (Γ_j). The estimates can also be used to calculate the marginal effects of each of the variables in Γ_j (Greene, 1993). These can be interpreted as the percentage point change in the probability to be in a particular labor market state given a unit deviation from the mean for continuous variables. They thus measure the change in the probability of choosing sector k due to a unit increase in the independent variable (Γ_j). A statistically positive estimate indicates that this variable increases this probability and a negative one that this variable decreases it. As a consequence, the marginal effects of the estimates of equation (1) provide evidence of the self-selection of workers on observable characteristics to individual sectors.

Furthermore, given self-selection on observables, workers' wages in the chosen sector are:

$$E\left(w_{jk} \middle| \pi_{jk} > \max_{k' \neq k} \pi_{jk'}\right) = X_j \beta_k + \mathrm{E}(u_{jk} | \pi_{jk} > \max_{k' \neq k} \pi_{k'}) \qquad (2)$$

Estimating these sector wage equations by ordinary least squares will therefore suffer from a selection bias arising from the term $\mathrm{E}(u_{jk} | \pi_{jk} > \max_{k' \neq k} \pi_{k'})$. As shown by Lee (1983) equation (2) can, however, be consistently estimated by a Heckman-type two-step selection model based on the multinomial choice modeled in equation (1). In this,

first, the "reduced form" multinomial logit selection equation (1) is estimated. Second, a wage equation (i.e. equation 2) including a selection correction for the choice derived from the first step estimates is estimated for each sector. In particular, Bourguignon et al. (2007) suggest controlling for self-selection by estimating wage equations of the form:

$$w_k = X\beta_k + \tau_{kk}\mathrm{m}(\mathrm{P_k}) + \sum_{k'\neq k} \tau_{kk'}\mathrm{m}(\mathrm{P_{k'}})\frac{\mathrm{P_{k'}}}{\mathrm{P_{k'}}-1} + \upsilon_k \qquad (3)$$

with $\upsilon_k = log(\mathrm{P_k}) + \eta_{jk}$, P_k the probabilities of choosing state k, $m(P_k)$ a complicated increasing function of these probabilities and β_k, τ_{kk} and $\tau_{kk'}$ a set of parameters to be estimated.[1]

Although the terms $\mathrm{m}(\mathrm{P_k})$ and $\mathrm{m}(\mathrm{P_{k'}})\frac{\mathrm{P_{k'}}}{\mathrm{P_{k'}}-1}$ in equation (3) can only be evaluated numerically, estimating this equation yields unbiased estimates of the returns to productive characteristics in the sector (i.e. β_k) as well as of a series of terms related to the self-selection on unobservable characteristics of workers (τ_{kk} and $\tau_{kk'}$). These have direct applicability to the self-selection of workers to the formal and informal sector based on observable and unobservable characteristics. For instance, with respect to self-selection on observable characteristics, workers should ceteris paribus chose to work in the sector in which the returns to their observed productive characteristics are highest. This implies that if estimates of β_k in a particular sector are larger than in the others (i.e. $\beta_k > \max[\beta_{k'}]$) then also workers with these characteristics should be selected into this sector. As a consequence, also the estimated marginal effects derived from equation (1) should be larger than in any other sector. Thus, by comparing estimates from the first stage equation (1) and the second stage equation (3) two situations can arise with respect to selection on observables. If $\hat{\beta}_i > \hat{\beta}_f$ and $\hat{\delta}_i > \hat{\delta}_f$ (or equivalently $\hat{\beta}_i < \hat{\beta}_f$ and $\hat{\delta}_i < \hat{\delta}_f$) this would be consistent with positive self-selection on observables and thus be compatible with the comparative advantage hypothesis. If by contrast $\hat{\beta}_i < \hat{\beta}_f$ and $\hat{\delta}_i > \hat{\delta}_f$ (or equivalently $\hat{\beta}_i > \hat{\beta}_f$ and $\hat{\delta}_i < \hat{\delta}_f$), this would be indication of negative self-selection on observables and thus incompatible with the comparative advantage hypothesis.

In addition, the estimated coefficients on the self-selection terms (i.e. τ_{kk} and $\tau_{kk'}$) have an intuitive interpretation (see Dimova and Gang, 2007). A statistically significant positive value of τ_{kk} indicates that persons who – after controlling for observable characteristics – are more likely to work in sector k (i.e. have a positive error term for sector k employment in equation 1) also have ceteris paribus higher expected wages in this sector. This would be consistent with positive self-selection on unobservable characteristics to a sector, as predicted by the comparative advantage hypothesis. A negative estimate of τ_{kk}, by contrast, would indicate that persons who are more likely to work in sector k have lower expected wages in this sector. Similarly, a positive coefficient on the selection terms $\tau_{kk'}$ indicates that individuals who are more likely to self-select to work in sector k' for unobserved reasons, would also

---

[1] Lee (1983), Dubin and McFadden (1984) and Dahl (2002) have all proposed methods to estimate equation (2). These differ in the details of the selection bias correction. Equation (3) is based on a modified version of the correction proposed by Dubin and McFadden (1984) that was found to be the most accurate in a Monte Carlo study by Bourguignon et al. (2007) and has since been widely used in the literature (see: John and Thomsen 2014, Wu and Shen 2013 and Dimova and Gang 2007).

receive a higher wage in sector k. This would contradict the comparative advantage hypothesis in particular if $\tau_{kk'}>\tau_{k'k'}$. A negative or statistically insignificant estimate of $\tau_{kk'}$, by contrast, would suggest that persons with a high probability to work in sector k', ceteris paribus have no wage advantage or lower wages in sector k and thus be consistent with the comparative advantage hypothesis.

3. Data

Our focus in the current paper is on Tajikistan. This may be considered a particularly interesting country as it is a small, landlocked country with a rapidly growing population[2], which is also one of the poorest countries in Central Asia as well as among the Former Soviet Union countries (Dietz at al., 2015). It is also a country where a large share of the population has experience with (mostly circular or return) migration to Russia (Danzer et al. 2013). According to Danzer and Ivaschenko (2010) 9% of the population worked abroad in 2009 and 28% of the households had at least one household member working abroad. As a consequence, the Tajik economy is also highly dependent on remittances. According to World Bank estimates remittances accounted for 49% of the Tajik GDP in 2013 (Danzer et al, 2013) and in 2007 27.4% of the households in received remittances within a year. As recently pointed out by Abdulloev et al. (2012) this may be of particular relevance for the study of the informal sector in Tajikistan as immigration and informal sector activity may well be substitutes or complements to each other in such a context.[3]

In addition, the Tajik labor market is also marked by rather low labor force participation, a high share of the informal economy in economic activity and substantial changes in employment structure since the beginning of transition (Staneva and Arabsheibani, 2014). Thus, despite very low official unemployment rates (of 2.3%) the participation rate in Tajikistan was only 51.7% and the informal sector accounted for more than 50% of employment in 2007 (and as much as 78% of total domestic employment in 2009 – ILO 2015). Thus, given both the high informal sector employment share as well as the low labor force participation rate, analyzing the self-selection of workers into inactivity, informal sector employment and formal sector employment may be of particular interest from a policy perspective in this country.

We therefore estimate the model described in equations (1) and (3) using data from the Tajikistan Standards of Living Survey of 2007 (TSLS). This is one of many Standards of Living Surveys that have become standard data sets for labor market analysis in developing countries (Nguyen et al., 2013). It contains a representative sample of more than 4,500 Tajik households and has been previously used in several papers on the Tajik labor market (Justino and Shemyakina, 2012; Staneva and Arabsheibani 2014 and Abdulloev et al., 2012). To focus on persons who are ready and able to accept a job we restrict attention to the working age population (aged 16 to 64) and exclude persons who are students, pensioners, handicapped and persons in military service or unwilling to work. Furthermore, we consider only the main job held

---

[2] According to the State Statistics Committee, the permanent population of Tajikistan amounted to 7,215,700 people as of the beginning of 2008 and between 1998 and 2007 the average annual population growth rate was 2.1% (see: http://www.stat.tj/en/database/socio-demographic-sector/).

