MINNESOTA POPULATION CENTER

UNIVERSITY OF MINNESOTA

Challenges of Designing and Implementing Three Sampling Frames Targeting Children in Tanzania and Nepal:
Proportional Stratified, Multi-Stage, and Geographically Dispersed Sampling Techniques

Anna Bolgrien†
University of Minnesota

Deborah Levison
University of Minnesota

May 2021

Working Paper No. 2021-02

DOI: https://doi.org/10.18128/MPC2021-02

†This paper comes out of the Animating Children's Views project, which is a collaboration between the two authors. Bolgrien took primary responsibility for sampling during the pilot projects; this working paper was part of Bolgrien's dissertation. The authors gratefully acknowledge support from the Human Rights Initiative Fund of the University of Minnesota; the University of Minnesota's Grants-in-Aid of Research, Artistry and Scholarship; and the Minnesota Population Center, funded through a grant (P2C HD041023) from the Eunice Kennedy Shriver National Institute for Child Health and Human Development NICHD.

1. Introduction

Calls for greater transparency in research practices and online publication of raw data are transforming research in all fields (Schooler 2014; Grahe 2018). Survey methods are also changing; moving from pen-and-paper surveys to digit tablets or online platforms allows for more data *about the survey* to be collected, stored, and analyzed (Hughes, Haddaway, and Zhou 2016). This data about the survey is categorized as *metadata* (data about the data) and *paradata* (data about the process of collecting the data). The accessibility of metadata and paradata allows researchers to validate the data collection process with greater integrity (Couper 2005). Utilizing paradata and metadata in sampling and survey data collection is one answer to the call for greater transparency, but it is not a panacea as processes of data collection and sampling are still marginalized in most social science literatures and practical strategies and best practices are rarely taught in classrooms.

Survey technology, with the ability to utilize metadata and paradata is now affordable and accessible to even small-scale researchers. But there are not clear guidelines on what to do with paradata (Lynn and Nicolaas 2010) The theoretical recommendations that do exist are often targeted at large-scale, nationally representative surveys. Detailed sampling frames and methodological reports published about these large projects outline cross-country comparability and high-level theoretical equations applied to large sample sizes[2]. But most researchers, particularly graduate students or early-career faculty, will not be involved in data collection on a large scale because of limitations of cost, budget, time, collaborators, expertise, or niche topics. Individuals aiming at independent field work and data collection struggle to apply literature and theory describing large-scale surveys to answer practical decision-making concerns in small-scope projects.

Small pilot projects are the bedrock of academic curiosity and exploration but lack the high-profile status of well validated and published large-scale quantitative research. Researchers learn many practical aspects of designing and implanting sampling frameworks via small-scale studies. New theories or hypothesis can be tested to provide justification and support when applying for grants or planning larger studies. Small survey projects also accompany qualitative

[2] Some examples of cross country surveys conducted in developing countries include the Demographic and Health Surveys (The DHS Program 2021), the UNICEF MICS surveys (UNICEF 2020), and the PMA2020 surveys (Zimmerman 2017)

work and mixed methods work (Onwuegbuzie 2007). Practitioners often learn from doing as they gain on-the-ground experience with the messy reality of fieldwork and adapt sampling methods to the local socio-political environment. Yet when it comes to publishing academic work, the messiness of the sampling process is often sanitized through the careful reporting of handpicked success rates that smooth over the details and complexity of the sampling processes. To discuss vulnerabilities in the data is to risk rejection by journal referees. As a result, first-time survey researchers doing small projects lack published examples. This creates a gap in the literature right at the intersection of validation and transparency.

This paper describes three small-scale pilot studies and highlights challenges of field research by taking a detailed look at the entire process of collecting survey data: from the decisions over what type of sampling strategy to use, through the implementation of the strategy in the field, and finally to the reporting of successes and failures of that process. This holistic view is aimed at development practitioners working at intergovernmental or nongovernmental organizations or independent researchers who intend to publish in academic journals.

Through three case studies, this paper describes the sampling processes used in the Animating Children's Views (ACV) project in rural and urban Tanzania in 2018 and peri-urban Nepal in 2019. The project's goal was to produce random sample of households in each of the case study locations[3]. However, due to the physical logistical limitations and social structures of geo-political life in each location, this proved to be complicated. Building on traditional methodologies for sampling frameworks in developing countries, the three case studies implemented three different sampling strategies: proportional stratified sampling (rural Tanzania), multi-stage sampling with two rounds of randomization (urban Tanzania), and geographically dispersed sampling with three stages of randomization (peri-urban Nepal). This paper will describe the logistical processes of sampling in each of these three case studies and consider how differences in social structure affects decisions related to the sampling and data collection (Parts 2-4). The results of the sampling process are summarized according to the Total Survey Error framework (Part 5). Then we compare the results of data collection for each of the cases studies through various reporting measurements common in academic publications, such as success rates, response rates, refusal rates (Part 6) and produce sample weights and population estimates (Part

[3] At each household, one adult (usually the mother) was surveyed in addition to all children age 12-17-years-old who resided in the household. This paper describes only the process of sampling and surveying households, not specifically the individuals within the household. Unless otherwise specified, a sampled and surveyed household refers to a household where the field team interviewed one adult and at least one child in the age range.

7-8) constructed using survey paradata. By constructing these measures, we evaluate the goal of creating comparable samples in three sites, though we do not claim to generalize to a broader target population nor intend to pool the data across the pilots. we suggest the measures described – success rates, response rates, refusal rates, sample weights, and population estimates – are commonly oversimplified in published research in ways that mask the complexity of the data and data collection process. Finally, we propose guidelines that indicate which measures allow for transparency in cases of sampling designs for small scale projects (Part 9).

1.1 Total Survey Error Framework

The field of survey research aims to improve sampling and data collection methodology and decrease errors in statistical measures. The Total Survey Error (TSE) framework established by Groves (Groves 2011; Groves and Lyberg 2010) provides a theoretical understanding of the process of conducting a survey and identifies areas of potential bias [Figure 1]. Briefly, conducting a survey has two major potential components: measurement and representation. The measurement side encompasses the validity of the survey instrument in capturing data that accurately represents the concepts it intends to measure. The representation side include the progression from the target population to the sampling frame, sample, and respondents. One major critique of the TSE is the lack of quantifiable measurement recommendations for the different types of error found along the process of conducting a survey (Groves and Lyberg 2010). Nonetheless, it can be a useful tool for survey practitioners to explicitly explore the possible errors in their survey.

[Figure 1]

In this paper, we use the TSE framework to conceptualize areas of potential bias and error in the ACV pilot studies specifically on the representation side. We do not address issues of measurement and instrument design. Within the representation side of the TSE, there are three main sources of potential error:

- *Coverage error* occurs when establishing a sampling frame from a target population
- *Sampling error* occurs when producing the sample from the sampling frame
- *Nonresponse error* occurs when identifying and collecting data on respondents from the sample.

We first discuss coverage and sampling error in reference to the sampling processes of three ACV pilot studies in the following sections. Nonresponse error will be addressed in greater detail in section 5 as we discuss different potential challenges of identifying, finding, and surveying respondents from the sample.

1.2 Background on Sampling Frames

A sample, in survey research, represents a population without conducting a census (Hubbard et al. 2016). It is important to first establish some general terminology. The *target population* is the population of interest for the survey. For example, the target population could be all households in a specific geographic area or all individuals with a shared characteristic, such as attending a specific school. In the ACV pilot, the target population is all households with at least one child age 12-to-17 living in specified geographic areas: a village in rural Tanzania, a specific urban area in Tanzania, and selected municipalities in Nepal. A *sampling frame* identifies all eligible units (i.e. households, individuals) of the target population. Ideally, this is a complete list of all members of the target population; for example, a list of all households and household members in the geographic area or a full roster of all students attending a school. The sampling frame is rarely a complete list and is often constructed via probabilistic selection or from multiple sources. From the sampling frame, a *sample* is drawn of individual units (households, individuals) who will be contacted to participate in the study. The sample is a representation of the target population drawn from the sampling frame using a probabilistic method[4].

When a sampling frame contains complete information about the target population, a simple random sample can be drawn from the sampling frame. However, this process is rarely as easy or as simple as it may seem. Random sampling in developing countries can be difficult, depending on availability of current population data, accurate spatial boundaries, and clearly organized and labelled households and communities. In areas that lack accurate or complete sampling frames, additional sampling techniques must be applied to create a representative and probabilistic sample of the target population. At the forefront of the research on developing sampling frames are epidemiologists, who generally use the Expanded Program for Immunization (EPI) framework developed by the World Health Organization in health studies; an early example of this method found in (Henderson et al. 1973). EPI sampling methods generally have a two-stage (or more) sampling process. First, communities or clusters are purposefully or randomly selected within a larger geographic area. For example, a sample of villages may be selected

[4] Summarized from Survey Research Center (2016).

within an entire country. Following the identification of communities, the EPI method requires either a complete list of households in the target population – which can be all households or households with a specific sub-population characteristic (e.g. children 0-5) – to create a random sample or another method of randomization can be used to determine which households to sample. Few communities have a such a list readily available; often, the only reasonable alternative is to conduct the census of households oneself, which can be expensive and time consuming. If this cannot be done, sampling methods such as a random walk or "spin the pen" method may be used to identify households, though these techniques are subject to criticism of their possible lack of probabilistic nature (Grais, Rose, and Guthmann 2007; Bauer 2016). Another option is to work with a national statistical office, which often requires complicated social relationships and a recent national census.

The EPI sampling method is considered the standard for sampling and has been adapted by many major cross country survey organizations in developing countries, including the Demographic and Health Survey Program (The DHS Program 2021). Modifications of this method have generated a rich diversity of sampling frames reminiscent of EPI as researchers adapt the method for the inclusion of new technologies such as GPS and satellite imaging (Haenssgen 2015; Wampler, Rediske, and Molla 2013; Kondo et al. 2014) or greater statistical specificity (Turner, Magnani, and Shuaib 1996; Milligan, Njie, and Bennett 2004). New technologies such as computer-aided personal interviewing (CAPI) have allowed researchers additional tools for sampling, tracking data collection, and verifying data quality (Savel et al. 2014; Hughes, Haddaway, and Zhou 2016; Caviglia-Harris et al. 2012; Abelsæth 2012).

While the use of technology moves the field of sampling and survey research forward, these modifications to accommodate more complex probabilistic sampling are challenged by the reality of conducting research in developing countries. The demands of each location are accompanied by limitations including lack of accurate and up-to-date data from government officials and cooperation or resistance of local leaders. Large data projects also tend to hire large field teams; it is difficult to completely account for differences among individuals doing the data collection and sampling, despite best efforts of training and streamlining survey procedures. The social and human element plays an important but underreported role in the success or failure of any field work project.

1.3 Animating Children's Views Project

The goal of the three pilots described in this paper was to produce three studies that were of similar size and deployed similar research designs. The Animating Children's Views (ACV) project is a mixed methods study that developed a new survey methodology; it uses cartoon videos to survey children about their views and perspectives on issues that are facing young people. The ACV pilots establish a methodology that could be expanded to a national or cross-national scale (Levison and Bolgrien 2020). The methodology is currently being tested in small pilot studies designed as household surveys, for eventual use by large-scale survey operations. Person-to-person interviews use the tablet-based survey software SurveyToGo (Dooblo, n.d.). Built into the SurveyToGo software are quality check measures that track time spent on each question, possible modifications to answers, or falsification of data. Although the ACV project is small in scope and overall budget, the project sought to mimic a large representative household survey in both design and sampling strategy using EPI and other sampling literature as the foundation for developing context specific sampling frames. The project sampled from the target population of households with 12-17-year-old members. These pilots provide realistic examples from which to examine and critique the process of applying textbook strategies in a complex and messy world, collecting data with small teams on limited budgets in three very different places.[5]

Many excellent textbooks and review articles have outlined different kinds of sampling frameworks (Johnson et al. 2019; Kish 1965; G. Kalton 1983). To use the language of Fottrell and Byaas (2008), the rural Tanzania pilot used a proportional stratified sample, the urban Tanzania pilot used a multi-stage sample where both stages included randomization, and the peri-urban Nepal pilot used a geographically dispersed sample with two stages of randomization and a random walk. For reference, the stages of sampling are outlined in Table 1. The literature relevant to each of these strategies will be outlined in more detail below. Each of these methods establishes a strategy that creates a sampling frame from a target population and then conducts a sample from the sampling frame.

[Table 1]

[5] This project is part of an ongoing protocol, approved by the Institutional Review Board at the University of Minnesota. The research was also reviewed for ethical and social appropriateness by COSTECH (Commission for Science and Technology) in Tanzania, district and municipality offices in Tanzania, and municipality offices in Nepal. Oral informed consent was obtained from local community leaders as they assisted in the sampling process in Tanzania.

2. Pilot 1- Rural Tanzania

Tanzania was selected as the first country to pilot the new ACV methodology. The East African country has a strong system of bureaucracy and local leadership within small communities of people. There are 30 major regions, divided into 169 districts that are divided into municipalities. In rural municipalities, villages are further divided into sub-villages. Urban municipalities are divided into wards, then “streets” (called *mtaa* (singular) or *mitaa* (plural) in Swahili), and finally ten-cells (originally groups of 10 households). Theoretically, there is a “ten-cell leader” who is responsible for knowing the identities of ten households living within a small area like a block; however, the size of cells varies greatly depending on the urban areas. The sub-village and ten-cell leaders are responsible for keeping lists of people in their cell or sub-village. Thus, Tanzania appears to be an ideal context to use a sampling strategy that relies on accurate household lists of current populations living in small spatial units even when national census data is out of date or not at a small enough spatial geography. The last census in Tanzania was conducted in 2012; since then the country has experienced a lot of population growth and migration, so the 2012 population figures were not likely to be accurate in 2018.

In Tanzania, the research team worked with a local survey research organization to hire a team of field researchers to conduct the sampling and the data collection. ACV conducted two pilot studies in Tanzania: one in a rural village and one in an urban city, both purposefully selected within Arusha Region in Northern Tanzania. The aim of these pilot studies was for the ACV study to mimic a large-nationally representative survey while being limited by a realistic budget and time constraints. We wanted to identify pilot areas diverse in ethnic groups, religions, and livelihoods located within the study area. By conducting pilots in the same region, we spent less time obtaining approvals and permission letters from regional officials. The target population of the ACV project is households with children 12 to 17 years of age. The intention was to survey one adult household member, preferable the mother, and at least one child in the 12-17 age range. We used probabilistic sampling; however, as described below, the project can only generalize within the geographic areas we worked in and not for Tanzania as a country. The authors were onsite to supervise the data collection process, with daily debriefings.

2.1 Village Selection in Rural Tanzania

In the rural Tanzania ACV pilot, we purposefully selected a single village for data collection. The village selected is adjacent to a small urban center along a major highway[6]. The village was selected based on previous knowledge of its population as being diverse in religions, tribes, and livelihoods (Ritter et al. 2010). It was also selected for logistical reasons: it was within a single day's drive from a major city where the team was based, the road to the village was passable in July, and the area had cell service to allow the team to communicate. Sampling and data collection were limited to a single 12-day time frame because of time and budget restrictions.

2.2 Sampling Process in Rural Tanzania

We used a proportional stratified sample based on the structure of rural villages in Tanzania. This strategy divides the village proportionally to the population of its sub-villages. The population of sub-villages varies; thus, a simple random sample would result in higher-population sub-villages having a higher probability of their households being selected. Proportional stratified sampling maintains the proportional share of participants by mandating a proportional number of households in each sub-village, the *stratum*, be included in the final sample. Randomization occurred within each of the sub-villages. Figure 2 show the process of sampling in rural Tanzania.

[Figure 2]

Each of the seven sub-villages in the village selected for the pilot was represented by a sub-village leader. After confirming cooperation from village leaders, we asked each sub-village leader to provide the research team with a written list of the persons and households living in their sub-villages. If a list did not exist already, we paid sub-village leaders for their time and help preparing the lists and identifying households. This process followed standard recommendations for EPI sampling to conduct complete enumeration of a selected area (Milligan, Njie, and Bennett 2004; Turner, Magnani, and Shuaib 1996). The success of this process will be discussed in further detail below.

