1 INTRODUCTION
This note introduces the reader to tracer studies. For the Let’s Work initiative, tracer studies are proposed to track and record or evaluate the effectiveness of interventions in terms of job creation and creating high quality jobs. The distinction between an assessment that tracks and records an intervention, as opposed to evaluating it, is largely a function of decisions on whether control groups are included, sampling is done to allow for representativeness and aims at statistical significance estimates, and on the timing and frequency of data collection. Assessments that trace and record are unlikely to have control groups, sampling is not an issue since the population assessed is determined by participants in an intervention and data collection is nearly always ex post. In comparison, evaluations would likely use control groups, sampling would be designed to obtain accurate estimates of change and to attribute outcomes and data collection would most probably take place ex ante (baseline), during the intervention and ex post.

The two primary goals of tracer studies under the Let’s Work initiative are:

(a) To document changes in employment, wages or other dimensions related to the quality of the job that beneficiaries experience; and
(b) To understand if and how the intervention contributed to these observed changes.

The two goals aim to provide the knowledge needed to improve the design of future interventions to create high quality jobs. To this end, this note poses questions about how, where and when tracers may be appropriately used for either assessments or evaluations of interventions designed to create high quality jobs.

2 DEFINITIONS
Tracer studies are intended to track and record the effect of interventions. Designed for the purpose of tracking and recording, tracer studies may yield information whether the variable the intervention meant to affect has indeed changed (and in the expected direction) after the intervention. For example, a tracer study designed this way can answer whether more credit is provided to firms after their managers have completed a business training course. The same tracer study may not be able to answer, however, whether improvements in access to credit lead to increased hiring in these firms or if firms would have received more credit even if business training had not been given at all. Designed this way, tracer studies provide a retrospective (ex post) look at how the situation of subjects changed once treated to a specific intervention or a defined set of interventions.

Alternatively, if a tracer study is intended to evaluate the effectiveness of the intervention on desired outcomes, it is designed to go beyond tracking and recording changes in the variable that the intervention is meant to affect. To do this, a tracer study begins collecting data before the intervention to get a baseline. This before-intervention data collection will likely include subjects that will not be treated in the intervention (a control group). In addition, depending on the nature of the intervention, data collection through tracer studies can happen once, twice or many times before, during and after the intervention. At minimum, data collection must happen before and after the intervention to approximate an evaluation.

If tracer studies are intended for evaluation, they should be designed cognizant of the change that is expected to take place in the course of the intervention. Identifying what variables the intervention sets out to influence and in what way observable changes can be attributed to the
proposed intervention is key to the effectiveness of a tracer study as an evaluation tool. In sum, to design a good evaluation with a tracer study, it is important to have a well-developed theory of change.

In fact, the development of a logic model or theory of change is important to illustrate explicitly the problems to be addressed by the intervention. The theory of change approach should include a series of assumptions, and include strategies on how to solve the identified problems. Links from the input to the outcome level are also included, followed by links to the short-, medium- and long-term results.

The value of using a theory of change in an impact evaluation is to identify how and why impact can be expected to occur. This theory of change begins by first outlining the goal of the intervention, identifies the variables the intervention aims to affect and in what direction, it specifies the subjects (firms, the informal sector, older workers, women, the young) the intervention aims to target, and it defines the approach and types of activities services or combination of activities that are planned to reach the goal. This logical sequence will most often be found in project documents, and may be in the form of a logical framework, a visual illustration or a narrative. It is important to identify what the project set out to do and how it planned to reach these goals.

Tracer studies designed to be evaluations must undertake two tasks: namely, (a) to detect changes in the variable of interest; and (b) to identify what accounts for these changes. Tracer studies designed to be assessments, not evaluations, are intended to accomplish the first task.

3 METHODS
Under this initiative, tracer studies should be designed to allow for comparisons across studies. For this reason, standardization of tracer studies is essential. To be able to compare across different studies, the methodology needs to be standardized to make valid comparisons possible. More quantitatively-inclined tracer studies allow for straightforward comparisons and yield clearer lessons on what works and what does not.

Finally, as mentioned previously in this note, whether the tracer study provides an assessment or an evaluation of the intervention largely depends on whether cohort or control groups are included, statistical sampling is used and on the timing and frequency of data collection.

