Formalize or Not to Formalize: Asking the Right Questions

Albert Berry

ABSTRACT

The dimensions and expansion of the informal sector of developing countries have generated a wide range of interpretations and policy prescriptions. Since much informality, however defined, is a natural and broadly efficient result of market forces, it should not automatically be treated as a problem to be solved. It is important to disaggregate the elements of informality and often treat them separately, since there is no reason that the optimal level of formality should be similar for each of the various elements involved. Some of the undesirable features of many informal firms, like low earnings and income instability, should for the most part be attacked directly rather than by focusing on formalization. The wide range of views as to the desirability of formalization qua policy is the result of the complexity of the issue together with the methodological weakness of many of the analyses, such as the failure to analyze the quality of the informal sector together with its size, or to treat agriculture and non-agriculture separately. The best source of understanding of both the implications of formalization and of the policies that may contribute to its effectiveness is the in-depth study of the process of formalization in actual cases.

JEL Classification: O11, O17, O20, P11

Keywords: Informality, Formality, Tax, Labour Regulations, Labour Allocation; Licensing, Registration, Inefficiency, Productivity, Exclusion, Exit, Mobility

Submitted: Nov. 4, 2016 Accepted: Jan. 2, 2017 Published: Feb. 28, 2017
1. Background to the formalization debate

The informal sector of developing countries has been a matter of heightened interest and attention for several decades; its dimensions and expansion have generated a wide range of interpretations and policy prescriptions – from forced formalization to *laissez faire*, with several of the more common views reflecting widespread misinterpretations or ignorance of the sector itself and of the likely impact of suggested policies. Like so many debates on economic policy, this one involves conflicting views on the normal effectiveness of government interventions as opposed to relatively free markets. It also reflects the re-awakened awareness of the importance of institutions as determinants of economic events that has emerged within the field of economic development during the last few decades. This interest has interacted in various ways with the long-standing attention to the importance of institutions as determinants of economic events that has emerged within the field of economic development during the last few decades. This interest has interacted in various ways with the long-standing attention to the importance of institutions as determinants of economic events that has emerged within the field of economic development during the last few decades.

The “informal sector” is typically defined as those non-agricultural enterprises that operate partially or wholly outside legislated or decreed norms in the areas of licensing, payment of taxes, contracting of inputs (labour, capital, other), and limitations on negative externalities. The fact that a share of enterprises act in these informal ways sometimes means that laws or norms are being flouted (as where taxes that “should” have been paid are not), sometimes that the regulations do not *de facto* apply to them, and sometimes simply that the enterprises are substituting informal patterns of behaviour for the “formal” ones. Thus in each of the labour, land, and capital markets there are alternative institutional patterns to those typically associated with the formal sector. In the labour market, a small enterprise may hire people on the basis of a verbal agreement rather than a written and legally binding contract. In the land market, informal substitutes for the standard Western pattern of exclusive and officially registered ownership may evolve when small enterprises make use of publicly owned (or even privately owned) land. In the capital market, informal credit arrangements typically bypass any regulations having to do with maximum interest rates or other conditions established in law. In what we may call the ‘justice market’ too, there are formal and informal ways of handling things; some social or ethnic groups are famous for being able to pressure members into honest forms of behaviour. There are thus informal enterprises and informal markets. Some of the arguments favouring formalization relate directly to the informal firms, though indirectly also to the informal markets in which they operate. Other arguments have focussed on specific markets, including those for each of land, labour and capital.

Most informal firms are quite small; often, in fact, small size is used as a proxy for or even a presumed equivalent to informality. While the overlap between the two conditions is high and can justify their interchangeable use for some purposes, for others it cannot. When the issue is the pros and cons of formalization, it is very important to distinguish the size spectrum from the formality spectrum. The same caveat is necessary with respect to the various component elements of formalization. Although there is again a relatively high correlation between compliance with one aspect of formalization, such as taxes, and another, such as labour regulations, it is important for policy purposes to disaggregate the concept of overall formalization into its component parts, since the benefits and costs of compliance may

---

2 For an extensive early overview see Lin and Nugent (1995); among more recent contributions, Acemoglu and Johnson (2005) is of interest. Institutions have figured prominently in the literature on the reasons for the “Natural Resource Curse”, e.g. Woolcock et al. (2001). This trend within development economics has in part reflected a similar one in economics as a whole (North, 1991: Williamson, 2000).

3 Usually the focus is on non-agriculture since the features of informality are standard and seen as unexceptional within most agricultural sectors and since the noticeable increase has taken place outside agriculture.
vary significantly among the elements of the full package and since some policy options apply to one element of formality and others to another.

The fact that there is neither a single “correct” definition of informality, nor a definition of universal application further complicates the task of analysis through cross-country and other types of comparisons. The relative size of the formal and informal sectors depends on the definitions used. For purposes of comparisons and aggregation across countries the main (and perhaps the only) practicable proxy for informality, luckily also a useful one, is that employed by the International Labor Organization (ILO); it embraces establishments of five workers or less except for businesses run by a professional or technical person.

Definition matters more than might at first appear for another reason. The differing criticisms and concerns expressed about informality tend to be directed at different features of the firms so defined. For example, those who see informality mainly as an abrogation of the social duty of sharing in tax payments tend to be thinking of countries where that aspect of informality involves not just tiny firms that would be informal by any definition and whose income levels seldom leave them liable for taxes, but also larger ones which according to the tax system should be paying, and/or not-so-small firms that should be subject to environmental regulations but evade them. Here the informality is often part of a pattern of non-compliance, which, in the aggregate creates severe problems for the public sector, both in achieving reasonable levels of public revenues, and in assuring horizontal equity among competing firms. Other commentators are much more concerned with whether the combination of small size and lack of advantages that may come with informality doom the firms in question to low productivity and associated low incomes for the people involved. These concerns are so different that they make it dangerous to talk about informality as if it were a more or less agreed upon and single-faceted phenomenon.

Confusion arises also from the heterogeneity of the firms lumped together into either category of a dichotomy such as formal/informal. For many purposes a two-sector framework of analysis is inferior to a three sector one. A number of authors have argued against the usefulness of the formal/informal dichotomy (e.g. Standing et al., 1996). The merits of such criticisms do not, however, imply that formality, however defined, is not worthy of analysis, but rather that a simple framework that groups firms into two categories, either for theoretical or empirical analysis, may easily be counterproductive. A better approach is to think in terms of formality spectra and size/modernity/labour productivity spectra and to group firms into whatever number of categories is needed to pick up such important thresholds as may exist in terms of the key variables in the analysis. In what follows we attempt to frame the discussion in this less constrained framework but since, as noted, the great majority of empirical analysis has opted for a simple two-sector framework, any overview must to a considerable degree proceed under this unfortunate limitation.

Another issue, if not exactly of definition, certainly of focus, relates to what it is that underpins the formality-informality spectrum (or dichotomy, if one views it that way). By common consent, the character of the firms differs along this spectrum, as do the way the labour market functions, the way the capital market functions, and the interface between the firms and the state (and perhaps private business associations as well). Each person naturally focuses his/her analysis on whatever he/she feels to be the factors underlying important differences in outcomes. Relevant thresholds in terms of compliance and/or size/modernity/labour productivity may well vary according to the analytical focus. We argue below that one of the issues still not firmly settled, much less agreed upon by a majority of

---

4 Thus Fine (1996) and Sender et al. (2005) put the focus on labour markets.
those interested in this area, is indeed what are the key markets or relationships that deserve focus. This is an empirical question; it is one of the reasons to look at all of the contending explanations simultaneously as one tries to sort out the facts of the matter.

Thinking about formalization policy entails consideration of the private and social benefit/cost ratios for each individual aspect of formality, for various combinations of aspects, and for the full package, however it is defined. Private and social benefits and costs will always differ from each other; for policy decisions the social benefits and costs are the directly relevant ones, but for formalization to occur voluntarily the private benefit cost ratio must exceed one.

The private costs of formality – associated with registration, compliance with labour, environmental and quality standards, and taxes may be too high for some small firms to bear. But while saving on some costs, it means the firm may have to pay bribes and forego the assistance of formal institutions in enforcing contracts. Advantages associated with compliance may include better access to credit or even to equity financing, to business associations, export markets, and government assistance programs. Then there are the benefits of economies of scale if growing into the size range where such benefits are reaped de facto requires some aspects of formality. The indirect benefits resulting from growth to a higher size range are one of the causal links between size and formality.

To the extent that formality helps to reduce vulnerability to idiosyncratic risks, this function can be pursued at least to some extent through private institutions as well as through public ones. An often neglected aspect of the formalization discussion is the role of membership in voluntary non-public associations. The possible benefits of such participation include networking with customers and suppliers, informal sanctioning of customers that renege on their contracts, and identification of and in some cases access to legal and technical services. Sometimes private associations substitute for the functions of the state, and sometimes they complement those functions, depending on the circumstances (Sukiassyan and Nugent, 2008). The relative attractiveness of membership in private associations would be expected to depend, among other things, on how well the official judicial system works and with what effectiveness. The two types of membership can be substitutes when judicial efficiency and cost of regulation are both high but complements at other times. Any consideration of the implications of formality must take account of the presence and functioning of these non-governmental associations.

