LCERPA

Laurier Centre for Economic Research & Policy Analysis

LCERPA Working Paper No. 2018-7

February 2018

Business Formalization in Vietnam

Brian McCaig,
Department of Economics, Wilfrid Laurier University

and

Jordan Nanowski,
Department of Economics, Wilfrid Laurier University

Business Formalization in Vietnam

Brian McCaig
Wilfrid Laurier University
75 University Avenue West
Waterloo, ON
N2L 3C5
Canada
bmccaig@wlu.ca

Jordan Nanowski
Wilfrid Laurier University
75 University Avenue West
Waterloo, ON
N2L 3C5
Canada
jordan.nanowski@gmail.com

February 2018

Abstract

We estimate the impact of business formalisation using nationally representative panel data on businesses in Vietnam. Our data allows us to observe businesses for two surveys prior to obtaining a license and hence to control for differential trends before formalisation. We find that obtaining a license is not associated with an increase in profits or other business outcomes such as revenue, expenses, and employment once we control for differential trends. Controlling for trends is crucial, as estimates that ignore trends consistently find a larger positive association between becoming licensed and business performance. Our results suggest that inducing more businesses to register is unlikely to bring about large-scale changes for these businesses.

Keywords: Informal, Formalisation, Asia, Vietnam, Household business

---

The authors would like to thank David McKenzie, John Rand, two anonymous referees and the editor for their helpful suggestions. The dataset used for the analysis was originally created for research funded by the Growth and Labour Markets in Low Income Countries Programme by the Institute of Labor Economics (IZA) and the U.K. Department for International Development. McCaig also acknowledges funding from the Laurier Centre for Economic Research and Policy Analysis and the authors wish to thank Nicholas Curtis, Kelsey Lise, and Katherine Ziomek for excellent research assistance in cleaning the data. The do files necessary for creating the dataset and the subsequent analysis can be found on McCaig's website.

1. Introduction

There is a large number of small, household-run, low-productivity businesses in most low-income countries (Banerjee & Duflo, 2007; Gollin, 2008; La Porta & Shleifer, 2014). These businesses are an important source of employment and household income. Hence, improved microenterprise business performance could have important household welfare impacts.[1] A common feature of these businesses is that they are informal, that is, not registered with the government. This raises an important question: is informality preventing these firms from growing?[2]

This paper estimates the impact of formalisation on business performance in Vietnam. We use nationally representative panel data on formal and informal businesses run by households. We use three rounds of the Vietnam Household Living Standards Survey (VHLSS) to construct a three-survey panel across four years.[3] We focus on businesses that are informal in the first two surveys. Hence, we are able to observe changes in business performance *prior* to formalisation. This turns out to be important, as businesses that will subsequently formalise are different from those that do not formalise not only at a point in time, but also in terms of growth prior to formalisation.

We find that after controlling for differential trends prior to formalisation, obtaining a license is not associated with an increase in profits. Indeed, the point estimate is very close to 0 and slightly negative in most regressions. In contrast, if we only used two surveys of data, one prior to and one after formalisation, we estimate an increase in profits of 8.6 to 10.5% in

[1] Improvements in productivity within these businesses or the reallocation of resources away from these businesses may also have implications for aggregate productivity (Hsieh & Klenow, 2009; Restuccia & Rogerson, 2017).
[2] Other constraints could also be hindering or preventing growth among these businesses. See surveys by Banerjee (2013) and Banerjee, Karlan and Zinman (2015) for the impacts of microcredit and McKenzie and Woodruff (2014) for business training.
[3] Few existing studies use nationally representative data. See section 2 for further details on existing literature.

association with formalisation. This suggests that existing non-experimental estimates based on two periods of data may suffer from correlation between the decision to become licensed and unobserved trends. When we examine the impact of formalisation on revenue and various expenditures by the businesses, we similarly find small, close to 0 effects for most outcomes. Lastly, there is some imprecise evidence that the number of workers in businesses grow in response to formalisation, but few businesses make the transition to start hiring paid employees as opposed to household members.

Our results suggest that becoming licensed does not generally alter the dynamics of the business relative to prior to becoming licensed. In short, the growth trajectory of the business is not changed by becoming licensed. These results are consistent with the growing body of evidence that suggests most informal microenterprises in low-income countries are unlikely to significantly expand in scale or become more productive (Banerjee, 2013; Banerjee, Karlan & Zinman, 2015; Bruhn & McKenzie, 2014; McKenzie & Woodruff, 2014). Businesses that stay informal and those that formalise generally stay quite small and there is not a significant difference (either economically or statistically) between the two groups. However, our results stand in contrast to many non-experimental estimates of the impact of formalisation on business performance.[4] Hence, we contribute to this literature by suggesting an explanation for the differences in results between estimates from randomised control trials (RCTs) and non-RCTs. Non-RCTs, particularly those based on two-survey panel data, may still suffer from correlation between the decision to formalise and unobserved trends among businesses.

[4] McKenzie and Sakho (2010) estimate an increase in profits of 88% among informal Bolivian firms and Demenet et al. (2016) estimate an increase in value added of 20% among informal Vietnamese firms. See section 2 for further discussion.

Our research makes important contributions to the existing literature. The use of three-survey panel data provides a more credible non-experimental way to identify the impacts of formalisation on business performance than existing two-survey panel based analysis. Additionally, our research helps to reconcile the differences in estimates between RCTs, which generally find very small effects, if any; and panel based results, which find impacts in the range of 15-20% increase in profits or value added. The latter studies may suffer from unobserved trends that are correlated with formalisation and business performance. Finally, we use nationally representative data that covers all industries, all locations, both urban and rural, and the full size range of informal firms including a large share of sole proprietor firms.

In section 2, we present a brief literature review that highlights conceptual issues on the choice to formalise and discusses existing empirical results. In section 3, we provide a detailed description of the survey data used in our analysis followed by descriptive statistics in section 4. In section 5, we describe our econometric methodology and then present our results in section 6. We briefly conclude in section 7.

2. Literature Review

The informal economy accounts for thirty to forty percent of total economic activity in the poorest countries and contains a large number of self-employed workers living at near-subsistence levels (La Porta & Shleifer, 2014). Existing literature generally associates firm informality with low profits and productivity, limited credit access, the absence of official employment contracts and limited or no social security for employees (Banerjee & Duflo, 2011; Rand & Torm, 2012).

Conceptually, the decision to formalise is endogenous, which makes estimating the causal effect of formalisation on business performance challenging. Informality is a choice by

firms based on the perceived costs and benefits of formality relative to informality. Firms may compare potential benefits of formalisation, such as better access to credit, infrastructure and other productive public goods, broader customer base, and lower risk of fines, with the potential costs, such as paying taxes, following regulations, and less employment and production flexibility (Bruhn & McKenzie, 2014; McKenzie & Sakho, 2010; Farrell, 2004; Almeida & Carneiro, 2009; Ulyssea, 2017). Given the large size of the informal economy, one may assume that many firms do not feel that it is beneficial to leave the informal economy via formalization, otherwise they would have done so. There is evidence of this perception in Vietnam as nearly half of informal household businesses surveyed in Ha Noi and Ho Chi Minh City believe that there is no benefit to obtaining a license (Cling, Razafindrakoto, and Rouband, 2012).

Empirical literature on estimating the benefits of formalisation can be divided into two main categories: experimental and non-experimental studies. Experimental studies involve the use of randomised control trials, where interventions attempt to randomly incentivise firms to formalise by providing them with information and/or lower costs of registering. The majority of experimental studies have found that few firms are induced to formalise in response to a variety of different interventions (e.g., registration subsidies, provision of information on how to register) and that formalisation has a relatively small effect on firm performance (De Giorgi & Rahman, 2013; de Mel, McKenzie, & Woodruff, 2013; Bruhn & McKenzie, 2014; de Andrade, Bruhn, & McKenzie, 2014; Benhassine, McKenzie, Pouliquen, & Santini, 2016; Galiani, Meléndez, & Ahumada, 2016).[5] Non-experimental studies can be grouped into two main types. First, there are those that use cross-sectional analysis and employ instrumental variable

[5] Campos, Goldstein, and McKenzie (2015) is an exception to this general pattern. They find a large share of businesses in Malawi are induced to register in response to randomised treatments that either reduce the cost of registration, provided business training, or provided tax services.

(McKenzie & Sakho, 2010) or regression discontinuity approaches (Fajnzylber, Maloney, & Montes-Rojas, 2011). Both of these types of studies find large impacts of formalisation on business performance. McKenzie & Sakho (2010) find an increase in profits of 88% while Fajnzylber et al. (2011) find an increase in profits of at least 45%, and in many specifications a much larger effect. These studies are estimating the local average treatment effect for a subpopulation of firms who are on the margin of formalisation. A second type of non-experimental studies use panel data to control for unobserved time invariant heterogeneity that may be correlated with both business performance and the decision to become licensed. These studies in general find more modest effects on business performance. Rand and Torm (2012) report an increase in profits of 12 to 16% while Demenet, Razafindrakoto, and Rouband (2016) report a growth of 20% in value added due to formalisation. The impact of formalisation estimated by these studies is for the entire subpopulation of informal firms that chose to formalise in a given period, not just those firms on the margin of formalisation.

When attempting to identify the causal impact of formalisation, endogeneity must be addressed. Household businesses that choose to formalise may differ from businesses that remain informal in certain observed and unobserved characteristics, which may cause them to be incomparable. Reverse causality may be of concern, since it is possible that businesses that experience higher performance may be more likely to register for a household business license. For example, a firm that experiences higher performance may be more visible to authorities, which may incentivise firms to formalise in order to avoid paying fines and/or bribes (Fajnzylber et al., 2011). Omitted variable bias may also be of concern, since other characteristics aside from firm performance may influence both a firm's decision to formalise and its performance. The

resulting selection effect might be explained by observable factors such as manager gender and education, as well as the industry/location in which the firm operates.

Our paper is most closely related to other studies using panel data. This approach allows for fixed effects models, which are able to remove time invariant unobserved factors. However, there may remain unobserved time-variant heterogeneity such as firm-specific time trends. For example, if the firm is experiencing productivity growth, this may be correlated with both business outcomes, such as profits or revenue, and the decision to become licensed. If businesses that decided to become licensed are experiencing faster productivity growth, then this trend is correlated with the decision to license and will not be adequately controlled for using business-fixed effects. Failure to account for this unobserved productivity growth would consequently bias results. We contribute to existing non-experimental literature by controlling for these differential trends across businesses by using a nationally representative three-period panel dataset, which allows us to control for pre-existing trends prior to registration.

A concern across the experimental and many of the non-experimental papers is the lack of nationally representative data. For example, McKenzie and Sakho (2010) use a survey of businesses from six industries in the 4 largest urban centres of Bolivia. Rand and Torm (2012) use businesses surveyed from 10 provinces in Vietnam where informal firms were selected through on-site identification (i.e., operating in close proximity to formal firms in their sample). De Mel et al. (2013) rely on a sample of informal businesses from the two largest cities in Sri Lanka. Similarly, Demenet et al. (2016) employ a dataset from Vietnam's two largest cities. Hence, even where internal identification is highly credible, such as in RCTs, there still remains a question about external validity. Our analysis suggests that after controlling for differential pre-existing trends, the benefits of formalisation are very small, which is consistent with the RCT

literature (Bruhn & McKenzie, 2014), but is based on a sample of nationally representative businesses, from urban and rural areas, that spans all industries.