3.1 Unit of intervention and analysis
While tracer studies are most commonly associated with measuring the effects of a particular intervention on individuals, tracer studies are not be confined to just evaluating changes in individuals. For the evaluation of jobs intervention, three types of tracers studies are considered; namely (a) tracer studies that assess or evaluate the effectiveness of interventions on individual beneficiaries; (b) tracer studies that assess or evaluate the effectiveness of interventions on firms targeted by a given intervention; and (c) tracer studies that assess or evaluate the effectiveness of users of a new service and/or infrastructure.

Standardization in the design of the tracer study is especially needed if samples are small and the unit of intervention is not the same across evaluations. When the tracer study is standardized,
**Evaluating Business License Simplification and Firm Performance**  
**Lima, Peru**

### 1 PROJECT SUMMARY:

Many microbusinesses in Lima, Peru, operate without municipal licenses because the licensing process is lengthy and brings them to the attention of tax authorities. An effort was made to gauge whether streamlining procedures would encourage more companies to get licenses and whether that would boost economic development and jobs.

### 2 OBJECTIVES:

- Analyze how having an operating license affects the performance of firms; and
- Identify institutional factors, in addition to high costs of registration that may negatively affect the performance and growth of firms.

### 3 METHOD:

Based on collecting data on a sample of firms twice a year for three years (six rounds), using the same survey each time, to construct a panel data set. The analysis compares the performance of firms that obtained their operating license during period of time with firms that continued to operate without a license.

A firm survey was used to collect information in each of the six rounds. The final data set contained information about the characteristics of firms and owners, as well as on the dynamics of different indicators of firm performance over the years.

The firm survey questionnaire was designed around theoretical insights of production theory. Providing incentives for applicants could encourage more to seek licenses and potentially to increase their output. Making it easier and less time-consuming to get a license was assumed to allow small firms to get credit more cheaply, gain increased legal protection, and have a better chance at expanding by teaming up with bigger firms.

### 4 SAMPLE:

The sample of firms in the initial round (the baseline) is composed by firms operating in Lima without a license. To assure comparability between firms that got licenses and firms that did not and to better identify the causal effect of getting the license on the outcome variables of interest, an encouragement approach was implemented such that to reduce selection bias by inducing those firms that otherwise would not have gotten the license to go through the process of registration.

### 5 RESULTS:

**No evidence linking licenses with stronger firm performance.** Unlicensed microbusinesses saw no benefit in obtaining a license because of their low operating scales and the costs of doing so. The small scale of their operations and their perception of the economic situation led them to make no connection between investment and increased performance as theory would expect. Some owners cited benefits such as avoiding tax payments and not paying the costs of health and safety regulations as benefits from being informal. A typical response from a microbusiness owner, operating at subsistence level, was to question what advantage would be gained by holding a license, since income was so low that the owner would never qualify for a bank loan anyway.

Further study found that while some small businesses that held licenses were good performers in terms of their incomes and profits per worker and their investments in machinery, there was no evidence to link the act of having a license with the stronger performance.

it does not introduce additional variation in what is already heterogeneous units of intervention and analysis. Most importantly, if tracer studies are designed to be compared, the interventions should be as identical as possible. Most assessments and evaluations are cognizant of the need for comparing across identical interventions, but less so about comparing across different types of subjects of intervention. Assuming identical interventions, the unit of intervention can be relatively more heterogeneous if samples are sufficiently large to control for that heterogeneity. This is addressed in the sampling section of this note.

Finally, especially for tracer studies designed to be evaluations, the theory of change will help define and pin down the unit of intervention and analysis. For example, interventions intended to create jobs in the smallholder agricultural sector may want to primarily target women since in many economies it is women that do most of this type of agricultural work. When comparing across evaluations of similar interventions, however, it will be important to understand if the results are affected by such targeting and how. Differences in some intervention target groups may not matter, such as an intervention on 25 year-olds versus the exact same intervention on 29 year-olds, but some may.

In sum, the unit of analysis, across tracer studies, must be as standardized as possible to make comparisons across studies relatively more rigorous. However, there are instances where standardization may not necessarily invalidate cross-study comparisons. In evaluations, where the intervention and the tracer study are being designed beforehand, the theory of change will help identify and target the most appropriate unit of intervention and analysis.