For most informal enterprises at any point of time, either the private costs of formalization outweigh the benefits or the latter are not fully recognized. From a societal perspective there are both good and bad reasons why businesses who could become more formal opt to stay informal; informality is a sort of natural state for some firms; for others that really should be formal it may be permitted by corrupt bureaucrats or by vote-seeking politicians, in what Tendler (2002) refers to as the ‘devil’s deal’. Other informal firms are directly complementary to formal sector ones, usually because they have access to lower cost labour than the larger formal firms.

2. Why the recent interest in the formalization question?

The current debate on formalization has been fuelled by three mutually reinforcing facts or beliefs: (i) the large size, and in some countries and periods the rapid growth of

---

5 Employing an unusually detailed data set on Mexican micro enterprises and state-level information on the quality of the judicial system and on regulatory costs Sukiassyan and Nugent (2008) find, as expected, that a weak judicial system discourages registration while encouraging participation in a non-governmental association.

6 Kenyon (2007) reports that many small and medium sized firms go out of their way to remain informal.
‘informal’ activities; (ii) the low labour productivity of those activities, which is presumed to contribute to poverty; and (iii) the (more controversial) idea that these activities interfere with the effective functioning of the economy as a whole.

On the first point, according to Kenyon (2007, 2) best estimates are that 30% of output and 70% of workers in the developing world are to some extent outside the scope of government regulation. In Latin America, as of 2000, about 40% of urban employment involved non-professional or technical workers in establishments of under five workers, and another 7% or so were domestic servants – sometimes treated as part of the informal sector (Berry, 2011a, Table 5).

Adding further importance to the formalization issue is a fear that current trends in the developing world do not point to a repetition of the historical experience of the now-developed countries, in which the share of output and employment that was informal fell as development proceeded beyond a certain level. For example Freeman (2009) argues that “The informal sector increased its share of the work force in the developing countries in the past two decades”8. Though the data used do not in fact support such a clear reading (Berry, 2011b) nor the conclusion that the informal share has been rising even in fast growing countries (Freeman, 2009, 36), there is certainly a significant possibility – more likely a probability – that that growth in many of today’s developing countries does not lower the share of people employed in the informal sector as much as was the case at comparable points in the evolution of the industrial countries.

If such is the case, several factors may have been playing a role. One, undoubtedly, is the very capital-intensive technologies now available to and used in developing countries, even when they suffer from a surplus of labour. Another possible factor at work, with more positive implications than the others, is that current day informal sectors may be relatively more productive than their historical counterparts, leading more people to wind up in them. If this were the case, the presence of microcredit, higher levels of education than in the past, or other factors could be at work. Sorting out what the actual trends have been and what they mean should be a high priority research objective.

3. The normative: points of view and conceptual frameworks for thinking about formalization

In part, it is the size, trends and income conditions associated with informality that fuel the interest in and concern with this phenomenon. At a more analytical level, though sometimes related to the above concerns with size and trends, are a set of what one might call microeconomic reasons to believe that informality should be curtailed or discouraged through public policy of some sort because either (i) informal firms compete unfairly with formal sector ones; (ii) they do not pay taxes or that they pay unduly low wages; and (iii) their informalities have a productivity cost and thus holds down the incomes of those working in the sector. Policy prescriptions vary from the view that a major impediment to formalization is bureaucratic red tape, whose removal would pave the way for a wave of formalization, to the idea that lack of collateralizable assets that can be used for access to credit constrains growth

7 This relationship remains clear in cross-country comparisons between more and less developed countries (e.g., La Porta and Shleifer, 2008, Table 1) but the question is whether the relationship between informality and development has been shifting over time such that a country’s informality trajectory cannot be predicted by the cross-section regression, with informality either falling less rapidly than that regression would suggest or perhaps not falling at all.

8 This conclusion is based on ILO data (ILO, 2002, 62-64) that are interpreted as showing an increase in this ratio between the 1980s and the 1990s from 26% to 32% for the world, with accompanying increases in each of Africa, Latin America and Asia (Freeman, 2009, 47).
and formalization, to the view that positive support policies for small informal firms are what matters the most, to the view that firms should be simply forced to formalize or otherwise disappear. Finally some observers either do not see a major economic inefficiency in this large informal sector and/or do not see any way that informality can be quickly reduced without doing more harm than good.

Most of the negative attention to informality relates to concerns that it hampers the core goals of economic development: economic growth, remunerative employment, and a reasonably equitable distribution of income. From the fact that early development thinking typically portrayed the informal sector as ‘backward’ and ‘traditional’, it seemed to follow that formalization would contribute to modernization and economic development, both by increasing tax revenues available for public spending and by ensuring the fulfillment of product standards and labour and environmental codes as these enterprises were brought into the realm where they could be better monitored. Along with being a symptom of a growth process gone somehow awry, informal activities have been viewed by some as creating unfair competition for formal ones which, after all, must pay taxes, satisfy labour legislation, etc. Others see the problem with the informal sector as the lack of access to productivity-enhancing services that are tied to formality, including access to capital which may be contingent on having property rights to land. Finally, others have argued that informality is an impediment to effective social policies, when these are tied to employment (Pagés, 2009).

In retrospect it is now clear that much of the initial speculation to the effect that a large informal sector represented a serious labor misallocation was either wrong or seriously exaggerated. The idea that people working there were, in general, isolated or marginalized from the formal sector (Vekemans and Silva, 1969) was fairly quickly disposed of (e.g. Perlman, 1976). Another negative interpretation had the informal sector systematically subordinate to or exploited by the formal one (Moser, 1978). At the other end of the spectrum many informal activities were seen in a positive light by the ILO in its 1972 Kenya report (ILO, 1972), one of the documents that put the term “informal” into the lexicon of development economists, questioned the presumption that informal activities were in any meaningful economic sense less productive than the formal sector, and emphasized their pivotal role in job creation. The ILO returned to this theme in a 2002 report entitled “Decent Work and the Informal Economy” (ILO, 2002), which emphasizes the long-run contributions of the informal sector and its role as a complement to the formal sector rather than as a substitute for it. The labor-intensity of most informal sector activities suggested to many (e.g. Liedholm and Mead, 1987) that its narrowly-defined efficiency levels were typically high when calculated properly, that is by using scarcity rather than market prices as the cost of each factor of production. Presumably its ‘broad social efficiency’ (allowing also for its impact on inequality and poverty) would be higher still, given its importance as a source of income for poorer families.

Related to the question of how informal (usually small) enterprises differ from more formal (usually larger) ones is the question of whether this distribution implies a significant degree of economic inefficiency. At one extreme is the view, still heard on occasion, that the informal sector is simply ‘inefficient’ (e.g. Masatlioglu and Rigolini, 2008), something of a throwback to the pre-1970s period. For holders of this view, the question is why so many

---

9 Such “unfair competition” has for example been linked to their having access to land or other property the rights to which are unclear, and sometimes to using (unrented) sidewalk space from which to compete with commercial establishments that do pay such rents.
10 Hart (1973) presented a similar story.
11 See Williams and Round (2007) for their study of informal enterprises complementing the formal sector in Ukraine.
resources are found in the informal sector and what policies could most effectively transfer them to the formal one. A related frequent interpretation of the large informal sector is bureaucratic (red-tape) ‘barriers to entry’ into the formal one, as popularized by de Soto (1989) and extended by Djankov et al. (2002). At the other end of the spectrum are those who see informality-formality as simply a manifestation of a plausible size structure of firms given the resource endowments and technology shelves of the countries in question; for this group there is no presumption of inefficiency attached to a large informal sector and the scope for formalization policy is usually seen as at best small. Just as the historical process of growth in the now industrialized countries has seen a secular upward shift in the distribution of enterprises by size, so might today’s developing countries.

Most analysts fall somewhere between the above extremes, recognizing that there are barriers which deter the formalization that some firms would otherwise undertake, but also that for many informal firms that step does not make sense and that, given the factor availability constraints they face, they are in any case relatively efficient. For reasons explained below, the author of this paper calls this the ‘empirical economists’ view. Within this perspective it is widely held that capital market imperfections go a long way to explain why so many firms are small and extremely labour intensive. But product market imperfections may work in the same direction, as where large-scale monopolists take steps to curtail potential competition from smaller producers. It is not easy to sort out how different would be the structure of firms by level of formality (and with it by size) in an economy with fully efficient markets, nor the relative importance of the various economic and political mechanisms that keep it from that ideal nor, hence, the optimal policy steps implicit in a full understanding of this aspect of the structure of an economy. Almost everyone would agree that more labour should be in the formal sector firms; many would agree that more capital should be in the (currently) informal firms, but there would be no consensus on whether an efficient economy would see more or less output in the informal sector than it now has.