[3] Abdulloev et al. (2012) find that in Tajikistan professional workers engage in informal activities to avoid migration, while low-skilled workers choose to migrate instead of working in the informal sector.

in the last 14 days, as the survey collects information on wages, hours worked and social security entitlement as well as contractual basis for this job only. This also ensures omission of a large number of very low income employment relationships. Setting these restrictions and dropping persons with missing variables we obtain information on 13,539 individuals.[4] Of these 5,043 are employed either in the formal or the informal sector and 7,496 are not employed.

3.1 Dependent variables

Previous research has shown that results may depend on whether a productivity based, legalistic or social protection definition of the informal sector is used (see e.g. Marcoulier et al. 1997, Henley et al. 2009, Nguyen et al. 2013 for the use of different measures and Perry et al. 2007 and Kanbur 2009 for discussions of different concepts) and how self-employment is treated in this definition. Recently Lehman (2015) compared different definitions of informal sector employment in Russia. He argued and presented evidence that in the context of transition economies legalistic and social protection based measures of informality are likely to minimize measurement bias (see also Henley et al. 2009).[5] He, however, also suggests that results may be sensitive to the inclusion or exclusion of self-employed. We therefore use two definitions of informal sector employment. In the first (the social protection definition) a person is considered formally employed if he or she is affiliated to the Tajik social security scheme in their main job. In the second (the legalistic definition) persons are encoded as employed in the formal sector if they signed a contract or written agreement with their employer and informally employed if no such agreement exists. Finally, in both instances persons are considered not employed if they do not have a primary job which generates income from work.

The top panel of table 1 shows the distribution of non-employed, informal and formal sector workers according to these definitions by different demographic groups. The overall share of non-employed in our sample is over 50 percent. These levels are markedly higher for females and young as well as less qualified persons. Also informal sector employment exceeds formal sector employment in both definitions of the informal sector and for most demographic groups. The only exceptions (in both definitions) are females, persons aged 45 or older and persons with tertiary education. For the youngest and persons with secondary education informal sector employment exceeds formal sector employment by most. The informal sector employment share is also somewhat higher when considering the social protection definition than when considering the legalistic definition both in aggregate and for almost all demographic groups. This is thus consistent with

---

[4] In total the data contain observations on 28,957 individuals. Of these 15,455 are dropped on account of our focus on groups close to the labor market and a further 179 due to missing information on education.

[5] Furthermore, Dimova et al. (2011) have proposed and proven the usefulness of a behavioral definition of informality based on revealed differences between household income and expenditures. Unfortunately, however, this measure is unambiguously defined on a household level, so that it cannot be used in the current paper.

Table 1 Formal and informal sector employment, non-employment and average log wages by different formal sector definitions and demographic groups

	Social protection entitlement			Legalistic definition		
	Not employed	Informally employed	Formally employed	Not employed	Informally employed	Formally employed
Share of observations						
Total	55.37	23.69	20.94	55.37	22.68	21.95
Share self-employed		37.22	2.40		38.39	2.79
Gender						
Male	36.30	37.39	26.31	**36.30**	33.85	29.85
Female	71.18	12.34	16.48	71.18	13.42	15.40
Age						
15-24	65.65	19.95	14.39	65.65	19.17	15.18
25-44	49.65	27.79	22.55	49.65	25.85	24.49
45-60	45.11	24.46	30.42	45.11	24.40	30.49
Education						
Basic	69.96	17.48	12.56	69.96	17.75	12.29
Secondary	56.01	26.00	17.98	56.01	23.95	20.03
Tertiary	25.11	22.19	52.70	25.11	24.61	50.28
Log wage excluding bonuses and in kind payments						
Total		1.50***	1.00		1.36***	1.16
Gender						
Male		1.77***	1.24		1.67***	1.42
Female		0.8***	0.68		0.81	0.73
Age						
15-29		1.31***	0.75		1.13**	1.01
30-44		1.64***	1.12		1.55***	1.25
45-60		1.54***	1.09		1.42***	1.18
Education						
Basic		1.12***	0.49		0.97***	0.72
Secondary		1.52***	0.87		1.41***	1.07
Tertiary		1.91***	1.47		1.70***	1.55
Employees vs. self-employed						
Employees		1.42***	0.98		1.20	1.15
Self-employed		1.62	1.53		1.64**	1.32

Source: TSLS, 16 to 64 years old. ***, (**), (*) show significance of a t-test for difference in mean log wages at the 1%, (5%), (10%) level.

comparative findings for Europe by Hazans (2011), who finds that in most countries the low-educated, the young and persons with disabilities are more likely to work informally, other things equal.

This top panel, however, also shows that virtually all self-employed are informally employed in Tajikistan. In the TSLS only 83 of the interviewed were self-employed in the formal sector according to the legalistic definition and 68 according to the social protection definition. By contrast, 1,179 respectively 1,194 interviewed were self-employed in the informal sector. This thus precludes a separate analysis of formal and informal sector self-employed as conducted for instance in Lehman (2015) or Mussurov and Arabsheibani (2015) as virtually all self-employed are informally employed. One explanation for this may be that the questions posed in the data to define informality focuses on the presence of a contract or affiliation with the Tajik social security scheme, which is unlikely to be the case for the self-employed.

For the second stage equation, the dependent variables are (log) hourly formal and informal sector wages. These are calculated from two questions in the questionnaire in which respondents were asked on the previous month's wages and weekly working hours in their main job.[6] The bottom panel of table 1 shows the average formal and informal sector log hourly wages in Tajikistan together with a t-test for difference in means between sectors. According to these tests – as also found by Staneva and Arabsheibani (2014) – informal sector wages were significantly higher than formal sector wages in all segments of the Tajik labor market (except for female workers, when considering the legalistic definition of the formal sector) in both definitions of formality. Furthermore, informal-formal wage differentials are slightly smaller in the legalistic definition of the formal sector than in the social protection definition and there seem to be even larger differences in the informal-formal sector wage differences of employees and self-employed according to the definition of informality used. Thus, in the social protection definition employees' wages are higher in the formal than the informal sector, while the self-employed earn similar amounts in both sectors (although this comparison is based on few observations only). By contrast, in the legalistic definition exactly the opposite applies. Informal employees earn about the same as formally employed employees, while the self-employed have significantly higher wages in the informal sector.

3.2 Explanatory variables

As explanatory variables in the wage equation we include the variables usually included in a Mincerian wage regression - (the log of) age, its' square and controls for human capital. These are dummy variables for the highest completed education of the individual (basic, secondary and tertiary education). In addition, we control for whether a respondent is not Tajik, female or lives in the capital city of Dunshabe through dummy variables. These variables are included because Karnite (2010) suggests that non-Tajik workers earn lower wages than Tajiks and to control for potential gender differences in wages as well as for the high urban-rural wage differentials found in many transition countries (Brülhart and Koenig, 2006). Furthermore, in the first stage multinomial logit regressions, as Z variables,

---

[6] We conducted the same analysis as below on wage measures including hourly in-kind and bonus payments as a robustness check. This led to very similar results that are available from the authors upon request.

additionally, a dummy variable for married persons, a dummy for children under the age of six in the household, and a set of dummy variables for the self-reported health status of respondents (which are very good, good or fair or bad) is included as we expect these variables to influence the non-pecuniary benefits of (and capability to) work in a sector without them having a (direct) impact on wages. In particular, here we expect that on the one hand side marital status and the number of children may impact on the labor supply decision without having a direct impact on wages. On the other hand, we expect that an individual's health status is a determinant of whether a person is employed in the formal or the informal sector, as the formal sector also comprises a number of branches (e.g. police, military service) in which good health conditions are a precondition for employment. As a consequence, the healthiest should be most strongly selected to the formal sector.

According to descriptive statistics for these variables (Table 2) in both definitions the non-employed on average are significantly younger, more often female and more often have children than those in formal employment. They are, however, also more often Tajik but are married and reside in Dunshabe less often than those working in the formal sector. Individuals not employed are also significantly less well educated than the employed in the formal sector. Those working in the informal sector, by contrast, are usually significantly younger, more often Tajik but are less often female, less often reside in the capital city, less often have very good health and also more often have children under the age of six than those in formal employment. Furthermore, these differences are often larger when considering the social protection definition of the formal sector than when considering the legalistic definition and differences in the mean characteristics of the employed in both sectors are also very similar irrespective of whether we focus only on employees or on all employed.