For each sub-village, we counted the number of households and people identified by the sub-village leader. To conduct our study, we needed to sample only among households with children ages 12-17 years old. To do this, we assigned each household on the list (those with and

[6] The village is unnamed for confidentiality reasons

those without children ages 12-17) a number. Then, using a random number generator, we identified households based on the order of the random number generator, keeping only those households with children ages 12-17 on our final list.

To implement the proportional stratified sample, the number of households per sub-village was capped based on a proportion of the total population of the village. These proportions were used as guidelines for the number of households successfully interviewed in each sub-village. The aim of our study was to survey approximately 100 households with children total in the village. On average, we needed 17 households per sub-village with smaller sub-villages needing a minimum of 13 and the larger sub-villages needing a maximum of 25 to be proportional to the overall population size of the village. We complied lists of between 20-30 total households with children in each sub-village in order to account for refusals, inaccurate reporting of children's ages on the household lists[7], and other problems such as not being able to find the physical location of the households or not being able to find the members of the household. In some sub-villages, listing 20-30 households with children 12-17 ended up being almost a census of households with 12-17-year-old children due to smaller overall population numbers. For all sub-villages, we achieved the desired number of households as proportional to the population of the village.

2.3 Social and Logistical Challenges in Rural Tanzania Sampling and Data Collection

Requesting and maintaining the assistance of local sub-village leaders during the sampling process was the first of many social challenges we faced as we conducted the sample and collected the data. The first step of creating a sampling list is to create an accurate sampling frame. Ideally to reduce coverage error, we hoped to create a sampling frame across the entire village before starting the sampling and data collection in any sub-village. We relied on the knowledge of the sub-village leaders to provide us with accurate information about the population. Some sub-village leaders were prepared and willing to share their lists openly with the research team. Others didn't have lists and took several days to go door-to-door to enumerate the households. One leader would bring a few handwritten pages of lists on one day and then the

[7] In the process of creating the sample frame from the sub-village lists, we knew that there would be households identified as being in the target population of those having a 12-17-year-old but that did not actually have a child of that age. It happened occasionally that the field team would arrive at a household and find that the intended child was actually 10, 11 or 18 years old. As we were able to anticipate this in advance, we were able to construct our sampling frame to accommodate this situation to reduce the potential of over-coverage affecting our coverage error.

next day, he would bring a few more. This resulted in lists being created while data collection in other sub-villages had already started. It was never clear if we had a complete list from this sub-village as the patience of the sub-village leader waned as the days went on and his enthusiasm for helping us diminished. There were also published figures of the population posted in the village office that provided a finer level of detail than figures from the last population census in 2012. We were not able to confirm exactly the date of publication for these posted figures as they varied substantially from the 2012 census numbers and from lists we collected from the sub-village leaders. These concerns fall into the Total Survey Error (TSE) framework potential for coverage error as they pertain to our ability to create a sampling frame from the target population.

In several situations, we completed the sampling frame but were missing several other pieces of information that would assist with the creation of sample weights and outcome responses during the analysis. Partial information varied across sub-villages. For example, for some sub-villages we recorded only the total number of households with children 12-17-year-old members but are missing information regarding the total number of households, and vice versa in others. Some of the sub-village leaders were only available to assist us on specific days, and we were unable to ask about the missing information or confirm our numbers within the data collection time frame. When we did not have proper enumeration of the target population results, overall population figures are estimated based on the information that was gathered when computing outcome measures such as response rates and survey metrics such as weights (discussed in detail in sections below). The social interactions needed to access this information, as described in the process of creating and generating the household lists, greatly depended on the sub-village leaders' interest and availability in working with the research team. The population figures and full enumeration counts used to establish the proportional number of households needed in each sub-village resulted in underestimating the total number of sampled households we expected to be able to find in some sub-villages and overestimating in others.

In the TSE framework, sampling error may occur in the process of establishing the sample from the sampling frame. In addition to social limitations, during the sampling process there were other practical and logistical challenges. First, though our aim was to find children ages 12-17-year-old, we were limited by scheduling conflicts such as school hours and extra tutoring sessions attended by many of the children. It was important for the success of the project to interview the children in a non-school setting. We had to work around the school schedule to find times when school-going children were at home. Additionally, when we were selecting the village, we were not aware of a policy that required all secondary school children in the municipality to attend boarding school. This is discussed in greater detail in sections below. Thus, the sample in the village pilot is missing many older children who were away from home and could not be interviewed in the 2-week period we were in the village. The limitation of school children and boarding school attendance was a logistical challenge in the survey process that potentially affects the sampling error.

Another logistical challenge arose due to transportation and budget issues. As we paid sub-village leaders to assist the data collection team in finding households, travelling between locations was time intensive and costly. Houses in the village were often spread out. The survey team was small and had only one vehicle. This also resulted in a lot of time that some members of the team spent waiting for others to finish interviews.

Concerns about safety also limited the team's activities. In order to keep the team safe, we encouraged all interviews to conclude by sunset, which was approximately 6:15 pm in July. This protected team members but also severely limited the time interviews could be conducted with children after they returned home from school and before they were expected to do chores and other work responsibilities at home. Most of the interviews with adults happened during the day while children were at school. This allowed the field team to prioritize the interviews with children outside of school hours, but it created the additional barrier of needing to return to a house multiple times to meet with the adult and then again with the child or children. Many households also lacked electricity, which made it difficult for the field team to do their job and also may have created an uncomfortable environment for children being interviewed by strangers at dusk or, occasionally, in the dark.

In this first case study in rural Tanzania, the success of the survey team to create a proportional-stratified sample and carry out quality data collection required adapting every step of the process to the local village context. The ability to use the sub-village enumerated lists greatly helped simplify the creation of the sampling frame. This approach was only possible based on the bureaucracy of the sub-village system in rural Tanzania villages. However, even with this seemingly straightforward approach, the social and logistical challenges of field work shaped the data collection process and any potential coverage and sampling error. The quality of the data, metadata, and paradata was entirely dependent on the dynamics between the field team, the sub-village leaders, and the respondents working together to make the sampling and data collection a success.

3. Pilot 2 – Urban Tanzania

3.1 Selection of the Urban Tanzanian Pilot Location

The second pilot was in the booming northern city of Arusha (population: 416,442[8]) that is growing rapidly as people migrate from rural areas. As in the village, the target population was households with at least one 12-17-year-old resident. Urban cities in Tanzania also operate within a hierarchical political system. This benefited our data collection, as it was not possible for us to sample the entire city on our budget. Unlike in the village pilot, it was not possible to conduct an enumeration of the entire city of Arusha. The first task of the ACV pilot was to identify a sampling unit that was small enough to have the household list that could be used as a sampling frame. Areas of Arusha were selected though a multi-stage sampling with two stages of randomization. This sampling method, like the multi-stage sampling technique in Fottrel and Byass (2008), is similar to the original EPI method of sampling which calls for the identification of clusters or strata and then the complete enumeration of the clusters in order to produce a household list; an example of this process is also described in Alves et al. (2012) in Brazil.

3.2 Sampling Process in Urban Tanzanian Pilot

First, it was determined that we would only sample within a purposefully selected municipality that represented the urban areas of the city. Like the rural village, the results from the urban Tanzania pilot can only be generalized to Arusha. There were many challenges in determining the study site because district and municipality lines do not exactly match with other geographic political borders, depending on the source. For example, the city of Arusha is located in Arusha Region, but we had to compare specific wards in order to determine which geographic area to sample: Arusha Municipality, Arusha Urban Municipality, or Arusha City. These three names represent similar, but not identical, geographic areas depending on political units that were not always clear to outside researchers[9]. Finding accurate lists of geographic areas required a local collaborator to make many trips to the municipality office. The sampling frame was based on the geographic boundaries of Arusha City (Arusha Mjini).

[8] From the 2012 census publications

[9] This is similar to the differences between counties, school districts, and congressional districts in the USA. It is imperative that the exact boundaries and units are known before proceeding with a sampling frame.

Within an urban municipality, the next geographic unit is the ward. Across the 19 wards in Arusha, there are 125 *mitaa* (singular *mtaa* in Swahili and generally translated as “street”). Within each of these *mtaa*, the smallest unit of geography is the cell; in an urban area the cells vary between having 10, 50, or even 100 households with one politically appointed cell leader. Figure 3 shows the sampling process in urban Tanzania.

[Figure 3]

We randomly selected 23 *mitaa* to visit. Two *mitaa* were excluded for being too rural and one was specifically used for training purposes. This left 20 *mitaa* in our sample. In each *mtaa*, local collaborators asked the *mtaa* leader to make a list of all of the cells and cell leaders in the *mtaa*. Then, we randomly selected one cell within the *mtaa*. Within this identified cell, we requested that the cell leader create a list of households with 12-17-year-old children in the cell. This resulted in 22 eligible households on average [min:7, max: 58] and we randomly selected 10 to be included in the sample. We allowed substitution of additional households if fewer than 6/10 of the originally selected households were found to be eligible[10]. Through this multi-stage process, randomization occurred at the *mitaa*, cell, and household level.

To make these results comparable with the pilot in rural Tanzania, we sought to interview about 100 households with children across the city. The sampling distribution was designed to be equal across each of the *mitaa*, instead of proportional like in the village study. We hired several local collaborators to go in advance to identify households and collect contact information prior to the start of data collection. We attempted to identify households willing to participate before sending the team, in an attempt to save the research team’s time. This resulted in high participation rates though it was time intensive and costly. We paid the cell leader to take us to residents’ homes and make introductions with sampled participants.

3.3 Social and Logistical Challenges in the Urban Tanzanian Sampling and Data Collection

This multi-stage process was time consuming and required multiple visits to different areas of the cities to allow cell leaders to create the households lists that were sampled from. While we have trust in our local collaborators, we don’t have a good understanding about the quality and completeness of the lists produced by the *mitaa* leaders and/or cell leaders. The city is expanding so rapidly due to migration that it is difficult to expect the leaders to know their

[10] If this did not lead to enough eligible households, we would have moved onto the next identified cell. But this step did not occur.

neighbors in the same way they can in a village or in a smaller urban area. Both of these issues could potentially increase coverage error in our pilot.

Even within one country that relies on the same hierarchical political structure, the process of sampling in rural and urban Tanzania required different sampling strategies. Social challenges of working with local leaders included travelling between *mitaa*, visiting offices, rescheduling appointments, explaining and reexplaining the purpose of the study, and following social norms of respect. All of these steps were necessary, but they required a significant amount of time, energy, and money before conducting a single interview with respondents. Every step influenced the success of data collection and affected the potential for error in the survey statistics. For example, we did not track metadata that recorded the number of attempts to contact each household or the number of visits to each household. It is possible that multiple visits to a household may impact the sampling error on the pilot based on unidentified differences between households that were available for surveying on the first attempted visit, households that were eventually identified after multiple visits, and households that were never identified.

Similar to the village pilot, in the city we faced logistical challenges of transportation, daylight hours, and safety described above. Field team members carried relatively expensive tablet computers, which led us to pay for private taxis instead of asking the team to use inexpensive, but somewhat erratic, public buses. In order to maintain social relationships with local cell leaders and respondents, it was important the field team arrived to scheduled meetings on time.

Respondents in the city tended to be busier and away from home for longer hours than respondents in the village. The team worked hard to accommodate schedules including school hours for the children. Unlike in the village pilot in Tanzania, the urban pilot was scheduled during a vacation time for many students. This benefited the team as it was easier to find children at home during the day, including secondary school students who had returned home from boarding schools for the break. Finally, one benefit of constructing a sampling frame in advance of data collection was that the field team had personal phone numbers for the respondents, given with permission in pre-data collection contact and consent processes. This allowed the team to call respondents in advance.

4. Pilot 3 – Peri-urban Nepal

Nepal was selected as the second country to be included in the ACV pilot projects. The success of the project in Tanzania needed to be replicated in a completely different context in order to show the viability of the ACV project globally. The cultural and social traditions of Nepal vary greatly from Tanzania while still providing a context where many children face difficulties in day-to-day lives. We selected Kathmandu, the country's capital and largest city, as the focus for our study.

4.1 Selection of Peri-Urban Pilot Location in Nepal

As the second Tanzanian pilot tested the ACV methodology in an urban setting, the third pilot aimed to identify a peri-urban or suburban area of Kathmandu. These peri-urban areas are home to a mix of people, both new migrants and multi-generational residents. The city is expanding into the hillsides of the Kathmandu Valley, and areas of jungle and rural villages are now booming with construction and people. The very shape of the Kathmandu Valley is conducive to peri-urban settlements. The central city of Kathmandu is enclosed by a circular road system, called the Ring Road, with road "spokes" that extend into the hillsides and into surrounding municipalities. In Kathmandu District, which includes Kathmandu City, there are 11 total municipalities. Our target population in the third pilot was households with children 12-17 years old within two purposefully selected municipalities outside of the Ring Road.

4.2 Sampling Process in Peri-Urban Nepal

While both pilots in Tanzania had a similar strategy of sampling from household lists, no such political organization exists in Nepal. Without these types of organization, we elected to use a geographically dispersed sampling frame with two levels of randomization. Geographic and spatial sampling has been building on the foundations of EPI sampling as GPS technologies and satellite imagery improve. Researchers can make a sampling frame using GPS and satellites by accessing open source remote sensing data on platforms such as Google Earth or OpenStreetMap. Typically, these maps are used to identify political boundaries and then identify random latitude and longitude within a given boundary in which to start the sampling process on the ground. Alternatively, specific units based on images, such as buildings or plots of land, can be identified to make a sampling list, and randomization happens among these units (Grais, Rose, and Guthmann 2007; Haenssgen 2015). This type of modification to a standard EPI framework has been used in a variety of formats across the world including Malawi (Escamilla et al. 2014), India (Kumar 2007; Montana et al. 2016), Iraq (Galway et al. 2012), Guatemala (Kondo et al. 2014), Lebanon (Shannon et al. 2012), and Haiti (McNairy et al 2019; Wampler, Rediske, and Molla 2013).

Because of budget constraints, we purposefully selected two municipalities with diversity in religions, migrant status, and overall variety in living standards based on recommendations from local collaborators. These municipalities are bounded on one side by the main Ring Road, but they extend far up the hillsides; areas that were formally rural are rapidly becoming peri-urban as the city expands. Once we selected two municipalities, we randomly sampled 50 percent of the wards within each municipality. One municipality had 10 wards and the other had 11; we selected 5 within each, randomly. Figure 4 shows the sampling process in Nepal.

[Figure 4]

Bolgrien downloaded ward boundaries files from the Nepal government municipal offices[11] and digitized them against existing municipality boundaries. Within the boundaries of the randomly selected wards, all buildings or structures were manually identified, and a list of the building latitude and longitudes was created using ArcGIS and satellite images from OpenStreetMap (Esri 2019; OpenStreetMap Contributors 2019). This process of identifying buildings is preferred to strategies that use randomly generated points within an area. Selection of random points can bias the sample weights as there is a potentially unlimited number of possible points to be selected within a geographic area; thus, it is difficult to tell the probability with which points might overlap actual selection of households (Grais, Rose, and Guthmann 2007). The selection of building structures increased the probability that the point selected would be a residence instead of a location in the middle of a forest or a field and decreased potential coverage error on the sampling frame. The identification of buildings also allowed for the construction of sample weights to control for varying population density across the wards by identifying all buildings and weighting based on additional information gathered in the sampling and data collection process (discussed more in later sections). Once all of the buildings in a ward were identified through this manual process, they were sorted into a random order to provide a basis for the sampling framework.

[11] Downloaded January 31, 2019 from the Nepal Federal Government Ministry of Federal Affairs and General Administration (copyright 2017).

