DRAFT – PLEASE DO NOT CITE

An experimental framework for continual improvement in survey research

Dennis Feehan*[1] and Matthew J. Salganik†[1,2,3]

[1]Office of Population Research, Princeton Univeristy
[2]Department of Sociology, Princeton University
[3]Microsoft Research, New York

September 27, 2013

*dfeehan@princeton.edu, Wallace Hall, Second Floor, Princeton, NJ 08544
†mjs3@princeton.edu, 145 Wallace Hall, Princeton, NJ 08544

Abstract

Surveys are an essential measurement tool for many of the most important theoretical and policy questions in the social sciences. Unfortunately, in order to measure the things we care about with surveys, we often have to make difficult decisions about exactly how we should collect information from our respondents. Our paper begins by describing such a situation that we encountered in a study of populations most at risk for HIV in Rwanda. We describe how we conducted a survey experiment and exploited known quantities to gather evidence about how to best measure unknown quantities. We then generalize our experience into a framework that would allow researchers in a wide variety of contexts to steadily accumulate evidence about best practices by embedding experiments in their surveys. Each new survey can be an opportunity to add more to the body of knowledge available, continually improving the quality of everyone's estimates.

1 Introduction and overview

Surveys are an essential measurement tool for many of the most important theoretical and policy questions in the social sciences. Unfortunately, in order to measure the things we care about with surveys, we often have to make difficult decisions about exactly how we should collect information from our respondents. For example, if we are interested in estimating how many visits each of our respondents made to a doctor in the past year, we might consider asking a respondent to tell us how many visits she made to the doctor in the past week, the past month, the past three months, and so forth. The tradeoff between these options is not immediately obvious: we might expect more recall error for longer time periods, but shorter time periods might be unduly influenced by seasonal patterns in illness, or other short-term factors that would make extrapolating to yearly totals difficult. Ideally, these decisions are based on extensive pre-testing, a good theoretical understanding of the quantity being measured, and lots of field experience. Unfortunately, this is not always an option and, even in cases where it is, the answer is rarely clear.

In this paper, we outline a new, experimental framework for continual improvement in survey data collection. For some quantities of interest, this framework would allow researchers to steadily build up evidence about best practices; each new survey would be an opportunity to add more to the body of knowledge available, with the aim of continually improving the quality of estimates over time. Our framework proposes randomizing respondents into different data collection procedures (questionnaire wording, interviewer training, etc). Under the conditions described below, we can estimate the total error under each condition. This estimate of the error is then a principled, pre-specified criteria that we can use to distinguish between the performance of the different data collection procedures.

Over time, this evidence would accumulate, allowing the community of researchers to make better design decisions. Each study, rather than being just a standalone effort, would also add to this larger pool of knowledge without sacrificing the quality of the estimates it is intended to produce. This added benefit is not completely free, however. It requires the ability to run survey experiments (which can be difficult to conduct), it requires that a specific set of assumptions be satisfied, and it requires slightly more complex data analysis.

The rest of this paper begins with a description of how we used a survey experiment to

learn about the design of a survey used to measure the sizes of groups most at-risk for HIV in Rwanda. Typically, we would introduce the new framework first and then illustrate it with an example. In this case, however, we think it is more illuminating to start with a concrete, motivating example. Measuring the sizes of groups at-risk for HIV is notoriously difficult, and we found it necessary to go beyond traditional survey research techniques to try and learn about what data collection procedures worked best. In the third section, we take the experience of the Rwanda study and generalize it into a framework that could be used in a wide variety of contexts. Next, we investigate how to make use of the full sample of respondents, even if different respondents have been randomized to different survey designs. Finally, we conclude with a discussion of next steps.

2 Motivating example: household survey from Rwanda

We begin with an example taken from a recent study that was conducted in Rwanda. The study was designed to produce estimates of several key populations at risk for HIV/AIDS, including injecting drug users (IDU), men who have sex with men (MSM), and clients of sex workers. Understanding the sizes of these populations is important to researchers and policymakers who need to design strategies to contain the spread of HIV/AIDS. In order to estimate the sizes of these populations, the Rwanda survey used the network scale-up method, a technique which requires respondents to tell us about the people they are connected to in their personal networks. We'll begin with a brief review of the network scale-up method, with the aim of highlighting one critical decision that has to be made in designing an appropriate data collection instrument. Then we will describe how we used a survey experiment to provide a principled, empirical basis for making that decision.

2.1 The network scale-up method

The network scale-up method is based on the assumption that we can learn about the general population by asking people about members of their social networks (Bernard et al., 2010). On a traditional survey, respondents are asked to report about themselves, and inferences about the general population are made from those responses. On a network scale-up survey, respondents report about members of their personal networks, and these

reports are used to estimate the sizes of hard-to-count groups. In the literature, these hard-to-count groups are often called target populations.

Figure 1 gives the intuition behind the network scale-up method. The members of the general population are represented by circles, and we wish to estimate the size of the target population which, here, is the circles colored grey. A quick count shows that, in reality, there are 30 members of the whole population, six of whom are grey. To produce an estimate, we begin by taking a sample of the population members. In this simple example, our sample has size 1 and the selected respondent is colored black. We ask the respondent how many of the people in her network are in the target population. In this case, the response is 2. We then ask the respondent a series of questions that estimate the number of people in her social network; here, this is 10. The result is that we estimate that the size of the grey population is $\frac{2}{10} \times 30 = 6$.