La Porta and Shleifer (2008, 1-3) draw a somewhat different taxonomy, which ranges across the ‘romantic’ view of those like de Soto for whom informal firms are actually or potentially extremely productive, waiting only to be freed from the shackles of bureaucratic regulation, taxes and the lack of secure property rights; the “parasite” theory under which these firms are inherently inefficient and unproductive but are threatening to the official ones and need to be displaced by efficient formal firms (“Wal-Mart” theory of development); and the “dual” view, which also holds that unofficial firms are inherently inefficient but recognizes that these firms provide the livelihood of millions (Tokman, 1992). The first two views portrayed by these authors overlap considerably with two of mine but their (favoured) “dual” view differs from my third (and favoured) one, in that it does not recognize the actual or potential economic efficiency of many small, currently informal firms. Among the three alternative views they present, none accepts such potential except for the

12 That is, barriers that cannot be defended on the basis of overall societal welfare (as opposed to that of individual groups with specific interests).
13 The threat can take the form of “unfair competition” with specific groups of formal firms or of the informal firms sharing in public services without helping to pay for them. Thus, Loayza (1996) argues that the presence of the informal sector, because it does not pay taxes but does use public services, lowers the economic efficiency of the system. Asea (1996) presents a very different view. As Schneider and Enste (2000, 89) point out, the hypothesis that the informal sector constitutes a drag on economic performance is not generally accepted. In fact, however, the discussion of the previous sections of this paper implies that asking the question in this very general way is in any case not very helpful, among other reasons because the informal sector is very heterogeneous.
14 One root of this view was the “big push” models of development thinking (e.g. Rosenstein-Rodan, 1943; Rostow, 1960).
“romantic” one, which they quickly discard as unrealistic. If one includes both their ‘dual’ view and my ‘empirical economists’ view, but omits the ‘exploited sector’ view and the ‘marginalized/isolated sector’ view, then there are four perspectives that need to be borne in mind.

A closer look at this wide range of views reveals that some do not deserve to be taken seriously and that, among those that do, there has been an evolution of ideas over time.

The easiest perspective to largely rule out from detailed consideration is the “parasite” or “unfair competition for formal firms” theory. It reflects an accurate perception that in some cases informal firms do engage in the sort of behaviour described, but a failure to reflect either the usual infrequency of such competition or any bigger picture of the economic and social contribution of the informal firms. At best, this perspective tells only a small part of the story and it tells it in an incomplete and usually distorted way.

Also easy enough to deal with is the “romantic” (La Porta-Shleifer’s term) or optimistic view that the informal sector can, under appropriate conditions, provide a significant or even the main dynamic to a country’s growth. This view is not necessarily the result of inaccurate observation or interpretation of current reality, since its main, forward-looking, contention is that this sector can under favourable conditions provide much dynamism to the economy and contribute greatly to poverty reduction. To know whether that appraisal is true, however, would probably require going through the experiment and waiting until the results were in.

The “empirical economists” view of this paper’s author typically involves a less optimistic judgment than the romantic one, but it does incorporate the idea that the informal sector can, at a point of time (e.g. now) contribute positively to the economy even if it does not have within it the potential for great dynamism extending into the future (though in some cases it might). The differences among those who hold these views are partly ones of degree (how much potential does the sector have?) and are sometimes linked to how innately cautious different researchers are.

For purposes of this discussion it is most important to understand the “small is inefficient” view that characterizes two of the perspectives distinguished by La Porta and Shleifer (2008). It seems mainly to derive from a failure to understand basic concepts of microeconomics, ideas that allow one to see why a modern technology may be inappropriate in a developing country setting, to the point where its use in preference to a less capital-intensive one can not only reduce the demand for labour and thereby worsen income distribution but also lower GDP. It is often accompanied by an oversimplified and/or erroneous view of what “efficiency” means in economics (e.g. its equation with labour productivity) and on how difficult it is to measure in very small firms, and by a simple intuition that the inefficiency of small informal firms is self-evident. The idea that bigger is better is of course widely held within business circles, especially big business ones. This presumption often combines the cited lack of understanding of economic concepts and an unfamiliarity with the writings of economic specialists in this area (e.g. Liedholm and Mead, 1999) to produce a strongly held view. Of course the fact that it is not a well-reasoned or

---

15 Although providing a very valuable review of much empirical evidence and an interesting take on the different views about formalization, La Porta and Shleifer (2008) systematically fail to distinguish between labour productivity (which they refer to as “productivity”) and total factor productivity, the economist’s standard definition of economic efficiency. This, unfortunately, leaves their conclusions both biased and uninterpretable in economic terms.

16 One need not go nearly as far as Schumacher (1974) in lauding the merits of smallness to recognize the basic economics that can make small enterprises efficient, whether or not they are “beautiful.”
An empirically informed view does not necessarily mean that it is wrong. The presumption that economies of scale are sometimes important is also incorporated into the "empirical economists' perspective but there it is something that must be measured in order to judge how important it is.

Those who feel the discussion ends with the alleged “fact” that larger formal firms are virtually always more efficient than informal ones and that a country should move in the former direction as quickly as possible are often also unaware of the possible income distribution implications of such a restructuring. Precisely because large firms use more modern technology and achieve higher labour productivity than small informal ones, the general equilibrium impact of the rapid replacement of the latter by the former could be a major increase in underemployment and poverty, as elucidated more than half a century ago by Eckhaus (1960). Thus, the other element of economic thinking that is essential in dealing with the formalization question is general equilibrium analysis – the ability to extrapolate the implications of major changes occurring at the firm level to economy-wide outcomes, especially in the labour market. Economic theory, in taking account of the fact that a resource in surplus supply (as with labour in many very poor countries) has a very low economic cost, not only explains why smaller firms with simpler technologies may be more efficient than larger ones, but also provides an efficiency basis for a better distribution of income.

The dialogue among those with a considerable empirical understanding of developing economies and of smaller firms by now excludes the "informal is necessarily inefficient" generalization and has come to a more refined and nuanced view of the potential contribution of smaller firms, including those in the informal sector, to economic growth and social well-being. As argued below, a useful discussion of the formalization question requires not only a grounding in how small firms work and how the entrepreneurs who operate them behave, but also in the economics of labour markets, the credit market, the tax system, and the provision of public services, all of which constitute elements of the debate on when, how often, and through what policy instruments formalization should be fostered by public policy.

The various views of informality just mentioned bring with them an overlapping intellectual divide between those who see the informal sector as a ‘problem to be solved’ though shrinking it as rapidly as feasible; those who see its presence, together with the heterogeneity of firm size, technology and productivity, as indicative of market or other inefficiencies that can be reduced; and those who see it as a natural feature of economic development, that does not necessarily imply an unusual degree of inefficiency nor, hence, call for any particular policy attention. Those who see informal firms as inherently inefficient look for ways to expand the formal sector as quickly as possible. Those who are inclined to see the reason for the existence of much of the informal sector as one of barriers to entry into the formal sector focus on lowering those barriers, such as red tape, taxes, artificially high labour costs, or lack of access to capital. Those falling in the third category often focus on the same sources of inefficiency but do not start with the presumption that formality is a universally necessary condition for satisfactory performance. This latter distinction involves differing views on the sources of low productivity. If a basic problem of informality is that economies of scale prevent firms from reaching their potential productivity until they attain a certain size and they are blocked from doing so by bureaucratic red tape, labour legislation, or other obstacles related to formalization, when the blockage is dealt with the firms should grow into efficiency, leading to a shrinkage both of the informal sector and the number of very small firms. If, in contrast, some aspects of formality that bring benefits are expanded in coverage in a situation where economies of scale are unimportant, then informality will shrink though the structure of firms by size will not change.
Behind the question of how informal (usually small) enterprises differ from more formal (usually larger) ones is the more general question of what determines the distribution of employment and other factors of production across the size-technology-formality-productivity spectra and whether this distribution implies a significant degree of economic inefficiency. Part of this general question is the starting point for any interpretation of information – what explains its existence in the first place?

4. Background on the Role of Smaller Enterprises in the Developing Economy

The context for the formalization question involves the differing roles of more formal (usually larger) and less formal (usually smaller) firms in a developing economy. For the developing world as a whole, the urban informal sector (overlapping heavily with microenterprise) probably produces 25-35% of total urban output and income but generates a much higher share of employment and of the earnings of the lower income groups, while large formal firms, at the other extreme, produce a sizeable chunk of total output but generate much less employment, especially for lower income people. The intermediate small and medium enterprise (SME) sector usually generates roughly comparable shares of output and employment; in SME-oriented economies like Taiwan that share is high, whereas in highly dualistic economies like most of those in Latin America it tends to be low, inviting the term ‘missing middle’ to describe the firm size structure. It is widely accepted that highly dualistic countries suffer more acute income inequality than do countries where the SME sector is prominent. In the former, the capital share of income in the large scale formal sector is usually high and workers (the so-called “labour elite”) are typically relatively well paid, while because so many people are forced into the informal sector the equilibrium wage in the economy is low, set by the productivity of the marginal firms in that informal sector and in small-scale agriculture. By helping to determine the equilibrium wage rate, the informal sector affects the earnings of many people, including some who work in other sectors where labour legislation is not operative to raise wages above equilibrium, such as most of agriculture and part of the SME sector.