Table 2 Descriptive Statistics of independent variables

	Overall	Social protection definition					Legalistic definition			
		Not Employed	Informally employed		Formally employed		Informally employed		Formally employed	
			All workers	Employees	All workers	Employees	All workers	Employees	All workers	Employees
ln(age)	3.47	3.41 ***	3.51 ***	3.50 ***	3.59	3.59	3.51 ***	3.51 ***	3.59	3.59
	(0.36)	(0.36)	(0.34)	(0.34)	(0.34)	(0.34)	(0.35)	(0.35)	(0.34)	(0.34)
Not Tajik	0.19	0.17 ***	0.21 **	0.21	0.23	0.23	0.22	0.23	0.22	0.22
	(0.39)	(0.37)	(0.41)	(0.41)	(0.42)	(0.42)	(0.41)	(0.42)	(0.41)	(0.41)
Male	0.45	0.30 ***	0.72 ***	0.73 ***	0.57	0.56	0.68 ***	0.67 ***	0.62	0.61
	(0.25)	(0.21)	(0.20)	(0.20)	(0.25)	(0.25)	(0.22)	(0.22)	(0.24)	(0.24)
Reside in capital	0.15	0.13 **	0.16 **	0.14 ***	0.18	0.18	0.16	0.14 ***	0.18	0.18
	(0.35)	(0.34)	(0.36)	(0.35)	(0.38)	(0.39)	(0.37)	(0.35)	(0.38)	(0.38)
Basic education	0.22	0.27 ***	0.16 ***	0.16	0.13	0.13	0.17 ***	0.17 ***	0.12	0.12
	(0.41)	(0.45)	(0.37)	(0.36)	(0.34)	(0.33)	(0.38)	(0.38)	(0.33)	(0.32)
Secondary education	0.66	0.67 ***	0.73 ***	0.73 ***	0.57	0.57	0.70 ***	0.68 ***	0.61	0.61
	(0.47)	(0.47)	(0.45)	(0.45)	(0.50)	(0.50)	(0.46)	(0.47)	(0.49)	(0.49)
Tertiary education	0.12	0.05 ***	0.11 ***	0.12 ***	0.30	0.30	0.13 ***	0.14 ***	0.27	0.28
	(0.32)	(0.23)	(0.31)	(0.32)	(0.46)	(0.46)	(0.34)	(0.35)	(0.45)	(0.45)
Very good health	0.11	0.12	0.10 ***	0.09 **	0.12	0.11	0.10 *	0.09 **	0.11	0.11
	(0.32)	(0.33)	(0.29)	(0.29)	(0.32)	(0.32)	(0.30)	(0.29)	(0.32)	(0.32)
Good health	0.74	0.73	0.76	0.78 ***	0.75	0.75	0.74 **	0.76 **	0.77	0.77
	(0.44)	(0.44)	(0.43)	(0.41)	(0.44)	(0.43)	(0.44)	(0.43)	(0.42)	(0.42)
Fair or bad health	0.14	0.15 *	0.14	0.12	0.14	0.14	0.16 ***	0.15 ***	0.12	0.12
	(0.35)	(0.36)	(0.35)	(0.33)	(0.34)	(0.34)	(0.37)	(0.36)	(0.32)	(0.32)
Married	0.72	0.69 **	0.74 *	0.72 ***	0.76	0.76	0.74 *	0.71 ***	0.76	0.76
	(0.45)	(0.46)	(0.44)	(0.45)	(0.43)	(0.43)	(0.44)	(0.45)	(0.42)	(0.43)
Children under 6	0.54	0.55 ***	0.55 ***	0.54 **	0.51	0.51	0.55 **	0.54	0.52	0.51
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Observations	13539	7496	3208	2014	2835	2767	3071	1892	2654	2889

Source: TSLS, 16 to 64 years old. Values in brackets are the Standard Deviations of the variables ***, (**), (*) show significance of a t-test for difference in means of the respective aggregates relative to the formally employed at the 1%, (5%), (10%) level

4. Results

4.1 Reduced form selection regressions

The marginal effects for the first stage (reduced form) multinomial logit regressions for the selection of workers into the three labor market states (non-employment, informal sector and formal sector employment) for both formal sector definitions are reported in Table 3. They suggest that irrespective of whether we include or exclude the self-employed, older workers have both a higher formal as well as informal sector employment probability, while younger ones are more likely not to be employed. The increase in formal sector employment probability with age is, however, substantially larger than the increase in informal sector employment probability. A one percent increase in age increases the formal sector employment probability by 0.21 percentage points, but increases the informal sector employment probability by 0.08 to 0.09 percentage points (depending on the definition of the informal sector) for all employed. For employees a one percent increase in age increases the formal sector employment probability of employees by 0.22 to 0.23 percentage points, but increases their informal sector employment probability by 0.06 to 0.07 percentage points. Non-employment probabilities also reduce with education irrespective of whether only employees or all employed are considered, while both the informal and formal sector employment probability increases with educational attainment. This increase is, however, much stronger for the formal sector employment probability than for the informal sector employment probability. In particular, among all employed with a tertiary education the informal sector employment probability is one to five percentage points higher than among all employed with only compulsory education but the formal sector employment probability is 27 to 29 percentage points higher, which is significant in both definitions of formality. Similarly, for employees the formal sector employment probability among tertiary educated is 30 to 31 percentage points higher than among those with compulsory education but their informal sector employment probability is 2 to 5 percentage points higher. This thus suggests a stronger selection to the formal sector based on age and education for employees than among all employed, although the differences in marginal effects between these two groups (as for all other variables) are too small to be statistically significant.

There is also some indication of selection by gender, marital status and ethnicity. Males have a significantly (29 to 33 percentage point) lower non-employment probability and their probability to work in the informal sector is (depending on the formal sector definition and group considered) by 15 to 26 percentage points higher than for males. Their probability to work in the formal sector, by contrast, is (depending on the formal sector definition and group considered) seven to 15 percentage points lower than that of males. Similarly, non-Tajiks are significantly more often employed in the formal and informal sector and less often not employed. This effect is also significantly stronger for formal than informal sector employment when the social protection definition of the formal sector is considered both when including and excluding the self-employed. Marriage, by contrast, increases the non-employment probability by between around nine to ten percentage points, but reduces both the formal and informal sector employment probability, with once more the effect

being stronger in the formal sector (where it amounts to six to seven percentage points) than the informal sector (with two to three percentage points).

Table 3 Results of first step selection equation (marginal effects)