Mathematically, the scale-up estimator can be written as

$$\hat{N}_t = \frac{\sum_i y_i}{\sum_i \hat{d}_i} \times N, \tag{1}$$

where N_t is the size of the target population to be estimated, N is the size of the total population, y_i is the number of members of the target population reported to be connected to the ith survey respondent, and $\hat{d}_i$ is our estimate for the network size of the ith survey respondent (Killworth et al., 1998).

Once we have estimated the size of each respondent's network, the application of the network scale-up estimator is straightforward, and requires only a question asking the number of members of the target population the respondent is connected to.

2.2 Estimating the size of respondents' networks

In order to use the scale-up method, we need to estimate the sizes of respondents' personal networks. There are several methods available for doing so, but the one we focus on here is called the known population method (Bernard et al., 2010). We begin by identifying several populations whose total size is known; for example, administrative records may give us a total count of the number of mailmen in the population. We then ask each survey respondent how many mailmen she knows; we expect the number she reports to be

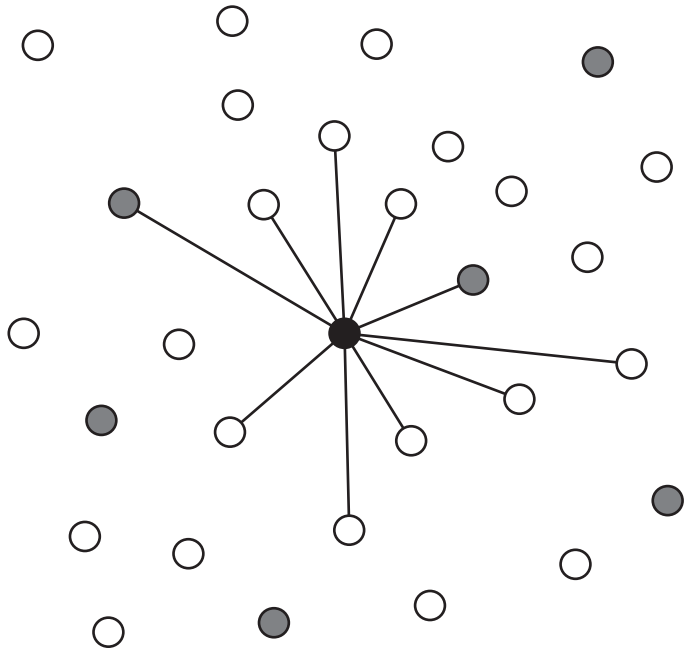

Figure 1: Illustration of the network scale-up estimator. Each circle represents a member of the population, and grey circles are members of the target population whose size we'd like to measure. The black circle is our survey respondent. The respondent reports being connected to two members of the target population. Her total network size is 10, and there are 30 people in the general popualtion, so the scale-up estimate of the target population's size is $\frac{2}{10} \times 30 = 6$.

proportional to the size of her network. If there are 10,000 mailmen in a country with a population of 1 million, and a respondent reports being connected to one of them, then we estimate that the size of her personal network is

$$\frac{1}{10,000} \times 1,000,000 = 100.$$

By asking questions about many populations of known size (typically about 20), we can obtain a more precise estimate of the size of each respondents network. Mathematically,

we compute our estimate of the size of respondent i's personal network, $\hat{d}_i$, with

$$\hat{d}_i = \frac{\sum_k y_{ik}}{\sum_k N_k} \times N, \tag{2}$$

where k indexes the known populations, y_{ik} is the number of people respondent i reports knowing in known population k, N_k is the total size of known population k, and N is the total size of the whole population (Killworth et al., 1998). Table 1 shows the known populations that we used in the Rwanda study.

Group name	Source
Priests	Catholic Church
Nurses or Doctors	Ministry of Health
Twahirwa	ID database
Mukandekezi	ID database
Nyiraneza	ID database
Male Community Health Worker	Ministry of Health
Ndayambaje	ID database
Murekatete	ID database
Nsengimana	ID database
Mukandayisenga	ID database
Widowers	RDHS (05, 07, 10)
Ndagijimana	ID database
Bizimana	ID database
Nyirahabimana	ID database
Teachers	Ministry of Educ.
Nsabimana	ID database
Divorced Men	RDHS (05, 07, 10)
Mukamana	ID database
Incarcerated people	ICRC 2010 report
Women who smoke	RDHS (05)
Muslim	RDHS (05, 07, 10)
Women who gave birth in the last 12 mo.	RDHS (10)

Table 1: The known populations used to estimate network sizes in the Rwanda study. RDHS denotes the Rwanda Demographic and Health Survey from the years indicated in parentheses, and groups from the ID database are names.

There are many advantages and disadvantages to using network approaches to estimate the sizes of hard-to-reach populations; a thorough discussion of them can be found in Bernard

et al. (2010).

2.3 Definition of a network connection: a survey experiment

In order to employ the network scale-up estimator, we must first establish what it will mean for a respondent to be connected to someone[1] (Bernard et al., 2010; Killworth et al., 1998). In terms of Figure 1, we must settle upon the definition that determines whether or not a pair of circles is connected with a line.

This is potentially a critical design decision, but at the time of the Rwanda study, there was no empirical evidence about what tie definition would produce the best estimates reported in the literature. Following the original scale-up study of Bernard et al. (1989), previous work has used a definition of a tie that is very *weak* (in the sense described in Granovetter (1973)). That is, respondents are asked "How many people do you know who are sex workers?" where "know" is typically explained as: you know them and they know you and you have been in contact over some fixed period of time (usually one or two years) (Bernard et al., 2010). This operationalization of a tie means that each respondent is essentially being asked to provide information about hundreds of other people. Such a weak definition means that we collect lots of information per respondent, but because the relationships between the respondent and the people they are reporting about are weak, the information we are receiving could be inaccurate. On the other hand, we might speculate that a definition of a tie that is stronger would lead to more accurate information about fewer people.