Buildings identified from satellite images provide a limited amount of information as they often just show the roof of a structure as a two-dimensional rectangle. It can't be determined if the identified structure is a residential building with people living there or another type of structure. Referring back to the TSE framework (Figure 1), the target population of households with 12-17-year-old residents make up *some* of the residents in the identified buildings, however a significant amount of work in order to identify the sampling frame and sort out ineligible households. In order to identify if the sampled buildings are residential buildings that also contain anyone age 12-17, the sampler enquired about all eligible and ineligible households in each visited building. This work to identify which buildings contained eligible households is a process where there is a potential for coverage error. In order to minimize coverage error and accurately produce a sample frame, and subsequently a sample, we used a second stage walk from the sampled building (Bennett et al. 1991). Using the list of sampled buildings, a team of samplers went to each of the identified structures based on their latitude and longitude and knocked on doors to see if the building was a residence and if there were any 12-17-year-old children living at that residence. The samplers had a protocol for identifying which building or structure was closest to the latitude and longitude; buildings identified from OpenStreetMaps contained few street names and no building numbers or addresses[12]. Samplers identified all of the households in that building and recorded the presence of children in the households. If there were no children residing at the identified households within the sampled building, the samplers were instructed to follow a protocol to identify other buildings close to the point until a predetermined number of eligible households within the vicinity had been identified.

4.3 Social and Logistical Challenges in Peri-Urban Nepal Sampling and Data Collection

In Tanzania, our teams faced social challenges of building and maintaining relationships with local leaders. However, leaders provided the team with legitimacy when interacting with households. In contrast, the field team in Nepal was not introduced to households by local government officials. The samplers and field team had to work hard to establish relationships with each family. This required many phone calls to schedule appointments, early mornings and late evenings to account for working schedules, and long distances traveled to meet with families in person.

[12] Nor were street names or building numbers typically available in neighborhoods across Kathmandu.

There were logistic challenges that came from using GPS coordinates to identify buildings. First, finding the sampled buildings based on a latitude and longitude from a two-dimensional map is very different from finding a three-dimensional building in a physical location. Identifying households required a great deal of perseverance. This was done without any formal street address. Samplers relied on mobile GPS apps like Google Maps to guide them to the specified building. Often buildings were not located on easy-to-access roads or were even located in private gated communities where the sampler was denied entrance. Finding the physical location of a building required samplers to have a strong sense of direction when Google Maps was misleading. These physical limitations took time. For example, a sampler reported a case where a Google Map showed a route to a building that took over 45 minutes of walking to reach, but upon arrival, the sampler realized there was a shortcut that would have taken 10 minutes. As the samplers did not have any information about the household, asking for directions at local businesses or other community members was not helpful, according to debriefing with the sampling team. If a household was not home or available, the team member sometimes relied on neighbors' knowledge of the household in question, particularly if the neighbors resided in the same building. It is uncertain how accurate this information was in some areas.

Once a sampler identified a building at the given latitude and longitude, it was often the case that the residents were not at home at the time, or the building was actually home to several households. Both of these concerns are related to sampling error. Samplers were instructed to try to identify if there were children in any of the families within a sampled building. They were supposed to ask neighbors about any households that were not home at the time. This strategy yielded accurate results in communities there were older, more established, or rural. But because in-migration to Kathmandu is increasing the number of renters, it was often the case that neighbors didn't know the other people living in the same building. Properly identifying households with children to be included in the final list of sampled households took time and some ingenuity. There was not enough time or money to send the samplers to a location many times.

4.4 Validity of Geospatial Sampling in Nepal

One major question affecting the validity of the Nepal sampling process is whether the samplers followed the protocol for identifying households. The protocol established guidelines for samplers to follow after identifying the originally sampled building from the given latitude and longitude. Our random walk protocol stated that the sampler should identify the building at the latitude and longitude given, and then move to the building to the right until the stated number of eligible households are identified. If there was no obvious building to the right (or the building was a non-residential building or open field), the sampler was instructed to move to the left. The intention was the sampler would not stray too far from the original point but also move in a systematic fashion.

We asked each sampler to keep detailed records of the buildings he visited, the number of households in each building, and information about the number of people living in each household he was able to interact with, including eligible and non-eligible households[13]. All of these visited households were reported at the cluster level, as identified based on the sampled GPS point. At each sampling point (i.e. the building identified by the latitude and longitude), we asked the sampler to identify approximately three to five eligible households while staying "within a reasonable distance from the originally sampled point". In some locations, this required the sampler to identify more than five buildings. In other locations, each building contained multiple families and required visiting a fewer number of buildings. Thus, each sampled point resulted in a cluster of households

This process used in the ACV pilot in Nepal varies from the non-probabilistic method of second stage sampling commonly known as a "random walk" or "spin-the-pen" methodology (Grais, Rose, and Guthmann 2007; Bauer 2016). In random walk sampling, the direction in which the sampler would turn after reaching a determined location is randomly decided. In the ACV process described above, the sampler was asked to turn in a systematic direction (to the right) unless he determined this was no longer a useful approach to finding households. The systematic protocol threatens the validity of the "random" walk as the sampler made decisions about which buildings to approach or skip in a way that could introduce bias (Chen et al. 2018). For example, if the sampler turned to the left instead of the right when the building on the left appeared to be better kept, we may have too few lower income families in our sample. Because the process of identifying points was done without any assumption that the building identified was a private residence, we also do not have a full picture of how often the originally-identified building was a nonresidential building such as a shop or a school[14]. The exact route of the walked sampling path was not tracked, so we do not know how far the samplers walked from the original point in order to identify the number of households requested. We have to rely on debriefing conversations with samples to assume that the protocol for identifying buildings were followed appropriately and any

[13] We hired two samplers who both happened to be male.
[14] Buildings labelled as nonresidential were skipped when tagging buildings in OpenStreetMaps

deviations were either the result of necessity (a dog or security guard prevented access to a household) or lack of a private residence.

Analysis Validity of Nepal Geospatial Sampling

In order to further explore the validity of the sampler's route, we leverage additional geographic information from the survey. The SurveyToGo software used to conduct interviews allows for GPS capture at the site of the survey. Most of the interviews were conducted in the home of the family. We compare the GPS location of the home of the respondent to the originally sampled point (building). This analysis provides a measure of the distance between the originally sampling point and multiple surveys conducted in different households that were identified based on the sampler's walk protocol.

Using the Generate Near Tool in ArcGIS, we identified the sampled point closest to each survey location. In 76 percent (n=118) of the interviews, the survey GPS point was paired with the expected sampled GPS point. Among these points, the average distance (as the crow flies) from the surveyed households to the point is 56.5 meters (min = 2 meters, max=307 meters). This indicates that the sampler's walk remained close to the sampled point and the interviewed households were associated with an area near the building.

The household surveys that were not paired with the expected GPS point fall into three possible categories. First, some households (5 percent, n=9) were interviewed in a different ward than the anticipated ward based on the sampling list. Most of these respondents, based on documentation from debriefing, are assumed to have been interviewed at a location different from their residences. For example, a respondent could request to meet at her place of work or a local café or community center. Another possibility for variation in the expected wards would be the cellular reception of the tablet computers used in the survey. Occasionally the tablet computers being used to conduct the interviews would not register the GPS location at a specific home but instead would use the GPS location at the next available point of cellular or Wi-Fi connection.

Second, half of the wards were very rural, spread out, or hilly based on the topography of the Kathmandu Valley[15]. Thus, the sampler may have had to walk longer distances to identify eligible households. The measurement of distance is as the crow flies instead of based on roads that often wound around hills or other steep or challenging terrain. In these more rural wards, twenty (n=20, 13 percent) of the households reported a GPS location in a ward that did not match the expected ward of the originally sampled point. Often, these households were just on the other

[15] This was true for two wards located in one municipality and three wards located in the other.

side of a ward boundary. These households might be a cause for concern as bigger distances between the originally-sampled buildings and the household may include more risk to validity. However, the representation of the exact ward boundary maps used in this study may have been slightly different from the socio-political reality on the ground. For example, one field researcher reported during debriefing that the head of the household claimed to reside in a ward that was different from the ward we had identified in satellite imaging due to recent local political changes in the area. It is unclear if our categorization or the head of household was more accurate in this particular situation. Regardless, this technicality does not factor into the analysis of any survey statistics.

Finally, the remaining eight (n=8, 5 percent) households were interviewed at a GPS location in the expected ward of the originally sampled point, but the distance between the survey GPS was closer to a different originally sampled point than the one the sampler identified. That is, the household may have been identified from the sampler's walk from one originally identified point, but the household was actually located closer to another point that the sampler would also visit. In some areas of Kathmandu, the houses were very dense and resulted in many buildings on OpenStreetMap identified in a very small area. This increased the probability that these buildings would be randomly selected in these areas. Therefore, it is possible that some of the households would have been associated with multiple sampled points. Because not every sampled building yielded eligible households with children 12-17 years old, the path the samplers walked to identify eligible households and the decisions they made resulted in overlap between sampled points within a ward that were close together. This is a threat to the validity of the sampling list because such households would have a higher probability of being included in the sample.

[Figure 5]

Figure 5 shows an example of the sampling process in Nepal using the semi-random walk. This figure is a stylized representation of the sample process and does not represent actual GPS coordinates of the sampled households or buildings in the ACV study. The red triangle represents a sampled building. The sampler would go to the location of these red triangles and start the pre-screening process. The blue circles represent locations where the surveys were conducted. In the ACV pilot, the majority of surveys occurred in the household of the eligible respondents. Therefore, the blue circles usually represent the location of the households identified in the sampling process. Cluster "A" shows a sampled building and six interviewed households that are close to the building. Cluster "A" represents a situation where the sampler followed the protocol and identified eligible households in close proximity to the particular building, as found in 76 percent of the sampled households. Cluster "B" and "C" demonstrate that a household selected from the second-stage sample process from original point “B” may have indeed been close to original building “C”, as found in 5 percent of the ACV sampled households. The household with an (*) represents an interview that challenges the assumption that households were identified from only one originally sampled building. Cluster “D” represents a sampled point that resulted in households that were spread out due to lack of eligible households, rural landscape, or difficult-to-identify buildings, as found in 13 percent of the sampled households. Cluster “E” represents a sampled point that may have crossed a local boundary, as found in 5 percent of ACV sampled households. The household with (**) falls to the outside of a major road, represented by the yellow line. If this road marked a boundary change, it is possible household (**) is located in a ward not sampled, despite being identified based from the sampled building “E”. Finally, the household (***) at the bottom of the map shows a household that was interviewed far from any originally sampled point. This represents households that were interviewed in cafes, community centers, or places of work upon request. This information was documented in the debriefing notes written at the end of each workday.

5. Metadata and Paradata Collected during the ACV pilots

The description of the sampling processes (in the above three sections) of the three ACV pilots provides transparent and realistic details not otherwise found in most published academic work. The social and logistical challenges faced in Tanzania and Nepal created situations that likely impacted the coverage and sampling error [Figure 1] in the ACV pilots, though, as mentioned before, it is not possible to quantify the extent. But establishing a sampling framework and creating a sample are only the first steps of research. Up until this point, we have described the process of establishing a sampling frame and conducting a sample within a target population. The actual work of identifying, contacting, and surveying sampled respondents takes up the majority of field work and is arguably even more prone to social and logistic challenges and potential for error. Data collected during field work can be used to quantify the third type of representation error established by the Total Survey Error framework: nonresponse error. Before we discuss nonresponse error, we provide detail on the types of data collected during the sampling and data collection processes. This data is the foundation for analyzing the nonresponse error on the Total Survey Framework.

The ACV pilots contracted with local survey organizations, but oversight and on-the-ground management was done by me and the project PI. Through daily debriefing and constant personal communication, we were directly responsible for the training, management, and coordination of the field team. We made decisions about challenges and issues that arose daily from the beginning of the sampling process through the end of the data collection[16]. In order to have proper records of every step of the process, we collected paradata about the sampling and surveying process. As a reminder, *metadata* is defined as data about the data. It includes the survey tools, sample design, training materials and research protocols developed in advance to be used by the field team to collect data. *Paradata*, as defined by Couper (2005), is data about the process of data collection. The paradata from the ACV pilots is documented in three main sources: sample lists produced by the sampling team, data collection tracking of complete and incomplete interviews, and daily debriefing notes.

The sample lists created by the sample team in rural Tanzania consist of handwritten lists of names produced with the assistance of the sub-village leaders. Other summary information about the population sizes of the sub-villages was documented during the conversations with the sub-village leaders and by manually counting the names written on the sub-village lists. Due to privacy and ethical reasons, we were not able to photocopy, take photos of, or replicate the sub-village lists for our records. As noted above, it was not necessary to refer back to the original lists except in the cases where some of the summary information was missing. Key pieces of information included the total number of households, number of eligible households, and number of sampled households.

In urban Tanzania, the sampling process was formalized to include worksheets to be filled out by the *mitaa* and cell leaders. On these worksheets, the total number of cells (along with an estimate of the number of households in each cell) and the number of eligible households in a cell provided information used in constructing sample weights and population estimates, described in later sections. The worksheets provided the sampling team with a consistent protocol to use in the *mitaa* and helped the sampling team verify that all of the summary data was collected properly. These worksheets were entered by hand into Excel. In Tanzania, we also obtained published population statistics from local government offices.

[16] In rural Tanzania, the field work period was 13 days for both the sampling and data collection. The urban Tanzania pilot scheduled 10 days for the sampling and 20 days for data collection (including rest days). In the Nepal pilot, sampling took approximately 3 weeks and data collection lasted for 20 days.

Similar worksheets were used in Nepal to document the sampling process. Sampling in Nepal required intensive documentation of the number of buildings visited, the number of eligible and ineligible households in each building and estimates of household as we were relying on the sampling process for all information about population size in the wards. This information was compiled by the local team into a spreadsheet.

Tracking the data collection followed a similar process in the three pilot locations. At the end of every day in the field, the field team gathered together to debrief. During these sessions, the team reported on the households visited. Households were marked as "complete" if both the adult interview and child or children interviews were completed. We recorded whether the team needed to return to the household to conduct one or more interviews. In subsequent debriefings, we amended the documentation to show if the household has been completed or not. Households that were contacted or unavailable were indicated as attempted contacts. Households that were not completed sometimes included details about why the interview was not conducted – child away at school or nobody answering the door—or if the household refused or otherwise indicated they did not want to participate. This spreadsheet for tracking the data collection was updated daily to include details about the overall progress of identifying, contacting, and interviewing households.

Finally, qualitive debriefing notes provided additional paradata about the data collection process. These notes included more details to compliment the data collection spreadsheet. Data collected from the survey process was validated based on the information from the debriefing notes. This included situations where interviews were eliminated from the data or data was edited based on new information. For example, the age of a child reported in the household survey was occasionally incorrect and modified in the data to reflect the age given by the child during the interview with the child. Another situation involved a girl who was interviewed twice by two different members of the field team, though nobody was entirely sure why she did not mention to the second interviewer that she had already participated. Her second interview was removed from the data. These are examples of information documented in the debriefing notes that was used to verify and edit the quantitative data and paradata.

The documentation of paradata is complex and nuanced. In a large-scale survey organization, paradata is standardized and collected automatically within most survey data collection software. For example, the SurveyToGo software used by the ACV project has the capabilities to track time spent on each question, patterns in responses, or retrospective changes to the data by the interviewer; any of which might suggest falsification or manipulation of responses in ways that threaten the validity of the data. In a small-scale survey such as ACV, the entire team could gather daily to discuss challenges and questions. Being on location with the field team meant that authors were on hand to answer questions and problem-solve immediately. The small nature of the pilots and the hands-on supervision by both authors generated a more organic and homegrown method of tracking paradata. That being said, a systematic way of documenting paradata would eliminate many of the challenges we faced when reconciling, cleaning, and properly classifying the paradata presented in this analysis.

In the ACV pilot, three categories of paradata were used collectively to construct measurements often found in publications: response rates and other outcome metrics, sample weights, and population estimates. All of these measurements require accurate data. Each of these measures will be described in a section below; section 6 discusses response rates, section 7 details the calculation of sample weights, and section 8 estimates population counts. We will describe how each of these measurements identifies and quantifies nonresponse error in the data collection process, the third type of representation error in the Total Survey Framework alongside coverage error and sampling error. We do not go into detail about sources of potential error resulting from the interview process itself such as interviewer effects, processing error, bystander effects, social desirability and acquiescence, and cultural power dynamics between respondents and interviewers, as there is a large literature on these topics in survey methods and psychology. Instead we focus on how the sampling process and data collection can yield measurable outcomes often reported in journals.