3. Business registration, context and data

3.1 Business registration in Vietnam

In Vietnam, private, domestic businesses may legally operate in one of three registration statuses. They may operate as an enterprise, a licensed household business, or as an unlicensed household business. In this subsection, we explain in detail the different types of business registration status and the legal determinants as to which status a business is required to have.

We begin with the most formal registration status, a private enterprise. Businesses that register as a private enterprise are subject to the same legal framework as state-owned, foreign-invested and collective firms, all of which must legally register as an enterprise.[6] This includes being required to follow formal accounting standards, making mandatory social insurance contributions on behalf of employees, and being subject to Vietnam's corporate income tax. All private businesses that regularly employ 10 or more workers or operate in more than one location are required to register as an enterprise.[7]

Smaller, single-location businesses have the option of not registering as an enterprise, but instead operating as what is called a household business in Vietnam. Unlike private enterprises, these businesses are not subject to formal accounting standards, nor are they required to make social insurance contributions on behalf of their workers. Furthermore, within the household business category, not all businesses are required to be a licensed household business. Low

[6] During our period of study, 2004 through 2008, the principle legislation is the Enterprise Law. See law No. 13-1999-QH10 Law on Enterprises.
[7] Decrees No. 02/2000/ND-CP of 3 February 2000 and No. 109/2004/ND-CP of 2 April describe household business and enterprise registration requirements during our study period.

revenue household businesses are not required to become licensed.[8]

3.2 Brief overview of the household business sector in Vietnam

During our period, 2004 to 2008, Vietnam experienced rapid growth of 5.7 percent per year in real GDP per capita (in PPP).[9] The fast growth was accompanied by a large shift of the workforce out of agriculture into services and manufacturing (McCaig & Pavcnik, 2017). Vietnam experienced large inflows of foreign direct investment and commensurate growth in the share of the workforce working in foreign-invested enterprises, particularly within manufacturing. The associated labour reallocations, both across and within sectors, were associated with large rural to urban migration. The distributive implications were largely favourable, or at least not negative. Absolute poverty continued to fall rapidly, from about 20 to 14.5 percent between 2004 and 2008 (World Bank, 2011), while inequality, as measured by the Gini coefficient, experienced relatively minor changes (World Bank, 2013; Benjamin, Brandt, & McCaig, 2017).

The structural transformation of Vietnam's economy occurred alongside a reallocation of workers from household businesses to enterprises (McCaig & Pavcnik, 2015, forthcoming). Nonetheless, a large number of household businesses remain and are an important source of employment. Table 1 presents estimates of the number of businesses captured by the VHLSSs. These estimates are based on sample weights to create national aggregates.[10] Note that the 2004 VHLSS did not separately identify private enterprises from licensed household businesses and

[8] It is very difficult to find information on the low revenue threshold as it varies across local administration units (either provinces or districts). Moreover, most household business operators seem to be unaware of registration requirements regardless of revenue (Cling et al. 2012).
[9] Based on data from the World Bank's World Development Indicators database.
[10] For a more recent analysis of the economic importance of the household business sector, see Pasquier-Doumer, Oudin, and Nguyen (2017) which uses nationally representative data on Vietnamese household businesses collected in late 2014 and early 2015. Additionally, see Cling, Razafindrakoto, and Rouband (2010) and Pasquier-Doumer et al. (2017) for analysis on the challenges faced by household businesses.

thus the third row includes both licensed household businesses and private enterprises. However, the number of private enterprises is relatively small in both 2006 and 2008. The estimates suggest that there were around 8 to 9 million household businesses operating in Vietnam during this period. In 2008, it is estimated there were 9.3 million household businesses, of which 7.0 million were unlicensed. These estimates are reasonably close to those derived from the 2007 Labour Force Survey. For example, Pasquier-Doumer and Pham (2017) estimate 9.1 million household businesses in 2007, of which 7.2 million were unlicensed. Similarly, the estimates from the VHLSSs of the number of private enterprises is consistent with the number of private enterprises reported in the annual enterprise survey conducted by the General Statistics Office (GSO) of Vietnam. For example, in 2008 there were 0.18 million private enterprises reported in the enterprise census.[11] Hence, the nationally representative VHLSSs produce consistent aggregate national estimates of the number of businesses relative to other available data sources.

Table 1: Number of businesses estimated in the VHLSSs

	2004	2006	2008
Total (millions)	8.3	7.9	9.5
Unlicensed	6.5	6.0	7.0
Licensed	1.8	1.9	2.5
Enterprises	n.a.	0.2	0.2

Author's calculations of the estimated number of businesses covered by the VHLSSs. All estimates are based on sample weights. The 2004 VHLSS business module did not separately identify enterprises from licensed household businesses and thus the row "licensed" includes both licensed household businesses and private enterprises.

Why do some household businesses choose to register while others do not? Based on surveys of licensed and unlicensed household businesses in Ha Noi and Ho Chi Minh City in 2007 and 2008, Cling et al. (2012) report that the vast majority of unlicensed household

[11] Authors' own calculation using GSO enterprise census.

businesses believe that registration is not compulsory and few know the legislative requirements (even among licensed household businesses). Furthermore, few report ever being asked to register by officials. They are generally ignored by authorities. Among those that are licensed, they report licensing helps them avoid corruption, access better locations, and gain contracts with larger firms. In contrast, among the unlicensed, the majority report no perceived benefits from becoming licensed. These results are confirmed by a nationally representative survey of household businesses conducted in 2014 and 2015 (Nguyen, Oudin, Pasquier-Doumer, & Vu, 2017).

3.3 Data on household businesses

We use the 2004, 2006, and 2008 Vietnam Household Living Standards Surveys (VHLSS), which were conducted by the GSO.[12] The VHLSS is a nationally representative survey based on a stratified sampling framework. First, rural communes and urban wards were separately stratified by province. Communes and wards were then randomly selected with probability proportional to population. In the second stage, three census enumeration areas per commune or ward were selected. Finally, in the third stage, households within an enumeration area were randomly selected (General Statistics Office, 2008). The total number of households surveyed is approximately 45,000 households in each of the three surveys.

Each survey contains data on household demographics, education, health, employment, income-generating activities, expenditures, and on businesses run by the household. The recall period is the past 12 months and the household was asked to provide information on all businesses that operated for any length of time during that period. Each survey collected

[12] We do not use the more recent 2010, 2012 and 2014 VHLSSs since these surveys did not collect as much information about the businesses run by households. In particular, they did not record the manager, start year, or number of workers.

information related to the operation of non-farm household businesses in a very consistent manner. In particular, the business modules asked the most informed member of the household about the industry, months of operation, revenue, license status, number of workers, location, start year, and numerous expenses, such as labour, materials, water, energy, taxes and fees, etc. for each business run by the household.[13] Note that some households report operating more than one business. We treat these as separate businesses and track businesses over time, not households running businesses over time. A key variable in our analysis is the license status of the business. The 2004 survey recorded whether the business has a license, but it did not distinguish between whether the license was a household business license or enterprise sector registration. In contrast, the 2006 and 2008 surveys asked whether the business had a household business license or was registered as an enterprise. Across the surveys, the share of businesses with a license rose from 0.21 to 0.24 between 2004 and 2006 and then to 0.27 by 2008.[14] Between 2006 and 2008, the share of businesses registered as an enterprise remained unchanged at 0.02. Since we cannot distinguish between a business that is a licensed household business versus a private enterprise in the 2004 survey and the share of private enterprises is very low in the 2006 and 2008 surveys, we focus on whether the business has any license (either a household business license or a private enterprise registration) versus no license.

The 2006 and 2008 surveys specifically asked the most knowledgeable household member to be recorded and we hereafter refer to this individual as the manager of the business. Since the 2004 business module did not directly record the most knowledgeable member of the household for each business we follow McCaig and Pavcnik (2016) and predict the manager of the business by matching information on the business, such as the industry, with individual

[13] The 2004 VHLSS only asked about the location and number of workers for one-fifth of the households surveyed.
[14] These estimates are weighted by sampling weights such that they are consistent estimates of the national average.

employment information reported in the employment module. The procedure works extremely well as testing the procedure using the 2006 survey leads to a correct prediction 92.9 percent of the time. The high success rate is driven by the fact that the typical household runs only one business and the typical business has only one worker: the manager. Thus, the majority of businesses in the dataset are easily matched to the one individual in the household reporting working in a household business in the same industry. See section A.2 and Table A.3 in Appendix A for complete details on the procedure.

An important characteristic of the surveys for our purposes is the inclusion of a household panel. It is a rotating panel in which approximately half of the enumeration areas in the 2004 survey, along with all of the households surveyed within them, were interviewed again in the 2006 survey. This is also true between the 2006 and 2008 surveys. In total, there are about 21,000 panel households between the 2004 and 2006 surveys and between the 2006 and 2008 surveys. Additionally, about half of the enumeration areas surveyed in both 2004 and 2006 were also surveyed in 2008. This produces a panel of 9,682 households that were interviewed in each of the three surveys. Considering all of the households surveyed in 2004 in enumeration areas that are part of the 2004-06-08 panel, 1,627 (14.4%) were not resurveyed by 2008 (some attrited between 2004 and 2006 and others between 2006 and 2008). In Table B.1 in Appendix B, we demonstrate that there is no evidence of selection of panel enumeration areas based on observable characteristics at the enumeration area level. Furthermore, we show that within panel enumeration areas the explanatory power of observable household head characteristics for whether the household was part of the panel was very low (R-squared of 0.004). However, female headed households were less likely to be part of the panel, while households headed by older and working individuals were more likely to be part of the panel. This is further supported

by the summary statistics presented in columns 1 and 2 of Table B.2. Column 1 reports summary statistics for all businesses in the 2004 survey while column 2 reports the same statistics only for businesses run by households that are part of the three survey panel. Those run by panel households have marginally lower revenue and profits, but are otherwise almost indistinguishable on average from the entire cross section.

In Table 2, we present a tabulation of panel households according to the number of businesses run in 2004 and 2008, the start and end of our period. For businesses reported in the 2008 survey, we only include businesses that reported starting in 2004 or earlier. Hence, this table is useful for understanding the potential number of businesses that can be matched over the three surveys. Over half of all households, 5,456, did not report a business in either 2004 or 2008. In total, there were 4,664 businesses reported in the 2004 survey and 3,388 businesses reported in the 2008 survey that started in 2004 or earlier.[15] One surprising feature of the table is the number of households that reported operating more businesses in 2008 than in 2004, despite our restriction that no businesses reported as starting later than 2004 were included from the 2008 survey. For example, among households that operated one business in 2004, 174 of them report operating two businesses in 2008 that reported starting in 2004 or earlier. There are a number of possible reasons. First, some of the businesses may have started in 2004 after the household was surveyed in 2004. Second, some of the businesses may have been started prior to 2005 by an individual that was not part of the household at the time of the 2004 survey, but was by the time of the 2008 survey. Third, there may simply be reporting and recording error in the start year reported in the 2008 survey. Fourth, the business may have been temporarily closed

[15] The number of businesses operated can be derived by summing the product of the number of businesses operated by the household by the number of households operating that many businesses. For example, in 2004 the total number of businesses run by these households is 1*2917 + 2*715 + 3*83 + 4*17 = 4,664.

during the 12 months covered by the 2004 survey. Fifth, the business may have simply failed to be enumerated in the 2004 survey for unknown reasons. Sixth, the manager's job may alternate over time between self-employment and working for other households. The difference can be subtle at times. The potential number of matches can be found by summing over the number of businesses represented by each cell. For example, 333 households reported operating two businesses in 2004, but only one business in 2008. If all of these businesses are matches, this represents 333 matched businesses. In total, there are a maximum of 2,661 possible matches between 2004 and 2008 based on this tabulation.