Figure 1 provides a simple picture of the interface between firm size and the labour market at a point of time, assuming no policy interventions in that market. The labour demand curve of the large firms (LL’) intersects the vertical axis high and has a steep slope, reflecting the high productivity of the first few workers in such firms (because of the large amount of capital and the modern technologies they work with) and the rapidly declining marginal productivity of labour as more workers are added to a fixed amount of capital. The SME sector’s labour demand curve (SMSM’) reflects its intermediate technology and capital/labour ratio, intercepting the vertical axis lower than LL’ and have a flatter slope. Finally the microenterprise demand curve (MiMi’) has a low vertical intercept (even the first workers in such firms have only modest labour productivity) and is much flatter since with only a little capital these firms can put a lot of people to work and because the monopolistically competitive character of many of the markets in which these firms operate facilitates entry. The total demand curve for labour (TT’) is the horizontal summation of those for the three size categories. The segment TA reflects only the large firm sector and thus coincides with the first part of that sector’s labour demand curve. From point A (i.e. for labour productivity and wage levels below point A) some SMEs also hire

17 As noted above (footnote 2) the treatment of agriculture is problematic and the interesting discussions almost all revolve around the non-agricultural component of the economy. However, many estimates of the informal sector share of GDP and employment do not allude to this complication. When the informal sector is defined to include underground activities in economies where these flourish, its share of GDP can get large, up to half or more. But such cases are not our focus here.

18 In the case of microenterprises, much of the demand is for own-employment so if the term “wages” implies paid employment then it may be more appropriate in this case to refer to labour earnings or implicit wages.
workers, so the middle part of TT' (the segment AB) reflects the hiring of both the large and the SME sectors. Finally, when wages are low enough the microenterprise sector becomes a participant; the segment BT' of the total labour demand curve reflects labour hired by the three sectors together as wages fall further, but primarily by the microenterprise sector since its demand curve is relatively elastic, as befits its easy entry character. Most of the marginal workers (the last ones to be employed as wages notionally fall in the neighbourhood of the equilibrium wage) are from the microenterprise sector. For the economy-wide wage (or labour earnings) to rise the overall labour demand curve must intersect the supply curve (SS') at a point higher than W", i.e. the labour demand curve TT' must shift upwards, more specifically the part of that curve which is near the supply curve must shift up.

![Figure 1: Interface between Firm Size and the Labour Market](image)

This illustrates the fact that inducing growth of the large scale sector may at best do little for overall wages in the economy as long as the sector hires few workers. Since output growth in that sector is often accompanied by a decline in the wage elasticity of its labour demand curve (such that it gets steeper), labour demand in the relevant range near its intersection with the supply curve may not increase at all or may even decrease. Realistically, it thus follows that for some period of time into the future there are only two ways to raise market-based labour earnings: (a) by increasing the productivity of microenterprise so that the segment BT' of the total labour demand curve shifts up, and the equilibrium wage with it and (b) through growth of the SME sector as more firms come into existence and the current ones grow, thereby shifting segment AB of the overall labour demand curve towards the right to the point where it eventually intersects the SS' curve, i.e. the SME sector expands until it has essentially absorbed those who previously worked in the microenterprise sector. Since SME firms have higher labour productivity than microenterprise firms, this shift of workers allows the overall wage rate to rise significantly.

A country’s structural features and policies determine what combination of these paths out of poverty it takes. All forms of policy support that help to raise the productivity of informal firms at reasonable cost are good; all supports for SMEs that either raise productivity or simply allow the sector to grow are also likely to be socially beneficial. It is
important that both the microenterprise or informal sector and the SME sector be allowed to fulfil this task as far as possible, given that the large firm sector is for the time being incapable of doing so. Thus an effective microfinance system is a good thing, as is a good system of technological support for SMEs. Where, among the instruments that are useful in raising the productivity of the small enterprise sector (including both informal firms and the often partially formalized SMEs) does formalization policy fit? Policies that successfully induce the large private sector to be more labour intensive are in principle helpful as well but have in practice been rare. ¹⁹ A better bet is to facilitate their subcontracting with smaller firms that create more employment. At the extreme where the growth of large firms will have a strongly negative impact on employment, as where such firms cut SMEs out of the markets in which the two groups compete, facilitating their growth is clearly questionable.

Apart from its central function of creating jobs, often for half or more of the urban labour force, and determining the equilibrium wage level for the economy, the informal sector constitutes a proving ground for new small firms, where they can benefit from a gestation or learning period without yet having to confront all of the costs and responsibilities of formality. The existence of this laboratory, in which nascent entrepreneurs find out whether they can survive in their chosen activity, allows more firms to explore their business skills than would otherwise be able to. Some eventually grow into medium or large operations. ²⁰

Looking beyond the present, of course, it is very important that the setting be one within which each category of firms can, through investment and productivity improvement, make its maximum contribution to growth, employment and other economic goals. In this connection, therefore, it is essential to ask how various types of formalization policy may affect these motors of growth. ²¹ It is important that any group of firms be in a position to invest at a rate that contributes to a satisfactory performance of the economy as a whole. The extent to which this goal is satisfied depends on the aggregate savings and investment rates in the economy and on whether the investment occurs in the socially optimal places, which in turn involves the efficiency of the capital market. In the case of productivity increase, the story is more complex and, at least in the context of developing countries, not well understood. The central actors in productivity increase are easy enough to identify:

---

¹⁹ Thus authors like Prahalad and Hammond (2002) who see the largest and most modern firms as the potential antidote to poverty in the short run really do deserve the term “romantic.”

²⁰ This “laboratory” argument is outlined by Nelson and Bruijn (2005), who emphasize that an economy with little space for experimentation by new small firms an economy loses out on some of its potential. Another loss is of the networks of associations among persons and firms that may work better than their counterparts in the formal sector (Sukiasyan and Nugent, 2008; Wong, 2008).

²¹ Explaining why informality itself or its determinants are important thus requires a focus on these dynamic motors of economic growth and how they affect investment and technology choice and hence economic growth and income distribution. Much of the country-specific analysis of this issue involves Latin American cases. Labour institutions (especially unions) have been blamed for slow long-run growth in Argentina, and for impeding adjustment and labour mobility in Brazil and other countries thus contributing to the painful nature of the 1980s recession and the resulting high levels of unemployment. It has often been suggested that the lesser degree of protection/intervention in favour of the formal sector workers in the East Asian than in the Latin American countries contributed to those countries’ superior growth performance, rapid labour absorption and lower levels of inequality. But many other factors may have been at work to create these differences. Latin America’s highly regulated (at least de jure) labour regimes have not prevented some countries (including most notably Brazil, but also Mexico, Venezuela and Costa Rica) from achieving high investment rates in the 1960s and/or 1970s nor the region as a whole from growing fast over 1945-80. A reallocation of capital across sectors could arguably have a bigger effect on GDP since the imperfections appear much greater there. Here, however, there is less empirical data to go on and it may be overly optimistic to think in terms of a very efficient allocation of capital across sectors.
researchers, inventors, entrepreneurs, imitators, and the set of policy-makers whose decisions define the context for the actions of the above economic agents. Some of the characteristics of good performers in each of these categories are also fairly well established, and though the process of technological change has been much more intensively studied in industrial than in developing countries, some of the resulting leads are of likely relevance to the present theme. Among stylized facts are the following:

i) Most revolutionary or breakthrough changes in technology come from small enterprises, some of which tend to be much more willing to take risk than their larger counterparts (Baumol, 2008, 576-77);

ii) The capacity to imitate effectively can be as important as the capacity to invent and innovate. Borrowing is very important, both between firms and between countries; and

iii) It is sometimes in the interest of large firms to block innovations pursued by smaller ones, when the latter would threaten their market position.

A variety of hypotheses posit causal links between formalization and technological advance/productivity change. Informal status is often seen as impeding the process of productivity enhancement. Alternatively SMEs or potential SMEs may be discouraged from risk-taking – and hence from experimenting with new (to them) technologies – by excessive or erratic imposition of taxes or labour regulations. Much productivity increase at the firm level involves learning by doing and by imitation. It is likely that both the risk-taking and the learning aspects of good entrepreneurship can be fostered by a positive interface between small and/or informal firms and the public sector and discouraged by a bad one. The question is whether various aspects of formalization are part of a good interface or a bad one.