Although we do not yet have evidence about the relationship between the strength of the tie defintion and the quality of the estimates produced by the network scale-up method, Figure 2 provides a speculative sketch of what the relationship might be. It illustrates our hypothesis that, as the tie definition moves from strong to weak, the sampling error will decrease, because we will learn about more people from each interview. On the other hand, we hypothesize that the non-sampling error will increase because the quality of the information we receive about weaker alters will diminish. Our prediction is that these two factors will trade off; if that is true, then it would be important to understand what kind

[1] In the network literature, connections are often referred to as 'ties'; we use both terms interchangeably here.

of tie definition is likely to produce the most accurate estimates. Ultimately, we can only learn about the relationship by accumulating empirical evidence.

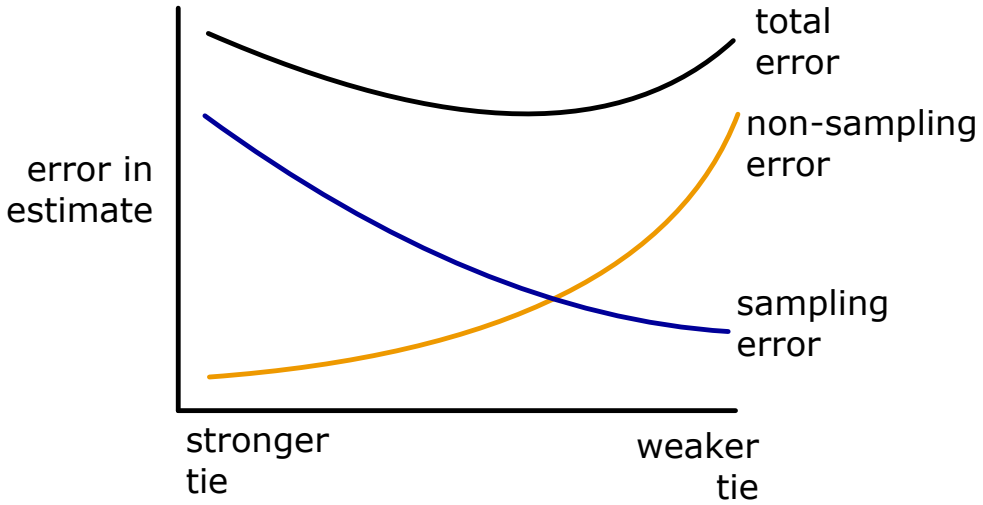

Figure 2: Our hypothesized relationship between the strength of the tie defintion and the quality of the estimates produced by the network scale-up method. We predict that, as the tie definition moves from strong to weak, the sampling error will decrease, because we will learn about more people from each interview. On the other hand, we hypothesize that the non-sampling error will increase because the quality of the information we receive about weaker alters will diminish. Our prediction is that these two factors will trade off; if that is true, then it would be important to understand what kind of tie definition is likely to produce the most accurate estimates. Ultimately, we can only learn about the relationship by accumulating empirical evidence.

In order to investigate this issue, our study in Rwanda included a survey experiment comparing two definitions of a tie. The first definition, referred to as the *basic* definition, is the one used most commonly in the scale-up literature (Bernard et al., 2010). It is described in the left-hand column of Table 2. The second defintion, described in the right-hand column of Table 2, is the *meal* definition. We constructed the meal defintion to be stronger, meaning that we predict that we will learn about fewer people with each interview, but that the information we receive will be more accurate. Each household in our sample was randomized to one of these two definitions of a tie.

Tie Definitions in Survey Experiment	
Basic ($n = 2,236$)	Meal ($n = 2,433$)
• people of all ages who live in Rwanda • people the respondent knows, by sight AND name, and who also know the respondent by sight and name • *people the respondent has had some contact with – either in person, over the phone, or on the computer in the previous 12 months*	• people of all ages who live in Rwanda • people the respondent knows, by sight AND name, and who also know the respondent by sight and name • *people the respondent has shared a meal or drink with in the past 12 months, including family members, friends, coworkers, or neighbors, as well as meals or drinks taken at any location, such as at home, at work, or in a restaurant.*

Table 2: The two definitions of a network tie that were used in this study. All of the conditions need to be satisfied in order for the respondent to consider someone a member of her network. Previous work on the network scale-up method has almost exclusively used the basic definition, which is described in the left-hand column. We expect the meal definition, described in the right-hand column, to be stronger; that is, we predict that it will result in smaller network sizes, but more accurate information from survey respondents' reports. We randomly assigned one of these two definitions to each household in our sample.

2.4 Sampling and data collection

Our study was intended to mimic a Demographic and Health Survey (DHS), so it used the same survey infrastructure as the 2010-2011 Rwanda DHS. In particular, we used the same interviewers, data entry protocols, training techniques and sampling frame as the Rwanda DHS. Our sample of approximately 5,000 respondents was drawn using a stratified, two-stage cluster design, and interviews were then conducted between June and August of 2011. As we explain above, each sampled household was randomly assigned to one of two possible definitions of a network tie. The full details of the sampling plan are described in RBC/IHDPC et al. (2012).

2.5 Estimating the error

It is reasonably clear that randomizing the definition of a tie used in the questionnaire for each household allows us to understand whether or not the estimates the two tie definitions

produce differ from one another: for example, using the meal definition, we estimated that the average respondent was connected to about 0.34 sex workers, while under the basic definition, the average was about 0.72 sex workers (RBC/IHDPC et al., 2012). We now describe how we can actually also learn about the more important question of which tie definition produces better estimates.