6. Response Rates and Outcomes Measures using ACV Paradata

In published research, the relationship between nonresponse error and the data collection process is most commonly reported as *response rates* as a way for researchers to provide evidence of validity of the survey results. This is such common practice that journals and reviewers adapted formal and informal interpretations of how high a response rates should be in order to signify a "good quality" study (Carley-Baxter et al. 2009; Johnson and Owens 2002; Bennett et al. 2011). Yet, few journals establish clear guidelines on what exactly defines a response rate and what threshold, if any, meets the standards of "high enough".

Broadly, *response rates* represent the proportion of respondents or participants out of the sampled or target population (Kviz 1977). For example, the number of paper surveys returned divided by the total number of surveys mailed or the number of respondents who answered their phone or door (and completed the survey) when contacted. Historically, reporting response rates when discussing survey research is a standard that was elevated by public opinion polling (Marton and Stephens 2001). Readers of public opinion polls demanded transparency in the validity and representativeness of the survey and the sampling process. An opinion poll that asked a few purposefully selected individuals could not be trusted compared to a poll that surveyed a large number of people across the entire population. Ideally, all individuals who are sampled for a survey would respond to the survey, thereby resulting in a 100 percent response rate. Achieving a 100 percent respondent participation preserves the assumptions of a probabilistic sample being draw randomly from the target population. The more people who responded, the less nonresponse error in survey statistics resulting from differences between the type of people who responded compared to the type of people who did not respond. This means that your sample is more likely to be representative of your target population.

Complete (100 percent) respondent participation is rarely accomplished in practice. But high response rates are seen by journals, reviewers, and readers as a proxy to indicate low nonresponse error[17]; though evidence suggests this may be a flawed assumption (Groves and Peytcheva 2008). *Nonresponse error* occurs when there is a nonrandom difference between respondents and non-respondents that were on the sampling frame [Figure 1], resulting in potential nonresponse bias in survey statistics (Galea and Tracy 2007). The pressure placed on authors and researchers to publish high response rates established the standard that only high response rates were acceptable[18]. In a paper by Carley-Baxter et al., the authors argue that there is not a clear understanding among academics of what it means to have “acceptably high standards”:

> Anecdotally, some of our colleagues hold fast to the perception that it is harder to get studies published if they fail to achieve acceptable response rate standards. However, these same individuals readily admit that they do not have an accurate picture of what, if any, standards regarding data quality or survey error are imposed by journal editors when considering manuscripts which report results based on data analysis of surveys. (2009)

[17] One example of an explicit expectation of response rates by the *Journal of Pharmaceutical Education* states that studies should have a 60 percent response rate, or 80 percent if the study is among college students (Fincham 2008).

[18] It has been well established that response rates for survey research are rapidly falling around the world leading to a number of dramatically titled research articles such as “Where Have All of the Respondents Gone? Perhaps We Ate Them” and “The End of the (Research) World As We Know It?”(Leeper 2019; Stedman et al. 2019). Response rates for mail in and phone surveys, the traditional method of data collection prior to the internet era, have fallen significantly since about the 1980s. Practitioners and academics race to find new methods, and the field of survey research is booming with new and innovative ideas for data collection, sampling, and measuring survey quality.

With an incentive to publish, author may seek to present their study favorably and find ways to report outcomes of the survey that depict a higher response rate and omits any concerns about data quality. For example, the inclusion of partially complete surveys or the exclusion of certain parts of the sampling frame (respondents for whom there was no additional information confirming eligibility) may inflate response rates.

In an attempt to standardize reporting response rates, the American Association of Public Opinion Research (AAPOR) published a report that provides a clear standard for survey methodologists to use for calculating and reporting four outcome measures: response rates, success rates, refusal rates, and contact rates (American Association for Public Opinion Research 2016). Each of these different outcome measures will be discussed and defined in detail in the following section. Such guidelines document what metadata and paradata metrics should be gathered during data collection regarding interviewed and non-interviewed units across different modalities of surveys, including phone interviews, web-based surveys, and household surveys. Using a standardized metric for reporting outcome measure allows comparisons between different surveys regardless of the sampling frames.

While not the first attempt to establish definitions (Smith 1999 as found in Carley-Baxter et al. 2009), AAPOR is the premier academic association for survey researchers working in the United States and a respected authority among survey research scholars[19]. The standards are clear, flexible, and adaptable to survey research in any field. Additionally, citing the AAPOR standards allows researchers to clearly communicate the validity and quality of the data collected to reader, reviewers, and journals. Several journals now require the AAPOR definitions to be explicitly stated in published work, including the *American Journal of Pharmaceutical Education* (Reierson Draugalis, Coons, and Plaza 2008) and the *Journal of the American Medical Association* (JAMA). While the AAPOR standards for response rates and other outcome measures are well accepted by survey researchers, they are rarely used by academics in other fields.

6.1 Adaptation of APPOR Guidelines for 3 pilots

We adapted the AAPOR *Standard Definitions: Final Disposition of Case Codes and Outcome Rates for Surveys Revision 9* from 2016 for the three ACV pilot studies. The AAPOR *Standard Definitions* are written as a tool that researchers can apply to any type of survey,

[19] While AAPOR is an American organization, the standards have been used in surveys conducted in other countries (Beerten et al. 2014)

regardless of the unit of observation, or sampling strategy. We specifically refer to the guidelines for “In Person Household Survey” (Page 23-27). The “household” is used as the unit of observation for consistency between the AAPOR definitions and the ACV pilots.

In this section, we summarize the standard definitions for four outcomes measures: response rates, cooperation rates, refusal rates, and contact rates. All of these outcome measures rely on a common understanding of household eligibility, contact, and completion of survey regardless of the sampling strategy used. We compare outcome measures across the three ACV pilots. Standardizing the response rate, and other outcome measures, allows for a standardized comparison potential nonresponse error in the ACV pilots.

AAPOR Definition of Eligibility

The foundation of all AAPOR outcome measures can be broken into the following categories: eligible households that were interviewed, eligible households that were not interviewed, households that were not interviewed and it is unknown if they would have been eligible, and ineligible households. These four categories are further divided into sub-categories, also described in Figure 6:

- Eligible Households that were interviewed
  - Completed Interview (I)
  - Partial Interview (P)
- Eligible households that were not interviewed
  - Refusals and break-off (R)
  - Non-contact (NC)
  - Not interviewed for other reasons (O)
- Households not interviewed and it is unknown if they would be eligible (UH)
- Ineligible households (IE)

Eligibility is determined based on the definition of the target population; the ACV pilots defined eligible households as households that included at least one 12-17-year-old who would be available to be interviewed. After the target population is defined and the sampling frame is established, the sample is drawn through any of the probabilistic or non-probabilistic methods available. Each sampled household must include one of the above categorizations at the end of the data collection process. The ACV debriefing notes document each interaction with the household and record the status of the household (eligible or ineligible) and the result of the final interaction with the household (interview completed or the reason the interview was not completed). If the household fell into the "unknown eligibility" category (UH), it may or may not have been possible to contact the household or no paradata was recorded about contacting the household, as depicted by dashed lines in Figure 6. We apply this categorization scheme to the ACV pilots, and present examples, in greater detail below.

[Figure 6]

AAPOR Definition of Response Rates

APPOR outlines four different types of rates that should be included in all survey research based on the categorization of sampled households described above: response rates, cooperation rates, refusal rates, and contact rates. Each of these four measures can be calculated multiple ways, as shown in Table 2.

[Table 2]

Response rates are the most familiar outcome metric to most researchers. Generally speaking, a response rate is defined as the number of interviews divided by the number of eligible households from the sample list. The AAPOR report provides six definitions of response rates; three include only completed interviews in the numerator and the other three are comparable but also include partial interviews in the numerator. As partial interviews were not relevant to the ACV project, we only consider three response rates: RR1, RR3, and RR5[20]. In Table 2, all three response rates have identical numerators: completed interviews (I). The variation across the three response rates depends on how households where there was unknown eligibility are included or excluded in the denominator. In RR1, *all* households that were unknown eligibility (UH) are assumed to be eligible but uncontacted. RR1 is the most conservative estimation of the proportion of households responding. Response rates calculated under the definition of RR1 will be lower than other definitions. RR3 adjusts the denominator to estimate that not all of the unknown households (UH) would have actually been eligible if they had been contacted. The proportion, represented by the *e* in the formula for RR3 on Table 2, is determined by the researchers (AAPOR, page 62). This estimation for *e* must be explicitly described if RR3 is to be reported by researchers. We describe below how *e* is estimated in the context of the ACV pilots. The final

[20] RR2, RR4, and RR6 are the official titles for the response rates that include partial interviews in the numerator. We omit these definitions as the ACV pilots did not have any cases with partial interviews. In order to properly follow the APPOR definitions, we maintain the discontinuous numbering system for clarity.

response rate, RR5, drops the unknown eligibility households (UH) from the denominator, assuming that *none* of the unknown households would have been eligible for the survey. RR5 is the least conservative response rate calculated and produces higher response rates than RR1 or RR3.

AAPOR guidelines offer more nuanced understandings of how respondents participate in a survey. The cooperation rate and the refusal rate assess which households actively refused to participate while the contact rate highlights which households were not able to participate due to a non-contact by the field team or an inability to participate in the survey during the field work time.

AAPOR defines *cooperation rates* as “the proportion of all cases [households] interviewed of all eligible units [households] ever contacted” (page 6). In the ACV pilots, we calculate the cooperation rates at the household level only, not for the individual child interviews conducted within the household[21]. The cooperation rates include all of the interviewed households in the numerator over all of the households that were contacted and were either interviewed or explicitly refused to participate. Cooperation rates are defined two ways (see Table 2)[22]. COOP1 includes that completed interviews (I) in the numerator divided by the sum of the completed interviews (I) plus the refusals (R), and other reasons for not completing the interview (O). The category of “Other reasons” include households that were contacted but were unable to participate due to reasons such as not having a proper translator, the participant not being in good health, or religious holidays preventing participation. Reasons specific to the ACV pilots are described below. The second measure of cooperation rate, COOP3, does not include households categorized as being not interviewed for other reasons (O). Thus, COOP3 will be greater than or equal to COOP1.

As a companion to the cooperation rates, *refusal rates* are defined as “the proportion of all cases in which a housing unit or respondent refuses to do an interview, or breaks-off an interview out of all potentially eligible cases” (page 7). That is, the numerator accounts for the number of households that refused to participate or ended an interview early and requested to be removed from the study[23]. As seen in Table 2, there are three versions of refusal rates that share

[21] Individual children can refuse to assent. But the overall household consent was needed for a complete interview of adults and children.

[22] COOP2 and COOP4 include partial interviews in the numerator; these are irrelevant to the ACV project.

[23] This request follows the IRB process of continuous informed consent. In the ACV pilot, there were no cases of partially completed interviews that were broken off.

the same denominator as the three denominators found in the response rates (RR1, RR3, RR5). The different denominators account for what proportion of the unknown eligible households to include in the estimate. The cooperation and refusal rates may be useful to report if a survey has a large proportion of refusals or was unable to conduct interviews for other systematic reasons such as language barriers. However, in the ACV project, there were so few refusals that the cooperation rates were very high, and the refusal rates were very low.

The final outcome measure defined by AAPOR is *contact rates.* Contact rates measure if a household was contacted by the field team. As shown in Figure 6, the flow chart of the AAPOR categorization of household eligibility, contacted households are categorized as resulting in a completed interview (I), a refusal (R), or another reason for an incomplete interview (O). The three variations of contact rates – CON1, CON2, CON3—are defined in Table 2 as the total number of contacted households over all eligible households in the samples. Variation between the three contact rate definitions again comes from differential inclusion of the unknown eligibility households (UH) in the denominator. Contact rates are best reported when there may be systematic concerns about households that were not contacted (NC). In the ACV pilots, the reasons households were not contacted vary greatly and will be described below.

Researchers should clearly state which definition of outcome metrics are being used in published papers. For example, using RR5 assumes unknown households would be ineligible and thus yields a higher response rate than RR3 or RR1. This is especially important if there is a large number of households with unknown eligibility; it is important for readers to be able to decern what assumptions were made about the unknown households when calculating and reporting the response rate. It is not always necessary to define and report all four outcomes measures but doing so provides the readers greater transparency of data quality and potential for nonresponse error [Figure 1].

In the next section, we apply the AAPOR *Standard Definitions* to the ACV pilots. We demonstrate how to apply the methodology to the paradata gathered during the sampling and data collections processes in rural and urban Tanzania and peri-urban Nepal.

6.2 Eligibility in the ACV pilots

Using three different types of sampling strategies, the sampling process in each ACV pilot resulted in a list of households that we assumed to be eligible for the survey. We made this assumption based on household information from the local leaders (in Tanzania) or pre-screening process (in Nepal) prior to the start of data collection. In the ACV study, the target population was households with an adult at least age 18 with at least one child age 12-17 living in the household. More specifically, in order for a household to be eligible for the study, the adult had to be *present* to give consent for the children to participate in the study and at least one child in the age range had to be *present* and available to be interviewed. Households that did not have an adult present or where all of the children in the age range were not available for interview were considered ineligible.

A benefit of the intense sampling process conducted by the ACV sampling teams [described in section 2, 3, and 4] was that ineligible households were often screened out prior to the start of the ACV data collection. Ineligible household do not count as a part of the denominator for any of the AAPOR outcome metrics, so this pre-screening process does not affect any of the rates[24].

Eligibility Concerning Boarding School Students

The original intention of the ACV pilots was to include households with an adult respondent age 18+ and resident children age 12-17. We anticipated that we would find a few households where there would be children who were away for part or all of the field work period and therefore not be available to participate in the study. In the first pilot study in rural Tanzania, we were unaware that we had selected a village located in a municipality where all secondary school students attended boarding schools. This led to a high number of older children in the village being away at school; we were unable to include many of these households in the study as there were often no other children in the age range at home. If the household included a boarding school student who was away but also at least one 12-to-17-year-old child who was at home, the household was eligible.

In the next study, we were more intentional in our screening process to only include households where children living in the household would be present during our field work period. In urban Tanzania, we also timed our field work period to be during school holidays, when we anticipated more students, even those who would normally have been away at boarding school, would be home. In Nepal, some children were visiting family, but fewer children attended

[24] AAPOR specifies four additional criteria for eligibility: (1) the selection of individuals within the household, (2) proxy respondents, (3) substitutions, and (4) status days. None of these criteria were applied in the ACV pilots in ways that changed the results of the outcome metrics of response rates, cooperation rates, refusal rates, or contact rates in a way that affects potential nonresponse error.

boarding school away from their place of residence. Careful pre-screening in the urban Tanzanian and Nepal pilots decreased the number of sampled households later deemed ineligible because all eligible children in the household were away, compared to the rural Tanzanian pilot.

In order to assess whether boarding school students and other children away from their family during the field work periods potentially affected the nonresponse error, we constructed the AAPOR outcome measures two ways. First, households with children who were away were treated as contacted households who were eligible but not interviewed in the field work period. This falls into the specific category of non-contact (NC) because it was determined that there was someone eligible at the household, but as all eligible children were away, the interview could not be completed[25]. This is similar to if an adult were contacted but the interview was not able to take place because the adult was at work at all times that the field team attempted to visit the household; these interviews were not considered refusals because there was still an attempt to contact the household but without confirmation of recruitment as determined in the process of continuous informed consent. In the calculated outcome measures, this inclusion of the boarding school students as eligible, but not contacted is noted with the addition of BS (to stand for "boarding school" in Table 4) to the outcome measure label. In the second set of definitions, households with children away were treated as not being in the sample as they no longer fit the criteria of having children present in the household during the field work period; that is, households with boarding school children were considered ineligible (IE) and therefore not considered in the construction of response rates, cooperation rates, refusal rates, and contact rates[26].

Other eligibility concerns presented in the APPOR Guidelines

In addition to the eligibility definition presented in Section 6.2, AAPOR specifies four additional criteria for eligibility criteria: (1) the selection of individuals within the household, (2) proxy respondents, (3) substitutions, and (4) status days. The *selection of individuals* within a household is a concern if a survey is designed to only interview one specific member of a household. *Proxy respondents* are defined as respondents who were not sampled, but who provide

[25] In ACV, we consider this non-contact (NC) instead of not participating in the survey for "other reasons" (O) because the child being away meant that the child could not be contacted in order to start the assent process with the child.
[26] For example, in Table 2.4 described below, RR1 for rural Tanzania is calculated with and without the boarding school children (BS).

the information needed in the survey on behalf of the sampled individual. *Substitutions* can happen when additional households are added to the originally sampled list when a household is found to be ineligible or eligible but does not participate in the survey. And *status day* refers to a set time frame within which the data collection occurs.