	P(inactive)	P(informal)	P(formal)	P(inactive)	P(informal)	P(formal)
	Social protection definition			Legalistic definition		
All employed						
ln(age)	-0.279***	0.072***	0.206***	-0.281***	0.076***	0.205***
	(0.016)	(0.013)[a]	(0.012)	0.0161	(0.013)[a]	(0.012)
Basic education			Base category			
Secondary education	-0.074***	0.037***	0.037***	-0.076***	0.020**	0.056**
	(0.012)	(0.010)	(0.010)	(0.012)	(0.010)[b]	(0.010)
Tertiary education	-0.299***	0.011	0.288***	-0.316***	0.045***	0.271***
	(0.019)	(0.016)[a]	(0.013)	(0.019)	(0.015)[a]	(0.013)
Not Tajik	-0.113***	0.035***	0.078***	-0.113***	0.052***	0.062***
	(0.012)	(0.009)[a]	(0.009)	(0.012)	(0.009)	(0.009)
Male	-0.330***	0.262***	0.068***	-0.333***	0.216***	0.117***
	(0.009)	(0.007)[b]	(0.007)	(0.009)	(0.007)[a]	(0.007)
Capital city	-0.035***	0.039***	-0.004	-0.035***	0.032***	0.003
	(0.013)	(0.011)	(0.010)	(0.013)	(0.011)[c]	(0.010)
Very good health			Base category			
Good health	-0.047***	0.056***	-0.009	-0.047***	0.037***	0.010
	(0.015)	(0.012)[a]	(0.011)	(0.015)	(0.012)	(0.012)
Fair or bad health	-0.002	0.045***	-0.043***	0.001	0.057***	-0.057***
	(0.019)	(0.016)[a]	(0.015)	(0.019)	(0.015)[a]	(0.015)
Married	0.085***	-0.020**	-0.066***	0.088***	-0.026**	-0.062***
	(0.013)	(0.011)[a]	(0.010)	(0.013)	(0.011)[b]	(0.010)
Children under 6	-0.007	0.011	-0.004	-0.007	0.008	-0.001
	(0.010)	(0.008)	(0.007)	(0.010)	(0.008)	(0.007)
No. Observations		13539			13539	
Log likelihood		-11853.926			-12018.012	
Employees only						
ln(age)	-0.283***	0.058***	0.226***	-0.286***	0.068***	0.218***
	(0.016)	(0.011)[a]	(0.013)	(0.016)	(0.011)[a]	(0.013)
Basic education			Base category			
Secondary education	-0.077***	0.031***	0.046***	-0.081***	0.012	0.069***
	(0.013)	(0.009)	(0.011)	(0.013)	(0.009)[a]	(0.011)
Tertiary education	-0.335***	0.024*	0.311***	-0.351***	0.056***	0.295***
	(0.019)	(0.013)[a]	(0.014)	(0.019)	(0.013)[a]	(0.015)
Not Tajik	-0.110***	0.025***	0.085***	-0.111***	0.044***	0.066***
	(0.012)	(0.008)[a]	(0.010)	(0.012)	(0.008)	(0.010)
Male	-0.292***	0.199***	0.093***	-0.298***	0.152***	0.146***
	(0.010)	(0.006)[a]	(0.008)	(0.010)	(0.007)[a]	(0.008)
Capital city	-0.049***	0.050***	-0.001	-0.051***	0.037***	0.014
	(0.015)	(0.011)	(0.012)	(0.015)	(0.011)	(0.012)
Very good health			Base category			
Good health	0.016	0.020	-0.036**	0.016	0.040***	-0.056***
	(0.019)	(0.014)[a]	(0.016)	(0.020)	(0.014)[c]	(0.016)
Fair or bad health	-0.020	0.012	0.007	-0.020	0.005	0.015
	(0.014)	(0.010)	(0.011)	(0.014)	(0.010)	(0.011)
Married	0.093***	-0.026***	-0.068***	0.097***	-0.034***	-0.063***
	(0.013)	(0.009)[c]	(0.011)	(0.013)	(0.009)[b]	(0.011)
Children under 6	-0.002	0.005	-0.003	-0.003	0.004	-0.001
	(0.010)	(0.007)	(0.008)	(0.010)	(0.007)	(0.008)
No. Observations		12,277			12,277	
Log likelihood		-10012.210			-10117.919	

Source: TSLS, 16 to 64 years old. Values in brackets are (heteroskedasticity robust) standard errors of the estimate ***, (**), (*) show significance at the 1%, (5%), (10%) level. [a], [b], [c] Null hypothesis of equal effects for formal and informal sector employment can be rejected at the 1%, (5%), 10% level.

Marriage, by contrast, increases the non-employment probability by between around nine to ten percentage points, but reduces both the formal and informal sector employment probability, with once more the effect being stronger in the formal sector (where it amounts to six to seven percentage points) than the informal sector (with two to three percentage points). Informal sector employment is also strongly associated with residence in Dunshabe and thus an urban phenomenon. Persons residing Dunshabe have a four to five percentage points lower non-employment probability than persons living outside Dunshabe, while the formal sector employment probability in Dunshabe is comparable to that outside Dunshabe in both definitions of the formal sector. As a consequence, residents of Dunshabe are by three to five percentage points more likely to be informally employed. Finally, health conditions mainly predict the formal and informal sector employment probabilities: When considering all employed, persons who state that their health conditions are fair or bad have a four to six percentage points lower formal sector and an equally higher informal sector employment probability. Persons stating to have good health have a significantly (four percentage points) lower non-employment probability but a significantly (six percentage points) higher informal sector employment probability than persons stating to have very good health. When considering employees only, however, persons with good health have an about equal non-employment probability but a significantly lower formal sector employment probability and, in the case of the legalistic definition, a significantly higher informal sector employment probability.[7]

4.2 Wage regressions

Comparing these results to the wage equation estimates (table 4), suggests that self-selection to a large degree accords with the differences in returns to observable characteristics in the formal and informal sector. For instance, the strong selection of tertiary educated workers into the formal sector accords with the substantially higher returns to tertiary education in the formal sector (when considering both definitions of informality as well as results including and excluding the self-employed) and the insignificant impact of tertiary education on informal sector wages. In a similar vein the returns for secondary education are only slightly higher in the formal than the informal sector, and secondary educated workers are also only slightly more likely to be employed in the informal sector according to the results in table 3. Furthermore, returns to age are in line with the self-selection of older workers into the formal sector. When excluding the self-employed the coefficients on the age term are significantly higher in the formal sector than in the informal sector irrespective of the formal sector definition, and when including them the same applies in the social protection definition of formality.

The differences in gender and ethnic wage differentials in the formal and informal sector are also consistent with positive self-selection on observable characteristics of workers into these two sectors. Non-Tajik minorities, all else equal, earn significantly lower wages than Tajiks in the informal sector if the social protection definition is considered. This is consistent with the higher probability of non-Tajiks to work in the formal sector in this definition of the formal sector. In the legalistic

---

[7] These results are consistent to those of Hazans (2011) and also justify the use of the health variable as an excluded variable to better identify the effects of self-selection in the wage equation.

definition, by contrast, there are no significant differences between these two groups in both sectors. This is consistent with the about equal employment probability of non-Tajiks in both sectors in this definition of informality. Men earn significantly more than women in both sectors. The male wage premium, however, is much larger in the informal sector. This is consistent with men more often working in the informal sector than in the formal one relative to women.

The only variables for which the wage equation estimates cannot be reconciled with positive self-selection of workers on observables are the capital city effects in regressions where the self-employed are excluded and returns to age when considering all employed and the legalistic definition. Thus, in the legalistic definition of formality when self-employed are included in the regression returns to education are slightly lower in the informal than the formal sector. Furthermore, location in Dunshabe, leads to slightly higher wage increases in the informal sector than in the formal sector when self-employed are included in the regression, but to oppositely signed effects when excluding the self-employed according to both definitions. This thus indicates higher returns to self-employment in Dunshabe than in other parts of the country. According to the results reported in table 3, however, both employees and all workers are more likely to be employed in the informal sector when residing in Dunshabe. This may thus indicate some segmentation for employees in Dunshabe.

Table 4 Wage regression (dependent variable log hourly wages)

	All Employed				Employees only			
	Social protection definition		Legalistic definition		Social protection definition		Legalistic definition	
	Informal sector	Formal sector	Informal sector	Formal sector	Informal sector	Formal sector	Informal sector	Formal sector
ln(age)	0.21	0.43***	0.49***	0.35***	-0.17	0.42***	0.03	0.37***
	(0.13)	(0.11)	(0.12)	(0.11)	(0.16)	(0.12)	(0.12)	(0.11)
Basic education				Base category				
Secondary education	0.23***	0.32***	0.21***	0.25***	0.29***	0.32***	0.18**	0.27***
	(0.07)	(0.06)	(0.07)	(0.07)	(0.08)	(0.06)	(0.08)	(0.07)
Tertiary education	-0.16	1.13***	0.25	1.17***	-0.06	1.18***	0.24	1.25***
	(0.25)	(0.19)	(0.19)	(0.17)	(0.34)	(0.21)	(0.21)	(0.18)
Not Tajik	-0.20**	0.10*	0.066	0.062	-0.25***	0.10	0.03	0.07
	(0.08)	(0.06)	(0.08)	(0.07)	(0.09)	(0.06)	(0.08)	(0.06)
Male	1.66***	0.91***	2.01***	0.93***	1.67***	0.84***	1.76***	0.98***
	(0.23)	(0.17)	(0.23)	(0.19)	(0.22)	(0.14)	(0.14)	(0.15)
Capital city	0.44***	0.39***	0.54***	0.35***	0.20***	0.39***	0.23***	0.37***
	(0.06)	(0.05)	(0.06)	(0.05)	(0.06)	(0.04)	(0.06)	(0.04)
m_0 (inactivity)	-4.47***	-2.33***	-6.94***	-3.23***	-5.04***	-2.15***	-6.85***	-3.22***
	(0.77)	(0.57)	(0.82)	(0.67)	(0.81)	(0.46)	(0.81)	(0.51)
m_1 (informal employment)	0.12	-0.04	0.58*	-1.73***	0.15	0.16	0.95***	-1.26*
	(0.32)	(0.54)	(0.31)	(0.63)	(0.34)	(0.61)	(0.34)	(0.68)
m_2 (formal employment)	-4.92***	-0.087	-5.78***	-0.19	-6.00***	-0.04	-7.15***	-0.12
	(0.84)	(0.23)	(0.80)	(0.22)	(1.10)	(0.24)	(0.94)	(0.23)
_cons	-4.45***	-2.72***	-8.01***	-3.16***	-4.26***	-2.65***	-7.71***	-3.23***
	(0.94)	(0.78)	(1.08)	(0.81)	(0.91)	(0.84)	(1.14)	(0.80)
N	3208	2835	3071	2972	2014	2767	1892	2889
R-sq	0.176	0.194	0.203	0.162	0.228	0.2	0.263	0.167
Condition Number	17.24	15.07	18.54	15.92	17.57	15.14	17.00	13.03