Table 2: Number of households by number of businesses for households observed in 3 surveys

Number of businesses run at the start of the panel	Number of businesses run at the end of the panel					
	0	1	2	3	4	Total
			2004-2006 household panel			
0	5391	519	37	3	0	5950
1	801	1885	211	18	2	2917
2	97	284	307	23	4	715
3	6	19	41	17	0	83
4	4	5	5	3	0	17
Total	6299	2712	601	64	6	9682
			2006-2008 household panel			
0	5496	499	42	3	0	6040
1	774	1862	218	11	0	2865
2	68	279	298	24	2	671
3	8	21	39	19	4	91
4	1	3	9	2	0	15
Total	6347	2664	606	59	6	9682
			2004-2008 household panel			
0	5456	468	26	0	0	5950
1	1144	1590	174	9	0	2917
2	171	333	197	13	1	715
3	12	38	25	8	0	83
4	6	4	5	1	1	17
Total	6789	2433	427	31	2	9682

We focus our analysis on businesses that can be observed across all three surveys, as this will allow us to control for pre-existing trends in the period prior to registration (see section 5 for further details on our methodology). Although the surveys contain a household panel, they were not designed to track businesses over time. In other words, there is not a unique business identifier. Hence, a critical first task is the construction of a business panel. We follow McCaig and Pavcnik (2014, 2016) and use information on the business that is unlikely to change over time for most businesses. In particular, we focus on the identity of the manager and the industry of operation. We start by matching businesses over time within a household according to whether the industry and manager both match. Subsequently, among remaining businesses within panel households, we match by either manager or industry. This produces a panel of 2,203 businesses across the three surveys. The panel was constructed in two waves: between 2004 and 2006 and then between 2006 and 2008. Between 2004 and 2006, 1,095 (79.5%) of businesses were matched by manager and industry, 96 (7.0%) were matched by manager only, and 186 (13.5%) were matched by industry only. Between 2006 and 2008, 1,041 (75.6%) were matched by manager and industry, 125 (9.1%) were matched by manager only, and 211 (15.3%) were matched by industry only. Across all three surveys 888 (64.5%) of businesses were matched by manager and industry. In our regression analysis, we present results using all matched businesses as well as only those that are matched by both manager and industry.

The sample used in our analysis consists of businesses that did not have a license during the first two surveys. This period, from 2004 to 2006, will serve as our pre-license period. As explained in section 5, we will use this period to capture pre-existing trends, which is a significant contribution of our research. Our sample consists of 1,377 businesses, of which 1,210

remain informal and 167 (12%) formalise between 2006 and 2008.[16] This allows us to create a sample of businesses that are similar in terms of license status for a period prior to their registration status in 2008.

It is worth noting that, like other papers that use business fixed effects, we are conducting our analysis on firms which survive the length of the panel. Hence, our results should be interpreted as the effects of licensing on surviving businesses. Business attrition is large, approximately 43 percent of household businesses operating in 2004 do not survive until 2008. Firms in our panel that survive this period are different from those who don't. McCaig and Pavcnik (2016) show that survival is more likely among initially larger, higher revenue firms. Given that larger and higher revenue firms are more likely to formalise and survive, higher attrition amongst firms which do not formalise may lead to an underestimation of the impact of formalisation. This selection issue is unavoidable in non-experimental studies using panel data. Column 3 in Table B.2 shows that by focusing on businesses that are present in all three surveys, we are constructing a sample of businesses that are larger in terms of profits and revenue, more likely to be licensed (26 versus 21%), and in general, larger and more successful. However, once we restrict the sample to those businesses that were unlicensed in 2004 and 2006, column 4, the mean business is slightly smaller than the sample of all businesses run by three-survey panel households in column 2.

4. Descriptive statistics

The businesses in our sample are small (see Tables B.2 and B.3 and Figure 1). Few of these businesses, have more than one worker, 29% and 37%, respectively, for businesses that

[16] The rate of formalisation is a few percentage points higher than in Demenet et al. (2016), but given the longer period and different data sources, the rates of formalisation seem relatively comparable.

continued to be unlicensed in 2008 and those that obtain a license by 2008. Note that this is very comparable to the share of unlicensed businesses with more than one worker in the 2004 cross sections (see Table B.2).[17] The small size of these businesses is very similar to that reported in Demenet et al. (2016), but significantly smaller than in McKenzie and Sakho (2010), where the average business has 3.2 workers, and in Rand and Torm (2012), where the average business has 5.6 workers. Moreover, on average, these businesses are not growing over time in terms of the number of workers. Among those that do not obtain a license, the share with more than one worker remains around 29% between 2004 and 2008. For those that do obtain a license, the share is consistently higher, 35-38%, but it barely changes over the period. The same pattern is observed for the ln number of workers. This suggests that obtaining a license is not associated with a change in the number of workers in these businesses. Similarly, profits within these businesses are low. Mean annual ln profits in 2004 among businesses that do not obtain a license is 8.53. This represents 5.07 million VND in January 2004 prices or about 1,045 USD PPP.[18,19] Initial profits among businesses that subsequently obtain a license is about 20 percent higher, but these are still generally small businesses. The small size is consistent with international evidence on the prevalence of small-scale, low productivity informal businesses in low-income countries (Emran, Morshed, & Stiglitz, 2011; La Porta & Shleifer, 2014).

Additionally, it is clear that businesses that eventually formalize are different from those that do not in a variety of ways. Across all three surveys, businesses that formalise generate

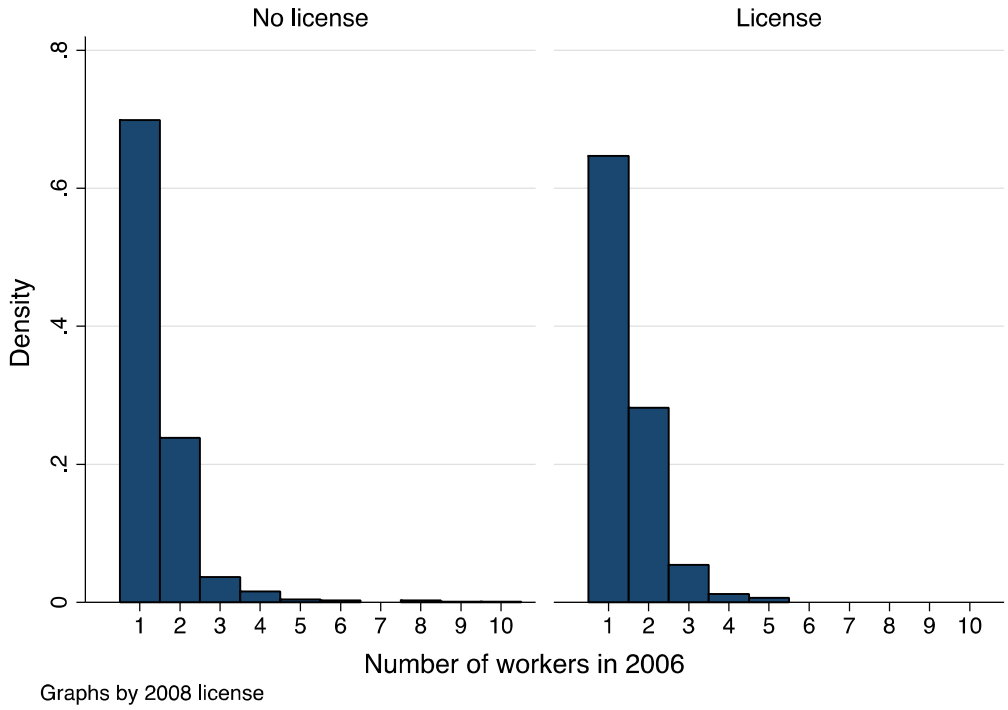
[17] In the repeated cross sections, 27% of unlicensed businesses have more than one worker, which is slightly lower than in our panel sample of businesses. The difference is possibly due to conditioning on survival. That is, our panel sample only includes businesses that survive across the three surveys and McCaig and Pavcnik (2016) show that survival is more likely among initially larger, higher revenue businesses.

[18] According to the 2005 International Comparison Program, 4,846.2 VND corresponds to 1 USD in PPP for actual individual consumption.

[19] Income from non-farm businesses account for about 22% of total household income during this period (Benjamin et al., 2017)

higher profits and revenue; employ more workers (although are only marginally more likely to report paying labour expenses and thus the extra workers are largely unpaid workers from the household); operate for more months of the year; are more likely to report paying taxes and fees, paying interest on a loan, to be the manager's primary job; and are run by better-educated managers. The differences between licensed and unlicensed businesses are even starker in the cross section (see Table B.2). The observable differences between businesses that become licensed versus those that do not are consistent with models such as Ulyssea (2017) and McKenzie and Sakho (2010) where larger, more productive firms find it optimal to choose to be formal whereas smaller, less productive businesses choose to be informal. Note that these differences already exist in our sample prior to obtaining a license.

Figure 1: Histogram of the number of workers in 2006 by license status in 2008



Furthermore, there is significant heterogeneity in terms of business performance within those that formalise and those that do not. Figure 2 plots the distribution of ln profit in 2008 for

businesses with and without a license. While the distribution of ln profit is shifted to the right for licensed businesses, the two distributions significantly overlap. There are some low profit businesses that have a license and some high profit businesses that do not. This is consistent with Ulyssea (2017) which predicts overlap in the distributions of formal and informal firms.