5. Judging the Informal Sector and its Components: Concepts and Analysis as a Backdrop for Policy Thinking

A convenient conceptual framework to aid in the analysis of the informal sector and of how policy towards it and the mechanisms which determine its size and characteristics should be designed involves distinguishing its static (narrow) efficiency effects, its dynamic (narrow) efficiency effects, and its broader socio-economic efficiency effects. In this taxonomy, “narrow efficiency” refers to the maximization of national income, whereas broader social efficiency includes also impacts on employment, income distribution, happiness and any other elements of the social welfare function.

A look at some of the relevant parameters of many developing countries provides some guidance as to the impacts of informality on the level and growth of GDP. The evidence suggests that we should not focus much on the possible benefits from static resource reallocation (e.g. movement of labour from the informal to the formal sector) but rather on those related to investment and productivity change – the dynamic motors of growth. For the typical developing country it seems unlikely that a point of time reallocation of labour among the formal, urban informal and rural sectors could raise GDP by a significant amount – say 3-4% of GDP. It thus seems clear that shrinking informality could only have a significant

---

22 An excellent review of the key ideas surrounding the process of technological changes is Baumol (2008).
23 Consider a case in which the formal sector employs 25% of the labour force and the informal (which includes most of agriculture) has 75%; there is a 40% earnings and marginal product of labour gap between the two sectors; the elasticities of demand for labour (and of the marginal product of labour (MPL) are -1 in the formal sector and -2 in the informal sector; and the initial labour share of GDP is 60%. Under these conditions, a reallocation of labour to the formal sector sufficient to equalize wages and MPL between the two sectors would involve an increase in formal sector employment to about 31.7% of the labour force and an increase in GDP of about three quarters of one percent. The combination of assumptions would have to be rather extreme for the
impact on GDP if it worked mainly through the sources of growth – investment and technological change. This, together with the role of informal enterprise in employment creation as described above, the major objective of policies directed towards it, and in particular towards that part of it that is small, newish and of low labour productivity, should be to raise productivity, stability and income generating capability, which will sometimes but not always involve fostering firm growth. Other objectives should be secondary. If discouragement of informality were to lead to a significant growth of formal sector employment, either by the previously informal firms or by others, the case for such policy would be clear. In fact, however, since the social necessity of the informal sector comes directly from the inability and/or unwillingness of existing formal sector firms to create enough jobs for the available labour force, it must be demonstrated that pushing formality will improve the performance of currently informal firms in this respect before such a policy can be taken seriously. Sometimes, certainly, this is the case, as noted by Tendler (2002), but how often this is so remains the big question.24

In brief, the merits of formalization depend mainly on (a) whether there are systematic efficiency differences between formal and informal firms which are causally related to formality and (b) whether the limitations of the formal sector in employment-generation are so serious as to make many informal firms desirable even if their total factor productivity is below that of formal firms. Other things equal, formalization is more likely to be beneficial from an aggregate output point of view if the productivity gap is large; however, since the gap’s existence does not per se imply that the relationship is causal, this must be established empirically before any such benefits may be presumed. One possibility is that total factor productivity is approximately the same in both sectors and that all or nearly all of the differences in labour and capital productivity are due to differences in factor proportions. In that case the marginal product of each factor would, like the average productivity, vary between the two sectors, causing a loss of aggregate efficiency and inviting analysis of what factor or product market imperfections have caused the imbalance. The main policy implication would be that more labour should be shifted to the formal sector and more capital to the informal one. If formal sector firms are generally more efficient (achieve greater total factor productivity), not only more labour and material inputs but perhaps even more capital should be absorbed by the formal sector. Whether formalization of very small firms should be pushed as a matter of policy thus depends both on the relative economic efficiency of informal enterprises vs. their formal counterparts and on the sources of those differences. Unfortunately for the goal of simplicity, defining and (especially) measuring economic efficiency is not easy.

A key distinction in thinking about the informal sector is that between constrained and unconstrained efficiency. Put another way, it may well be true that, in a perfectly efficient economy (no imperfections in any markets) there would be many less small firms with such low labour productivity as typical informal sector firms have, i.e. these firms might not be part of an optimally functioning economy. At the same time it may be true that unless and
gain to reach 2% of GDP, a gain that in reality would be spread out over, however, many years this adjustment process required. Adding on an output gain related to reallocation of capital to the informal sector (where its marginal product would initially be higher) could add some further gain.

24 In the 1980s, Prime Minister Lee Kwan Yew of Singapore adopted a policy of pushing wages up with a view to pressuring firms to raise productivity. The policy was much debated at the time; if in fact it was a good one, this was because Singapore had so much growth momentum that the equilibrium wage was rising fast and destined to continue to do so. For firms which did not foresee further rapid wage increases, this form of signaling may have helped them to make wise technology decisions. But Singapore, with its extremely fast growth, was a very unusual case.
until certain impediments to that economy-wide efficiency can be removed, these firms are, for the most part, ‘efficient’; they are doing as well as possible given the resources available to them and the setting in which they operate. So until the reasons for resource misallocation are dealt with, these firms remain desirable – they raise GDP, employment and the incomes of the poor. At bottom, informality is not what anyone should be worried about, but rather low productivity, at both the firm and the economy-wide levels. So whether informality defined by degree of compliance to rules and regulations or by small size declines or not is not the important thing. The main reason to be worried about informality is that it may be not just a correlate of low productivity but a cause thereof. This approach implies that the policy question should not be “what do we do about informality?” or “how do we shrink informality?” but rather “is a reduction in informality part of the solution to low productivity and/or low incomes?”

So what, if anything, should public policy do about informality? One approach, consistent with some of the more positive interpretations of informality, is to do little or nothing about it directly, on the grounds that firms will evolve naturally into formality and/or seek it when the time is ripe, e.g. when enough capital is available in the economy to allow them to raise labour productivity, pay taxes, etc. For growing firms formalization is a natural process that comes with size, age, complexity of structure, and visibility. Many features of very small, recently created enterprises suggest that informality of various sorts is their natural state. Thus labour legislation may be largely or wholly irrelevant to a one-person firm or a family firm. A tiny new firm may not be making any profits so the income-profit tax system is also irrelevant. Large, established firms, in contrast, naturally function in the ‘formal’ economy. This laissez faire approach to informality, in the sense of not designing policies specifically to raise formality per se (as opposed to policies that will lead to an increase in formality by making firms more productive) has its greatest logic when formalization tends to occur on its own under proper circumstances and/or when it does not matter much anyway for grey area firms whether they do or do not formalize, as well as when government may not realistically be expected to apply policies well. Some recent evidence tends to support the view that neither the benefits nor the costs of certain components of formalization are very large.

Helping formality along. Some of those who see the key objective as raising the productivity of informal sector firms hold that because formality can help to achieve that goal, incentives (whatever combination of penalties and rewards best achieves the objective) should be provided to facilitate the process. Thus, if collateral improves access to credit and the titling of informally held property provides that collateral, such titling should be encouraged. The benefits of such formalization may be seen as an appropriate trade-off from the firm’s perspective against such costs as having to pay taxes, meet more labour obligations, etc. Formality and its potential benefits can also be made more attractive by reducing its disadvantages. Curtailing red tape and other unnecessary bureaucratic requirements is seen as good for business in general, but in fact its advantages are likely to be greatest for those small firms that are or might be subject to them. The World Bank’s studies of the cost of such red tape are revealing. Labour regulations and taxes are the other two most commonly cited disincentives to formalization. While the details must be worked out in each case, it is

25 A more detailed discussion of the concept of economic efficiency, as it applies to the subject of this paper, is presented in Section 2.
26 USAID (2005) emphasizes that the basic objective of policy should be improvement in the quality of life of the informal sector workers and Maloney (2006) considers that it is appropriate to tolerate informality.
27 Levenson and Maloney (1998) note the systematic links between informality and the age and size of firms.
28 See Berry, 2011b.
generally understood that small firms should not be subject to as severe requirements in these areas as are large profitable ones. It is an empirical question how much each type of regulation or imposition affects willingness to formalize in general and, with that, the likely willingness to grow.

Care not to proceed too fast. Those who see informality as having enough positive features that early exit should not necessarily be encouraged argue that care should be taken not to damage functioning networks that provide a basis for the productivity that informal sector firms do achieve. Some thinking has gone into how these supporting networks and the beneficiary firms might be made more productive, including the roles of microcredit (especially that not based on collateral), education and training, and the curtailment of harassment by public officials, but the literature so directed remains rather small, except for that relating to microcredit.\(^{29}\) Most people holding these views do not go so far as to maintain that the informal sector should stay large forever, but rather than it will be an important contributor to societal goals for long enough into the future to make it important to retain and build on its current strengths.