Source: TSLS, 16 to 64 years old. Values in brackets are (heteroskedasticity robust) standard errors of the estimate ***, (**), (*) show significance at the 1%, (5%), (10%) level. m_0, m_1, m_2 are the equivalents of the inverse mills ratio associated with self-selection into inactivity, the selection into the informal sector and the formal sector, respectively (see equation 3 for a formal definition).

The selection terms in the bottom of table 4 (indicated by the variables m_0 to m_2), which are the equivalent of the inverse mills ratios included in the standard Heckman selection model (but are calculated as the terms $m(P_k)$ and $m(P_{k'})\frac{P_{k'}}{P_{k'}-1}$ in equation 3), by contrast, suggest that in both sector workers are also positively selected on unobservable comparative advantages. Thus, the significant negative sign of the variable m_0, that is the equivalent of the inverse mills ratio associated with self-selection into inactivity, suggests that persons who, after controlling for observables, have an above average probability to be in non-employment also have below average expected wages in both the formal as well as the informal sector, irrespective of the definition of informality. Persons who supply their labor to neither the formal nor the informal sector therefore also have low earnings capacities in both sectors, which is consistent with the comparative advantage hypothesis.

The coefficients of the term m_1, which is the equivalent of the inverse mills ratio associated with the selection into the informal sector, and the m_2 term, which is the equivalent of the inverse mills ratio associated with selection into the informal sector, are also compatible with self-selection according to comparative advantages. The coefficients on the m_1 term are insignificant for the social protection definition of informality irrespective of whether self-employed are included in the sample or not, but the significant coefficients for the legalistic definition imply that persons, who (after controlling for observable differences) have a higher probability to work in the informal sector also earn (ceteris paribus) higher wages in the informal sector. By contrast, they could expect or significantly lower wages in the formal sector. The estimates for the m_2 term suggest that persons, who (after controlling for observables) have an above average probability to work in the formal sector, do not receive above average wages there, but would receive significantly below average wages in the informal sector in both definitions of the informal sector and irrespective of whether the self-employed are included in the sample or not.

Finally, the test statistics reported in the bottom of table 4 suggest that with our wage regression – depending on specification – 18% to 26% of the variance in individual level wages can be explained. More importantly they also suggest that the model is also well specified in other respects. In particular Leung and Yu (1996) in the context of standard Heckman-models show that co-linearity between the inverse mills ratio and other variables, which arises from badly specified exclusion restrictions and/or a large number of missing observations on wages for the non-employed may invalidate the use of such models. Further they argue that the condition number is a reliable indicator of such co-linearity and suggest that a condition number below 20 should indicate the absence of co-linearity issues. Belsley, Kuh and Welsch (1980) by contrast consider 30 to be the maximum which can be allowed in the context of a Heckman model (see Puhani, 2000 for a discussion). While these rules of thumb were developed for the Heckman model and we are not aware of studies focusing on the issue of potential co-linearity in the multinomial selection model we are using, the condition numbers in our application are well below the more conservative level suggested by Leung and Yu (1996) and thus indicate that co-linearity at least does not seem to be a major problem in our specification

4.3 Selectivity Corrected Oaxaca decompositions

The results of the wage regression in table 4 can also be used to analyze informal-formal sector wage differential by the familiar (selectivity corrected) Oaxaca-Blinder decomposition. In particular the average wage in sector k can be written as $\bar{w}_k = \bar{X}_k \hat{\beta}_k + \hat{\tau}_{kk} \bar{\mu}_{kk} + \sum_{k'} \hat{\tau}_{kk'} \bar{\mu}_{1i}$ where bars over variables represent means, hats indicate estimated parameters and the $\bar{\mu}_{kk}$ and $\bar{\mu}_{kk'}$ are the means of the selectivity terms $m(P_k)$ and $m(P_{k'}) \frac{P_{k'}}{P_{k'}-1}$ in equation (3). Thus, the difference in average informal and formal sector wages can be written as:

$$\bar{w}_i - \bar{w}_f = \left[(\bar{X}_i - \bar{X}_f)\hat{\beta}_i \right] + \left[(\hat{\beta}_i - \hat{\beta}_f)\bar{X}_f \right]$$
$$+ [\hat{\tau}_{i0}\bar{\mu}_{ii} - \hat{\tau}_{0f}\bar{\mu}_{0f}] + [\hat{\tau}_{ii}\bar{\mu}_{ii} - \hat{\gamma}_{if}\bar{\mu}_{if} + \hat{\gamma}_{fi}\bar{\mu}_{fi} - \hat{\gamma}_{ff}\bar{\mu}_{ff}] \tag{4}$$

This equation decomposes the informal-formal sector wage differences into four separate terms (see Yun, 2005 and Madden, 2000). The first of these $((\bar{X}_l - \bar{X}_f)\hat{\beta}_i)$ is due to the differences in mean characteristics of workers in these two sectors. It therefore estimates the share of wage differences due to selection on observables of workers in the two sectors. The second term in square brackets $((\hat{\beta}_i - \hat{\beta}_f)\bar{X}_f)$, by contrast, is due to difference in parameters between the two sectors and measures the share of sector wage differentials due to differences in returns to productive characteristics in these sectors. Finally, the terms in the third and fourth square brackets $([\hat{\tau}_{i0}\bar{\mu}_{ii} - \hat{\tau}_{0f}\bar{\mu}_{0f}] + [\hat{\tau}_{ii}\bar{\mu}_{ii} - \hat{\gamma}_{if}\bar{\mu}_{if} + \hat{\gamma}_{fi}\bar{\mu}_{fi} - \hat{\gamma}_{ff}\bar{\mu}_{ff}])$ are the selectivity corrections in this decomposition. They measure the contribution of self-selection on unobservable characteristics to inactivity and between sectors to the difference in average wages between the formal and the informal sector.[8] This decomposition thus provides insights as to what factors are most closely associated with the higher informal than formal sector wages in Tajikistan, after correcting for the selectivity of participation of either one of these sectors.