Figure 2: Distribution of ln(profit) for businesses by license status in 2008

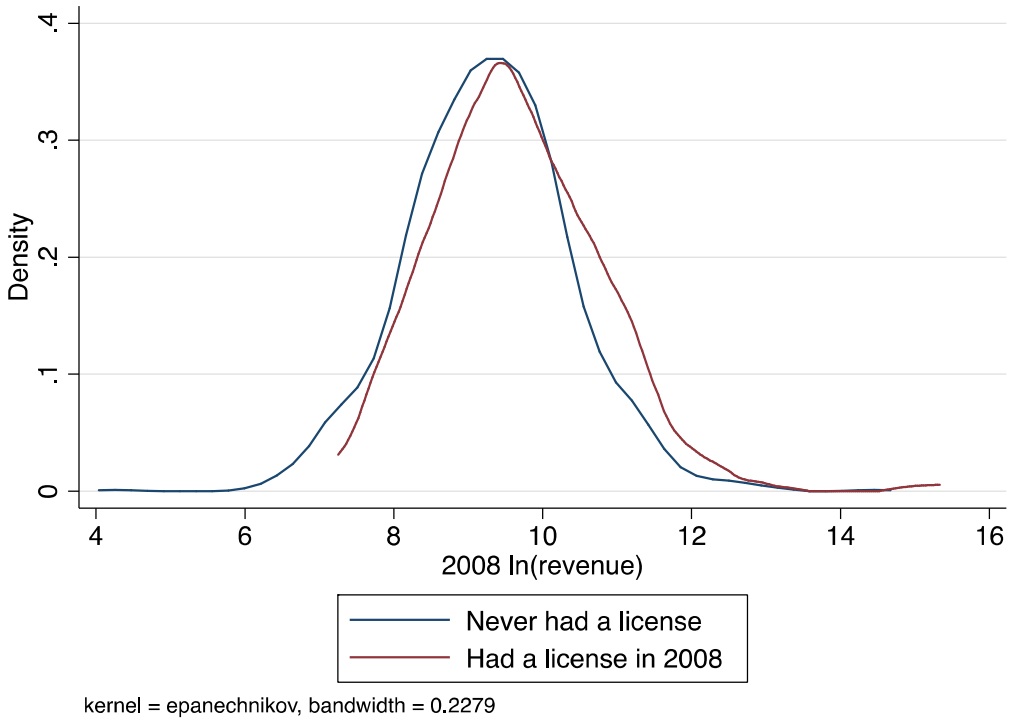


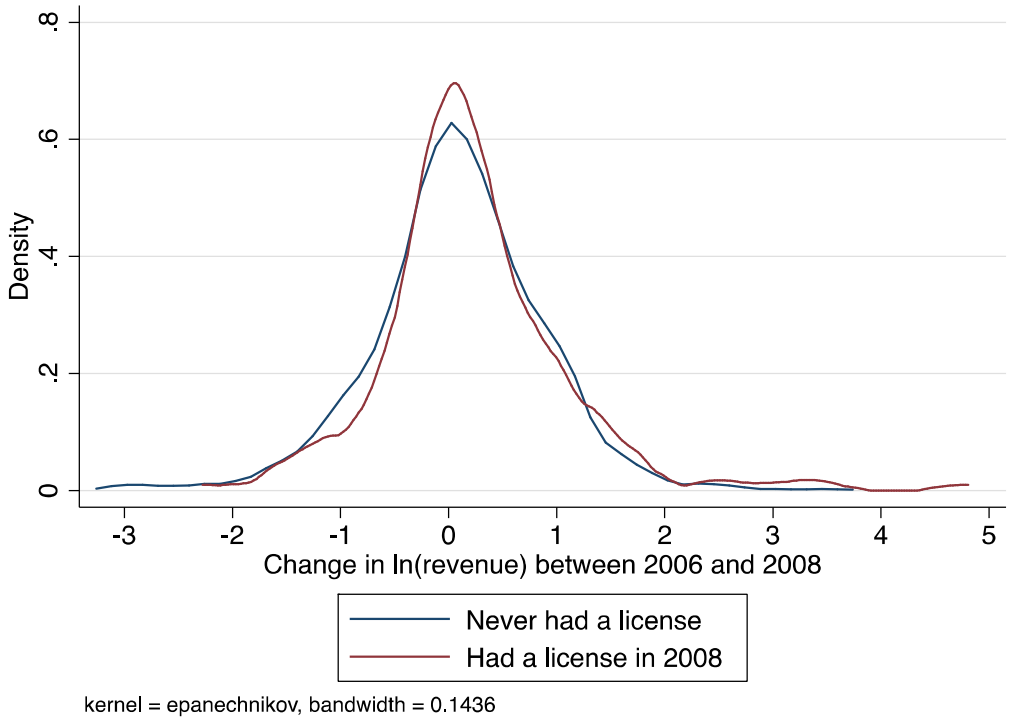
Table B.3 also shows the importance of having panel data to estimate the impact of becoming formal on business performance. Consider a conventional two-period analysis, similar to Demenet et al. (2016) and Rand and Torm (2012), by looking at 2006 and 2008. A standard difference-in-differences estimate would compare the average change in ln profits for firms that formalised to firms that did not and arrive at an estimate of a 0.075 log point increase in profits associated with formalising. This estimate is similar in magnitude to those of previous non-experimental literature using panel data (Rand & Torm, 2012; Demenet et al., 2016) and noticeably smaller than those using cross-sectional analysis (Fajnzylber et al., 2011; McKenzie

& Sakho, 2010). Again, there is substantial heterogeneity in the change in profit coinciding with becoming licensed. Figure 3 displays the distribution of the change in ln profit between 2006 and 2008 for businesses that become licensed and those that do not. The distributions largely overlap, but the tails have become longer relative to the ln profit in levels distribution, which suggests that outliers could play an important role in the analysis (i.e., businesses with unusually large increases or decreases in reported profit).[20] If however, the same difference-in-differences estimate is applied to the 2004-2006 period, a period in which all businesses in our sample are still informal, it would suggest that *future* formalisation is associated with a 0.082 log point increase in profits. Thus, controlling for pre-existing trends is potentially important as the firms that formalised between 2006 and 2008 were already growing more quickly, on average, than firms that did not formalise in our study period. Removing the pre-existing trend leads to a lower estimate of the gains from formalisation, -0.007 ln points. This is a significant point to note as it suggests that the presence of pre-existing trends may be correlated with the subsequent decision to formalise and thus studies that do not address this concern may produce biased results. In the next section, we use econometric analysis to test whether this pattern remains once we condition on initial conditions, but the intuition remains: controlling for pre-existing trends may be important for correctly estimating the causal effect of formalisation on informal businesses.

Hence, a key feature of our data that distinguishes our contribution from the existing literature is that it is nationally representative and our dataset contains observations on the same businesses three times over a four-year span, including a period prior to formalisation. Our dataset is well suited to provide nationally representative results and to control for pre-existing differences between businesses that do and do not formalise.

[20] De Mel et al. (2009) note that formal record keeping is rare among microenterprises and consequently measurement error for business outcomes such as revenue is common.

Figure 3: Distribution of the change in ln(profit) for businesses by license status in 2008



5. Econometric Model Outline

We denote Y_{it} as our outcome variable, where *i* denotes a firm and *t* a survey. To measure the effect of formalisation we introduce a dummy variable for whether or not the household business is licensed in the given survey, denoted by l_{it}. We include year fixed effects (θ_t) as well as firm fixed effects (α_i):

$$(1)\ Y_{it} = \alpha_i + \beta l_{it} + \theta_t + \varepsilon_{it}.$$

The above model is similar to the approach taken in many recent non-experimental papers. It identifies the impact of business formalisation based on comparing the mean change in outcomes among businesses that formalised to the mean change in outcomes among businesses that did not formalise. While able to control for unobserved time-invariant heterogeneity, unobserved time-variant heterogeneity may remain, such as firm-specific time trends and these trends may be

correlated with changes in business performance and the decision to formalise. Indeed, the summary statistics in Table B.3 strongly support this conjecture. Our main contribution to existing literature is that we are able to control for pre-existing trends, whereas previous non-experimental papers do not.

The identification assumption in equation (1) is that no time varying heterogeneity is correlated with both business performance and the decision to formalise. We relax this assumption by introducing a firm-fixed effect interacted with a time trend, where ρ_i captures unobserved firm-specific time trends:

$$(2)\ Y_{it} = \rho_i t + \alpha_i + \beta l_{it} + \theta_t + \varepsilon_{it}\ .$$

In equation (2), identification comes from differences in the rate of change of the outcome and its association with formalisation. In other words, in the example of ln profit as the outcome variable, identification is based on whether the growth rate of profit increases following formalisation. An intuitive way to see this is to take the difference of equation (2) across two consecutive surveys and introduce Δ to represent changes:

$$(3)\ \Delta Y_{it} = \rho_i + \beta \Delta l_{it} + \Delta\theta_t + \Delta\varepsilon_{it}$$

Equation (3) demonstrate that first differencing removes unobserved time-invariant heterogeneity (α_i) and is comparable to existing approaches in non-experimental literature. However, it also makes clear that other forms of unobserved heterogeneity, specifically those that vary over time within a firm (ρ_i), may still induce a bias in estimating β. Our approach allows us to control for one specific form of time varying unobserved heterogeneity: firm-specific trends. If these trends are correlated with the decision to license, as suggested in Table B.3, then estimates of β from equation (1) will be upward biased.

A key advantage of our data is the presence of three surveys. This allows us to construct changes across two periods, not just one period as is conventional. Taking the difference in changes produces:

(4) $\Delta Y_{i3} - \Delta Y_{i2} = \delta + \beta l_{i3} + \Delta \varepsilon_{i3} - \Delta \varepsilon_{i2}$

where we define δ to be equal to $\Delta\theta_3 - \Delta\theta_2$ and simplify $\Delta l_3 - \Delta l_2$ to l_3 since the indicator for having a license is 0 for all businesses in our sample in the 2004 and 2006 surveys. In equation (4), the causal effect of formalisation is identified through changes in changes. In other words, it captures whether the rate of change in the outcome variables increases after formalisation relative to prior to formalisation. We additionally add controls to our double difference model by including covariates X_{i1} from our first survey, 2004, to help control for differential trends induced by variation in initial conditions. These covariates include industry and province fixed effects as well as manager characteristics such as gender, age, education, and area of operation (urban or rural).

6. Results

We begin our econometric analysis by presenting results on the association between becoming licensed and profits. We focus initially on profits since this is the key metric for a manager when deciding on whether becoming licensed is optimal (McKenzie & Sakho, 2010). We present results using both the first differenced and double differenced frameworks. The first differenced results provide comparability with existing literature and allow us to draw subsequent comparisons with our double differenced framework, which differences away unobserved trends. In Panel A of Table 3, we present first differenced regression results for the period 2006 to 2008. Our simplest specification includes no controls and all businesses in column 1. We find that becoming licensed is associated with a 0.082 ln point increase in profits

or 8.6%. This mirrors the results from the summary statistics in Table B.3. Note as well that the R-squared is very low. Having a license explains very little of the variation in profits across businesses. In column 2, we remove outliers, businesses with a change in ln profits in the top or bottom 1 percent, and find very similar effects. In columns 3 and 4 we add additional control variables (industry fixed effects, province fixed effects, an urban indicator, a female indicator, and education of the manager) to the specifications in columns 1 and 2 and continue to find similar results. Overall, our first differenced results suggest that becoming licensed is associated with an 8.6 to 10.5% increase in profits, but only when outliers are removed in our simplest specification are the results statistically different from 0 at conventional test levels.

Table 3: Profits and informality

	No Controls	No controls, trimmed	Controls	Controls, trimmed
	(1)	(2)	(3)	(4)
Panel A: First differenced				
License indicator	0.0822	0.0998*	0.0882	0.0944
	(0.0680)	(0.0581)	(0.0708)	(0.0623)
R^2	0.001	0.002	0.077	0.085
N	1,313	1,285	1,308	1,280
Panel B: Double differenced				
License indicator	-0.00467	0.0412	-0.0602	-0.00954
	(0.120)	(0.0975)	(0.125)	(0.103)
R^2	0.000	0.000	0.070	0.084
N	1,313	1,285	1,308	1,280

The table reports the coefficient on an indicator for having a license at the end of the period. In Panel A, the dependent variable is the change in ln profits. In Panel B, the dependent variable is the differenced change in ln profits. In columns 1 and 2, no additional controls are added to the regression. In columns 3 and 4, control variables include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. In columns 2 and 4, the top and bottom 1 percent of observations in terms of the dependent variable are trimmed. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

These results are slightly lower than those reported in other recent non-experimental studies that observe businesses for two periods (i.e., are able to control for time invariant heterogeneity that may be correlated with license status). For example, Rand and Torm (2012) report an increase in profits of between 12 and 16 percent while Demenet et al. (2016) find an increase in value added of slightly more than 20 percent in association with obtaining a license.