### 3.2 Comparison across groups and time

An evaluation of an intervention essentially answers whether the change can be attributed to the intervention or not. Whether this can be answered with confidence largely depends on how comparators are designed and used. Comparators refer to cohorts to benchmark changes that may have occurred to entities subject to the intervention.

Attribution is not possible when the tracer study focuses only on the treated subjects *ex post* (after treatment). This case of a tracer study applied to only treated subjects is represented in the first row of the table. While possibly most expedient and least costly way to do a tracer study, it is also likely to be the most modest in terms of answering questions about effectiveness of the intervention. In this way, the tracer study becomes a monitoring, not so much an evaluation tool.

<table>
<thead>
<tr>
<th>Tracer applied to:</th>
<th>Ex ante</th>
<th>Ex post</th>
<th>Comparator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treated only</td>
<td>no</td>
<td>yes</td>
<td>No comparator</td>
</tr>
<tr>
<td>Treated only</td>
<td>yes</td>
<td>yes</td>
<td>Comparisons of subjects before and after treatment</td>
</tr>
<tr>
<td>Treated and untreated</td>
<td>no</td>
<td>yes</td>
<td>Comparisons between treated &amp; untreated</td>
</tr>
<tr>
<td>Treated and untreated</td>
<td>yes</td>
<td>yes</td>
<td>Comparisons between treated &amp; untreated &amp; before &amp; after</td>
</tr>
</tbody>
</table>

Applying the tracer study *ex ante* and *ex post* to subjects that were treated makes for a tracer study that can better track changes in that population, but attribution is still difficult. The subject of an intervention can be their own cohort if there is *ex ante*, baseline data about the variable before the treatment to be able to measure changes after (*ex post*) the treatment. To make a determination that the change observed between the before and after subjects is statistically significant, a good sample of the treated population has to be included. Sampling issues are discussed in more detail in the sampling section. Second, based on good sampling, change in the variable of interest is sought to be statistically significantly different *ex post* as
compared to *ex ante*. What has been demonstrated, however, is that the intervention affected the variable of interest. To get attribution, that variable has to be associated with an outcome (getting a job and/or obtaining a better quality job). Since this is the case, it is unlikely not to be cost effective to design the tracer study to collect outcome data—changes in the number of jobs data, for example—since attribution is unlikely to be possible. If such data were to be collected, it may be included in an evaluation report, but conclusions about causality are, at best, suggestive.

Attribution becomes clearer as control groups are added as part of the tracer study. The table suggests the possible combination with respect to cohort groups that can be included in a tracer study. The tracer study can track, after the intervention, a sample of subjects on which the intervention was applied and a sample of subjects that were not subject to the intervention.

The fourth scenario, where tracer studies are applied to treated and untreated subjects before and after, is close to a randomized control trial (RCT) — though not quite. Depending on how the sampling between treated and untreated groups is designed, external factors are controlled for, the intervention is applied and how the variables of interest are quantified, it can indeed be a RCT.

In a nutshell, it is nearly impossible to attribute the changes (or lack thereof) to the intervention since untreated subjects may have undergone the same changes (or have had fewer or lesser changes) in the variables of interest. For example, if a firm creates X number of jobs during the period of time that it was subject to an intervention, a tracer study only applied *ex post* to that firm cannot answer whether that same firm would have created X+ or X- jobs during that same period of time or whether other firms produced just as many though they were not subject to the treatment. In sum, the counterfactual does not exist.

### 3.3 Sampling

A tracer study evaluation based on findings with statistically relevant estimates requires that a certain number of former participants/beneficiaries be traced and interviewed. While it may not necessary to trace and interview all former beneficiaries, statistically reliable conclusions cannot be drawn on the entire target population from interviewing a few, former beneficiaries. With sampling, a group may be selected from the target population and defined as the sample. If the sampling sizes follow certain rules, the characteristics from this smaller group are said to reflect the larger group.

Sample sizes are determined by the size of the population of interest, but most importantly by the variance of the variable(s) of interest. If the change variable that the tracer study is focused on has a relatively small and bounded variance—such as a variable that measures whether or not an individual gets a job (no=0 and yes=1)—then sample sizes can be substantially smaller than if the variable of interest is a change in the salary before and after the intervention. Theoretically, the change in salary can be nothing to something rather large, like infinity if the subject had no salary before the intervention. If changes in salary or wages is the variable of interest, then the sample sizes needed for statistically significant estimates are likely to be larger than bounded, proportions variables.