Selection of individual: In many household survey designs, the target population includes only one adult per household (Smyth, Olson, and Stange 2019). In some surveys, like many national censuses, the household is represented by a self-appointed head of household. In other surveys, the adult is selected through a pre-determined protocol. Regardless of the protocol, nonrandom or random, individual members of the households are not known in advance. The ACV pilots sampled the household as a whole and did not attempt to identify a specific household member. The intended interviewee representative of the household was an adult women who was the mother or primary guardian of the children residing in the household (see below), but this role was occasionally filled by a father, grandmother, or other relative if the mother of the household was unable or unwilling to be interviewed.

Proxy respondents: In the ACV pilots, the predetermined protocol for interviewing household members specified that the mother or primary female guardian of the household's children should be interviewed, if possible. This was because adult females are likely to know more about the situation of children living in the household (Galdo, Dammert, and Abebaw 2019). In rural Tanzania, the mother was interviewed 69 percent of the time, and in both urban Tanzania and Nepal, the mother was interviewed 60 percent of the time. In the remaining cases, a father, grandfather, grandmother, older sibling, or aunt was interviewed. This is not considered a proxy because the sampling unit was the household, not the individual adult in the household[27].

Substitutions: The AAPOR guidelines also recommend that it is important to report any substitutions, e.g., when a sampled household cannot be found and another is included in the survey instead. In urban Tanzania, one *mtaa* required substitution from the additional households on the sample frame when the originally selected households were deemed to be ineligible because the children were the wrong ages. In rural Tanzania, the quota system used within the sub-villages to determine proportional representation is not considered substitution. In Nepal, no

[27] Each household completed at least two surveys. One that was answered by an adult in the household, described above, and at least one answered by each eligible child. Each child (age 12-17) in the household was the respondent for the survey intended specifically for children age 12-17. If an adult had answered the child survey on behalf of a child, that would have been considered a proxy response. But this was not allowed in the ACV pilots; therefore, proxy reporting is irrelevant for the adult survey and not possible for the child survey.

substitution occurred; however, several households were added to the list when they were discovered in a previously selected building. The protocol when establishing the sample list was to include all eligible households in a building. Therefore, the discovery of additional eligible households in a building was not substitution but rather the addition of a household that *should have* already been on the list.

Status Day: The ACV study did not have a set *status day* that determined eligibility of households. The final categorization of completion, contact, or unknown eligibility in all three pilots was determined by the last contact, per AAPOR recommendations (page 11).

6.3 Categorization of Contact and Completion in the ACV pilots

We used paradata collected during the sampling and data collection process to assign each household in the three ACV studies an AAPOR categorization of eligibility, contact, and completion as found in Figure 6. For some households, the appropriate category was obvious. For example, if a household completed both the adult interview and at least one eligible child interview, the household was considered to be a complete interview (I). For other categories such as non-contact (NC), we highlight several detailed examples that resulted in the NC categorization. In this section, we present examples of the four main categorizations of eligible households that were interviewed (I), eligible households that were not interviewed (R, NC, and O), households with unknown eligibility (UN), and several specific cases of ineligible households (IE).

The easiest to measure were households that had complete interviews (I). This meant that a household had a completed adult survey and a completed survey at least one eligible child interview. Households were considered to be complete once they achieved this status even if there were other eligible children who were not surveyed[28].

Households that were determined to be eligible but resulted in no survey being conducted are further categorized as refusals (R), non-contact (NC), or other (O). In the ACV pilots, a household could be considered eligible if the field team was able to confirm the ages of at least one of the children residing in the household fell in the age range. In most cases, this determination was the result of a short phone call to the household, a brief visit and conversation

[28] The AAPOR definitions also specify “partial interviews” (P). In the ACV study, respondents were allowed to skip questions, but this situation is considered a complete interview despite the missing information.

with some member of the household, or a conversation with a knowledgeable neighbor or community leader.

Following IRB protocol, the main adult participant consented to the study and gave consent for all eligible children in the household. Refusal (R) to participate could happen in one of two ways. First, an adult member of the household refused over the phone during the initial contact and the household was not visited at all. Second, an adult in the household refused only after the field workers were at the home. We did not have any situations where a respondent, adult or child, refused to continue once already starting with the survey process.

Defining non-contact (NC) in the ACV pilots was more nuanced. Sometimes, a phone call or visit was made, but the household was not available to participate in the survey. Reasons included being too busy, not being home on the specific day that the interviewers could visit, or the adults were at work and wouldn't be home within a reasonable time for the field research team to visit. It is possible that respondents may have invented excuses to avoid flatly refusing to participate due to cultural norms against direct refusals. There is a sizable literature in survey research about how different cultural norms about social expectations vary between countries (Johnson et al. 2002; Lalwani, Shavitt, and Johnson 2006). The nuance of these conversations with household members was not recorded during field work, so it is impossible to accurately categorizes these potential households as refusals (R) instead of non-contact (NC). Therefore, they are considered non-contact (NC) in the following analysis.

Other situations arose where a household was confirmed to be eligible, but the household did not participate in the interview (O for other). For example, in one case the sole eligible child in a household was experiencing mental health issues and could not knowingly assent to participate. Another situation involved a family that was in mourning and the field researcher determined that it would not be appropriate to ask them to participate. An interview that could not occur due to language barriers would also fall in this category, but this did not happen in the ACV pilots.

There were some cases, in all three pilots, where the field team completed the household survey with the adult but were unable to complete a survey with an eligible child in the household. These cases are included as being incomplete in "other reasons" (O) as this situation generally occurred when the child was too busy to be available during the field work period. We removed the completed adult survey from the data as a household needs both the adult survey and at least one survey of an eligible child in order to be considered a complete interview (I). This also applied in one situation where the child was interviewed with the consent of the adult, but the adult interview could not be scheduled within the field work period.

The next major categorization of households in the ACV pilot was when the field team was not able to determine whether the sampled household was eligible (UH for unknown household eligibility). The documentation in the paradata was not standardized across the three pilots, though similar language was used. In Table 3, we describe some of the common situations found in the paradata. These included "no attempted", "Household unsafe or unable to reach", "Unable to locate", and "Unable to make contact via phone". All of the households described in Table 3 are considered in the analysis and construction of AAPOR outcomes measures as unknown eligibility (UH).

Finally, the paradata documentation from the original sampled households also recorded specific households being eliminated from the study for being ineligible (IE). The most common reasons for excluding households from the sample was that the children in the household were the wrong age. We attempted to mitigate this situation through the pre-screening process but came across many households with 11-year-olds and 18-year-olds due to the local leader or other members of the households not knowing the child's exact age during the pre-screening process[29]. If an adult in the household was interviewed and it was later discovered that the intended child respondent was ineligible, the household was excluded from the data.

6.4 Results of ACV Response Rates, Cooperation Rate, Refusal Rate, and Contact Rate

Once all of the ACV households are categorized, we applied the AAPOR methodology and definitions to construct the four outcome measures: response rates, cooperation rates, refusal rates, and contact rates. The standard categorizations allowed comparison of the outcome measures across the three pilots. For each of the three pilots, we calculate each of the four outcome measures according to all of the definitions presented in Table 2. This allows me to compare within-pilot variations resulting from different definitions of outcome measures.

As mentioned before, we also calculate all of the outcome measures including and excluding the boarding school students; first, excluding households where the child or children in

[29] When the field team made the initial contact with the households, we had a predetermined birthdate period that determined eligibility. In one case in Nepal, we made a single exception where a household had multiple children and one of them was turning 12 on the day we conducted the interview, and we felt it was unfair to exclude her from participating when she was officially the age we had asked for and her siblings were participating.

the age range were absent from the household throughout the data collection period, thus categorizing the household as ineligible (IE) and then including households with boarding schools kids to be considered eligible but not contacted (NC) or interviewed – indicated by BS in the columns in Table 4. In the situation that considers boarding school children eligible, households with *only* boarding school children designates the boarding school children as non-contacts and therefore these households always included in the denominator of the outcome measures[30].

Table 4 reports the percentages for the three response rates (RR1, RR3, RR5), two cooperation rates (COOP1, COOP3), three refusal rates (REF1, REF2, REF3) and three contact rates (CON1, CON2, CON3). The methodology for calculating these rates is found in Table 2. The outcome measures are aggregated for the entire pilot due to small samples sizes for the finer geographic units: sub-villages in rural Tanzania, *mitaa* in urban Tanzania, and wards in Nepal.

In all of the formulas that require an estimate of how many unknown households would have been eligible (RR3, REF2, CON2), we calculate *e* as the inverse probability of being ineligible among the households in the original sample frame. In rural Tanzania, 25 percent of the households on the original sample list of households established with sub-village leaders were deemed to be ineligible upon contact. Using this information, we assume that 75 percent of the households with unknown eligibility due to non-contact would have been eligible (*e*=0.75). In urban Tanzania we estimate that 98 percent (*e*=0.98) of households would have been eligible and in Nepal 80 percent (*e*=0.8) of households would have been eligible. The higher estimates of eligibility in urban Tanzanian and Nepal are partially due to improved sampling strategies and pre-screening process employed in these pilots.

[Table 4]

Across all pilots, response rate 1 (RR1) is lower than response rate 3 (RR3) and response rate 5 (RR5). RR1 assumes all households with unknown eligibility (UH) would be eligible and thus are included in the denominator [Table 6.1.2]. Across all three pilots, response rates (regardless of the exact definition used) were between 64 percent (RR1 in rural Tanzania when including households with boarding school students) and 91 percent (RR5 in rural Tanzania excluding households with boarding school students). The inclusion of households with boarding school students in the denominator reduces the response rates across all definitions and across all pilots.

[30] If a household had at least one child at home who was able to participate in the survey, even if other children were at boarding school, the survey was considered complete (I).

Rural Tanzania has the largest range of possible response rates largely driven by the status of households with boarding school students and a large proportion of households that were unknown eligibility (UH). The ranges of response rates in urban Tanzania and Nepal are narrower than in rural Tanzania; urban Tanzania response rates ranged between 73 percent (RR1 with boarding school students) and 84 percent (RR5 without boarding school students) and in Nepal the range was 76 percent to 86 percent for the same minimum and maximum definitions.

The APPOR *Standard Definitions* report (2016) does not make specific recommendations about which response rates to report in published works so long as the authors are explicit about which response rate is reported. The purpose of reporting a response rate is to communicate the potential for nonresponse error in the survey. Overall, the response rates in the ACV pilots are high and fairly constant across all definitions (RR1, RR3, RR5). If required to select only one definition to report, we would recommend using the RR3 definitions as the results fall in the middle of the extremes of RR1 and RR5. RR3 utilizes the paradata to estimate of the proportion of unknown households that *would have been* eligible. This data driven approach best captures the nuances the sampling and data collection process and the social and logistical challenges faced by the field team to identify households.

Deciding whether to report the response rates that exclude or include households with boarding school students (BS) would depend on if there could be nonresponse error correlated specifically to households that sent children to boarding school and households that did not. For example, an analysis of family wealth or educational attainment may be sensitive to nonresponse error of the households that were not included in the survey because all eligible children were away at school. But for most analysis of the ACV pilots, the high and generally consistent response rates suggest that nonresponse error may be minimal.

To support the results presented in the response rates, we also report the cooperation rates (COOP1 and COOP3) and refusal rates (REF1, REF2, and REF3) in Table 4. The *cooperation rates* report the number of interviews over the number of households contacted and the *refusal rates* report the number of refusals over the number of households [Table 2]. Across the ACV pilots, the cooperation rates were very high; households that were contacted were very likely to participate in the study.

Rural Tanzania had near universal cooperation and no refusals. The refusal rates in urban Tanzania were also very low; regardless of the definition used (REF1, REF2, or REF3) the refusal rates in urban Tanzania were 4 percent. It is possible that the high cooperation rates and low refusal rates in Tanzania resulted from the team being accompanied by a local leader who conducted the introduction between the field team and the household. The partnership with a local leader may have increased the legitimacy of the field team, so households were more willing to participate in the survey (Groves, Cialdini, and Couper 1992). Alternatively, households may have felt more social pressure to participate because of the presence of the local leader. In the debriefing notes recorded daily, such social pressure was not reported by the field team; the field team followed the informed consent protocol that assured adult household members that participation was voluntary.

In Nepal, the cooperation rates and refusal rates were not at all affected by the inclusion or exclusion of households with boarding school students. The cooperation rates (particularly COOP1) was lower than in the Tanzanian pilots—88 percent compared to 94 percent (urban) and 99 percent (rural). In Nepal, the sampling team and data collection team conducted work without the involvement of local government officials. The entire team had name tags clearly identifying them as being part of a local organization, but the higher refusal rates and lower cooperation rates reflect the challenges and extra effort the team had to make to explain the project and engage participants in informed consent interactions compared to the Tanzania pilots.

In the ACV pilots, the cooperation rates and refusal rates calculated are functional inverses of each other. They each convey a similar message that household participants overwhelmingly cooperated in the ACV study if the field team was able to contact the household. High cooperation and low refusal rates suggest that nonresponse error caused by a potential difference between households that refused to participate and households that did participate is small. As with the response rates, AAPOR recommends that researchers communicate to readers which definition of the cooperation rate and refusal rate was used if such a rate is reported in a published article. For the ACV pilot, the preferred measure reported would be REF2 as it again uses the paradata driven approach to appropriately account for households with unknown eligibility. The preferred cooperation rate would be COOP1 as it accounts for households that did not complete an interview for reasons other than refusals (O); this was a very rare occurrence in the ACV pilots as described in section 6.3.

The final outcome measure recommended by AAPOR is the contact rate. The *contact rates* convey the success of the field team in contacting and determining eligibility of sampled households, as represented in the flow chart in Figure 6 (I + R+O). In Table 4, the contact rates (CON1, CON2, CON3) for all three pilots report similar percentages to the response rates. The contact rates indicate that between 77 percent and 96 percent of eligible households were contacted. In the ACV pilots, the contact rates reflect the hard work of the field team to find sampled households from the sample lists. In Tanzania, the help of local leaders was an essential element in identifying eligible households and making introductions. In Nepal, the contact rates were slightly higher – 90 percent (CON1 without boarding school students) to 96 percent (CON3 without boarding school students) – compared to Tanzania pilots. The Tanzanian rural pilot had contact rates of 77 percent (CON1) to 90 percent (CON3); and they were 82 percent (CON1) to 89 percent (CON3) in the city. The extensive sampling and pre-screening process in Nepal functioned as the first contact with household; therefore, when the field team called or visited the household during data collection, the sampled household had already talked to a member of the field team. In many households, the adult respondents remembered the interaction with the sampling team and had been waiting for a member of the field team to call to set up a time to conduct the survey. These high contact rates in Nepal suggest that pre-contact with households during the sampling process could increase contact rates during data collection[31].

The inclusion of households with boarding school students (BS) influenced the contact rates differently than the other outcome measures. For each of the pilots, CON1 and CON2 were higher when boarding school students were included (columns with BS) compared to columns that excluded the households with boarding school students. This is different from all three definitions of response rates; the inclusion of boarding school students decreased the rates for RR1, RR3, and RR5. The mathematical mechanisms at work in the formula balance the number of households with unknown eligibility and the non-contact households (NC) in the denominator. In Table 2, the formula for CON3 does not include these unknown households as it considers that none of the unknown households would have been eligible if they had been contacted (similar to RR5 and REF3). The inclusion of boarding school students decreases the CON3 rates in all three pilots.