Recall that the summary statistics presented in Table B.3 indicate that businesses that are initially informal, but formalise in the future experienced faster profit growth *prior* to formalisation than businesses that stayed informal. This suggests that the results in Panel A of Table 3 may be biased due the existence of pre-existing trends that are correlated with both profit growth and license status. Hence, in Panel B of Table 3, we display regression results from our double differenced framework, equation (4). Relative to the results in Panel A, the association between profits and being licensed is weaker. For example, in column 1, which includes all businesses and has no additional control variables, the estimate is now essentially 0, although the standard error has increased. The largest estimate is in column 2, which still includes no additional covariates but trims the double differenced profits distribution of the top and bottom 1 percent of observations. For this sample, becoming licensed is associated with a 0.041 ln point increase in profits.

Our results from the double differenced framework suggest that becoming licensed is not associated with a statistically significant change in profits. Furthermore, they suggest that the increase observed using a first differenced framework (Table 3, Panel A), are largely due to differential pre-existing trends between businesses that subsequently formalise and those that do not. This highlights the importance of being able to observe businesses for repeated periods prior

to formalisation, which our data allows.[21] The results are consistent with theoretical models that predict that the decision to formalise is endogenous and that the marginal firms that become licensed are not expected to experience a significant change in profits from formalising (Ulyssea, 2017). In Appendix A, Table B.5, we report results based on the subsample of businesses that were consistently matched by both manager and industry across all three surveys. We similarly find that becoming licensed is not associated with an increase in profits.[22]

We next use our data to explore various dimensions along which businesses may respond to becoming licensed. First, we look at revenues and expenses, along with a breakdown of expenses into various categories. Second, we examine labour input, location, and whether the business reports paying back a loan (i.e., evidence of access to credit).

In Table 4, we explore how revenue and expenses change in response to becoming licensed. As in our analysis of licensing and profits, we examine differences across two periods, i.e., a first differenced framework, as well as controlling for pre-existing trends using all three periods. In Panel A of Table 4 we present first differenced results for the association between becoming licensed and ln revenue, ln expenses, and the share of total expenses by various expense items. The results suggest that becoming licensed is associated with an increase in revenue of 0.14 ln points or 15%. We find that expenses also increase in conjunction with becoming licensed, by 0.17 ln points or 18%. In terms of the composition of expenses, the share

[21] Rand and Torm (2012) add the growth rate in the previous period as a control variable instead of using a double differenced framework. This will not perfectly control for the unobserved trend in our econometric model unless the coefficient on the previous change is -1. Indeed, when we use the previous growth rate as a control we find that the estimated coefficient on the license variable increases in magnitude relative to the results in Panel A of Table 3. These results are available in Table B.4.

[22] Note that we do not know the exact timing of when a business obtains a license. It could vary from immediately after the 2006 survey to just before the 2008 survey. However, the uncertainty over the exact timing of becoming licensed is common to other studies, see Demenet et al. (2016) which also features two years between surveys, and when we do not control for pre-existing trends, we find evidence of a positive effect. Thus, we do not believe the exact timing of becoming licensed is driving the results, but rather controlling for pre-existing trends.

Table 4: Impact of formalization on revenue and expenses

	ln revenue	ln expenses	Materials share	Labour share	Energy and water share	Non-durables, repair, and depreciation share	Rent share	Taxes and fees share	Other expenses shares
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: First differenced results									
License indicator	0.141*	0.168	0.0311	-0.0118	-0.000680	-0.0392*	-0.0122	0.0242	-0.00446
	(0.0757)	(0.109)	(0.0249)	(0.0112)	(0.0204)	(0.0212)	(0.0103)	(0.0150)	(0.0205)
R^2	0.084	0.081	0.082	0.075	0.088	0.076	0.113	0.071	0.092
N	1,330	1,367	1,367	1,367	1,367	1,367	1,367	1,367	1,367
Panel B: Double differenced results									
License indicator	0.0579	0.111	0.0889**	-0.0204	-0.0414	-0.0613	-0.0190	0.0402	-0.00364
	(0.123)	(0.186)	(0.0448)	(0.0187)	(0.0379)	(0.0398)	(0.0203)	(0.0248)	(0.0359)
R^2	0.081	0.088	0.089	0.078	0.084	0.082	0.141	0.083	0.090
N	1,330	1,367	1,367	1,367	1,367	1,367	1,367	1,367	1,367

The table reports the coefficient on an indicator for having a license on the indicated outcome. In Panel A, the dependent variable is the first difference and in Panel B it is the double difference. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

of expenses on various items does not change much as the coefficient estimates are all very small in magnitude and not statistically different from 0 in general. In Panel B, we report results from double differencing estimation. As we found in Table 3 for profits, the growth in revenue and expenses in connection to becoming licensed is lower once we account for differences in pre-existing trends. The estimated coefficient on revenue falls from 0.14 to 0.058 and the coefficient on expenses falls from 0.17 to 0.11 when we moved from the first differenced to the double differenced estimation framework. The composition of expenses remains relatively similar in the double differencing framework, although the estimated coefficient for materials increases and gains statistical significance at the 5 percent level. For the labour share of expenses, both the first and double differenced results indicate that obtaining a license is not associated with a change in the share of labour expenses. In Table B.6, we report results using the subset of businesses matched by both manager and industry across all three surveys. We find no evidence of an increase in revenue or expenses in association with becoming licensed.

In Table 5, we evaluate whether becoming licensed is associated with a change in labour inputs, adopting fixed premises, receiving loans, and becoming the primary job of the manager.

Becoming licensed is associated with an increase in the number of workers (the coefficients are 0.087 and 0.38 for the first difference and double difference results), although the results are not statistically significant. Furthermore, the magnitude of the association suggests that becoming licensed does not increase employment by very much, only by about 1/3 of a worker in the double differenced results. Additionally, the increase in the number of workers in column 1 is largely coming from an increase in employment in businesses that already have more than one worker. In column 2, the coefficients for having more than one worker are smaller, around 0.004 and 0.16, suggesting that relatively few businesses that become licensed

Table 5: Impact of formalization on labour inputs, location, loans, and job of the manager

	Number of workers	Indicator for having more than one worker	Indicator for hiring outside workers	ln days worked by manager	Indicator for a fixed premise	Indicator for a fixed premise outside of the home	Indicator for repaying a loan	Indicator for businesses being the manager's primary job
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: First differenced results								
License indicator	0.0871	0.00423	-0.00462	0.0313	0.0793***	0.0388	0.0425	0.0493*
	(0.0572)	(0.0384)	(0.0218)	(0.0430)	(0.0257)	(0.0358)	(0.0285)	(0.0270)
R^2	0.106	0.082	0.062	0.123	0.087	0.091	0.096	0.075
N	1,372	1,372	1,372	1,157	1,372	1,372	1,372	1,164
Panel B: Double differenced results								
License indicator	0.378	0.156	-0.0224	0.0807	0.200	0.172	0.0136	0.0635
	(0.283)	(0.148)	(0.0365)	(0.0727)	(0.132)	(0.114)	(0.0507)	(0.0517)
R^2	0.416	0.369	0.060	0.091	0.356	0.293	0.090	0.078
N	286	286	1,372	1,148	286	286	1,372	1,164

The table reports the coefficient on an indicator for having a license on the indicated outcome. In Panel A, the dependent variable is the first difference and in Panel B it is the double difference. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

begin to employ more than one worker if they were not already doing so. Moreover, becoming licensed is not associated with a change in the incidence of hiring workers. Thus, most of these initially informal businesses continue to solely employ household labour even after becoming licensed. We do find evidence of becoming licensed being associated with an increase in the incidence of having a fixed premise, particularly one outside of the home, but the results are imprecisely estimated. We find no evidence that becoming licensed is associated with an increase in the incidence of repaying a loan, our proxy for access to credit. Lastly, we find that becoming licensed is associated with an increase in the likelihood that the business is the manager's primary job. In sum, there is fairly weak evidence that becoming licensed changes labour input and access to credit, but there is some evidence that these businesses become more likely to operate in a fixed location outside of the home and of the manager devoting more time to the business. In Table B.7, we report consistent results based on the subsample of businesses that were matched by both manager and industry across all three surveys.

In general, our results suggest that becoming licensed does not have large effects on most of these businesses. In other words, lacking a license is not a major constraint for firm performance. Our results are consistent with models such as by Banerjee and Duflo (2005), Gollin (2008), and Emran et al. (2011) that provide explanations for why there are so many small, low productivity businesses in low-income countries. Banerjee and Duflo (2005) stress the shape of the production function, Gollin (2008) focuses on the distribution of entrepreneurial talent and its interaction with aggregate demand, while Emran et al. (2011) highlight the role of imperfect labour markets. Additionally, our results are consistent with models such as Ulyssea (2017) which highlight that formalisation is an endogenous decision and a marginal decision. That is, the marginal firm is indifferent between becoming formal or remaining informal. As our

sample focuses on firms that were initially informal and subsequently chose to formalise, it is perhaps not surprising in light of these predictions that business performance does not improve dramatically.

In Tables B.8 through B.10, we provide results based on splitting the sample according to the number of workers in the business. In particular, we distinguish between businesses that had more than one worker in 2006 versus businesses where the only worker is the owner/manager. Previous studies have found evidence of differential effects across this margin (Demenet et al. 2016; Fajnzylber et al., 2011; McKenzie & Sakho, 2010). We find that first differencing results suggest greater profit growth for businesses that initially had more than one worker (although only our simplest specification with no controls yields results that are statistically significant), but the difference with one-worker businesses significantly lowers once trends are accounted for. Similarly, we find that revenue and expense growth is faster for businesses that had more than one worker using first differences, but again the results are attenuated once trends are accounted for. Finally, businesses with more than one worker initially were more likely to experience an increase in the number of workers following licensing, but the results are not statistically different from 0. Overall, these results suggest that initially larger businesses, those that already employ at least one worker other than the manager, possibly another household member, experienced more positive changes in association to becoming licensed, but the estimated magnitudes are still quite small.

7. Conclusion

Using a nationally representative, three-survey panel dataset on informal and formal businesses in Vietnam, we estimate the impact of formalisation on business performance while controlling for differential trends that existed *prior* to formalisation. We find that after

controlling for differential trends, obtaining a license is not associated with an increase in profits. By comparison, we estimate that formalising is associated with approximately a 9 percent increase in profits when using only two periods of data, which is similar in magnitude to many non-experimental studies of the benefits of formalisation. This suggests that time-varying unobserved heterogeneity may be inducing an upward bias in many non-experimental results. We similarly find that obtaining a license is not associated with increases in revenue, expenses, and employment once we control for pre-existing trends. We find marginal evidence that becoming licensed is associated with an increase in the likelihood of operating out of fixed premises outside of the home and in the likelihood of the manager operating the business as his/her primary job.

Our results are consistent with models that predict that becoming formal is an endogenous decision and that the benefits for the marginal firm that obtains a license are predicted to be small (McKenzie & Sakho, 2010; Ulyssea, 2017). Furthermore, our results are in line with recent experimental results where few informal firms are induced to formalise and those that do experience small benefits on average (Bruhn & McKenzie, 2014). The consistency of our results with previous experimental studies suggests that earlier non-experimental studies may overestimate the benefits of formalisation. These non-experimental studies typically either control for unobserved heterogeneity through the use of fixed effects or employ an instrumental variable strategy. However, these approaches do not control for unobserved trends that may be correlated with the decision to become licensed. Hence, our approach highlights the importance of using data that allows researchers to observe firms repeatedly before formalising.

References

Almeida, R., & Carneiro, P. (2009). Enforcement of labor regulation and firm size. *Journal of Comparative Economics*, *37*(1), 28-46.