The results (in table 5) suggest that the self-selection of workers on unobservable characteristics is the main factor contributing to higher informal sector wages in Tajikistan. In both formal sector definitions the selection between sectors effect (i.e. the term $\hat{\tau}_{ii}\bar{\mu}_{ii} - \hat{\gamma}_{if}\bar{\mu}_{if} + \hat{\gamma}_{fi}\bar{\mu}_{fi} - \hat{\gamma}_{ff}\bar{\mu}_{ff}$) contributes significantly positively and sizably to the wage differences between the sectors in all specifications. Similarly, the selection into work effect (i.e. the term $\hat{\tau}_{i0}\bar{\mu}_{ii} - \hat{\tau}_{0f}\bar{\mu}_{0f}$) contributes significantly to higher informal sector wages, with the size of this effect exceeding the total log wage differences between the two sectors in all specifications. The differences in returns to observed characteristics (differences in parameter effect) as well as the differences in characteristics effect, by contrast, both work to statistically significantly reduce wage differences between the sectors in all specifications. The coefficient of the difference in characteristic effect, however, is rather small, while the difference in parameters effect attains a higher value only account of the impact of the constant term on this effect. In addition, the difference in characteristics effects for individual characteristics suggests that the self-selection of highly educated and older workers to the formal sector contributes significantly to reducing wage differences between the two sectors

---

[8] To avoid sensitivity of detailed decompositions to the choice of base categories for dummy variables, we follow Yun (2008) and parameterize estimates such that coefficients sum to zero.

as does the increased share of informal sector workers in Dunshabe. The self-selection of women to the formal sector, by contrast, works to significantly increase sector wage differentials. With respect to the differences in parameters effect, by contrast, only the lower returns to education in the informal sector increase wage differentials between the two sectors in all specifications, while in the social protection definition, when considering employees only, the higher returns to age in the formal sector statistically significantly reduce sector wage differentials, but the higher ethnic and gender wage differential in the informal sector increases them.

Table 5 Selectivity corrected Oaxaca decompositions of informal-formal sector wage differentials (hourly wages)

	All employed						Employees only					
	Social protection definition			Legalistic definition			Social protection definition			Legalistic definition		
	Diff.		S.E.	Diff.		S.E.	Diff.		S.E.	Diff.		S.E.
Total difference	0.497	***	0.028	0.212	***	0.029	0.434	***	0.033	0.047		0.033
Selection to work	1.126	***	0.474	1.976	***	0.526	1.615	***	0.497	2.074	***	0.514
Selection between sectors	1.682	***	0.485	2.074	***	0.506	2.241	***	0.594	3.484	***	0.605
Diff. in characteristics	-0.078	***	0.039	-0.121	***	0.026	-0.085	**	0.041	-0.136	***	0.026
Age	-0.035	**	0.010	-0.026	**	0.009	-0.038	***	0.011	-0.028	***	0.009
Education	-0.163	***	0.025	-0.145	***	0.018	-0.168	***	0.028	-0.145	***	0.018
minorities	-0.002		0.001	0.000		0.001	-0.002		0.001	0.001		0.001
Gender	0.133	***	0.027	0.056	***	0.016	0.139	***	0.026	0.050	***	0.016
Capital	-0.010	***	0.003	-0.006	***	0.002	-0.017	***	0.003	-0.013	***	0.003
Diff. in parameters	-2.232	**	1.197	-3.717	***	1.287	-3.338	***	1.308	-5.376	***	1.370
Age	-0.771		0.593	0.517		0.577	-2.069	**	0.706	-1.189		0.575
Education	0.252	***	0.074	0.173	***	0.052	0.256	***	0.096	0.160	***	0.056
Minorities	0.087		0.041	-0.001		0.041	0.099	**	0.046	0.011		0.042
Gender	0.161		0.112	0.190		0.110	0.192	*	0.104	0.129		0.091
Capital	-0.018		0.032	-0.067		0.034	0.068		0.042	0.048		0.031
Constant	-1.943	*	1.030	-4.529	**	1.143	-1.885	*	1.091	-4.535	***	1.237

Source: TSLS 2007 16 to 64 years old. Table reports contributions of the individual variables to total informal-formal sector wage differential based on results in table 5, ***, (**), (*) show significance at the 1%, (5%), (10%) level. S.E.=heteroskedasticity robust standard error of the estimate

4.4 Results for structural form multinomial logit regressions

Finally, the estimates of equation (3) can also be used to estimate structural multinomial logit models for the impact of relative wages in the two sectors on the choice of sector employment. Such an estimate would be biased if observed wages were used as an explanatory variable, but based on the estimates for equation (2) unbiased predictions of the unobserved expected wages of individuals in each sector (including those in which they are not employed) can be derived. As demonstrated by Lee (1979) these predicted wages can then be used to estimate an (unbiased) structural (multinomial) selection equation. This provides consistent estimates of the impact of wages in different sectors on the choice of sector of employment.

Table 6 reports the results of the structural form multinomial logit regression for choosing a particular sector, based on predicted wages derived from the results in

table 4.[9] These results suggest a rather strong impact of relative wages on formal and informal sector labor supply. An increase of informal sector wages by one percent significantly reduces the non-employment probability by 0.24 percentage points and the formal sector employment probability by 0.11 percentage points but increases the informal sector employment probability by 0.35 percentage points, when considering the social protection definition of the formal sector and including the self-employed in the sample. Similarly, in this specification an increase in the formal sector wage by one percent, statistically significantly reduces the non-employment probability by 0.19 percentage points and the informal sector employment probability by 0.21 percentage points, but increases the formal sector employment probability by 0.40 percentage points.

These marginal effects are, however, slightly lower when either considering the legalistic definition of the formal sector. Thus, when focusing on the legalistic definition and including the self-employed an increase of informal sector wages by one percent significantly reduces the non-employment probability by 0.17 percentage points and the formal sector increases the informal sector employment probability by 0.19 percentage points, while it has no significant impact on the formal sector employment probability. An increase in formal sector wages by one percent, by contrast, statistically significantly reduces the non-employment probability by 0.24 percentage points but increases the formal sector employment probability by 0.24 percentage points without having a statistical significant impact on the informal sector employment probability.

When excluding the self-employed in case of the social protection definition these marginal effects are somewhat smaller but are largely consistent with those when including the self-employed. A one percent increase in informal sector wages, reduces the non-employment probability by 0.16 percentage points and the formal sector employment probability by 0.06 percentage points and increases the formal sector employment probability by 0.22 percentage points. A one percent increase in formal sector wages reduces the non-employment probability by 0.23 percentage points and the informal sector employment probability by 0.15 percentage points and increases the formal sector employment probability by 0.38 percentage points.

By contrast the effects increase slightly in the legalistic definition when excluding the self-employed although and in addition also the impact of an increase in formal sector wages on the non-employment probability is statistically insignificant. Here a one percent increase in informal sector wages, reduces the non-employment probability by 0.26 percentage points and increases the formal sector employment probability by 0.22 percentage points and a one percent increase in formal sector wages, reduces the informal sector employment probability by 0.14 percentage points and increases the formal sector employment probability by 0.20 percentage points.

---

[9] In this regression, all variables in the wage equation are excluded to avoid co-linearity with wages.

Table 6 Marginal effects of structural multinomial selection equation estimates wages)

	P(non-employed)			P(informal sector)			P(formal sector)		
	Coeff.		S.E.	Coeff.		S.E.	Coeff.		S.E.
All employed									
Social protection definition									
Predicted wage in informal sector	-0.240	***	0.020	0.347	***	0.016 a)	-0.107	***	0.016
Predicted wage in formal sector	-0.186	***	0.030	-0.210	***	0.024 a)	0.396	***	0.023
Very good health	Base category								
Good health	-0.109	***	0.015	0.117	***	0.013 a)	-0.008		0.012
Fair or bad health	-0.070	***	0.019	0.122	***	0.016 a)	-0.051	***	0.015
Married	0.015		0.011	0.007		0.009 b)	-0.021	**	0.009
Children under 6	-0.009		0.009	0.034	***	0.008 a)	-0.025	***	0.007
Observations				13539					
Log Likelihood				-12245.563					
Legalistic definition									
Predicted wage in informal sector	-0.173	***	0.030	0.186	***	0.024 a)	-0.013		0.024
Predicted wage in formal sector	-0.243	***	0.036	0.005		0.028 a)	0.238	***	0.029
Very good health	Base category								
Good health	-0.087	***	0.015	0.072	***	0.013 a)	0.015		0.012
Fair or bad health	-0.068	***	0.022	0.132	***	0.019 a)	-0.064	***	0.018
Married	0.020	**	0.011	-0.037	***	0.009 a)	0.016	*	0.009
Children under 6	0.002		0.009	0.023	***	0.008 a)	-0.025	***	0.008
Observations				13539					
Log Likelihood				-12436.208					
Employees only									
Social protection definition									
Predicted wage in informal sector	-0.164	***	0.014	0.219	***	0.010 a)	-0.055	***	0.012
Predicted wage in formal sector	-0.225	***	0.027	-0.152	***	0.019 a)	0.378	***	0.021
Very good health	Base category								
Good health	-0.104	***	0.015	0.093	***	0.011 a)	0.011		0.013
Fair or bad health	-0.035	*	0.019	0.073	***	0.014 a)	-0.038	**	0.016
Married	0.006		0.012	0.017	**	0.008 a)	-0.023	**	0.010
Children under 6	0.012		0.009	0.009		0.007 a)	-0.021	***	0.008
Observations				12277					
Log Likelihood				-10332.482					
Legalistic definition									
Predicted wage in informal sector	-0.262	***	0.030	0.221	***	0.021 a)	0.041		0.026
Predicted wage in formal sector	-0.062		0.041	-0.140	***	0.029 a)	0.202	***	0.034
Very good health	Base category								
Good health	-0.119	***	0.016	0.087	***	0.012	0.0318	**	0.014
Fair or bad health	-0.127	***	0.025	0.162	***	0.019 a)	-0.035		0.022
Married	0.004		0.011	-0.023	***	0.007 a)	0.0193	**	0.009
Children under 6	0.013		0.009	0.012	*	0.007 a)	-0.025	***	0.008
Observations				12277					
Log Likelihood				-10465.725					