Given the enormous heterogeneity of the enterprises grouped under the ‘informal’ umbrella, and given the different and discrete aspects of ‘being formal’, any serious discussion of ‘formalization’ policy needs to be carried on at a significant level of disaggregation.\(^{30}\) Some “informal” firms are engaged in criminal activities and others have negative environmental externalities; such groups should be reined in regardless of size, though if smallness facilitates the crime in question the smaller ones will require special attention. Some enterprises are partly formalized, e.g. they have licences to operate and pay some taxes but do not meet the requirement of the country’s labour legislation; some have just been started and need a little breathing space before these impositions are levied on them; some exist only temporarily anyway, and need the flexibility to begin and end production at the drop of a hat. Some are potential beneficiaries of support programs provided by the government but contingent on a degree of formalization. Others are able to access support from private associations.

The real policy question. While this heterogeneity implies that there could be no broad, simple answer to whether more ‘formalization’ is desirable nor to how it should best be brought about, it remains an important question whether a preponderance of informal establishments would make a bigger social contribution if more formalized than at present or, more specifically, for how many categories of firms would adopting a given element of formalization (say, getting a licence to operate) bring a net social benefit and for how many a net social loss. Putting the question in these broad terms allows, among other things, for the possibility that formalization should be increased in some dimensions (e.g. licensing) but decreased in others (e.g. liability to certain types of taxes), increased for some groups of informal enterprises but decreased or left unchanged for others, etc. Perhaps, for example, adherence to labour legislation should not be pushed (i.e. the informal segment of that market should not be tampered with) while the land market should be formalized (as often suggested). Countries differ in the extent to which one aspect of formality is linked to the others, e.g. the extent to which liability for taxes is tied to receipt of services. Where the links are tight, the decision to formalize or not tends to have an ‘all or nothing’ character. When this is not the

---

\(^{29}\) A somewhat related view is that regulation is necessary to protect some informal enterprises against flooding of the market for their product to the point where the activity no longer provides subsistence for anyone.

\(^{30}\) References in this discussion to the informal sector or to formalization should thus for the most part be taken as shorthand references to phenomena that, for purposes of serious policy discussion, warrant such disaggregation.
case firms can cherry pick to some extent and/or upper level policy makers may push one aspect of formalization but not others.

Both age and size of firm (and probably profits) would be likely correlates of possible benefits from formalization. The key groups are those that one might call ‘marginally informal’ either in respect of a specific element of informality or more generally. Focussing on these groups in the case of taxes, for example, directs attention to those whose incomes are high enough to imply that paying taxes would not constitute an undue burden and which can also be induced to do so at reasonable cost. The same goes for labour legislation.

A final broad consideration in the framing of policy in this area, as in any other, is administrative competence, probity and political neutrality. Only strong administrative systems can successfully apply complex policies. Weaknesses in these areas mean that the best feasible approach will often entail substantially less regulation and other forms of intervention than would in principle (i.e. with perfectly efficient implementation) be desirable. They create, in other words, a general case against intervention – in the present context against using policy to force formalization. Apart from suggesting upper limits as to what can be expected of policy, the qualities of an administrative system sometimes suggest that relative efficiency will be greater when the policy involves positive inducements rather than negative penalties, or vice versa, so this consideration will also weigh into the design of an optimal policy package.

To summarize, a higher level of formalization, whether in general or for a certain element of formality (like paying taxes or abiding by labour legislation) could be desirable for various reasons and could be pursued through the use of negative pressure (penalties for non compliance) or through positive incentives, the latter of which could either raise the benefits to formalization (e.g. by improving credit access of smaller formal enterprises) or lower the costs (e.g. softening the applicable labour legislation). Some combination of these types of policy would presumably be the norm.

6. How to Study the Formalization Issue

Given the prominent gaps in our understanding about formality/information and whether and what to do about it, a key property is to advance that understanding. What methodological approaches can help us to do so?

One way to get at the meaning of differences between the formal and informal sectors (or along the formality and size spectra) is to measure average differences in productivity and earnings. Another is to observe movement of people and firms between the sectors, check for queuing up of informal workers to move into the formal sector. In this connection, it is important to distinguish between paid workers in the informal sector and the self-employed. As clarified by Perry et al. (2007) and others, many of the former would prefer to have formal sector jobs, whereas many of the latter are happy to operate in the informal sector, in fact they have deliberately chosen to do so (This distinction produces the contrasting concepts of

---

31 Even when policy is thought through and designed to benefit society in a neutral fashion, the administrative apparatus can still steer its application in a way that is political and/or non-neutral, in the sense of favouring some groups over others, and this can happen even if it is not the deliberate intention of those involved in that administrative process.

32 Many studies of informality have made valuable contributions though none, to my knowledge, has pulled together evidence from the full range of relevant literatures. This is hardly surprising, since so many aspects of economic functioning do bear, either directly or indirectly, on the issue. One outstanding contribution, both for its breadth and its depth is Perry et al. (2007). This study should put to rest many of the pervading myths about informality and how people choose it or wind up in it.
“exclusion” and “exit”). Concerns about labour market segmentation are mainly relevant to the first group; concerns about productivity are mainly relevant to the second one.

6.1 What exactly are we trying to understand?

The main reason is the low labour productivity and earnings, lack of job security and stability, and poor working conditions that characterize many of its enterprises, this implies either that informality slows growth, good job creation, and poverty reduction, the main goals of economic policy, or at least that it is the setting within which the reasons for low productivity and incomes need to be understood. It follows that its presence and size might matter little if the earnings and productivity gap with large firms were of modest proportions. Hence, any serious analysis of the size of the informal sector should in principle also focus on the magnitude of differences in productivity and earnings vis-à-vis the formal sector. It would be inappropriate to conclude that a country with a larger informal sector is suffering from ‘excessive informality’ relative to another country if the first one also has a relatively high-productivity informal sector. Among other things, there would in that case be less pressure or desire on the part of workers to move to the formal sector and this might by itself explain the larger than usual informal sector. Along these same lines, the implications of informality may be deduced in part from observable features like the extent of mobility of labour and other factors between the two sectors. If such mobility is high and relatively balanced between the two directions, this suggests little segmentation and may also imply less efficiency loss than would otherwise be the case. Unfortunately, as evidenced below, the guideline that the size and the relative productivity of the informal sector should be analyzed simultaneously seems seldom to have been followed in the literature, partly no doubt because of the difficulty in organizing the statistical evidence, but probably also due to failure to recognize the importance of the issue.

If one accepts that formalization is not of interest primarily in and of itself, but rather because it may be causally related to outcome variables of inherent interest – growth, decent employment, income inequality and poverty – then the level of formality is mainly an intermediate variable, through which hypothesized determinants of these variables of ultimate interest may act – including certain features of the capital market, labour regulations, or adequacy of land titling. It may follow that the more promising research approach is to test for the impacts of such posited determinants of the key outcome variables directly on those variables rather than on and through the level of informality. There are arguments for both approaches; their relative merits depend on the circumstances, including the quality of data on the different variables. In any case three things are clear: first, that unless both quality and quantity of employment in the informal sector are taken into account together, the value of studying the level of informality is seriously diminished, and second, that ceteris paribus it is better to have evidence from both approaches than just from one. Finally, the fact that

---

33 Although their overall approach is quite different from this paper, Loayza et al. (2009, 6) make a parallel point when they note that one would not expect the same impact on growth from two countries with the same sized informal sector if in one case this was achieved by “drastic enforcement” and in the other by “light regulations and appropriate public services.”

34 Consider the hypothesis that overly draconian labour regulations lead to slower growth and to greater income inequality, with some of the mechanisms involving greater dualism and a larger informal sector. Ideally one would like to both correctly identify the impact of such legislation on the outcome variables – growth and inequality, and also understand the causal mechanisms linking them, including the role of informality. But the ideal is usually not possible. Where analysis of how labour legislation affects outcome variables through informality is very difficult, as it certainly is, one may learn more about the overall effects of the labour legislation by studying them directly (as many people do), i.e. without trying to understand the role of informality as an intermediate variable. If analysis suggests that such legislation raises inequality, this strengthens any argument to the effect that informality may be a mechanism whereby that negative impact is felt.
formalization is mainly desirable when it contributes, on balance, to desirable final outcomes implies the need for a distinction between good formalization and bad formalization, or more precisely for a measure of the efficiency of formalization. That process should go just as far as it is socially useful, but no farther. The task of related analysis is to sort out how to achieve that goal and what related policies can contribute to it.

Another essential aspect of the analysis of informality, as noted above, is a general equilibrium approach, needed both because of the large size of the formal and informal sectors and because of the many aspects of formality/informality that should be analysed simultaneously. The pattern of (at least partially) offsetting cost advantages between formal and informal firms makes it theoretically possible that the optimal size of the informal sector could be either larger or smaller than it is in a given country. With capital market imperfections favouring the formal sector and labour market imperfections favouring the informal sector it is clear that there is an inefficient allocation of each factor between the two sectors – *ceteris paribus* more capital should be attached to the firms now in the informal sector and more labour to those in the formal one. But it is not clear which sector (if either) would grow as a reduction of those imperfections brought a better allocation of resources between the two categories. Put another way, it is not clear whether the resource transfer that raised aggregate economic efficiency would mainly involve a shift of capital from the formal to the informal sector or a shift of labour in the opposite direction. In the former instance, many of the now informal firms might be sufficiently larger as to no longer be informal, e.g. they would have grown into small formal firms, but this cannot be taken for granted. In any case, the objective of policy should be the reduction of inefficiency; this might or might not be accompanied by a reduction in the size of the informal sector, however defined.