Banerjee, A. V. (2013). Microcredit under the microscope: What have we learned in the past two decades, and what do we need to know? *Annual Review of Economics*, *5*, 487-519.

Banerjee, A. V., & Duflo, E. (2005). Growth Theory through the Lens of Development Economics. In P. Aghion & S. Durlauf (Eds.), *Handbook of Economic Growth* (Vol. 1A, pp. 473-552). Amsterdam, Netherlands: North Holland.

Banerjee, A. V., & Duflo, E. (2007). The economic lives of the poor. *The Journal of Economic Perspectives, 21*(1), 141-168.

Banerjee, A. V., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York, NY: PublicAffairs.

Banerjee, A. V., Karlan, D., & Zinman, J. (2015). Six randomized evaluations of microcredit: Introduction and further steps. *American Economic Journal: Applied Economics*, *7*(1), p. 1-21.

Benhassine, N., McKenzie, D. J., Pouliquen, V., & Santini, M. (2017). *Can enhancing the benefits of formalization induce informal firms to become formal? Experimental evidence from Benin*. (World Bank Policy Research Working Paper No. 7900). Washington, D.C.: World Bank Group.

Benjamin, D., Brandt, L. & McCaig, B. (2017). Growth with equity: Income inequality in Vietnam, 2002-14. *Journal of Economic Inequality*, *15*(1), 25-46.

Bruhn, M., & McKenzie, D. (2014). Entry regulation and the formalization of microenterprises in developing countries. *The World Bank Research Observer, 29*(2), 186-201.

Campos, F., Goldstein, M. & McKenzie, D. (2015). *Short-term impacts of formalization assistance and a bank information session on business registration and access to finance in Malawi*. (World Bank Policy Research Working Paper No. 7184). Washington, DC: World Bank Group.

Cling, J. P., Razafindrakoto, M., & Roubaud, F. (2010). *The informal economy in Vietnam*. Vietnam: ILO Office in Vietnam.

Cling, J. P., Razafindrakoto, M., & Roubaud, F. (2012). To be or not to be registered? Explanatory factors behind formalizing non-farm household businesses in Vietnam. *Journal of the Asia Pacific Economy, 17*(4), 632-652.

de Andrade, G. H., Bruhn, M., & McKenzie, D. (2014). A helping hand or the long arm of the law? Experimental evidence on what governments can do to formalize firms. *The World Bank Economic Review*, *30*(1), 24-54.

De Giorgi, G., & Rahman, A. (2013). SME's Registration: Evidence from an RCT in Bangladesh. *Economics Letters*, *120*(3), 573-578.

de Mel, S., McKenzie, D. J., & Woodruff, C. (2009). Measuring microenterprise profits: Must we ask how the sausage is made? *Journal of Development Economics, 88*(1), 19-31.

de Mel, S., McKenzie, D., & Woodruff, C. (2013). The demand for, and consequences of, formalization among informal firms in Sri Lanka. *American Economic Journal: Applied Economics, 5*(2), 122–150.

Demenet, A., Razafindrakoto, M., & Rouband, F. (2016). Do informal businesses gain from registration and how? Panel data evidence from Vietnam. *World Development*, *84*, 326-341.

Emran, M. S., Morshed, A. K. M., & Stiglitz, J. E. (2011). *Microfinance and missing markets*. (MPRA Paper No. 41451). Munich, Germany: University Library of Munich, Germany.

Fajnzylber, P., Maloney, W. F., & Montes-Rojas, G. V. (2011). Does formality improve micro-firm performance? Evidence from the Brazilian SIMPLES program. *Journal of Development Economics*, *94*(2), 262-276.

Farrell, D. (2004). The hidden dangers of the informal economy. *The McKinsey Quarterly,* 3, 26–37.

Galiani, S., Meléndez, M., & Ahumada, C. N. (2017). On the effect of the costs of operating formally: New experimental evidence. *Labour Economics*, *45*, 143-157.

General Statistics Office. (2008). Operational handbook: Vietnam household living standard survey 2008.

Gollin, D. (2008). Nobody's business but my own: Self-employment and small enterprise in economic development. *Journal of Monetary Economics*, *55*(2), 219-233.

Gollin, D., Lagakos, D., & Waugh, M. E. (2014). The agricultural productivity gap. *Quarterly Journal of Economics*, *129*(2), 939-993.

Hsieh, C. T., & Klenow, P. (2009). Misallocation and manufacturing TFP in China and India. *Quarterly Journal of Economics*, *124*(4), 1403-1448.

La Porta, R., & Shleifer, A. (2014). Informality and development. *Journal of Economic Perspectives*, *28*(3), 109-126.

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

McCaig, B. & Pavcnik, N. (2014). Export markets and household business performance: Evidence from Vietnam. Unpublished manuscript.

McCaig, B., & Pavcnik, N. (2015). Informal Employment in a Growing and Globalizing Low-Income Country. *American Economic Review*, *105*(5), 545-550.

McCaig, B. & Pavcnik, N. (2016). Out with the old and unproductive, in with the new and similarly unproductive: Microenterprise dynamics in a growing low-income country. Unpublished manuscript.

McCaig, B. & Pavcnik, N. (2017). Moving out of agriculture: Structural change in Vietnam. In M. McMillan, D. Rodrik, & C. P. Sepulveda (Eds.), *Structural change, fundamentals, and growth: A framework and case studies* (pp. 81-124). Washington, DC: International Food Policy Research Institute.

McCaig, B., & Pavcnik, N. (forthcoming). Export markets and labor allocation in a low-income country. *American Economic Review*.

McKenzie, D., & Sakho, Y. S. (2010). Does it pay firms to register for taxes? The impact of formality on firm profitability. *Journal of Development Economics*, *91*(1), 15-24.

McKenzie, D. & Woodruff, C. (2014). What are we learning from business training and entrepreneurship evaluations around the developing world? *The World Bank Research Observer*, *29*(1), 48-82.

Nguyen, T. K. T., Oudin, X., Pasquier-Doumer, L., & Vu, V. N. (2017). Characteristics of household businesses and the informal sector. In L. Pasquier-Doumer, X. Oudin, & T. Nguyen (Eds.), *The importance of household businesses and the informal sector for inclusive growth in Vietnam* (pp. 63-92). France: Vietnamese Academy of Social Sciences and the French National Research Institute for Sustainable Development.

Pasquier-Doumer, L., Oudin, X., Nguyen, T. (2017). *The importance of household businesses and the informal sector for inclusive growth in Vietnam*. France: Vietnamese Academy of Social Sciences and the French National Research Institute for Sustainable Development.

Pasquier-Doumer, L., & Pham, M. T. (2017). Evolution of the informal and household business sectors in Vietnam in a time of growth and trade liberalization. In L. Pasquier-Doumer, X. Oudin, & T. Nguyen (Eds.), *The importance of household businesses and the informal sector for inclusive growth in Vietnam*. France: Vietnamese Academy of Social Sciences and the French National Research Institute for Sustainable Development.

Rand, J., & Torm, N. (2012). The benefits of formalization: Evidence from Vietnamese manufacturing SMEs. *World Development*, *40*(5), 983-998.

Restuccia, D., & Rogerson, R. (2017). The causes and costs of misallocation. *Journal of Economic Perspectives*, *31*(3), 151-74.

Ulyssea, G. (2017). Firms, informality and development: Theory and evidence from Brazil. Unpublished manuscript.

World Bank. (2011). *Poverty reduction in Vietnam: Achievements and challenges*. Washington, DC: World Bank

World Bank. (2013). *Vietnam Poverty Assessment: well begun, not yet done - Vietnam's remarkable progress on poverty reduction and the emerging challenges*. Washington, DC: World Bank.

Supplementary Materials

for

Business Formalization in Vietnam

(for online publication only)

Appendix A: Dataset Construction

This appendix explains how we prepared our data for analysis. It is based on the discussion in McCaig and Pavcnik (2016) for which the dataset was first constructed. We describe three data preparation steps. First, we describe how we verify that the correct individual within the household has been reported as the manager for the respective business. Second, we detail how we predict a manager for all businesses in the 2004 VHLSS. Third, we explain how we create a household business panel.

A.1 Verifying the manager of the business

In both the 2006 and 2008 VHLSSs, the business module asked the household to identify the most knowledgeable household member for each business, which we refer to as the manager. For both surveys we checked whether the reported manager reports information in the employment module that is consistent with managing the household business. The reported manager should report working, should report working in a household business, and report working in the same industry as the business if sufficient detail is provided for the job. Both the 2006 and 2008 VHLSSs collected detailed information on the primary and secondary job.

We find that the vast majority of reported managers, 97.2 percent, provided consistent information in the labour module in 2006 and 2008 respectively (see Table A.1). For the businesses in which the reported manager did not provide consistent information, we conducted a search within the household for which individual is most likely to be the correct manager, including the originally reported manager. For example, the originally reported manager might have indicated being self-employed in a household business in the labour module, but a recording error led to the wrong industry code being recorded in the labour module data. Or the wrong manager might have been mistakenly recorded in the business module. That is, the

inconsistency could have been introduced in either the labour module or the business module. Given the small number of inconsistencies, we employed visual inspection of the business information combined with the labour module information for all household members. Where available, we also employed panel information for the household from the preceding or ensuing survey or both. We find that most instances are due to the wrong manager being recorded in the business module (Table A.2). However, in other instances the inconsistent information is due to the industry of the business being recorded incorrectly or the industry or ownership in the labour module.

A.2. Predicting the manager of a household business in 2004 VHLSS

The 2004 VHLSS household business module did not ask for the most knowledgeable person for the business. Hence, we predicted the manager of each business in 2004. This is useful for two reasons in our context. First, we include manager characteristics as control variables in our regression analysis. Second, knowing the manager of the business helps to facilitate the matching of businesses over time.

We combine data from the employment and business modules of the 2004 VHLSS, which can be matched. In particular, from the employment module we identify individuals that reported being self-employed in a household business for either their primary or secondary job during the past year. For these jobs, we use information on the industry, the number of months worked during the past years, the number of days per month usually worked, and the number of years the individual has been doing the job. From the business module, we use information on the industry, the number of months operating during the past year, the average number of days per month operating, and the year the business started.

In Table A.3, we provide a summary of the matches by the step within the manager prediction algorithm at which the match was made. The table is organized sequentially such that the first step of the algorithm was to identify the manager for businesses in which only one household member reported being self-employed in the industry of the business and then only businesses remaining without a predicted manager would proceed to the next row. The first step of the algorithm matches a manager for 70.5% of all businesses in the 2004 VHLSS. The corresponding rate of success using the 2006 VHLSS is 98.9%. Thus, for a large share of businesses we have a very high degree of confidence in our predicted manager. Next, we identified a manager for any remaining businesses when there was only one household member that reported being a manager of a business in the same industry in the 2006 VHLSS and so on down the rows of the table. In sum, the algorithm correctly identified the manager for 92.9% of businesses in the 2006 VHLSS. Thus, our manager prediction algorithm is highly adept at identifying the manager of the business.

A.3 Creation of a Panel of Businesses

In this section, we explain how we match businesses in the two- and three-survey VHLSSs panels. The surveys are household-level panels. The household surveys were not designed to directly follow businesses and thus we use characteristics of the business that should not change for most businesses in order to match them over time. We use the longitudinal dimension of our data at the household and individual level.