An advantage of the estimator we applied is that it is possible to use it to estimate size of the groups whose total sizes are known (see Table 1). Recall, from Equation (2), that the known population method uses several groups whose total size is known to estimate the network size of each respondent. This means we can check to see how accurately we can estimate the size of each of the known groups. To do so, we can take each of these known populations in turn, pretend it is not known, use the remainder of the known populations to estimate the respondents' network sizes, and then the scale-up method to estimate the size of the held out group. We refer to this process as an internal validation check. Mathematically, if we are holding out known population j, the modified version of Equation (2) would be:

$$\widehat{d_{i,-j}} = \frac{\sum_{k \neq j} y_{ik}}{\sum_{k \neq j} N_k} \times N, \tag{3}$$

and the scale-up estimate of the size of known group j, $\widehat{N_j}$, would be

$$\widehat{N_j} = \frac{\sum_i y_{ij}}{\sum_i \widehat{d_{i,-j}}} \times N. \tag{4}$$

An example of this internal validation exercise using the data we collected in Rwanda is shown in Figure 3 (RBC/IHDPC et al., 2012). Since our study included a survey experiment that randomized households to one of two different definitions of a network tie, this internal validation check also provides a criterion for evaluating the different tie definitions. The left-hand panel shows results from the most common tie definition, while the right-hand panel shows results from the meal definition. The relationship between the true subgroup sizes (x axis) and the network scale-up estimates (y axis) are shown in Figure 3. If the scale-up estimate is exactly correct, it would lie on the solid line. Points below the dashed line are underestimated, while those above are overestimated. We see that the estimated values are clustered around the true ones, with a few notable exemptions.

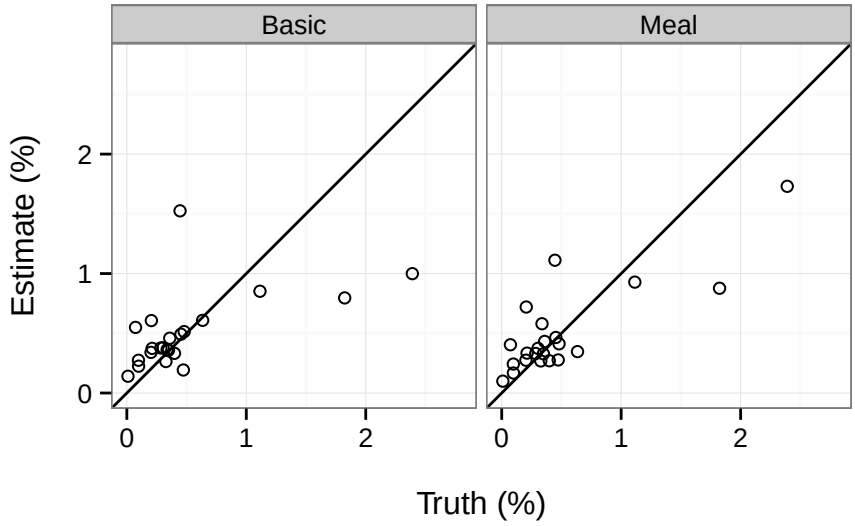

Figure 3: An example, from RBC/IHDPC et al. (2012), of an internal validation check that the known population method for estimating network size allows us to perform. In this case, since the Rwanda study used two different definitions of a network tie, this internal validation check also provides a criterion for evaluating the different tie definitions; the left-hand panel shows results from the most common tie definition, while the right-hand panel shows results from the meal definition. Each point is a subgroup of known size in the general population. Each survey respondent was asked how many of the members of each known subgroup she was connected to, and the responses were used to estimate the size of each respondent's personal network. By removing each of the known subgroups, one at a time, and treating it as unknown, we can use the network scale-up approach to estimate its size (see Equations 3 and 4). The relationship between the true subgroup sizes (x axis) and the network scale-up estimates (y axis) are shown here. If the scale-up estimate is exactly correct, it would lie on the solid line. Points below the dashed line are underestimated, while those above are overestimated. We see that the estimated values are clustered around the true ones, with a few notable exemptions. RBC/IHDPC et al. (2012) shows that three different error metrics suggest that the meal definition produces more accurate estimates than the standard one.

In order to quantify the quality of the predictions illustrated in Figure 3, we computed three summary measures of the error: average relative error, root mean squared error, and mean absolute error. Figure 5 shows that the meal definition outperforms the basic one by all three measures (RBC/IHDPC et al., 2012).

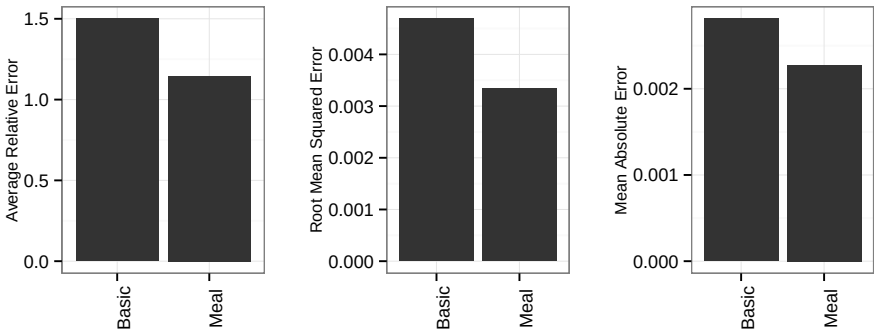

Figure 4: In order to summarize the quality of the predictions illustrated in Figure 3, we computed three summary measures of the error: average relative error, root mean squared error, and mean absolute error. The meal definition outperforms the basic one by all three measures (RBC/IHDPC et al., 2012).