Sample sizes are also a function of the level of desired precision. Intuitively, the more precise the requirements of estimates, the bigger samples must be larger. A high level of precision is not always necessary if the change that the intervention is designed to initiate is large. On the other hand, if the change is relatively small, then a high level of precision is desirable and sample sizes must be larger to achieve these levels.
Stratification of the sample can be a way to cut down on the size of samples. Stratified random sampling is a type of probability sampling technique that involves first dividing a population into subpopulations and then applying random sampling methods to each subpopulation to form a test group. With the stratified random sample, there is an equal chance (probability) of selecting each unit from within a particular stratum (group) of the population when creating the sample.

When the evaluation is interested in understanding the effects of an intervention on a particular strata within the population (e.g., males vs. females; small firms vs. large firms, female-headed households vs. households with two parents, etc.), then stratified sampling may be more appropriate than simple random sampling.

Finally, stratified sampling offers several of the following advantages over random sampling:

- Greater precision than random sample of the same size;
- Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money;
- It mitigates against an ‘unrepresentative’ sample (e.g., an all-male sample from a mixed-gender population); and
- Ensures obtaining sufficient sample to support separate analysis (sufficiently high levels of precision) of any subgroup.

A stratified sample may require more administrative effort than a simple random sample. Researchers must identify members of a population being studied and classify each of them into one, and only one, subpopulation. Finding an exhaustive and definitive list of an entire population is a challenge. In some cases, it is downright impossible. In addition, the evaluator must have some theory of change that varies across strata. In other words, unless it is clear that females and males are expected to experience substantially different effects from the intervention, it is not clear that a stratified random sampling technique should be used. Either from past empirical experience, theory or (even better) both, there must be a reason why a sample should be stratified in the first place; namely, the variable of interest is expected to vary across strata.

Finally, samples can be designed using propensity-score matching (PSM). PSM is a quasi-experimental option used to estimate the difference in outcomes between beneficiaries and non-beneficiaries that is attributable to a particular program. Note that this is used when doing an evaluation that includes participants and non-participants in an intervention and the aim is to compare across these two groups to evaluate the purported impact of that intervention.

PSM reduces the selection bias that may be present in non-experimental data. Selection bias exists when units (e.g. individuals, households, or firms) cannot or have not been randomly assigned to a particular program, and those units which choose or are eligible to participate are systematically different from those that are not.

A propensity score is an estimated probability that a unit might be exposed to the program; it is constructed using the unit’s observed characteristics. The propensity scores of all units in the sample, both beneficiaries and non-beneficiaries, are used to create a comparison group with which the program’s impact can be measured. By comparing units that do not participate in a program, but otherwise share the same characteristics as those units which have participated, PSM reduces or eliminates biases in observational studies and estimates the causal effect of a program on an outcome or outcomes.
3.4 Implementation

Standardized tracer studies require standardize data collection. Collection of the data for tracer studies are most often done through questionnaires. Issues on the implementation of the survey, excluding discussions of sampling which was mentioned above, can influence the results of the tracer study. Implementation issues include questionnaire design, how the questionnaire is applied, how many times it is applied and when it is applied.

The substantial and key issues of finding and tracking the subjects of an intervention are not discussed in this note. While tracking past subjects of an intervention is not discussed, the considerable challenges this task represents should not be minimized. Whether easy or difficult to track subjects depends a lot on data collection efforts before the intervention begins and immediately after it is completed; the period of time that has passed between the end of the intervention and the time the tracer study begins—to be addressed in another section—and how transitory the subject population may be.
**Questionnaire design**

A standardized tracer study begins with a uniform questionnaire. The value of the questionnaire will largely depend on its design. The questionnaire must be concise and coherent, easy to apply, focused on obtaining information on the variables and outcomes of interest and nothing more and minimize *ad hoc* responses. Questions should be designed to use yes/no answers as a gateway to other, more complicated questions or to skip questions altogether. Where possible, the questionnaire should also minimize questions that require responses based on memory or recall. Information on how to put together an appropriate survey instrument can be found in any good reference on how to design questionnaires: there is nothing particularly different in a tracer study questionnaire where the lessons of general survey questionnaires do not apply.