Source: TSLS 2007, 15 to 60 year olds. ***, (**), (*) show significance at the 1%, (5%), (10%) level. Coeff.=marginal effects, S.E.=heteroskedasticity robust standard error of the estimate. a), (b), (c) Null hypothesis of equal effects for formal and informal sector employment can be rejected at the 1%, (5%), 10% level.

Furthermore, the control variables largely reconfirm the results found in the reduced form regressions. The only major differences are that the structural form equations suggest that in all specifications persons with less than 6-year-old children in their household have a statistically significantly higher informal sector employment probability and lower formal sector probability when including the self-employed in both formal sector definitions and in the legalistic definition of the informal sector when excluding the self-employed. Furthermore, persons who report that their health is fair or bad now have a significantly lower non-employment and a higher informal sector employment probability than persons stating to have very good health in all specifications.

5. Conclusions

This paper uses Tajikistan as an example of a poor transition economy, where, informal sector wages are higher than formal ones, and in which the informal sector, according to World Bank (2007) estimates, accounts for more than half of total employment, to analyze the self-selection of workers into inactivity as well as formal and informal sector employment. The results suggest that stylized facts broadly accord with self-selection on comparative advantages both with respect to observable and unobservable characteristics. As a rule, selection of workers on age, education, gender and ethnicity to the sectors follows the estimated differences in relative returns to these observable characteristics in the two sectors. Also the results with respect to selection on unobservable characteristics are broadly consistent with self-selection on comparative advantages. The only exception are employees working in Dunshabe for whom the selection to the informal sector despite higher returns to working in the formal sector is difficult to reconcile with the comparative advantage hypothesis.

We also find a number of differences between formal and informal sector wage setting, such as that returns to education as well as to experience are higher in the formal than the informal sector and that wage penalties for women and non-Tajik workers are higher in the informal than the formal sector. In total, however, Oaxaca-Blinder decompositions suggest that the self-selection of workers on unobservable characteristics is the main reason for higher wages in the informal than the formal sector.

Furthermore, the results indicate a rather strong impact of relative wages in the two sectors on the decision of workers to work in the formal or informal sector. An increase in informal sector wages by 1 percent increases the probability of working in the informal sector by 0.22 to 0.35 percentage points but reduces the probability of working in the formal sector by up to 0.13 percentage points. An increase in formal sector wages by 1 percent increases the probability to work in the formal sector by 0.20 to 0.40 percentage points but reduces the probability of working in the informal sector by up to 0.21 percentage points when the social protection definition of informality is applied but has no significant impact on informal sector employment in the legalistic definition.

These results are also largely robust to using alternative definitions of the formal sector and to excluding the self-employed from the analysis. They thus suggest that policies that increase wages in the formal sector (or reduce those in the informal sector) would in all likelihood also be effective in reducing the extremely high share of the informal sector employment in Tajikistan in aggregate. The results with respect to differences in formal and informal sector wages in Dunshabe, however, suggest that for certain sub-segments of the labor market segmentation may be important. Future research could focus on taking a more disaggregate approach to identifying labor market segmentation between the formal and informal sector and could also extend the analysis of this paper to later time periods or other countries to be better able to assess the generality of the results.

APPENDIX

Table A1 in this appendix provides some additional specifications of the wage equation. In particular, first of all we ran wage equations for the overall sample (without selectivity control) but including a dummy for informal sector employment. This was done to see whether informal-formal sector wage differences also persist after controlling for other observable characteristics. The results (reported in the columns labeled all observations in table A1) suggest that this is indeed the case. According to the results after controlling for observable differences the informal-formal sector log-wage differential is between 0.49 (in the social protection definition) and 0.23 (in the legalistic definition). This compares rather well to the wage differential before controlling to for observable characteristic, which is 0.50 and 0.20, respectively (see table 1)

Second, we also estimated sector specific wage regressions without controlling for selectivity, to assess the impact of omitting the selectivity bias on results. Comparing these results (reported in the columns headed social protection definition and legalistic definition in table A1) suggest that not controlling for selectivity does indeed bias coefficients substantially. In particular, returns to age and education are much higher in the formal sector after accounting for self-selection, while the gender differences in the informal sector are much smaller. This thus justifies our approach as it is well known that estimates of wage regressions without controlling for self-selection are biased and inconsistent.

Table A1 Alternative specifications of the wage regression (dependent variable log hourly wages, sample all employed)

	All observations		Social protection Definition		Legalistic definition	
	Social protection definition	Legalistic definition	Informal sector	Formal sector	Informal sector	Formal sector
ln(age)	0.22***	0.17***	0.31***	0.14***	0.32***	0.019
	(0.040)	(0.041)	(0.060)	(0.050)	(0.060)	(0.055)
Basic education	Base category					
Secondary education	0.23***	0.24***	0.19***	0.26***	0.24***	0.22***
	(0.039)	(0.040)	(0.055)	(0.051)	(0.055)	(0.056)
Tertiary education	0.57***	0.48***	0.39***	0.71***	0.35***	0.56***
	(0.046)	(0.046)	(0.075)	(0.056)	(0.073)	(0.060)
Not Tajik	-0.12***	-0.14***	-0.19***	-0.047	-0.20***	-0.082*
	(0.033)	(0.034)	(0.051)	(0.041)	(0.049)	(0.047)
Male	0.67***	0.75***	0.91***	0.46***	0.86***	0.63***
	(0.028)	(0.028)	(0.045)	(0.034)	(0.042)	(0.037)
Capital city	0.32***	0.33***	0.29***	0.34***	0.32***	0.34***
	(0.030)	(0.030)	(0.043)	(0.041)	(0.044)	(0.042)
Informal sector dummy	0.49***	0.23***				
	(0.027)	(0.028)				
Constant	-0.49***	-0.23	-0.43**	-0.17	-0.54***	0.37*
	(0.140)	(0.140)	(0.210)	(0.170)	(0.200)	(0.190)
N	6043	6043	3208	2835	3071	2972
R-sq	0.203	0.170	0.165	0.189	0.182	0.152

Source: TSLS, 16 to 64 years old, all employed. Values in brackets are (heteroskedasticity robust) standard errors of the estimate ***, (**), (*) show significance at the 1%, (5%), (10%) level.

REFERENCES

Abdulloev I, Gang IN, Landon-Lane J (2012): Chapter 6 Migration as a Substitute for Informal Activities: Evidence from Tajikistan. In Lehman H, Tatsiramon K (eds): *Informal Employment in Emerging and Transition Economies*. Emerald Group, Bingley, UK, 205-227.

Amaral P, Quintin E (2006): A competitive model of the informal sector, *Journal of Monetary Economics,* 53(7):1541–1553.

Bargain O, Kwenda P (2015): The Informal Sector Wage Gap: New Evidence Using Quantile Estimations on Panel Data. *Economic Development and Cultural Change*, 63(1):117-153.

Belsley DA, Kuh E, Welsch RE (1980): *Regression Diagnostics, Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, New York.