2.6 Summary

In the Rwanda study, we were faced with an important challenge in designing the questionnaire to be used in estimating our quantities of interest: which definition of a network tie would yield the most accurate estimates? There was no empirical guidance available in the literature, so we randomized our responding households to one of two different definitions. Since we could use our estimator to produce estimates of the sizes of known populations, we were able to obtain evidence about how well each estimator recovered the known population sizes. This allowed us to estimate the total error from each tie definition. This estimated error was then a principled, quantified criterion that allowed us to evaluate the two tie definitions and to conclude that the meal definition outperformed the basic one.

3 General framework

3.1 Overview

Although the details of the experiment in Rwanda were specific to the subject matter of that study, the approach we employed could be of use in many other contexts. We now turn to expanding our experience from the Rwanda study into a general framework for continual improvement in survey research. The situation we encountered in Rwanda had two key elements that are the foundation of our method: first, we had an estimator that we expected to perform similarly in terms of accuracy when it was used to estimate several different quantities; and, second, we had both estimands whose true values were unknown, and estimands whose true values were known. We'll explain this in more detail below.

Given data from a sample, and a target quantity of interest, what we call an estimator is a method for producing an estimate of the target quantity. If we call our estimator f, we will write

$$f_{X_1}(s) = \hat{X}_1, \tag{5}$$

to denote the estimate that f produces for the target quantity X_1 from the data collected in sample s.[2] In the Rwanda example, f is the network scale-up method and X_1 could be one of the known populations; for example, it could be the number of priests.

Now suppose that the data we collected from our sample s contains information about two different types of estimands: known and unknown. In the Rwanda example, the purpose of the study was to estimate sizes of groups whose true size is unknown and of considerable interest, like injecting drug users. But we also used the same method to estimate the sizes of groups whose true size we already knew from other sources like administrative records and the census. At first, this may seem like a waste of resources, but it is actually crucial: the fact that we have the data necessary to produce estimates for groups of known size allows us to estimate the total error.

Why is being able to estimate the total error so useful? When we use a survey to try and measure some quantity, we typically care about how accurate our estimate is; that is, we

[2] This is a slight abuse of notation: in order to avoid complication, we're using s as a shorthand for all of the data collected from sample s.

want to understand how close we think our estimate will be to the true value. This is the total error of our estimate. The total error of a survey estimator can be decomposed into sampling and non-sampling error (Groves and Lyberg, 2010). Sampling error arises because we do not see a census of the population we are interested in studying; it typically decreases with sample size. Non-sampling error, on the other hand, comes from everything else that can lead to inaccuracies in our estimate; this includes, for example, incomplete sampling frames, nonresponse, interviewer effects, and many other factors. Unfortunately, there is no general reason to suspect that non-sampling error decreases with sample size in the same way that sampling error does. Although we typically have techniques for estimating the magnitude of sampling error, we do not usually have similar techniques for estimating total error. By using estimands whose true value is known, we are able to estimate total, rather than just sampling error.

Future versions of this paper will present a more finely developed formalization of these concepts. For now, to be more concrete, suppose we have k known estimands $X_1, \ldots, X_k$, which we estimate from our sample with $f_{X_1}(s), \ldots, f_{X_k}(s) = \hat{X}_1, \ldots, \hat{X}_k$. The idea is that we can compare the estimates, $\hat{X}_1, \ldots, \hat{X}_k$ to the true values, $X_1, \ldots, X_k$ to learn about how accurate f is as an estimator. For example, if we were interested in the mean squared error in estimates made by f, $\text{mse}(f)$, we might use

$$\widehat{\text{mse}}(f) = \frac{1}{k} \sum_{i=1}^{k} (\hat{X}_i - X_i)^2. \tag{6}$$

This expression reveals that our strategy is to use the evidence we have about how accurate f is when estimating the known estimands $X_1, \ldots, X_k$ to try and learn something about how accurate f is when using it to estimate any quantity of interest. In the Rwanda example above, we use the evidence about how accurately we are able to estimate the sizes of the known populations to try and understand which tie definition will do a better job of estimating population sizes in general. This highlights the fact that any estimator of the total error of f from our sample and our known estimands will have to make assumptions about the relationships between the errors from using f to estimate each individual known estimand, which we observe, and the errors of f used to estimate other quantities. The simplest assumption that would justify Equation 6 would be that $\text{mse}(f_X)$ is identical for any estimand X; future versions of this paper will articulate the exact assumptions that we must make more precisely.

To recap, if we have an estimator f and estimands whose true value is known, then we can develop estimators for the total error that results from using f to predict those known quantities. Under additional assumptions, we can also estimate the total error that results from using f to estimate other quantities, including the unknown estimands we are most interested in.