An adequately general equilibrium approach also calls for an interpretation of the overall determinants of the size of the informal sector and, if different, the factors that may make that size excessive. An early model that could explain an excessive growth of the informal sector was that of Harris and Todaro (1970), who posited that a large gap in earnings between the urban formal (protected) sector and the rural sector made it profitable for migrants to seek urban jobs even if the likelihood of having such a job over a given period was not high. Later versions have tended to assume these migrants would be working in what are now called informal sector activities (Reynolds, 1969).

One concern with informality is possible indirect effects. Authors like Pagés (2009) have argued the informal sector has negative externalities either on the formal sector itself and/or on the effectiveness of certain public policies, in particular those in the area of social protection. The hypothesized mechanisms constitute one more of several reasons why the productivity aspect of this issue is not the only one warranting attention and analysis. The impact upon and interaction with social protection policies is of particular concern since the informal sector, in fact, constitutes the main mechanism of “social protection” for many of the people who work in it. Thus, a good way to look at this challenge is as a need to redesign labour and social policies to make them “informality-proof” as advocated by Pagés (2009). Whatever the extent to which the informal sector complicates certain elements of social assistance policy, those complications probably only strengthen the argument that policy on informality needs to be undertaken with great care, since a protection policy which might function “better” if the informal sector could in advance or simultaneously be “shrunk” would stand a considerable risk of backfiring if the gains from a better social protection were offset by the loss of earnings from informal activities. The negative externalities from

---

35 The existence of the informal sector is commonly recognized as making the widespread use of unemployment insurance very difficult or impossible, since it blurs the delineation between having a job and not having one.
informality to the effectiveness of social policies are thus unlikely to constitute an argument for involuntary formalization, but rather for trying to make voluntary formalization more attractive.

6.2 The Range of Methodologies

Given the complexities involved and the currently very incomplete level of our understanding, reaching judgments on the merits of furthering formalization and the best ways to pursue it calls for as wide a range of methodological approaches as possible, of which we here distinguish five. The best, in principle, would be careful ex-post evaluations of the overall formalization experiences of a number of countries. But no such evaluations have, to my knowledge, been done. Another source is the few careful quantitative ex-post evaluations of policies or changes that promote selected aspects of formalization. Thus, to better predict whether having collateral will improve credit access of currently informal firms, one may try to identify what happened after changes to property rights were achieved.

Closely complementary to the ex-post analysis of changes in areas like property rights or labour legislation is a feeling for how the relevant markets work. Thus a good understanding of the capital market as it affects informal and small firms helps to predict and understand the impact of better property rights. Similarly, an understanding of labour market functioning as it affects those firms in the grey area where formalization decisions occur, helps to predict and understand the effects of labour legislation, and a knowledge of whether there is a marked segmentation between the formal and informal workforces is clearly important. Among the modest number of analyses of individual aspects of formalization, several have focused on property rights to land. Other helpful literatures look at the impacts of the legal system and contracting, the credit market and the labour market/labour legislation.

A fourth category, complementary to and interfacing to some extent with the issue-specific literatures just mentioned is the overall body of evidence on the dynamics of small firm creation, growth, and productivity enhancement. A good understanding of the microeconomic context and behaviour of informal and small formal firms includes evidence on their savings and investment patterns, the way the capital markets affect their operations, the way the public sector performs in applying rules and regulations to them, the structure of private associations, etc., all analysed from an overall and hopefully integrated firm-level perspective. Such information includes evidence on the firm-level effects of certain policies designed to encourage or impose formalization. For example, quantitative measures of the costs of bureaucratic red tape have provided evidence on one component of the private and social costs of formalization. As for labour market regulations, one can draw on an extensive literature, though unfortunately analysis has rarely focused precisely on the size/formality range relevant to formalization issues. While a substantial existing theoretical literature and a modest empirical one has expounded on the benefits of formalization generally and on specific aspects like the clarification of property rights to urban land or improvements in the financial system, it is unfortunate that much less effort has been directed at understanding small firm dynamics as they touch on these questions.

Fifth, we have aggregate analyses of the relationships (in both directions) between the size and characteristics of the informal sector and aggregate outcome variables for the economy such as the growth rate, income inequality, the wage rate, etc. These provide a

---

36 Firms referred to by Kenyon (2007, 10) as “amphibians”, with one foot in the informal camp and one in the formal camp.

37 For a brief review of the evidence in each of these areas see Berry (2011b).
different window on the issues. With luck they may throw light on the overall desirability of an increase in formality and, with even greater luck, also on what instruments may foster desirable formality. This inherently general approach has a ‘black box’ character in that the main objective is not to identify and understand how higher or lower levels of informality affect or are affected by specific microeconomic and macroeconomic processes but rather to measure the structural and policy determinants and/or the overall impacts of those levels. A major weakness of the first three approaches – that not all indirect effects are likely to be detected – is the strength of this one; but it suffers from its own severe methodological challenges. Since it is normally cross-country in character, it is vulnerable to the various weaknesses characteristic of such studies, which have led many analysts in recent years to place little confidence in them.³⁸

The fact that these five types of analysis can, from different but complementary angles, shed light on issues of formalization means that the total body of information should be greater than the sum of the parts. Apparent inconsistencies between the conclusions drawn from the credit market literature and those drawn from the small firm dynamics literature can provide new hypotheses for analysis, whose probing then gets us closer to an adequate understanding. For purposes of the narrow discussion focussed on the merits of land titling, for example, the key questions about small firm dynamics per se are (a) whether enough of these small firms have the potential to become more productive (either through growth of inputs or of factor productivity) to make the pursuit of that goal worthwhile policy-wise; (b) whether better access to credit can unleash significant potential for that growth in productivity; and (c) whether land titling facilitates access to credit enough to make it an important policy goal. In relation to the broader analysis of the implications of formalization, question (a) is complemented by analysis of which aspects of formalization can unlock productivity/growth potential.

Comparing and contrasting the tentative conclusions emerging from each of the above methodological alternatives to understanding informality is the final logical component of any defensible attempt to reach policy conclusions on formalization. Reliance on one approach alone, an all too common practice, is extremely risky. Worse, in too many discussions of formalization no reliable evidence from any of them is brought adequately to bear. Had it been, many of the wilder and more naïve propositions could never have been advanced. With respect to many of the key features of informal firms and their role in an economy, reliable evidence remains in short supply. Data accuracy is often weak, and the meaning of such figures as are available is often theoretically ambiguous. Probably the best example of this relates to the much-cited evidence on relative productivity or economic efficiency of firms by size or level of formality, a veritable hornet’s nest of error and ambiguity. In addition, careless and faulty analysis of causation has been prevalent in this area, one reason that randomized experimental analysis, albeit relatively recent in this as in other areas, is already helping to undermine some of the shibboleths.

The next section attempts, by way of example, to give a flavour for the potential and the pitfalls of the last cited approach, the aggregate analysis.

³⁸ Cross-country analysis has been heavily relied on in the literature on the impact of trade barriers in developing countries, but many critics have concluded that not much can be learned that way. This is a sobering judgment given the relative econometric sophistication of many of the cross-country studies in question (e.g. Rodriguez and Rodrik, 2001).
6.3 Analysis of Informality Using Aggregate Data

Attempts to identify the economic causes and implications of informality through the use of aggregate variables involves probing the statistical relationships, either over time or across countries or regions, between overall economic performance indicators and the size and other characteristics of the informal sector. Such analysis is important both in itself and because it is complementary to other approaches. Microeconomic analysis cannot capture some of the indirect and general equilibrium effects that can, in principle, be picked up in aggregate analysis.\(^{39}\)

Unfortunately the contribution of such aggregate analyses to the understanding of informality/formalization is limited by several general weaknesses. First, without exceptions of which the author of this paper is aware, the existing analyses have not dealt simultaneously with the ‘quality’ or performance of the informal together with its ‘quantity’ (size), even though the economic and social implications of informality clearly depend on both. Worse, it is plausible that a relatively productive informal sector (\textit{vis-à-vis} the formal sector) would tend to be causally related to a larger one since the pressure for resources to shift from informal to formal would in that case be less; this rules out the assumption of independence between the sector's size and its 'quality', an assumption that would provide some reassurance that failure to take account of the latter in statistical analysis might not be too damaging. Second, the size of the informal sector is usually treated as an indicator of a problem and the policy variables included in the analysis are ones that would be expected to discourage formalization, like a weak legal system, rather than things that could actively encourage “better” informality (and small firms) through higher productivity (micro credit, say). Since some of these latter factors contributing to a large informal share of employment could also affect overall economic performance positively, their exclusion implies a potentially one-sided analysis. The ideal would be to include factors that might contribute to desirable as well as to undesirable informality, together with data on both the size of the sector and its relative productivity and welfare impacts. Implicit in the above points and also generally missing in this literature is a recognition that in each developing country there is an optimal level of informality at a given point of time and under the circumstances of that time, i.e. the relationship between level of informality and economic performance is a quadratic one. A statistical specification that allows for that fact should be utilized unless there are solid grounds for believing that all countries are on the same side of the optimal level. These failings mean that this literature can at best be seen as a source of hypotheses for deeper probing and as an auxiliary source of information that may provide modest additional strength to conclusions based mainly on other types of analysis. If and when it is demonstrated that size and quality of the informal sector are either unrelated or negatively related to each other, the studies that do not include relative performance of the informal sector will take on greater value.