Why does this even matter? The inclusion or exclusion of households with boarding school students, even in small samples like the ACV pilots, can change the outcome measures to look more or less favorable. This variability could be utilized to manipulate the results of a small study to report highly favorable results in order to increase the possibility of getting published. This manipulation, while technically accurate, masks the complexity of the fieldwork and decisions of eligibility criteria, recruitment, and sampling. When the study is small, which

[31] The sampling team was trained to use materials that reflected the IRB process of recruitment and informed consent. The informed consent process started with the first contact with the household and was continued during the data collection process.

outcome measures are reported may not affect the survey results or increase potential error (coverage, sampling, or nonresponse error as established by the TSE framework). In a large study or a study where there is a large sub-population of participants that could be defined in multiple ways, such decisions could be highly influential in the results. Therefore, it is important for authors and researchers to clearly state how these sub-populations are or are not being included in the calculation of outcome measures, including response rates and contact rates. The importance of considering the sub-population is necessary if there is a potential for nonresponse error affecting specific variables of interest.

The primary purpose of reporting outcome measures is to quantify the sampling data collection process in a way that indicates that the results of the survey are not biased by nonresponse error. It is unlikely that all four of these types of rates would be included in a published paper as they are all intricately related. This application of the APPOR *Standard Definitions* to the ACV project demonstrates how different outcome measures and different definitions of outcome measures can be constructed through varying interpretations of the end result of household visits and interviews in paradata and documentation of the data collection process. Reporting outcome measures such as response rates without properly defining the method of calculating the rate is misleading. Researchers should be transparent and specific when reporting outcome measures and highlight any potential nonresponse error due to refusals, noncontact, or unknown eligibility households.

Journals and reviewers should also reconsider rejecting manuscripts based solely on small response rates or high refusal rates. Small sample sizes in the ACV pilots result in variability of each outcome measures. A single household refusal may increase the refusal rates substantially in a small-scale survey without contributing nonresponse error to the overall study. Nonresponse error is an issue only if there is a correlation between households that do not respond and the variables of interest. Low response rates do not *prove* there is nonresponse error just as high response rates do not suggest a perfectly bias-free survey. The best practice for any survey is to report clearly which outcome measures were calculated and provide other analyses of nonresponse error, such as constructing weights and population estimates, described in the next two sections.

7. Constructing Weights from ACV Paradata

Weighting in an important part of any probabilistic sample as it allows users to adjust the results of outcomes measured by the survey to represent the target population by accommodating sample design and nonresponse (Kalton and Flores-Cervantes 2003; Yansaneh 2003; Solon, Haider, and Wooldridge 2015). Sample statisticians calculate weights in order to adjust the results of sample statistics so that they more accurately reflect population parameters. The construction of weights for the ACV pilots allows for a detailed look at the usefulness of weighting data.

Samples weights have two main functions. First, weights adjust the sample to reflect the descriptive size and composition of the target population. This can be useful for researchers working with raw frequencies. Second, weights can be applied to analysis to adjust the specific sample statistics to reflect the size of the underlying target population (Gelman 2007). It is only necessary to apply sample weights to an analysis if there is a concern that coverage, sampling, and nonresponse error may impact the results of the particular variable of interest (Makela, Si, and Gelman 2014). In this section, we describe the process of creating weights for each of the three pilots in the ACV study. We demonstrate the usefulness (or not) of applying sample weights to adjust the sample to represent the descriptive target population size. Finally, we conduct an analysis showing the application of sample weights for a variable of interest in the ACV pilots that is potentially related to nonresponse error. We conclude this section by identifying potential reasons that researchers should or should not utilize survey weights.

Typically, there are three different elements to sample weights. A *base weight* (also called a sample weight or a design weight) takes into account the sampling process and allows the sample to be scaled to the size of the target population. For example, a random sample may include approximately 10 percent of the total households in a population and the base weights can be applied so that each household represents 10 other (unsampled) households when researchers present descriptive statistics. A *non-response weight* adjusts for the nonresponse bias possible in the sample. For example, households within a geographic area that have responses can represent households in that same area that didn’t respond if the researcher assumes that all households in the area share homogenous characteristics, and therefore would have answered similarly. Finally, a *post-stratification weight* allows researchers to re-calibrate the sample to look more similar to the established target population. For example, if immigrant households are underrepresented in a sample, these weights would “scale up” or overrepresent the sampled immigrant households to accurately represent their proportion of the target population. In this exercise, we have created base weights and nonresponse rates. The product of the base weight and the nonresponse rate equal the *final weight*. We do not create post-stratification weights because we do not have accurate enough distributions of the target population of households with children 12 to 17 to make any adjustments. Additionally, we did not make any sample design choices that resulted in purposeful over- or under- sampling of a particular group[32].

The base weight (W_B) is calculated as the inverse probability of being selected (*p*):

$$W_B = \frac{1}{p}$$

We calculate a different base weight for each primary sampling unit (PSU) in each of the three samples: sub-village, *mtaa*, and ward. The process of calculating the probability of an eligible household being selected depends on the sampling process that happened within each pilot.

To calculate the nonresponse weight, we use the response rates calculated in the previous section. In order to explore the construction of weights fully, we calculate a maximum of six different nonresponse rates for each PSU in each sample based on the three different response rates (RR1, RR3, and RR5) and the inclusion or exclusion of the boarding school children for a total of six nonresponse rates. Overall, the nonresponse weight (W_{NR}) is the inverse probability of the specific response rate (*RR*)

$$W_{NR} = \frac{1}{RR}$$

Nonresponse weights for each of the different definitions of response rate are written as:

- W_{NR1} to refer to the nonresponse rate that uses the RR1 response rate
    - W_{NR1BS} to refer to the nonresponse rate that uses the RR1 response rate that includes boarding school students
- W_{NR3} to refer to the nonresponse rate the uses that RR3 response rate
    - W_{N31BS} to refer to the nonresponse rate that uses the RR3 response rate that includes boarding school students
- W_{NR5} to refer to the nonresponse rate that uses the RR5 response rate

[32] The missing boarding school student households is not something that can be fixed with post-stratification weights as we do not have population figures that show the number of children in boarding schools within our geographic areas. Thus, the boarding school students will be adjusted for in the non-response weights following the logic presented in the construction of non-response rates.

- W_{NR5BS} to refer to the nonresponse rate that uses the RR5 response rate that includes boarding school students

In all cases, the final weight (W) is the product of the base weight and the nonresponse weight:

$$W = W_B * W_{NR}$$

For each of the pilots, the combination of base weights (W_B) and six variations of the nonresponse weights (W_{NR}) defined above produces six different possible weights for each of the PSUs for each of the pilots. In order to make a recommendation of which particular weights to use, we present figures that show the distribution of the magnitude of the weights. Each weight represents the inflation factor to be used when conducting analysis about a particular parameter of interest. Thus, the weights can be interpreted as the number of eligible households in the PSU that are represented by each individual respondent household present in the sample.

7.1 Rural Tanzania

The sample in the rural Tanzanian village was almost a complete census of eligible households in some sub-villages. An accurate census with full participation would not require weights. As we almost achieved this, the weights for each household in each of the sub-villages are not very large. For each sub-village, we construct the base weight from the number of households in the sampling list divided by the known or estimated number of households with 12-17-year-old children in the sub-village.

In three of the seven sub-villages, we did not have an accurate count of the target population households due to social and logistical challenges described in the sampling process. In order to construct weights, we estimated the total number of eligible households based on the proportion of households with 12-17-year-olds out of the total households in the remaining four districts where both figures were known; on average 29 percent of households in a sub-village included at least one 12 to 17 year old child. we established the denominator of target households, actual or estimated when missing, in order to calculate the probability of an eligible household being sampled within each sub-village. Thus, the base weight is the inverse of p, where

$$W_B = \frac{1}{p} \; where \; p = \frac{Sampled\ HH}{Total\ eligible\ HH\ in\ PSU}$$

These base weights (W_B) were combined with nonresponse weights (W_{NR}) calculated using the three different types of APPOR response rates described above ($W_{NR1,}$ $W_{NR3,}$ W_{NR5}). In addition, we calculated nonresponse weights at each of the three AAPOR rates ($W_{NR1BS,}$ $W_{NR3BS,}$ W_{NR5BS}) to account for the inclusion and exclusion of boarding school children for a total of six possible candidates for final weights (*W*). Figure 7 represents the spread of the six iterations of the final weights for each of the sub-villages. The orange dots represent the median final weight of the twelve variations calculated while the blue represents the minimum and the grey represents the maximum possible final weight. The range of possible final weights shows the extent to which the base weight and nonresponse weights vary because of differences in the original sampling frame and the response rates in each sub-village. Through these weights, we are able to generalize our sample findings to the target population of households with 12 to 17-year-old children in the selected village in Tanzania. We cannot generalize to other areas of Tanzania.

In the Tanzanian village pilot, the sample of eligible households was close to a census of the eligible households in each sub-village. Thus, the weights attached to each respondent household are barely larger than one [Figure 7]. In sub-village 1, each household has a weight of 1.18 (the overall lowest median) and the highest overall median (1.78) is in sub-village 7. Most of the sub-villages, particularly sub-villages 1, 2, and 5, have a narrow range. The sub-villages with large spreads (6 and 7) were also the sub-villages with higher than average proportions of eligible households in the sub-village (that is, more households with children ages 12-17) and higher than average numbers of households that were unknown eligibility due to non-contact. The important lesson here is that when facing a small eligible population size, small deviations from the mean can result in great divergences in weights and non-response rates.

For this pilot, using the weights provides only marginal added value to the overall results from data collected from the sampling frame. But the overall adjustment to the sample in order to reflect the target population is fairly minimal and thus would not add great complexity or additional concern for error in the descriptive statistics. As we will see below, applying the weights increases the standard error in variables of interest compared to not using any weights at all.

[Figure 7]

7.2 Urban Tanzania

Weights constructed for urban Tanzania followed a process similar to the construction of weights for rural Tanzania as each stage of the multi-stage sampling process recorded the known probability of selection of the particular unit. The primary sampling unit in urban Tanzania is the *mtaa* and every *mtaa* had an equal probability of being selected. In total, 23 *mitaa* were sampled but two were later excluded for being too rural and one was used for training the field team.

According to our sampling process, cells are a secondary sampling unit. One cell was randomly selected from a complete list of cells for each *mtaa*. Remember, a cell theoretically represented ten households that were all known by a single local leader known as the cell leader. In practice, the cells in urban Arusha were larger and many had between 20 to 50 families. Finally, a complete list of eligible households in the cell was sampled from to create the sample list. The base weight is one over the product of the probability of the *mtaa* (P_M), cell (P_C), and household (P_H) being sampled in each *mtaa*.

$$W_B = \frac{1}{P_M * \ P_C * \ P_H}$$

Where

$$P_M = \ \frac{Randomly\ sampled\ Mtaa}{All\ Mitaa\ in\ Arusha\ Urban} = \frac{20}{122} = 0.164$$

$$P_C = \ \frac{Randomly\ sampled\ cell}{All\ cells\ in\ particular\ mtaa} \quad e.g. = \frac{1}{7}$$

$$P_H = \ \frac{Sampled\ HHs}{Total\ eligible\ HHs\ in\ Cell} \quad e.g. = \ \frac{10}{40}$$

In most cases, we knew the number of households only from the sampled cell. Thus, the calculation of weights assumes that other cells in the *mtaa* would have similar numbers of households in the other listed cells that were not sampled. This is a major assumption as it was not always clear if the lists created with the help of *mtaa* and cell leaders were complete. This potential issue will be addressed further in the discussion on population estimates below where we compare the assumptions of population size and distribution with the established population reports.

The base weights (W_B) were combined with the different six nonresponse weights (W_{NR}) calculated in exactly the same fashion as the rural Tanzania pilot to create the final weight (W). There are six total final weights candidates for urban Tanzania. Figure 8 shows the distribution of the minimum, median, and maximum values of the final weights for each of the *mitaa*. These weights allow us to generalize our sampling findings to all households with at least one 12-17-year-olds in Arusha. In comparison to the weights for the rural Tanzania pilot, the weights for urban Tanzania are larger; the average median final weight is 85.5 (min = 2.8; max= 234) and the spread of all of the final weights ranges from 2.6 to 341.6. A larger weight indicates that each sampled household represents a greater number of total households.

A significant factor in the size of the final weights is the small proportion of *mitaa* included in the sample out of the total *mitaa* in Arusha; this increased the size of the base weight for all *mitaa* which in turn increased the size of the final weight. The *mitaa* across Arusha vary greatly in population size. For example, *mtaa* 11 and *mtaa* 16 were both small neighborhoods of ethnic minority families that were very different from other areas of Arusha. In each of these two *mitaa*, there was only one cell, and the ACV sampling resulted in close to a census of the eligible households with 12-17-year-olds. Thus, respondent households in these *mitaa* have small final weights (median final weights at 6 and 2.6 respectively) because the *mitaa* had a small number of cells and an overall small population size compared to other *mitaa*.

In contrast, some *mitaa* have larger numbers of cells and large cell sizes (i.e. many households per cell). For example, *mtaa* 21 reported 14 cells and within the single cell that was randomly sampled, there were 29 eligible households. *Mtaa* 2 reported 11 cells, with 40 eligible households in the sampled cell. These two *mitaa* both have high median final weights (W = 164 and W = 122 respectively) as each household included in the study is weighted to represent large populations in these *mitaa*.

The calculation of the base weights remained stable for each of the *mitaa* as the estimates of base weights are produced in the sampling process, not the data collection process and identification of AAPOR categorizations. The differences in the range of possible final weights is driven by differences in the nonresponse weights and in the inclusion or exclusion of boarding school children in the calculations. *Mitaa* with consistent measures of response rates across the different definitions (RR1, RR3, and RR5) produce final weights that are identical across all twelve calculated final weights; for example, *mitaa* 4, 8, 9, 15, 16, 17, and 21 have no spread of final weights as shown by the minimum, median, and maximum values overlapping in Figure 8. The calculation of the base weight affects the size of the final weight – i.e. the final weight in *mtaa* 4 is 14 and the final weight in *mtaa* 8 is 116 —but the lack of spread indicates stability across the six nonresponse weights calculated. For these *mitaa*, the selection of final weights to use in analysis does not matter as they are all the same. In contrast, *mtaa* 23 has a wide range of final weights. This *mtaa* was highly influenced by several boarding school children and unknown eligible households (described in the above section as being households were the field team was unable to establish the eligibility status of the household). With these factors influencing the response rate calculation, the range of final weights is wide (minimum = 136.6; maximum=341.6). In mitaa with wide ranges, the selection of final weights used in the analysis depends greatly on the definition of response rate used.

Given the assumption made about the calculation of the base weights and the wide variation in nonresponse weights, it is difficult to know which final weights should be used in analysis of the urban Tanzanian households. Theoretically, the small sample (n=145 households) of the pilot could be weighted to reflect the population of over 100,000 households of Arusha[33]. But it is important to clearly define the calculation of weights and assumptions in order to assess if such weighting is valid and appropriate. The large magnitude of the weights and many assumptions required in the construction of the weights suggest it would be unwise to present descriptive statistics that adjust the sample size to the size of the target population. More on this will be discussed in the section on population estimates. Finally, just as in rural Tanzania, the use of weights should be carefully considered based on specific parameters of interest that might be affected by nonresponse error, as demonstrated below.

[Figure 8]

7.3 Nepal

Due to the nature of sampling in Nepal, the process for calculating a base weight was vastly different than the previous pilots. The fundamental challenge in Nepal was that there was no way to accurately create a sample frame based on the target population. The sampling of households via satellite imaging of OpenStreetMaps inevitably included eligible and ineligible households. In the sampling process and response rate calculation, these ineligible households are ignored completely. Ineligible households would not be included in the sample at all and therefore are not considered respondents or potential respondents. However, when constructing weights, it is important to treat the eligible households as inherently different from ineligible

[33] Number of households (n=104,093) provided by Arusha District office.

households. We must consider the proportion of buildings identified in OpenStreetMap that *would have contained* at least one eligible household *if* they had been sampled. Of course, not all buildings would have an eligible household, given that the target population includes only households with 12-to-17-year old residents and not every building will include households meeting this requirement.

In order to demonstrate this difference, we have constructed base weights (W_B) three different ways. All base weights start with the same probability of a primary sampling unit (PSU) – in Nepal this is the ward – being included in the sample (P_W). These are stratified by municipality. We aim to be able to generalize to all households with a 12-17-year-old in two specific municipalities in the Kathmandu District of Nepal.