3.2 Using survey experiments to improve quality

Now we consider how estimating the total error can be of use to us in distinguishing between different possible ways of desiging our survey instrument or data collection procedure. We extend the situation described above to consider an estimator f that depends on some design parameter, which we will call θ. This is just convenient shorthand for different ways that the estimator f could be operationalized; for example, in the Rwanda study, the network scale-up estimator depended upon the definition of a network tie that we chose. In that situation, we might denote the basic definition by θ_A and the meal definition by θ_B. As we saw in the Rwanda study, we can randomize which respondents receive which version of the questionnaire, θ_A or θ_B. Our sample s then has two disjoint parts, s_A and s_B. We adjust the notation introduced above to write the estimator for X_1 using θ_A from sample s_A as

$$f_{X_1}(s_A, \theta_A) = \hat{X}_1^A. \tag{7}$$

The fact that we can estimate the total error, $\widehat{\text{err}}$ in s_A and s_B gives us an objective, pre-specified criterion that we can use to compare the performance of θ_A and θ_B. After our survey has been conducted, we can to compare the estimated total errors from s_A and s_B to draw conclusions about how different estimates made using θ_A and θ_B are, and also whether θ_A is preferable to θ_B, or vice-versa. In the Rwanda case, we used three different pre-specified error criteria to conclude that the performance of the meal definition was superior to the basic one. For example, using the basic definition we estimated that about 0.39% percent of the population were clients of sex workers, while the meal definition produced an estimate of about 0.5% (RBC/IHDPC et al., 2012). Which of these is to be preferred? Based on the estimates for groups of known size, we conclude that the meal definition seems to produce estimates with lower error.

Our proposal is illustrated in Figure 5. Conceptually, we begin with the sample drawn by

our survey design, s. We randomly assign the elements we sampled to be in condition A or B. We apply the data collection technique denoted by θ_A to the sample in condition A, and the technique denoted by θ_B to the sample in condition B. Within each condition, we use our estimator f and either θ_A or θ_B to produce estimates $X_1^A, \ldots, X_k^A$ and $X_1^B, \ldots, X_k^B$ of the known estimands, $X_1, \ldots, X_k$. We then compare the estimates to the known values to produce an estimate of the total error within each experimental condition, $\widehat{\text{err}}_f(\theta_A)$ and $\widehat{\text{err}}_f(\theta_B)$. These errors form the basis for our comparisons of θ_A and θ_B.

3.3 Combining estimates to use the entire sample

Using a survey experiment to learn about the effectiveness of different data collection procedures might seem like it comes at a high price. Using the Rwanda study as an example, recall that we assinged half of the sample to one tie definition and half to the other. This was useful because we learned something about which of the tie definitions appeared to perform better in producing estimates, but it might appear that this came at the cost of halving our sample size. Fortunately, for many estimators the choice between conducting an experiment like the ones we describe and producing precise estimates of the unknown quantities of interst need not be so stark. In this section, we show how we can blend the estimates from each experimental condition to produce an overall estimate that makes use of our entire sample. We can also produce a standard error for the overall estimate.

Figure 6 illustrates the approach we suggest. We have randomized our survey respondents to conditions θ_A and θ_B as described in Figure 5. For an unknown quantity of interest Y, we produce estimates $\hat{Y}^A$ and $\hat{Y}^B$ within each condition, and also estimates of the standard errors, $\hat{\sigma}_{\hat{Y}^A}$ and $\hat{\sigma}_{\hat{Y}^B}$. We then combine the estimates from within each experimental condition into an overall estimate that uses the entire sample using a blending estimator, g. One simple example of such an estimator is derived in Appendix A.

We are actively developing the details of how to blend different estimates together; Appendix A shows some preliminary results of our work. The final paper will include an expanded section on this important matter.