Bennett J (2011): Informal Production and Labour Market Segmentation. *Journal of Institutional and Theoretical Economics*, 167(4):686-707.

Bennett J, Estrin S (2010): Informal firms in developing countries: entrepreneurial stepping stone or consolation prize? *Small Business Economics*, 34(1):53-63.

Bosch M, Maloney WF (2005): Gross Worker Flows in the Presence of Informal Labor Markets: Evidence from Mexico. 1987-2002. *IZA Discussion Papers*, no. 4864.

Bourguignon F, Fournier M, Gurgand M (2007): Selection Bias Corrections Based on the Multinomial Logit Model: Monte Carlo Comparisons. *Journal of Economic Surveys*, 21(1):174-205.

Brülhart M, Koenig P (2006): New economic geography meets Comecon. *The Economics of Transition*, 14(2):245-267.

Cunningham W, Maloney W (2001): Heterogeneity in the Mexican micro-enterprise sector: an application of factor and cluster analysis. *Economic Development and Cultural Change*, 50(1):131–156.

Dahl GB (2002): Mobility and the returns to education: testing a Roy model with multiple markets. *Econometrica*, 70(6):2367-2420.

Danzer AM, Dietz B, Gatskova K (2013): Migration and Remittances in Tajikistan: Survey Technical Report. *Institut für Ost- und Südosteuropaforschung Working Papers*, no. 327.

Danzer AM, Ivaschenko O (2010): Migration patterns in a remittances dependent economy: Evidence from Tajikistan during the global financial crisis. *Migration Letters*, 7(2):190-202.

Dickens W, Lang K (1985): A test of dual labor market theory. *The American Economic Review*, 67(2):792–805.

Dietz B, Gatskova K, Ivlevs A, (2015): Emigration, Remittances and the Education of Children Staying Behind: Evidence from Tajikistan, *Institut für Ost- und Südosteuropaforschung Working Papers*, no. 354.

Dimova R, Gang IN, Landon-Lane J (2011): Revealed Informal Activity. *IZA Discussion Papers*, no. 5607.

Dimova R, Gang IN (2007): Self-selection and wages during volatile transition. *Journal of Comparative Economics*, 35(3):612-629.

Dubin JA, McFadden DL (1984): An econometric analysis of residential electric appliance holdings and consumption. *Econometrica*, 52(2):345-362.

Gindling T (1991): Labor market segmentation and the determination of wages in the public, private-formal and informal sectors in San-Jose, Costa-Rica. *Economic Development and Cultural Change*, 39(3):585–603.

Greene WH (1993): *Econometric Analysis*, 2nd Edition, Prentice – Hall, London

Günther I, Launov A (2012): Informal employment in developing countries: Opportunity or last resort? *Journal of Development Economics*, 97(1):88-98.

Hazans M (2011): What explains prevalence of informal employment in European countries: the role of labor institutions, governance, immigrants, and growth. *World Bank Policy Research Working Paper Series*, no. 5917.

Heckman J, Sedlacek G (1985): Heterogeneity, aggregation and market wage functions: an empirical model of self-selection in the labor market. *Journal of Political Economy*, 93(6):1077–1125.

Henley A, Arabsheibani R, Carneiro FG (2009): On Defining and Measuring the Informal Sector. *World Development*, 37(5):992-1003.

ILO (2015): *Decent Work Country Program of the Republic of Tajikistan 2015-2017*. International Labour Organization, Geneva.

John K, Thomsen SL (2014): Heterogeneous Returns to Personality - The Role of Occupational Choice. *Empirical Economics*, 47(2):553-592.

Justino P, Shemyakina O (2012): Remittances and labor supply in post-conflict Tajikistan. *IZA Journal of Labor & Development*, 1(1):1-28.

Kanbur R (2009): Conceptualising Informality: Regulation and Enforcement. *IZA Discussion Papers*, no. 4186.

Karnite R (2010): Analysis of the Current Situation of the Labour Market of Tajikistan, *manuscript*, Economic Prognosis Institute, Dunshabe.

Lee F (1979): Identification and Estimation of Binary Choice Modes with Limited (Censored) Dependent Variables. *Econometrica*, 47(5):977-996.

Lee F (1983): Generalized econometric models with selectivity. *Econometrica*, 51(2):507-512.

Lehmann H (2015): Informal Employment in Transition Countries: Empirical Evidence and Research Challenges. *Comparative Economic Studies*, 57 (1):1-30.

Lehmann H, Razzolini T, Zaiceva A (2012): Job Separations and Informality in the Russian Labor Market. *Research in Labor Economics*, 34, 257-290.

Lehmann H, Pignatti N (2007): Informal Employment Relationships and Labor Market Segmentation in Transition Economies: Evidence from Ukraine. *IZA Discussion Paper*, no. 3269.

Lehmann H, Muravyev A (2014): Labor Market Institutions and Informality in Transition and Latin American Countries. In Froelich M, Kaplan D, Pages C, Rigolini J, Robalino D (Eds.): *Social Insurance and Labor Markets: How to Protect Workers while Creating Good Jobs*, Oxford University Press, Oxford, 375-410.

Leung SF, Yu S (1996): On the choice between sample selection and two-part models. *Journal of Econometrics*, 72(1–2):197–229.

Loayza N, Rigolini J (2006): Informality trends and cycles, *The World Bank Policy Research Working Paper*, no.4078.

Madden D (2000): Towards a Broader Explanation of Male-Female Wage Differences. *Applied Economics Letters*, 7(12):765-70.

Magnac T (1991): Segmented or competitive labor markets, *Econometrica*, 59(1):165-187.

Maloney W (1999): Does informality imply segmentation in urban labor markets? Evidence from sectoral transitions in Mexico. *The World Bank Economic Review*, 13(2):275–302.

Maloney W (2004): Informality revisited. *World Development*, 32(7):1159–1178.

Marcouiller D, Ruiz de Castilla V, Woodruff C (1997): Formal Measures of the Informal-Sector Wage Gap in Mexico, El Salvador, and Peru. *Economic Development and Cultural Change*, 45(2):367-92.

McFadden DL (1973): Conditional Logit Analysis of Qualitative Choice Behavior. In: Zarembka P (eds): *Frontiers in Econometrics*, Academic Press, New York, 104-142.

Mussurov A, Arabsheibani GR (2015): Informal self-employment in Kazakhstan. *IZA Journal of Labor & Development*, 4(1):1-19.

Nguyen HC, Nordman CJ, Roubaud F (2013): Who Suffers the Penalty? A Panel Data Analysis of Earnings Gaps in Vietnam. *Journal of Development Studies*, 49(12):1694-1710.

Pagés C, Stampini M (2009): No Education, No Good Jobs? Evidence on the Relationship between Education and Labor Market Segmentation. *Journal of Comparative Economics*, 37(3):387-401.

Paulson A, Townsend R (2005): Financial constraints and entrepreneurship: evidence from the Thai financial crisis. *Economic Perspectives*, 3Q/2005:34–48.

Perry G (Ed.) (2007): *Informality: Exit and Exclusion*, World Bank, Washington DC.

Pratap S, Quintin E (2006): Are labor markets segmented in developing countries? A semiparametric approach. *European Economic Review*, 50(7):1817–1841.

Puhani P (2000): The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys*, 14(1):53-68.

Staneva A, Arabsheibani G (2014): Is there an informal employment wage premium? Evidence from Tajikistan. *IZA Journal of Labor & Development*, 3(1):1-24.

Tansel A, Kan EO (2012): The Formal/Informal Employment Earnings Gap: Evidence from Turkey. *IZA Discussion Papers*, no. 6556.

The World Bank (2007): *World Bank, Informality: Exit and Exclusion*, The World Bank, Washington D.C.

Wu MW, Shenb CH (2013): Corporate social responsibility in the banking industry: Motives and financial performance. *Journal of Banking & Finance*, 37(9):3529–3547.

Yun MS (2005): A Simple Solution to the Identification Problem in Detailed Wage Decompositions. *Economic Inquiry*, 43(4):766-72.

Yun MS (2008): Identification problem and detailed Oaxaca decomposition: A general solution and inference. *Journal of Economic and Social Measurement*, 33(1), 27-38.