Not all businesses run by a panel household should be matched over time. For example, any household that reports running a different number of businesses across the two years has experienced net entry or exit of businesses and thus at least one business within the household should not be matched. Thus, for any given household the maximum number of matched

businesses is the minimum of the number of businesses run in either year. Table 2 summarises the number of businesses run by panel households in each of the two-survey panels where businesses in the end survey must report having started as of the previous survey. For example, businesses reported in the 2006 VHLSS that started in 2005 or 2006 are not included in the upper panel when calculating how many businesses the household operated in 2006. A little over half of the households did not operate a business in either the start or end year of the respective panel. The number of businesses that can potentially be matched is 7,076 between 2004 and 2006 and 6,661 between 2006 and 2008.

We start by matching businesses using information on the industry of operation and the manager of the business. We match 5,186 and 4,506 businesses based on these matching criteria between 2004-06 and 2006-08 (Table A.4) or 73 and 68 percent of the maximum possible number of matches, respectively. We subsequently relax the matching criteria and consider matching the remaining unmatched businesses first by industry (allowing the manager of the business to change over time) and then by manager (allowing the industry of the business to change over time). Matching by industry leads to an additional 564 and 732 matches, while matching by manager leads to 993 and 1,106 matches. In Table A.5, we summarise how the matches were made for each of the two-survey panels. Matching across the two two-year panels leads to a panel of 2,203 businesses across 2004, 2006, and 2008.

Table A.1: Number of business with inconsistent manager information

	2006	2008
Total	20,458	20,465
Manager provides consistent employment information	19,878	19,887
No manager reported	3	1
Manager did not report working	0	105
Manager did not report working in a household business	388	281
Manager worked only one job and it is not consistent with the business	71	69
Manager worked two jobs and neither is consistent with the business	118	122

Table A.2: Resolving inconsistent manager information

	2006	2008
Total number of businesses with inconsistent manager information	580	578
Of which, changed		
Manager	449	482
Industry of business	26	20
Indicator for self-employment in a business	63	6
Occupation of primary job	1	4
Industry of primary job	14	13
Ownership of primary job	1	0
Occupation of secondary job	0	0
Industry of secondary job	20	10
Ownership of secondary job	0	0
Indicator for working a third job	8	14
No changes	10	40

Table A.3: Manager prediction results

Match criteria	Number of matches in 2004	Share of matches in 2004	Share of correct matches 2006
No job matched the business	131	0.006	0.000
Businesses matched to a primary or secondary job			
Only job that matched the business by industry	15,122	0.705	0.989
Only manager in same business in subsequent survey	1,595	0.074	0.755
Only job that matched by year, months and days	232	0.011	0.912
Only job that matched by months and days	1,250	0.058	0.794
Only job that matched by months	191	0.009	0.755
Highest number of years in the job	742	0.035	0.789
Highest number of days worked in the past year in the job	180	0.008	0.659
Only one of the head or spouse matched	186	0.009	0.831
Highest number of hours per day in the job	313	0.015	0.681
Highest ranked individual within household	968	0.045	0.703
Only primary job	3	0.000	1.000
Businesses not matched to a primary or secondary job			
Only third job that matched business	421	0.020	0.952
Only manager in same business in subsequent survey	33	0.002	0.667
Highest ranked individual within household	91	0.004	0.443
Total	21,458	1.000	0.929

Table A.4: Number of business matched by manager and industry, manager only, or industry only

	2004-06 panel	2006-08 panel
Manager and industry	5186	4506
Manager only	564	732
Industry only	993	1106
Total	6743	6344

Table A.5: Number of business matched by manager and industry, manager only, or industry only

	2004-06 panel		2006-08 panel
Manager and industry	1769	Manager and industry	1416
		Manager only	142
		Industry only	211
Manager only	143	Manager and industry	76
		Manager only	53
		Industry only	14
Industry only	291	Manager and industry	149
		Manager only	20
		Industry only	122
Total	2203		2203

Appendix B: Supplementary regressions results and summary statistics

B.1 Panel enumeration areas and households

Our analysis focuses on businesses run by households that were surveyed three times across the 2004, 2006, and 2008 VHLSSs. The selection of the household panel was based on the selection of enumeration areas to be surveyed across all three surveys. Then, all households originally surveyed in 2004 in the selected enumeration areas were to be surveyed again in 2006 and 2008. Hence, the representativeness of the panel relies on representative enumeration areas being selected. In this appendix, we provide regression analysis on the representativeness of enumeration areas selected to be part of the three-survey panel.

There are 3,062 enumeration areas in the 2004 VHLSS with 15 households per enumeration area.[23] Of these, 1,570 were part of the panel between the 2004 and 2006 VHLSSs and 754 were part of the panel between 2004, 2006, and 2008. The total number of households that are part of the three-survey panel is 9,682 or about 13 households per enumeration area. This suggests that resurveying households within panel enumeration areas was imperfect.

We begin by exploring the representativeness of panel enumeration areas. We estimate a linear probability model at the enumeration level using all enumerations areas in the 2004 VHLSS. The dependent variable takes the value 1 if the enumeration area is part of the three-survey panel and 0 otherwise. We regress this indicator on a vector of enumeration area characteristics derived from the head of each household. Specifically, we include as controls the share of heads that are female, the mean age and highest grade achieved, the share that worked, the share that worked in manufacturing, the share that worked in services, and the share based on

[23] There are two enumeration areas with only 14 households.

different ownership categories. We present the results in column 1 of Table B.1. We find that the regression has almost no explanatory power. The R-squared is only 0.002 and none of the controls have a meaningful or statistically significant relationship with the probability of the enumeration area being selected for the panel.

Next, we explore selection of households *within* panel enumeration areas using data on all households in these enumerations areas in the 2004 VHLSS. We estimate a linear probability model at the household level. The dependent variable takes the value of 1 if the household is part of the three-survey panel and 0 otherwise. We regress this indicator on a vector of household characteristics (based on the head of the household again) and enumeration area fixed effects. The inclusion of enumeration area fixed effects mean that it is the variation of household characteristics within enumeration areas, not between, which we will be exploring in terms of household selection. Note that this procedure drops all enumeration areas for which all 15 of the households surveyed in the 2004 VHLSS were surveyed in each of the 2006 and 2008 VHLSSs. We present the results in column 2 of Table B.1. We find that the regression has very low explanatory power as the within R-squared is only 0.004. However, a small number of household characteristics have a statistically significant relationship with whether the household was part of the three-survey panel. Households were more likely to be part if the head was male, older, better educated, and worked.

Overall, we view the evidence of non-random selection into the household panel as relatively minor. There is no evidence of selection bias at the enumeration area level and within enumeration areas the explanatory power of household head characteristics is very low.

Table B.1: Selection of panel enumeration areas and households

	Indicator for panel enumeration area	Indicator for panel household
	(1)	(2)
Female	0.0365	-0.0257***
	(0.0587)	(0.00894)
Age	-0.000451	0.00109***
	(0.00196)	(0.000334)
Highest grade completed	-0.00372	0.00253*
	(0.00457)	(0.00133)
Worked	0.0133	0.0425***
	(0.0795)	(0.0125)
Manufacturing	0.0184	-0.00264
	(0.0802)	(0.0134)
Services	0.0828	-0.0159
	(0.0533)	(0.0105)
Other households	-0.0584	-0.0108
	(0.0748)	(0.0117)
State sector	-0.0289	0.00516
	(0.0807)	(0.0139)
Collective sector	-0.209	0.00468
	(0.345)	(0.0361)
Private sector	0.0250	-0.00743
	(0.188)	(0.0267)
Foreign sector	-0.322	-0.0108
	(0.370)	(0.0662)
Constant	0.261*	0.761***
	(0.150)	(0.0267)
Observations	0.002	0.004
R-squared	3,060	11,309

Notes: *** p<0.01, ** p<0.05, * p<0.1. In column 1, the unit of observation is an enumeration area and the explanatory variables are shares calculated over all household heads within each enumeration area. In column 2, the unit of observation is a household and the explanatory variables are based on the household head. Furthermore, in column (2), the sample is based on households within panel enumeration areas.

B.2 Supplementary analysis of the impact of becoming licensed

Tables B.2 through B.10 provide summary statistics and additional regression analysis discussed in the text.

Table B.2: Summary statistics for various samples of businesses in 2004

Variable	All businesses	Those run by households that are part of the 3-survey panel	Those that are present in all three surveys	Those that were unlicensed in both 2004 and 2006
	(1)	(2)	(3)	(4)
ln(revenue)	9.12	9.06	9.33	9.02
ln(profit)	8.63	8.59	8.84	8.55
License	0.22	0.21	0.26	0.00
ln(workers)	0.30	0.28	0.32	0.25
Pays labour expenses	0.09	0.08	0.08	0.04
Paying taxes and fees	0.36	0.38	0.46	0.28
ln(months of operation)	2.23	2.24	2.33	2.29
Paying loan expenses	0.09	0.08	0.10	0.07
Has more than one worker	0.32	0.30	0.37	0.30
Business is manager's primary job	0.73	0.73	0.82	0.77
Urban	0.33	0.32	0.35	0.29
Operates in a fixed location	0.84	0.84	0.89	0.85
Operates in a fixed location (not home)	0.28	0.31	0.32	0.33
Manufacturing	0.28	0.28	0.26	0.31
Services	0.72	0.72	0.74	0.69
Manager characteristics:				
Female	0.59	0.59	0.62	0.65
Age	40.5	40.4	41.2	40.8
Did not complete primary	0.18	0.18	0.17	0.20
Completed primary	0.30	0.29	0.28	0.29
Completed lower secondary	0.35	0.35	0.37	0.38
Completed upper secondary	0.17	0.18	0.19	0.13
Number of observations	21,458	4,664	2,203	1,377

Table B.3: Summary statistics for our sample of businesses

	2004			2006			2008		
Variable	Never has a license	Becomes licensed	Difference	Never has a license	Becomes licensed	Difference	Never has a license	Becomes licensed	Difference
ln(revenue)	8.99	9.20	0.21	9.17	9.42	0.25	9.28	9.65	0.37
ln(profit)	8.53	8.72	0.20	8.85	9.13	0.28	9.29	9.64	0.35
License	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00
ln(workers)	0.25	0.31	0.06	0.26	0.28	0.02	0.25	0.31	0.06
Pays labour expenses	0.04	0.04	0.00	0.06	0.08	0.02	0.06	0.08	0.02
Paying taxes and fees	0.26	0.43	0.17	0.23	0.40	0.17	0.28	0.53	0.25
ln(months of operation)	2.28	2.35	0.07	2.31	2.37	0.06	2.31	2.41	0.10
Paying loan expenses	0.07	0.10	0.02	0.06	0.08	0.03	0.04	0.12	0.08
Has more than one worker	0.29	0.38	0.09	0.30	0.35	0.05	0.29	0.37	0.07
Business is manager's primary job	0.76	0.87	0.12	0.78	0.88	0.10	0.82	0.94	0.13
Urban	0.29	0.33	0.04	0.29	0.33	0.04	0.31	0.35	0.05
Operates in a fixed location	0.85	0.91	0.06	0.86	0.86	0.00	0.86	0.92	0.06
Operates in a fixed location (not home)	0.32	0.38	0.05	0.30	0.32	0.02	0.29	0.34	0.05
Manufacturing	0.31	0.25	-0.06	0.32	0.22	-0.10	0.32	0.23	-0.08
Services	0.68	0.74	0.06	0.68	0.78	0.10	0.68	0.77	0.08
Manager characteristics:									
Female	0.66	0.58	-0.08	0.65	0.57	-0.08	0.65	0.57	-0.08
Age	40.68	41.93	1.26	43.09	44.12	1.03	44.95	45.83	0.88
Did not complete primary	0.21	0.17	-0.04	0.21	0.20	-0.02	0.21	0.16	-0.04
Completed primary	0.29	0.28	-0.01	0.29	0.26	-0.03	0.30	0.26	-0.04
Completed lower secondary	0.38	0.37	-0.01	0.37	0.36	-0.01	0.37	0.37	0.00
Completed upper secondary	0.12	0.18	0.06	0.12	0.19	0.06	0.12	0.20	0.08
Number of observations	1210	167		1210	167		1210	167	

The sample is all businesses observed across the 2004, 2006, and 2008 VHLSSs that do not have a license in 2004 and 2006, but may or may not have a license in 2008.