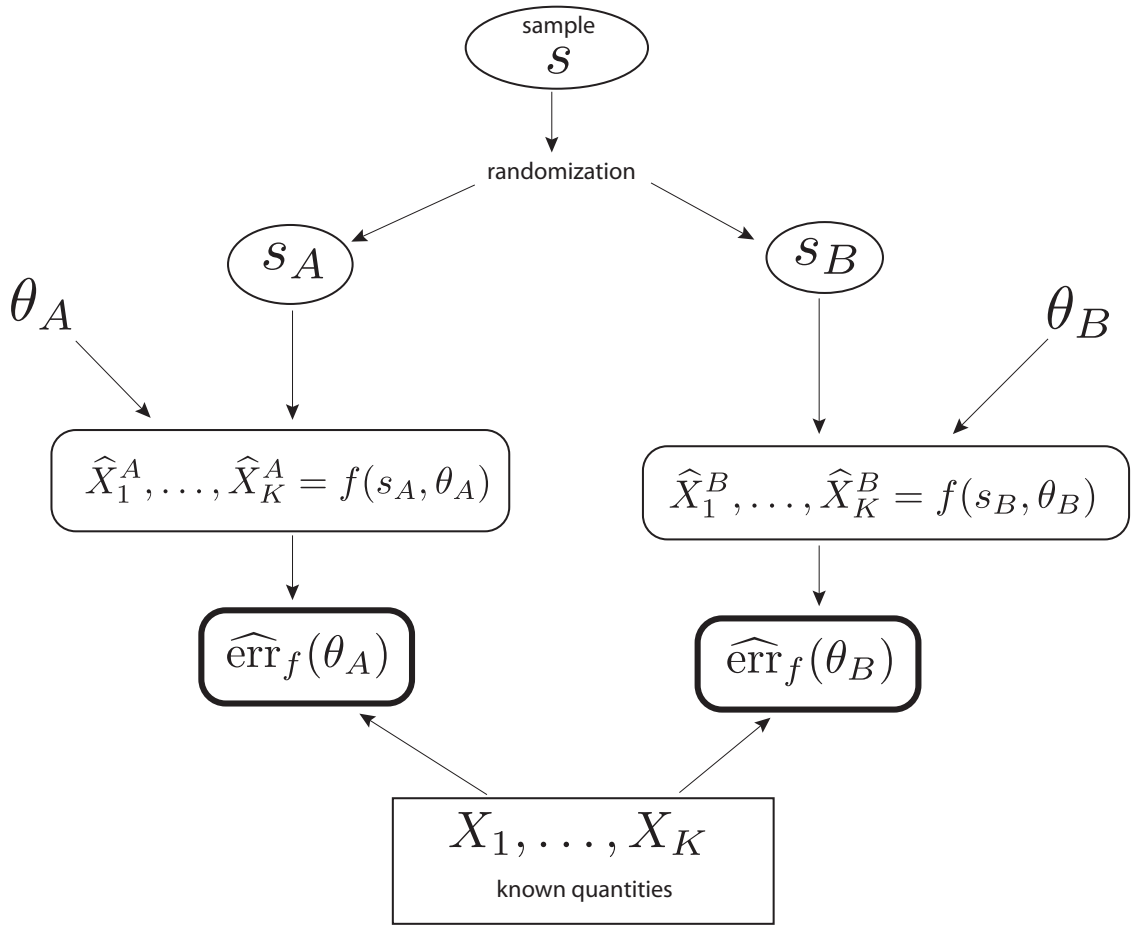

Figure 5: Our proposal for estimating total error due to two different possible data collection strategies, θ_A and θ_B. Conceptually, we begin with the sample drawn by our survey design, s. We randomly assign the elements we sampled to be in condition A or B. We apply the data collection technique denoted by θ_A to the sample in condition A, and the technique denoted by θ_B to the sample in condition B. Within each condition, we use our estimator f and either θ_A or θ_B to produce estimates $X_1^A, \ldots, X_k^A$ and $X_1^B, \ldots, X_k^B$ of the known estimands, $X_1, \ldots, X_k$. We then compare the estimates to the known values to produce an estimate of the total error within each experimental condition, $\widehat{\text{err}}_f(\theta_A)$ and $\widehat{\text{err}}_f(\theta_B)$. These errors form the basis for our comparisons of θ_A and θ_B.

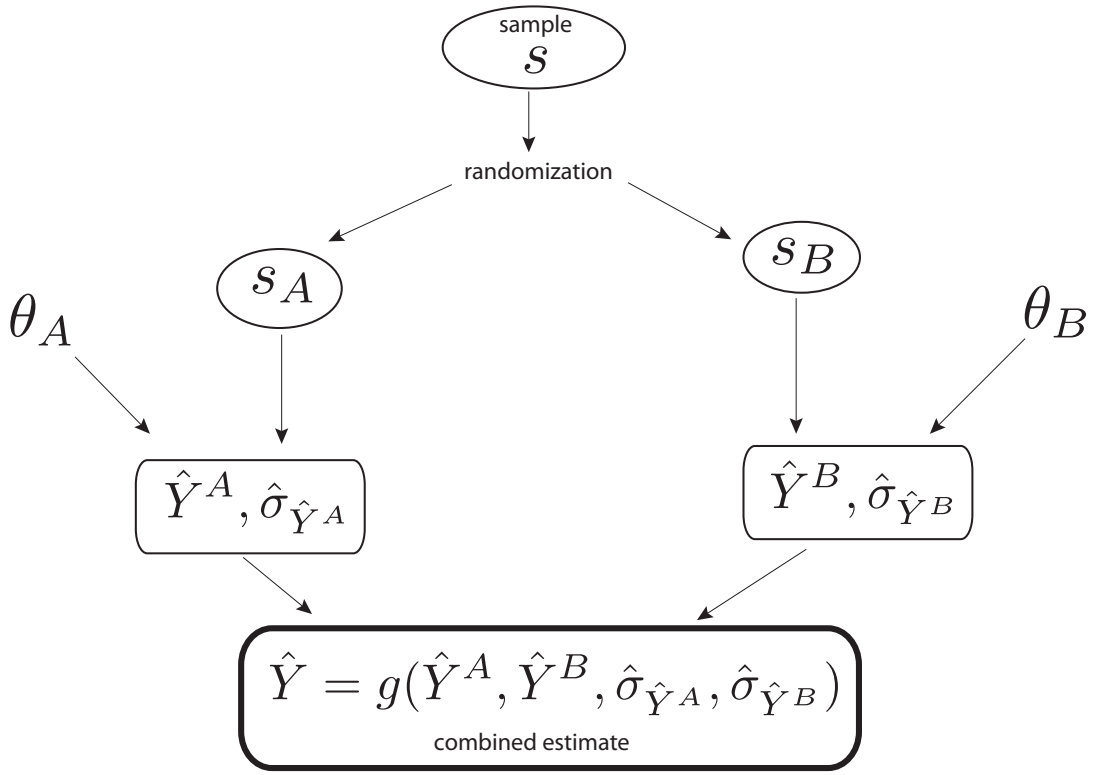

Figure 6: The approach we suggest for combining estimates from different experimental conditions. We have randomized our survey respondents to conditions θ_A and θ_B as described in Figure 5. For an unknown quantity of interest Y, we produce estimates $\hat{Y}^A$ and $\hat{Y}^B$ within each condition, and also estimates of the standard errors, $\hat{\sigma}_{\hat{Y}^A}$ and $\hat{\sigma}_{\hat{Y}^B}$. We then combine the estimates from within each experimental condition into an overall estimate that uses the entire sample using a blending estimator, g. One simple example of such an estimator is derived in Appendix A.