\(^{39}\) Thus if a policy, such as microcredit, undertaken in the hope of improving the performance of informal firms/microenterprise and encouraging their formalization, is revealed by microeconomic analysis to be effective in raising their productivity but the price elasticity of demand for the output of that sector as a whole is low, good microcredit programs will not necessarily lead to either a more prosperous microenterprise/informal sector or to growth into formal status; productivity will indeed increase but will show up in lower relative prices and in benefits for the buyers of the products. Since it would reflect all such general equilibrium effects, a perfectly successful aggregate level analysis could provide a more accurate picture of the overall effects of microcredit on formalization – it would detect the resulting output increase and the fact that informal sector incomes were not raised.
Another general limitation on the implications of such studies for policy questions surrounding formalization is that even if it can be demonstrated that certain factors have impeded growth by encouraging informality, this does not mean that taking some arbitrary action against informality will undo the damage; it could in principle just as easily make the damage greater. For example, an association between a large informal sector and slow growth and a high level of dualism is interpretable in many different ways. One of the most likely is that capital and product market (and perhaps other) imperfections have both fostered informality and slowed growth but that, in the context of those imperfections, informality is a way to prevent growth and inequality from being even worse than they are. To attack informality would be a classic and potentially very damaging case of attacking the symptom rather than the root cause of the problem, not to mention a confusion between correlation and causation.

The above caveats notwithstanding, this literature has produced several interesting results. The existing studies include on the one hand cross-country econometric tests of the relationships between the level of informality and a variety of proposed determinants and economic outcome variables, and on the other hand modeling efforts to study how certain changes in policy or structural variables would affect outcome variables under plausible assumptions about how the formal and informal sectors function as part of the economy. These two approaches feed off each other – part of standard scientific methodology – with the assumptions built into models drawn from the results of the empirical analysis and with refinements to the latter being based on the implications of the models. The analysis can draw on cross-section or time-series variability, or the two merged. Thus far the cross-section work has dominated; the widely recognized weaknesses of this approach will hopefully lead to a greater balance between the two in future. In fact, aggregate analysis will probably only become persuasive when cross-country (or cross regional) and temporal variance corroborate each other. Cross-country analysis has the advantage that the relative size of the informal sector does vary considerably across the countries, giving promise of a successful search for correlates, but faces the difficulty that so many other things also vary across countries. Over time analysis allows one to greatly diminish the variability of many factors that could lead to errors in cross-country work. But it may take some time for the relative size of the informal sector to change enough to facilitate the temporal analysis of its determinants and effects, and over that time many other factors will have changed as well. It will hopefully be possible soon, if it is not already, to undertake some cross-country time series analysis such that both forms of variability can be drawn on simultaneously. Over time variability is likely to be more important in this sort of analysis since, from a policy perspective, one is trying to effect change in specific countries and that is what over time data reflect directly, whereas cross-section data reflect it only indirectly if that is the case.

Many causal mechanisms linking informality to aggregate income levels and/or their growth have been posited. Thus, Loayza (1996) argues that the presence of the informal sector, because it does not pay taxes but does use public services, lowers the economic efficiency of the system. Asea (1996) criticizes that view and puts forward a very different one. Such disagreement reflects the fact that there is nothing close to an empirically-based consensus on how the presence of informality affects overall income and employment, nor is the view that the informal sector constitutes a drag on economic performance generally accepted (Schneider and Enste, 2000, 89). Asking the question in this very general way is in any case not very helpful, among other reasons because the informal sector is so heterogeneous and because there are various different aspects of informality but, more fundamentally, because of the difficulty in distinguishing the effects of economic performance on informality from the effects of informality on economic performance and
either of these from the undoubtedly important causal processes in which both development and informality are products of underlying ‘third’ factors. The matter of how informality is related to development takes on real interest when it includes a specification of how the level of informality may be changed through policy and what the effects will be. Correlations are less interesting if they do not suggest possibly useful policies. Clearly, if informality is seen to result from bureaucratic red tape that raises costs to formal firms and drives some of them into informality while preventing informal firms from complying with the conditions of formality, the removal of that red tape should lead to a decline of socially inefficient informality and an increase in overall economic performance. Authors focusing on more complicated interpretations of informality naturally tend to have more difficulty in pinpointing positive policies.

Thus far it appears that our understanding has not been advanced a great deal by this approach and perhaps not much can be expected, given the many methodological problems confronting such studies\(^\text{40}\) and the major challenge of aggregating the contributions of each study into a coherent picture.

How much might low productivity informality be reduced through formalization?

The statistical literature designed to explain the size of the informal sector (with the implication that large size is bad) and/or to identify its impacts on overall economic performance provides a very modest contribution.\(^\text{41}\) Among the studies that find confirmation of the proposition that regulations hurt, a few have hazarded estimates of the ‘cost’ of informality. Given the data and methodological constraints on such exercises, none could at this time be taken as at all definitive, even if they did not contradict each other. In a conceptually interesting modelling exercise, Galal (2005) concludes that under present conditions (i.e. the current regulatory system) formalization would lower GDP by 1% in Egypt whereas under a reformed system the impact would be a 1.3% gain.\(^\text{42}\) Clearly we could expect more useful conclusions on this issue had there been a number of careful ex-post analyses of real-life attempts to curtail informality. I know of no studies of this kind at an adequate quality level for the results to be helpful.

As discussed earlier, the values of relevant parameters do not suggest much potential output gain from static resource reallocation between formal and informal sectors. If this is the case, studies which accurately identify significant output or growth gains from formalization or from the removal of barriers thereto are presumably picking up dynamic mechanisms involving investment and productivity growth. Complementary to such aggregated analyses, therefore are those studies that try to pinpoint at the microeconomic level the causal processes that link formality to these dynamic variables.

---

\(^{40}\) An illustrative case is Loayza and Rigolini (2006) who report that the share of the labour force that is self-employed (the variable used as a proxy for the size of the informal sector) is correlated with most of their measures of the density of regulation. But since the authors do not appear to have been able to exclude the agricultural sector, where the self-employment ratio is typically much higher than in non-agriculture and where it has a quite different meaning, this result is suspect.

\(^{41}\) A small theoretical literature (e.g. Dabla-Norris et al., 2008) has the same objective but its value will only become clear when the ideas it contains are given some econometric confirmation.

\(^{42}\) At a less serious level, there are studies like those of Palmade (2005), who concludes, on the basis of studies undertaken by the McKinsey Global Institute (including Capp et al., 2005, and Farrell, 2004) that informality has reduced annual growth by 1-2 percentage points in a group of countries studied.
7. Conclusions

1. Much informality, however defined, is a natural and broadly efficient result of market forces, so that part should not be treated as if it were a problem to be solved; the efficient or desirable level of informality in most developing countries appears to be fairly high.

2. It is important to disaggregate the elements of informality and often to treat them separately; there is no reason that the optimal level of formality should be similar for the various elements involved.

3. Given the absence of any general presumption to believe that the level of informality is too high or too low, *ceteris paribus*, any attempts to treat informality as the direct target of policy need to be based on an in-depth understanding of the relevant aspects of the economy in question.

4. To the extent that informality is associated with such undesirable features as low earnings and income instability, the appropriate policy stance is to attack those underlying problems rather than informality, which is normally no more than a symptom of those problems.

5. The wide range of views as to the desirability of formalization *qua* policy is the result of the complexity of the issue together with the low quality and/or incompleteness of many of the analyses. A common feature of one set of analyses is a misunderstanding of such basic economic concepts as economic efficiency and total factor productivity. Another is excessive reliance on specific statistical methodologies that suffer many often-inevitable shortcomings. An endemic weakness has been the failure to analyze the quality of the informal sector together with its size.

6. Although all approaches to the analysis of informality and formalization have actual or potential value, and it is important to integrate them, the best source of understanding of both the implications of formalization and the policies that may contribute to its effectiveness is the in-depth study of the process of formalization in actual cases. Unfortunately, little organized evidence of this sort is available, one of the main reasons that overall understanding is so deficient.
References