The next step of the sampling process involved the identification of buildings based on the satellite imaging. Each building has a known, non-zero probability of being selected. Remember that the sampling was conducted where all buildings in a ward were identified and randomly selected. The samplers visited the selected buildings and determined the eligibility of households. If there were not eligible households in the identified building (the starting point), the samplers continued to an adjacent building.

A preliminary base weight (base weight 1 or W_{B1}) establishes a weight of the number of buildings identified as a starting point divided by the total number of buildings in the ward. This maintains the probability sample of the starting buildings as being randomly selected. Of the eligible households (with 12-17-year-olds) in the selected building, they all have a 100 percent probability of being included in the sample as per the instructions to the sampler. The probability of a building being selected as a starting point over the total buildings possible to be selected as a starting point (P_{B1}) is the second part of the base weight 1 equation.

$$W_{B1} = \frac{1}{P_W * \ P_{B1}}$$

Where

$$P_W = \frac{Randomly\ sampled\ wards}{Total\ wards\ in\ municipality}$$

$$P_{B1} = \frac{Buildings\ selected\ as\ a\ starting\ point}{Total\ buildings\ in\ PSU}$$

However, not every building has a non-zero probability of being included in our sample because not every building was home to a household with a 12-17-year-old resident. The base weights constructed above highly inflate the weight of each respondent within a ward. On average, the sampler visited five buildings for every one randomly-selected starting building. This was done through a random walk. An additional complication is that the process of doing a random walk from a probabilistically selected starting point is no longer a probability sample (Bauer 2016).

In order to address the proportion of buildings that would be randomly selected as a starting point but would not yield an eligible household, we calculate two additional variations on the base weights. First, we account for all of the buildings visited by the sampler over the total number of buildings in the ward (base weight 2 or W_{B2}). This effectively inflates the numerator by five without changing the denominator[34].

$$W_{B2} = \frac{1}{P_W * P_{B2}}$$

Where

$$P_{B2} = \frac{Sampled\ building + All\ additional\ buildings\ visited\ by\ sampler}{Total\ buildings\ in\ PSU}$$

Alternatively, we estimate the proportion of buildings that *would have* an eligible household based on the known proportion of buildings with an eligible household visited by the sampler (base weight 3 or W_{B3}). This strategy reduces the denominator and numerator to only include the probability of a randomly identified starting building with an eligible household.

$$W_{B3} = \frac{1}{P_W * P_{B3}}$$

Where

$$P_{B3} = \frac{Sampled\ HHs}{e(Total\ buildings\ in\ PSU)}$$

[34] The samplers visited approximately 5 buildings per originally sampled building in order to find the needed number of eligible households. This is reflected in this calculation of the base weight 2.

where e is an estimated proportion of eligible households per building in the PSU based on a calculation of the number of eligible households identified divided by the total households identified in the all visited buildings in the PSU. We then create an estimated average of eligible households per building across the entire ward based on the observed data.

Both of these methods (W_{B2} and W_{B3}) attempt to consider only the target population when creating base weights and as an added confirmation, the results from each of these corrections are very similar (and very dissimilar from the base weights constructed with no regard for eligibility). We calculate all iterations of the product of the base weights 2 and 3 with the nonresponse weights calculated using the same methods as in the other two samples, but only report base weight 3 here due to the high similarity with base weight 2.

The Nepal final weights (W) were calculated using the same process as in Tanzania to combine the base weight and the nonresponse weights (W_{NR}) calculated using the different AAPOR response rates ($W_{NR1,}$ $W_{NR3,}$ W_{NR5} and $W_{NR1BS,}$ $W_{NR3BS,}$ W_{NR5BS}). Similar to the Tanzanian pilots, this combination yields six different potential final weights that explore the three response rates with and without boarding school students. In Figure 9, we compare the six final weights produced using both base weight 1 [Figure 9A] and base weight 3 [Figure 9B]. The main difference between the two figures is the final weights produced using base weight 1 (W_{B1}) are significantly higher than the final weights produced using base weight 3. If we did not account for the proportion of eligible households per building in the sampling process, we would be applying weights that inaccurately describe the population of households with 12-to-17-year old residents. The utilization of paradata about the sampling process greatly improves the calculation of sample weights. Therefore, we recommend using final weights calculated by base weight 3 (W_{B3}) for the Nepal pilot analysis.

In Figure 9B, the final weights calculated for Nepal are similar in magnitude to the Tanzanian urban weights. This provides support for the validity of the weight calculations as the geographic area and starting population sizes of the two pilot sites are similar. The average median final weight across the wards is 67.8 (min = 30.9; max = 140). Like in urban Tanzania, the magnitude of the final weight is predominately driven by the base weight, in this case calculated by base weight 3. Similar to our recommendations for the urban Tanzania pilot, the large magnitude of the sample weights, instability of the results based on definitions, and assumptions made in calculating the weights, we would not recommend using weights to adjust the sample to reflect descriptive statistics of the target population in the Nepal pilot. This will be discussed in more detail in the section on population estimates.

The spread of the minimum, median, and maximum final weights is driven by the nonresponse weights (*W_{NR}*) in each ward. The Nepal final weights benefit from a lack of variation in the response rates because of the pre-sampling contact process where a team member confirmed eligibility in advance. This sampling process, as described above, led to high response rates and high contact rates. In two of the wards, there is no difference between the multiple response rates within the wards in Nepal resulting in fewer than 6 unique final weights. The wards with wider variation in final weights – i.e. T6, T7, and N3 – were wards that are located closer to Kathmandu City center. These wards had more varied nonresponse weights as these wards also had lower response rates due to refusals and noncontact of households as described in the above section. We explore the relationships between response rates and specific variables of interest in the next section.

[Figure 9 A & B]

7.4 Application of Weights to Social Desirability Index

As demonstrated above, the calculation of final sample weights depends on the proper specification of base weights and nonresponse rates[35]. The base weight adjusts for the inflation of the sample size to the target population. Nonresponse weights adjust for the potential that survey respondents and non-respondent may have answered a specific question or set of questions differently. Transparency in this calculation is another way that researchers can provide evidence that their results are valid. The base weights produced in the two Tanzanian pilots used similar methodology of calculation. In Nepal, we recommend using the base weights 3 (W_{B3}) produced using the estimated eligible buildings as a proportion of total buildings. For all three pilots, we recommend using the nonresponse weight based on response rate 1 (W_{NR1}). Response rate 1 (RR1), as described in above, is considered the "minimum response rate" and is the most conservative estimation of response rate among households of unknown eligibility. Therefore, the nonresponse weights constructed using this response rate will provide a conservative weight. The product of the base weight and the nonresponse rate equals the final weight reported in the above section, though each of the weights could be applied to a sample independently.

However, weighting data is only meaningful if the desired outcome is to inflate the respondents to reflect the population figures or in the context of parameters of interest. In the first situation, the base weight will correct for the sample size to reflect the underlying target

[35] And post-stratification weights if this applies to a particular survey. It did not in the ACV pilots.

population. Given the ACV pilots' small sample sizes, it is unlikely that we would apply the base weights given the large magnitude of the weights, particularly in the urban Tanzanian and Nepal pilots. The magnitude of the base weight (and as a result, the final weight) leads to each household in the study representing an average of 85 households in the urban Tanzania pilot and 65 households in the Nepal pilot while each of these pilots only surveyed 145 and 155 households respectively. With all of the social and logistical challenges of the sampling processes, the coverage and sampling error threatens the validity of the sample weights representing such large populations. We conducted only one round of data collection in each of the populations. It is also not possible to test the validity our single sample in the overall sampling distribution.

Instead, we can turn to the second situation of applying the nonresponse weights to specific parameters of interest. The nonresponse weights are necessary if a question of interest may have been answered differently by respondents and non-respondents; this is often referred to as *unit non-response bias*. We test for nonresponse bias in the ACV data by applying different nonresponse weighting schemes and comparing the estimated means and standard deviations for unweighted and weighted estimates on a composite variable, described below. Specifically, we are testing the three nonresponse weights: $W_{NR1,}$ $W_{NR3,}$ W_{NR5}. This methodology, adapted from Blom (2009), purposefully selects variables that may be correlated to nonresponse. If the nonresponse weights are calculated accurately, then the application of weights to the survey can add value to the analysis of the specific parameter[36].

In the ACV pilots, we asked the adult respondent a series of questions to construct a social desirability index. The social desirability index used in the ACV pilot was constructed from the Marlowe-Crowne Index (Crowne and Marlowe 1960). Scales presented positive and negative personality traits, and social desirability was defined as existing when the respondent claimed socially desirable personality traits and denied socially undesirable personality traits to him- or herself (Edwards 1960 as cited in Helmes and Holden 2003). Such scales have been translated and adapted for use in other countries (Verardi et al. 2010; Vu et al. 2011). The ACV pilots use a variation of the Marlowe-Crown Index to measure tendencies toward social desirability of the adult respondent in the household. Thirteen individual questions were translated into Swahili and Nepali. An index of responses had a range of 0, reporting all socially undesirable answers, to 13, answering every question in a socially desirable way. Of the variables

[36] Blom (2009) applies this method to the European Social Survey which details nonresponse and post-stratification weighting schemes in multiple European countries. Our analysis applies this method but lacks the sample size and underlying data to construct comparable analyses.

present in the adult survey in the ACV project, the social desirability index (SDI) score is likely to be associated with nonresponse. Agreeing to participate in t survey is a socially desirable action in itself (Harling et al. 2017; Johnson and van de Vijver 2003; Gosen 2014). Thus, respondents may have higher SDI scores than non-respondents, for whom we do have any data. We calculate the mean score of the SDI for each pilot without weights and then apply each of the different nonresponse weights to see if the mean score varies due to the application of weights[37].

[Figure 10]

Figure 10 shows the mean score on the SDI plus and minus the standard deviation for respondents in the three pilots (the red bar represents rural Tanzania, the blue bar represents Nepal, and the black bar represents urban Tanzania). The first column shows the raw, unweighted scores. In all three pilots, the average scores ranged between 8 and 9 with respondents in rural Tanzanian scoring the lowest (or least socially desirable) and urban Tanzania scoring highest (most socially desirable) on average.

The following three columns show the SDI scores when applying the three versions of the nonresponse weights. We conducted several sensitivity analyses using the three nonresponse weights that include boarding school children ($W_{NR1BS,}$ $W_{NR3BS,}$ W_{NR5BS}). Additionally, we tested the six final weights (*W*) – nonresponse weights (with and without boarding school children) combined with the base weights – presented in the section above (with the base weight three being using in the Nepal pilot as the preferred base weight). The results for these nine weights are nearly identical to the results from the nonresponse weights (without boarding school children) and are not included in the figure.

There are virtually no differences between the SDI scores when applying the three different nonresponse weights calculated from the three different response rates. There is no statistically significant difference in mean SDI score when using any of the weights compared to not using weights. In rural Tanzania, the standard deviation of weighted SDI scores is larger than the unweighted scores. This suggests that using nonresponse weights to look at SDI scores does not benefit the user of the ACV data. In some cases, weights might even increase the standard deviation.

[37] We apply only the nonresponse weights to correct potential nonresponse bias in this analysis as the base weights would be the same for all households in the same primary sampling unit (PSU).

8. Population Estimation from ACV Paradata

A third way to quantify potential nonresponse error, a potential source of error in the Total Survey Framework found at the transition between the sample and the respondents, is to construct population estimates from the paradata documentation of the sampling and data collection process. The sample sizes, response rates, and overall data about eligible and ineligible households in the paradata can be used to construct population estimates that can be compared to population figures found in other sources of data. The idea of estimating the population from the sample survey data presents the ultimate paradox. The sample is created as a representation of the target population because the sample was created from the target population. Therefore, how can the sample tell us anything about the population if the information about the population was unknown at the start of the sampling process? We argue that despite the ACV pilot samples being created from population data, after field work is complete, we have more information about the target population than we did prior to sampling.

The target population of the ACV pilots was households with 12-to-17-year-old residents. The sampling strategies employed in the three pilots sought to identify this subset of households from the broader populations. There are no accurate population level data sources that could provide the foundation for a sampling frame for such a specific type of household. Population figures for the rural Tanzanian village, the Arusha urban municipality, and the two municipalities in Kathmandu were only available as aggregate numbers of individuals or households. The process of sampling with the assistance of the local leaders in Tanzania and satellite imaging in Nepal was a way to create a probabilistic sample in the absence of information about the target population.

One way to assess whether the sampling and data collection represent a probabilistic sample of the target population is to compare the ACV sample paradata and calculated outcome measures and sample weights to established population statistics. In this section, we discuss the feasibility of this comparison and, when possible, we construct population estimates based on the sample data. The population estimates are constructed from paradata about the sampling process that estimated the total size of the sub-village, *mtaa*, or ward based on information from the local leaders or the process of geospatial identification of buildings. We also utilize paradata from the data collection process including identifying eligible and ineligible household and estimates on household sizes[38]. We compare our estimates to the 2012 Tanzania Population Census and the 2011 Nepal Population Census from the IPUMS-International census microdata[39]. This analysis is aimed at validating the paradata produced and collected about the sampling and data collection process by the local government officials and hired sampling teams.

8.1 Rural Tanzania

The pilot village in rural Tanzania was small. Posted on the wall of the village office was a table that stated there were 3,800 individuals living in 1,259 households in the village in 2015. We calculated the proportional stratified sampling methodology on the proportion of households in each sub-village from the 2015 sample. Based on the paradata from the household lists provided by the sub-village leaders, we estimate that the village had a total population of just over 3,300 individuals in 878 households during the field work period in July 2018. The estimate of the population for the village comes from the number of households that were counted on the household lists provided by the sub-village leaders. In all seven sub-villages, we were able to access the entire sub-village list and the total number of households and individuals was verified by two members of the field team[40]. Despite the three-year difference, the similarity in these numbers is reassuring.

In addition to checking the population estimates based on the posted population figures, we compare the proportion of eligible households in the village to estimates of the proportion of eligible households produced by the 2012 Tanzanian census. The smallest geographic unit in the IPUMSI microdata in Tanzania is the district. The village in the pilot is located in Monduli District (total 2012 population 158,929). It is not possible to identify individuals or households from the ACV pilot village in the 2012 census, but we compare the estimates for all of Monduli to the pilot village.

[38] We do not apply the base weights constructed in section 7 in these population estimates as the base weights only represent the target populations of households with 12-to-17-year-old resident and it would be inaccurate to apply these weights to the ineligible populations.

[39] Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [dataset]. Minneapolis, MN: IPUMS, 2019. https://doi.org/10.18128/D020.V7.2 and the Census Bureau of Statistics of Nepal and National Bureau of Statistics of Tanzania.

[40] In one sub-village, the two counts varied by approximately 20 households and we were not able to revisit the list to count a third time. In the analysis, the average of the two counts was used.

Using the census data, we estimate that 41 percent of households in Monduli District would have been eligible for the ACV pilot[41]. Based on the population counts in the village, we estimate that 29 percent of the total households on the sub-village lists would have been eligible for the sample. One reason for the 12 percent difference between the district as a whole and the village could be changes in the population between 2012 and 2018, such as changes in fertility and family size or migration of families out of the district. Second, it is possible that other areas in Moduli are very different from the specific village in the ACV pilot.

A final, and most likely, reason for the difference comes from the ACV population estimates. In three of the seven sub-villages, we did not record the total number of eligible households. In these three sub-villages, we estimated the number of eligible households using the proportion of eligible households from the total number of households of the four known sub-villages. The average proportion among the four known sub-villages is 29 percent (minimum = 23 percent; maximum = 39 percent). Therefore, if the three unknown sub-villages actually looked more like the sub-village with the maximum 39 percent eligibility of the total households, then the overall proportion of eligible households in the village would be closer to the 41 percent estimation found in the 2012 census data.

The incompleteness in the paradata for the rural Tanzania pilot results in discrepancies between the 2012 census data and the sample. The small size of the village also makes comparative population estimates impossible because most nationally representative datasets will rarely provide data with individual household detail at such a small geography due to confidentiality concerns.