4 Conclusion: continual improvement as a philosophy

We introduced an experimental framework for continually improving survey research. We started by describing the results of a survey from Rwanda that had an experiment embedded within it. We were able to use the results of the experiment to learn something important about collecting data using the network scale-up method. We then generalized that experience into a framework that highlights the two key conditions required to enable us to produce estimates of total error, rather than just sampling error. We also described how this framework, combined with experimental designs, can allow us to obtain evidence about which survey practices produce the most accurate estimates. This framework could potentially be useful in a wide range of different survey contexts. Finally, we described how to make the best use of all of the data collected by combining estimates from the different experimental conditions.

Our hope is that continual improvement in survey techniques can become standard practice in many fields. Surveys are conducted to help shed light on specific empirical questions. But we argue that each survey is also an opportunity to learn something about the science of surveys in general. Our experimental framework for continual improvement is intended to illustrate one way that continual progress might be made each time a new survey goes into the field.

There is much work left to be done. Future versions of this paper will improve all of the sections, with particular attention paid to the formal description of the framework and the results needed to combine estimates from the different experimental conditions.

A Combining estimates from a survey experiment

In this section, we illustrate how estimates from two experimental sub-samples can be combined to form an overall estimate. Later versions of this paper will consider more complicated situations; for now, we will address the simplest case of linear combinations of two unbiased estimators.

Suppose that $\hat{Y}^A$ and $\hat{Y}^B$ are estimators for a population quantity Y based on s_A and s_B. We assume that $\hat{Y}^A$ and $\hat{Y}^B$ are unbiased, and that we estimate their standard errors to be σ_{Y^A} and σ_{Y^B} respectively. We will also assume that the estimates are independent of one

another, that is, that $\text{cov}(\hat{Y}^A, \hat{Y}^B) = 0$. (We will explore relaxations of this assumption, which might be neccesary if, for example, the same interviewers conducted interviews in s_A and s_B, in future versions of this paper.)

We'll consider possible linear combinations of these two estimates

$$\hat{Y} = w\hat{Y}^A + (1-w)\hat{Y}^B, \tag{8}$$

where $w \in [0, 1]$.

Our goal is to determine what the optimal value of w should be. The mean squared error of $\hat{Y}$ is

$$\mathbb{E}[(\hat{Y} - Y)^2] = \text{var}(\hat{Y}) + \text{bias}(\hat{Y})^2, \tag{9}$$

where the expectations are taken with respect to the different possible samples (in a design-based sense). Since we have assumed that the estimators are unbiased, this is reduced to simply $\text{var}(\hat{Y})$. Note that

$$\begin{aligned} \text{var}(\hat{Y}) &= \text{var}(w\hat{Y}^A + (1-w)\hat{Y}^B) && (10)\\ &= w^2\sigma^2_{\bar{Y}^A} + (1-w)^2\sigma^2_{\bar{Y}^B}. && (11) \end{aligned}$$

So we have concluded that, under these assumptions,

$$\text{MSE}(\hat{Y}) = w^2\sigma^2_{\bar{Y}^A} + (1-w)^2\sigma^2_{\bar{Y}^B}. \tag{12}$$

Now we wish to know which value of w will minimize this error. Taking derivatives, we see that

$$\begin{aligned} \frac{\partial \text{MSE}(\hat{Y})}{\partial w} &= 2w\sigma^2_{\bar{Y}^A} - 2(1-w)\sigma^2_{\bar{Y}^B} && (13)\\ &= 2w(\sigma^2_{\bar{Y}^A} + \sigma^2_{\bar{Y}^B}) - 2\sigma^2_{\bar{Y}^B}. && (14) \end{aligned}$$

Since we are after a minimum, we set this equal to 0 and solve for w to obtain

$$w = \frac{\sigma^2_{\bar{Y}^B}}{\sigma^2_{\bar{Y}^A} + \sigma^2_{\bar{Y}^B}}. \tag{15}$$

We can double-check that this is a minimum by differentiating Equation 14 again to obtain

$$\frac{\partial^2 \text{MSE}(\hat{Y})}{\partial w^2} = \sigma^2_{Y^A} + \sigma^2_{Y^B}. \tag{16}$$

This is indeed greater than 0, meaning that Equation 15 is a minimum.

In this simple case, then, the best linear estimator uses the weight given by Equation 15. In future versions of this paper, we will explore optimal estimators under relaxations of the assumptions outlined above.

References

Bernard, H. R., Johnsen, E. C., Killworth, P. D., and Robinson, S. (1989), "Estimating the Size of an Average Personal Network and of an Event Subpopulation," in *The Small World*, ed. M. Kochen, pp. 159–175, Ablex Publishing.

Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., Mahy, M., Salganik, M. J., Saliuk, T., and Scutelniciuc, O. (2010), "Counting hard-to-count populations: the network scale-up method for public health," *Sexually Transmitted Infections*, 86, ii11ii15.

Granovetter, M. S. (1973), "The strength of weak ties," *American Journal of Sociology*, pp. 1360–1380.

Groves, R. M. and Lyberg, L. (2010), "Total survey error: Past, present, and future," *Public Opinion Quarterly*, 74, 849879.

Killworth, P. D., McCarty, C., Bernard, H. R., Shelley, G. A., and Johnsen, E. C. (1998), "Estimation of seroprevalence, rape, and homelessness in the United States using a social network approach," *Evaluation Review*, 22, 289–308.

RBC/IHDPC, SPH, UNAIDS, and ICF (2012), *Estimating the Size of Key Populations at Higher Risk of HIV through a Household Survey*, Rwanda Biomedical Center/Institute of HIV/AIDS, Disease Prevention and Control Department (RBC/IHDPC), School of Public Health (SPH), UNAIDS, and ICF International, Calverton, Maryland, USA.