Table B.4: Profits and informality, with previous trend as a control

	No Controls	No controls, trimmed	Controls	Controls, trimmed
	(1)	(2)	(3)	(4)
License indicator	0.112*	0.115**	0.143**	0.130**
	(0.0596)	(0.0530)	(0.0623)	(0.0568)
2004-2006 change in ln(profits)	-0.345***	-0.289***	0.119**	0.107**
	(0.0293)	(0.0241)	(0.0571)	(0.0512)
R^2	0.137	0.114	0.219	0.200
N	1,313	1,285	1,308	1,280

The table reports the coefficient on an indicator for having a license at the end of the period. The dependent variable is the change in ln profits. In columns 1 and 2, no additional controls are added to the regression. In columns 3 and 4, control variables include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. In columns 2 and 4, the top and bottom 1 percent of observations in terms of the dependent variable are trimmed. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.5: Profits and informality (Matched by manager and industry)

	No Controls	No controls, trimmed	Controls	Controls, trimmed
	(1)	(2)	(3)	(4)
Panel A: First differenced				
License indicator	-0.0236	-0.00201	-0.0661	-0.0534
	(0.0773)	(0.0669)	(0.0840)	(0.0718)
R^2	0.000	0.000	0.111	0.131
N	852	834	852	834
Panel B: Double differenced				
License indicator	-0.109	-0.000629	-0.206	-0.104
	(0.131)	(0.116)	(0.142)	(0.125)
R^2	0.001	0.000	0.109	0.122
N	852	834	852	834

The table reports the coefficient on an indicator for having a license at the end of the period. In Panel A, the dependent variable is the change in ln profits. In Panel B, the dependent variable is the differenced change in ln profits. In columns 1 and 2, no additional controls are added to the regression. In columns 3 and 4, control variables include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. In columns 2 and 4, the top and bottom 1 percent of observations in terms of the dependent variable are trimmed. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.6: Impact of formalization on revenue and expenses (Matched by manager and industry)

	ln revenue	ln expenses	Materials share	Labour share	Energy and water share	Non-durables, repair, and depreciation share	Rent share	Taxes and fees share	Other expenses shares
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: First differenced results									
License indicator	-0.0850	-0.167	0.00843	-0.00994	0.0111	-0.0141	-0.0297**	0.0211	0.0211
	(0.0796)	(0.110)	(0.0294)	(0.0112)	(0.0237)	(0.0273)	(0.0146)	(0.0211)	(0.0190)
R^2	0.126	0.156	0.161	0.169	0.133	0.116	0.092	0.103	0.133
N	864	885	885	885	885	885	885	885	885
Panel B: Double differenced results									
License indicator	-0.245*	-0.428**	0.0233	-0.0254	-0.0179	-0.000747	-0.0435	0.0528	0.0276
	(0.131)	(0.197)	(0.0487)	(0.0210)	(0.0426)	(0.0498)	(0.0293)	(0.0350)	(0.0377)
R^2	0.121	0.162	0.144	0.159	0.121	0.137	0.124	0.116	0.144
N	864	885	885	885	885	885	885	885	885

The table reports the coefficient on an indicator for having a license on the indicated outcome. In Panel A, the dependent variable is the first difference and in Panel B it is the double difference. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.7: Impact of formalization on labour inputs, location, loans, and job of the manager (Matched by manager and industry)

	Number of workers	Indicator for having more than one worker	Indicator for hiring outside workers	ln days worked by manager	Indicator for a fixed premise	Indicator for a fixed premise outside of the home	Indicator for repaying a loan	Indicator for businesses being the manager's primary job
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: First differenced results								
License indicator	0.277*	0.0844	0.00330	-0.0277	0.0221	0.127*	-0.00818	0.00740
	(0.155)	(0.0776)	(0.0257)	(0.0492)	(0.0685)	(0.0763)	(0.0321)	(0.0324)
R^2	0.474	0.348	0.133	0.184	0.401	0.410	0.118	0.133
N	201	201	888	747	201	201	888	755
Panel B: Double differenced results								
License indicator	0.297	0.146	-0.0275	-0.0396	0.136	0.154	-0.0247	-0.0156
	(0.198)	(0.148)	(0.0427)	(0.0759)	(0.132)	(0.127)	(0.0609)	(0.0608)
R^2	0.587	0.363	0.116	0.145	0.423	0.388	0.112	0.107
N	201	201	888	747	201	201	888	755

The table reports the coefficient on an indicator for having a license on the indicated outcome. In Panel A, the dependent variable is the first difference and in Panel B it is the double difference. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.8: Profits and informality by initial size of employment

	No Controls	No controls, trimmed	Controls	Controls, trimmed
	(1)	(2)	(3)	(4)
Panel A: Businesses with one worker				
First differenced				
License indicator	0.0420	0.0445	0.0703	0.0536
	(0.0879)	(0.0715)	(0.0941)	(0.0781)
R^2	0.000	0.000	0.095	0.114
N	902	882	898	878
Double differenced				
License indicator	-0.0328	0.0833	-0.0809	0.0540
	(0.157)	(0.119)	(0.170)	(0.130)
R^2	0.000	0.000	0.091	0.119
N	902	882	898	878
Panel B: Businesses with more than one worker				
First differenced				
License indicator	0.177*	0.143	0.191	0.141
	(0.107)	(0.104)	(0.119)	(0.115)
R^2	0.007	0.005	0.198	0.237
N	411	401	410	400
Double differenced				
License indicator	0.0823	0.0102	0.116	0.0622
	(0.180)	(0.165)	(0.183)	(0.173)
R^2	0.001	0.000	0.237	0.253
N	411	401	410	400

The table reports the coefficient on an indicator for having a license at the end of the period. In Panel A, the sample is all businesses with 1 worker in 2006 while in Panel B the sample is all businesses with more than 1 worker in 2006. The dependent variable is the change in ln profits for "First differenced" results and the change in the difference in ln profits for "Double differenced" results. In columns 1 and 2, no additional controls are added to the regression. In columns 3 and 4, control variables include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. In columns 2 and 4, the top and bottom 1 percent of observations in terms of the dependent variable are trimmed. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.9: Impact of formalization on revenue and expenses by initial size of employment

	ln revenue	ln expenses	Materials share	Labour share	Energy and water share	Non-durables, repair, and depreciation share	Rent share	Taxes and fees share	Other expenses shares
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Businesses with one worker									
First differenced results									
License indicator	0.0710	0.0634	0.0232	0.00833	0.00656	-0.0386	-0.00448	0.0228	-0.0297
	(0.0951)	(0.135)	(0.0300)	(0.00720)	(0.0279)	(0.0293)	(0.0135)	(0.0184)	(0.0288)
R^2	0.108	0.102	0.116	0.150	0.121	0.096	0.099	0.097	0.112
N	913	943	943	943	943	943	943	943	943
Double differenced results									
License indicator	0.0534	0.104	0.0975*	0.0117	-0.0424	-0.0661	-0.00447	0.0435	-0.0601
	(0.164)	(0.242)	(0.0506)	(0.00864)	(0.0496)	(0.0533)	(0.0262)	(0.0308)	(0.0494)
R^2	0.108	0.121	0.143	0.178	0.117	0.114	0.118	0.099	0.128
N	913	943	943	943	943	943	943	943	943
Panel B: Businesses with more than one worker									
First differenced results									
License indicator	0.223*	0.251	0.0523	-0.0365	-0.0117	-0.0528*	-0.0227	0.0253	0.0266
	(0.116)	(0.174)	(0.0499)	(0.0264)	(0.0325)	(0.0309)	(0.0153)	(0.0261)	(0.0300)
R^2	0.276	0.289	0.260	0.301	0.231	0.287	0.414	0.226	0.210
N	417	424	424	424	424	424	424	424	424
Double differenced results									
License indicator	0.111	0.156	0.115	-0.0615	-0.0325	-0.0878	-0.0378	0.0175	0.0654
	(0.173)	(0.296)	(0.0957)	(0.0491)	(0.0644)	(0.0589)	(0.0302)	(0.0456)	(0.0518)
R^2	0.266	0.254	0.202	0.279	0.215	0.219	0.431	0.239	0.197
N	417	424	424	424	424	424	424	424	424

The table reports the coefficient on an indicator for having a license on the indicated outcome. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.

Table B.10: Impact of formalization on labour inputs, location, loans, and job of the manager by initial size of employment

	Number of workers	Indicator for having more than one worker	Indicator for hiring outside workers	ln days worked by manager	Indicator for a fixed premise	Indicator for a fixed premise outside of the home	Indicator for repaying a loan	Indicator for businesses being the manager's primary job
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: Businesses with one worker								
First differenced results								
License indicator	0.0359	0.0379	0.0308	0.00305	0.0849	0.0906	0.0231	0.0595*
	(0.0858)	(0.0854)	(0.0201)	(0.0538)	(0.109)	(0.0931)	(0.0299)	(0.0345)
R^2	0.412	0.420	0.152	0.148	0.473	0.424	0.127	0.105
N	188	188	946	789	188	188	946	801
Double differenced results								
License indicator	0.232	0.228	0.0382	0.0374	0.348*	0.0575	0.0102	0.103
	(0.173)	(0.171)	(0.0244)	(0.0896)	(0.194)	(0.189)	(0.0544)	(0.0648)
R^2	0.381	0.379	0.179	0.113	0.480	0.414	0.116	0.102
N	188	188	946	789	188	188	946	801
Panel B: Businesses with more than one worker								
First differenced results								
License indicator	1.148	0.344	-0.0847	0.0283	-0.0307	0.0376	0.0820	0.0216
	(1.067)	(0.241)	(0.0520)	(0.0724)	(0.0898)	(0.115)	(0.0580)	(0.0546)
R^2	0.597	0.733	0.217	0.310	0.851	0.780	0.212	0.182
N	98	98	426	359	98	98	426	363
Double differenced results								
License indicator	1.627	0.238	-0.145	0.118	-0.0340	0.125	0.0540	0.0208
	(1.736)	(0.314)	(0.0940)	(0.127)	(0.113)	(0.223)	(0.105)	(0.103)
R^2	0.666	0.792	0.228	0.317	0.859	0.740	0.224	0.235
N	98	98	426	359	98	98	426	363

The table reports the coefficient on an indicator for having a license on the indicated outcome. Both panels include industry fixed effects, province fixed effects, urban indicator, gender of the manager, and education of the manager. Heteroskedasticity robust standard errors in parentheses. ***, **, and * denote significant at 1, 5, and 10 percent level, respectively.