Complicated questionnaires require substantially more training of enumerators. This and other aspects of questionnaire design can be largely how a survey instrument should be implemented.

**Implementing the survey instrument**

Decisions such as whether the survey can be done on the phone, face-to-face, in one visit, with one person answering all questions or with a few answering specific questions, etc. are aspects of survey implementation that are often left as second or third order considerations but can affect results in drastic ways. Phone conversations are the least expensive and fastest, but there are problems with selection bias, confirmation that the actual subject is responding to the survey and where there is a survey with a lot of answer options, it may be best to implement the survey face-to-face. In that way, a card can be shown to the respondent so that they can review all of the options before choosing and answering the question. Also, a survey that may have confidential or personal information may be best implemented face-to-face to gain the confidence of the respondent. The same is true if the survey instrument is relatively long. Having something that seems more like a face-to-face conversation may further induce the respondent to continue responding and responding accurately to the questions posed. Finally, the gender of the interviewer may matter.

In sum, more complicated questionnaires are most effectively applied in face-to-face interviews.

**Timing of data collection**

The timing when a tracer study is implemented can also affect results. This is especially the case if the evaluation or assessment will be based on substantial amounts of recall data about conditions before and after the intervention. Concerns about compromising the accuracy of the data through recall has to be balanced with the recognition that data collection have to take place after sufficient time for expected outcomes to become apparent. If it is believed that the intended
effects of the intervention require a relatively long gestation period before these are evident (time after $t_1$), then it may be best to allow some time before the tracer study is implemented (time $t_2$). If there are concerns about waiting too long for results to appear and compromising data accuracy, the tracer study may be designed to collect data more than just before the intervention and sometime afterward. However, more frequent data collection will jeopardize survey participation rates, as participants may tire of answering questions on several different occasions, and there are concerns about costs.

Finally, finding former participants of a program or some other sort of intervention is likely to be harder the longer the wait, after the intervention, there is to begin the tracer study. Tracing is the task of physically locating the selected members of the sample. Finding the former beneficiaries may be relatively easy and straightforward or may require much effort and time. This, of course, depends on specific conditions, which can be the amount of time elapsed since the intervention was completed, among many other factors.

4 CONCLUSION

Once a tracer study is determined to be the best form of measuring the impact of an intervention on the number and/or quality of jobs, the first decision is whether the tracer study will serve as an evaluation or an assessment. In an assessment aims to measure changes, as much as possible, that already occurred; it is an \textit{ex post} evaluation where the subjects have already been identified and the major task is to collect information from them. While a tracer study for assessment may seem to be a pre-determined task, there are many parameters that still need to be decided on. For example, there is the opportunity to sample cohorts that have not been subject to the intervention or subscribers to a new service. There are also decisions to be made about whether all of the participants, or a representative sample, will be traced. Finally, there are decisions to be made about the frequency and timing of the data collection activities. The \textit{Let’s Work} initiative aims to standardize questionnaires, survey implementation methods, and some of the decisions about the timing, frequency of data collection activities as well as the use of cohorts and sampling issues, but in the early stages much has yet to be determined.

If the tracer study is to be used as an evaluation of interventions on jobs and jobs quality, the number of decisions multiply rapidly. It would be difficult (though not impossible) to do a robust evaluation of a program if the intervention has already taken place before the tracer study is designed. Ideally, a tracer study is designed and implemented as part of the intervention. Key is obtaining baseline data before the intervention begins. The mention of baseline data brings to mind questions about sampling, especially of sampling the control or cohort group that will not be subject to the intervention or will not be serviced. While there could conceivably be questions about sampling the participant population, whenever possible, it may be best to take baseline data from all participants at first since it is unclear what drop rates will be like at the beginning of a project. The timing and frequency of data collection activities is also important. There will be data collection activity after the intervention. How soon after is going to be a function of the expected gestation of the intervention before outcomes are realized and how much recall data will be required as part of the evaluation.

Finally, there is a need for standardization in methods. Standardization will likely be based on the intervention. Interventions of different types will likely have different methods, but interventions with the same aim will best serve for cross-study comparisons when standardized. Standardization will include the use of cohorts or control groups, the questionnaire and its implementation and sampling. Whether tracer studies assess or evaluate the effectiveness of jobs interventions, standardizing these is more likely to yield lessons on what works and what does not. In the end, this is the goal of carrying out a tracer study.