8.2 Urban Tanzania

Data for the sampling process in urban Tanzania was provided by local government officials in the Arusha district office. These population figures were obtained by our Tanzanian partners after numerous visits to the offices. The local government data was based on the 2012 census but provide aggregate population figures (by sex) for a much smaller geographic unit than the IPUMS-International microdata. IPUMS-International microdata for Arusha District can be used to compare the ACV population estimates for the specific target population of households

[41] The IPUMS-International census data for Tanzania is a 10 percent sample. Thus, statistics using weights are estimates of the total population and not actual population counts.

with 12-to-17-year-old residents, similar to the rural village. The local government aggregate data can be used to assess the quality of the ACV population estimates for the entire district.

As described in section 3.2, there are ambiguities between the exact municipality boundaries in Arusha District[42]. In the 2012 census data, an estimated 42 percent of households in Arusha Urban District would have been eligible for the ACV pilot. This exactly matches our calculation that 42 percent of households in the selected cell of the selected *mtaa* were eligible. This comparison offers some evidence of validity in the sampling process in the ACV urban Tanzanian pilot.

We obtained actual *mitaa* population estimates by sex from the local government offices. With these numbers, we are not able to construct anything about household composition or eligibility, but through several assumptions, we can create population estimates at the *mitaa* level.

The sampling process in urban Tanzania, outlined in Figure 3, included three levels of randomization: *mtaa*, cell, and household. At each level, the paradata recorded estimated numbers of cells per *mtaa* and households per cell. We estimate the overall population size of the *mtaa* based on this paradata and assumptions about the average household size based on the collected data. We compare our estimate of the *mtaa* population size to the provided population estimates from the local government office. The local government data includes only total population sizes, but no information about how these individuals are grouped into households or the ages of the population.

The sampling team worked with each *mtaa* leader in each of the 20 selected *mitaa* to identify all of the cells in the *mtaa* and provide an estimate of the size of the cell. Cells are traditionally ten households but in the urban areas they can vary in size; if the *mtaa* leader did not know the exact number of households in the cells, we asked the sampler to probe for the leader's best estimate of size. The 20 selected *mitaa* had 7.6 cells on average (range: 1-17) with an average of 67 households per cell (range: 25-150). In each *mtaa*, we sampled one cell. In this sampled cell, we asked the cell leader to further verify the number of households and estimate the number of people living in the cell.

[42] In the ACV pilot, we sampled from Arusha City (Arusha Mjini) but the 2012 Tanzania census microdata indicates residences within Arusha Urban District (Arusha Vijijini). These two terms do not identically define the same geographic area. Prior to the start of the sampling process, our team compared which wards were located in Arusha City and which were in Arusha Urban District and found considerable overlap in the vast majority of the wards.

Based on this information collected during the sampling process, we constructed estimates at the *mitaa* level based on the following assumptions. First, that the data provided by the *mtaa* and cell leaders is accurate and complete. The first assumption is flawed as 7 of the 20 *mitaa* do not have estimates for the size of all of the cells reported by the *mtaa* leader. This leads to the second assumption that the cells within a *mtaa* are homogenous in size so that the average number of known households per cell can be applied across all cells within a *mtaa*. This allows me to estimate the number of households in cells that do not have complete data based on the average cell size (number of households) from known cells in order to produce estimates of the total households in each *mtaa*.

The population estimates we calculate based on these assumptions end up underestimating numbers of households in each *mitaa* by approximately 50 percent of the published figures. The average difference is an underestimate of 452 households. This average is highly skewed by a single ward (Ward #2) where the published number is 4,289 households, but our calculations estimate 720 households. If this outlier ward is excluded, the estimates of the number of households per *mitaa* are underestimated by an average of 279 households.

We constructed estimates of the population of the *mitaa* based on the average number of people living in each household based on the paradata gathered from the cell and *mitaa* leaders. Similar to the estimates above, the paradata estimates of population numbers were incomplete and we made assumptions that the known average number of persons per household reported by the cell leader can be applied to all households in all cells in the *mtaa*. The estimates of population size were also underestimated by 32 percent on average, compared to the official figures from 2012.

Both the calculation of households and population estimates rely on faulty assumptions that cell sizes are homogenous, and the average cell size and household size can be applied to all cells in a *mtaa*. It is also possible that the composition of the city has changed between 2012 and 2018. Certain *mitaa* may be smaller if migrants are moving to other areas of the city or peri-urban suburbs. However, the city of Arusha is expanding rapidly, and it may be difficult for a cell leader to know what is happening among each family across the cell. Traditionally, the cells would established to represent 10 households, but increases in rural-urban migration have expanded the population of Arusha. Even if local leaders were aware of the results published in the 2012 census, they may still underestimate exactly the number of households in their *mitaa* or cell[43].

Given these caveats, the construction of population estimates of the *mitaa* in Arusha City from the ACV paradata in urban Tanzania is not recommended. The small sample size and limited paradata about unsampled cells do not have enough accurate detail support accurate population estimates. This does not suggest that our results in the ACV pilot are not valid or inaccurate. The multi-stage probabilistic sample was implemented to generalize to the broader population of Arusha City's 12-to-17-year-old children; the population estimates for the total population was not the purpose of aim of the study, but merely a post-survey experimental analysis of the collected paradata.

8.3 Nepal

In Nepal, we used the estimated household and population sizes from the sampling frame complied by the sampling team through the process of identifying buildings and eligible households to estimate a total population size across all of the sampled wards in the two municipalities. The paradata collected by the samplers included estimates of the number of households in building visited and the average household sizes for eligible and ineligible households. We used these numbers to extrapolate estimates of the number of people and number of households in all of the buildings in the wards identified from the satellite images. Similar to the Tanzania pilots, we obtained aggregate population data for the wards provided by the municipality offices based on the 2011 Nepali Census. However, unlike Tanzania, the population figures do not group people into households.

In order to calculate population estimates for each sampled ward in the two municipalities, we make several assumptions similar to those made in urban Tanzania. First, we assume the data collected by the sampler is accurate and complete. Second, that the estimates of the number of households per building visited is representative across the ward. Third, that the number of people in each known household is, on average, the same across the households that were not visited in the ward. Finally, it is unknown if the buildings identified by the satellite image are residential homes. Obviously labeled buildings such as schools, hospitals, or offices in

[43] We tested the theory that cell leaders would better estimate the number of households in the cell than the *mtaa* leader by applying the number of households in the sampled cell to the other cells in the *mtaa*. The estimates were essentially the same for all *mitaa*.

OpenStreetMaps were avoided, but the analysis of population estimates assumes that the buildings identified in OpenStreetMaps were residential.

We compare our estimates of total population size to the published aggregate figures. The average published ward size is approximately 7,000 people in both municipalities. In one municipality, we overestimate the population size by 65 persons on average across the 5 wards, though the estimates for the five individual wards had a standard deviation of 1,200 people. Our local collaborators report that this area is a well-established suburban area of Kathmandu and has been residential for several decades. The stability of the neighborhood and the largely residential areas contribute to the accuracy of the population estimates compare to the published results.

In the other municipality, our population estimates are overestimates by almost 5,000 persons on average across the 5 wards (s.d.= 4,200). Based on local knowledge of the area, this second municipality is a growing area. We saw many multi-unit apartment buildings newly constructed or under construction in this municipality during the field work. The published figures from 2011 are likely to be outdated and not representative of the current population.

From the 2011 census, 34 percent of the households included a child age 12-17 across the entire Kathmandu district (which includes 11 municipalities including the 2 sampled). In the ACV pilot, approximately 23 percent and 25 percent of the households identified in the sampling process were eligible in the two municipalities in which we worked. The discrepancies in the ACV estimate and the census figures could be the difference in geographic units or the nine years between the 2011 census and the 2019 ACV data collection. Alternative explanations include the high rates of migration to peri-urban areas of Kathmandu may be predominantly adults looking for work (Graner 2001); other areas of Kathmandu may have different family compositions.

9. Recommendations and Conclusions

Survey data collection in developing countries is a complicated and difficult process. Researchers with limited budgets, small teams, and little formal training face challenges at every stage of the survey process. The potential coverage, sampling, and nonresponse errors found in the Total Survey Error framework are amplified in small size projects due to daily decisions made by individuals on the project. Discussing potential errors is disincentivized in publications, thus early-career researchers and students have few applied examples of how to design, collect, and reflect on the quality and shortcomings of data. This paper aims to provide guidance to researchers who find themselves designing research that falls into a gap in the survey literature.

The ACV pilots exemplify the gap in the literature between small- and large-scale survey designs. The sampling processes of the ACV pilots mimic large-scale surveys, but the resulting sample size of the pilots is small. In large samples, individual decisions of samplers and field workers that occur randomly often do not impact a sample systematically. Therefore, a large-scale survey can reduce coverage error, sampling error, and nonresponse error throughout the establishment of a representative and probabilistic large sample. The broad representation, either in geography or population, eases the process of creating sample weights and population estimates that compare to existing source data and sampling frames. The rise of technology such as GIS imaging for sampling and tablets using CAPI (computer-aided personal interviewing) software for interviewing in survey work allows teams to identify and correct quality control issues faster than traditional methods. The standards, methods, and infrastructure created by large-scale survey organizations can be applied and adapted for use by small studies, such as ACV.

The ACV pilots also demonstrate the practical, social, and logistical challenges of establishing a sampling frame of atypical households—those in which at least one 12-to-17-year-old child resides. Most large-scale household surveys seek to interview adults; therefore, the literature assumes that probabilistic samples of households will yield an adult respondent. Literatures on sampling "hard to find" or hidden populations other methods of sampling including snowball sampling or convenience sampling (Watters and Biernacki 1989; Salganik 2006; Magnani et al. 2005; Sadler et al. 2010). Potential coverage and sampling errors that arise during the sampling process of a specified target population can also amplify potential nonresponse errors demonstrated in the process of creating outcome measures such as response rates, sample weights, and population estimates.

Based on the descriptive accounts of the ACV pilot studies in rural Tanzania, urban Tanzania, and peri-urban Nepal, we offer the following recommendations for adapting and applying sampling methods and data collection techniques found in text books and literature based on large-scale survey collection in developing countries[44]. Many of the recommendations here do not contradict the teachings from the field of survey research but instead highlight specific areas that may be more relevant for small-scale studies. We organize these

[44] An excellent resource is Survey Research Center (2016).

recommendations by the two areas where the ACV pilots experienced the most challenges: social and logistical.

9.1 Social Recommendations

- Work with local survey organization partners to build relationships with key stakeholders such as local leaders, community leaders, and field team members.
- Build legitimacy with local government officials. This process may be time consuming and bureaucratic. Local government officials may be key in providing existing population figures needed to create a sample, and they may be able to assist in making key introductions to local leaders. Local leaders don't care about your research and your research is not a priority for them.
- Establish training and field work protocols for all members of the sampling and data collection teams. Work with local partners to learn and incorporate norms and cultures of survey collection into the training and protocols.
- Hire field team members that are trained in human-subjects and ethical research practices. Field team members should be respectful of the ethical protocols while also friendly and efficient when interacting when interacting with participants.
- The members of the sampling or pre-sampling screening procedure team may not be the same individuals as being hired to do the data collection. The sampling process may require a different set of skills than the data collection.

9.2 Logistical Recommendations

- Be creative about sources of data to construct sample frames. Identifying information about the eligible and ineligible populations as well as the sampled and unsampled populations can aid in the production of sample weights and population estimates to analyze the validity of the data collection and sampling processes.
- Budget enough money and time to properly conduct the sample. Take into consideration the travel time between locations and the time spent in each location.
- Whether you have a complex or simple sampling design, clearly communicate to the sampling and data collection team the information to be collected. Systematize the collection of paradata and documentation of the daily process.
- If possible, use digital technology such as CAPI to track paradata in the sampling and data collection process. Some survey software includes features in the program to make

this easier. This includes interactions with eligible and ineligible households, results of each contact with a household, and comments about each household.

- Over-document everything. One benefit of a small-scale study is that the sample size may be small enough to have a discussion about each household individually. Daily debriefing notes provide justification and clarification for categorization of interactions with households.

These social and logistic recommendations are particularly important for small-scale studies that aim to make generalizable claims about the validity and statistical accuracy of the survey results. The general techniques of identifying and correcting nonresponse error—outcome measures such as response rates, sample weights, and population estimates—are data intense calculations that require additional data and paradata beyond the results of the survey. To assess nonresponse error, it is not enough to know about the participants included in the survey, researchers must know about the households that did not respond, were not identified, and were not eligible.

It is important for researchers to properly specify definitions of outcome measures reported. As discussed in section 6, outcome measures can be subtly manipulated in a way that is not incorrect or unethical but allows researchers to portray the quality of their data in different ways. We therefore recommend researchers follow the AAPOR recommendations of proper definitions and clarification in all publications and reports. This does not suggest that researchers should not be strategic in reporting measures that accurately reflect the sampling and data collection process, but overall transparency will strengthen the validity and reliability of the data and results.

It may be that the results of the analysis of nonresponse error do not contribute to the overall interpretation of the survey results. In the ACV pilots, the high response rates and low refusal rates suggest the interpretation of the survey results may not be influenced by nonresponse bias. The accurate comparison of the proportion of the total population eligible in existing census data compared to collected ACV paradata about eligible and ineligible populations suggest minimized coverage and sampling errors in the ACV pilots.

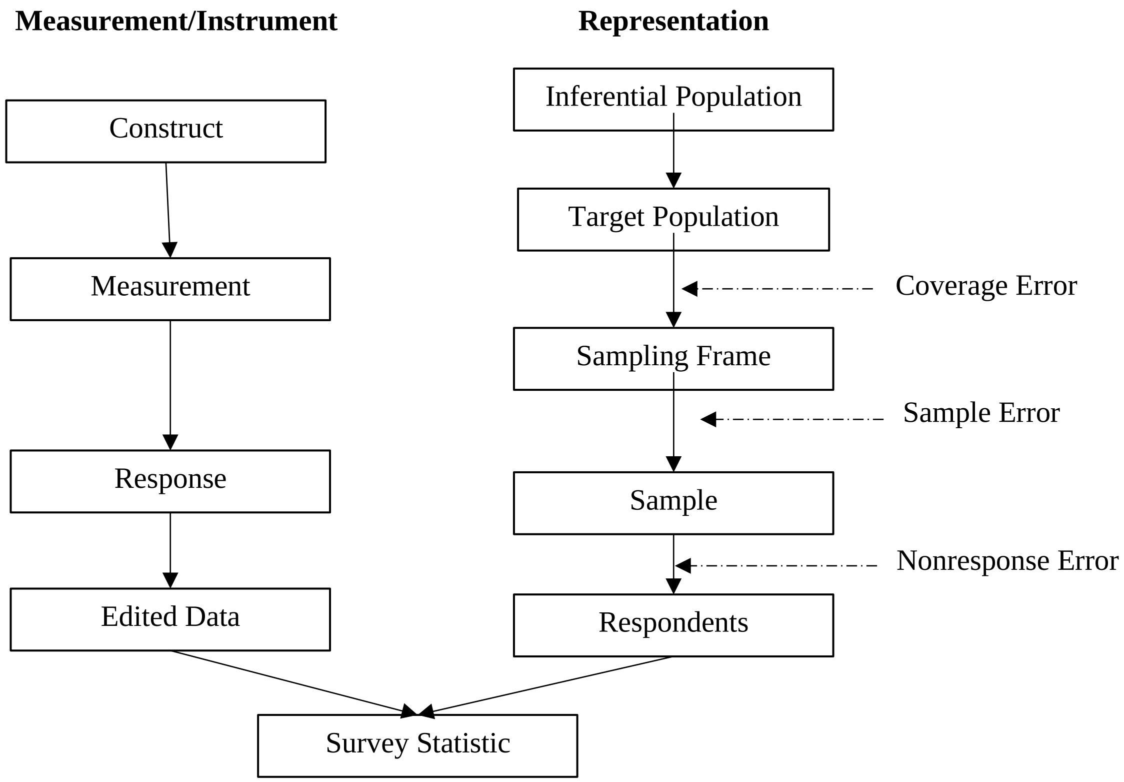
On the other hand, the calculated base weights (weights designed to inflate the sampled population to the target population), and population estimates of the target population were less accurate. Specifically, the urban Tanzania and peri-urban Nepal pilots were designed to be generalizable to large populations and geographies[45]. The team collected paradata about eligible households in the sampled communities. But the paradata required too many assumptions about unsampled and ineligible households and populations to produce accurate base weights and population estimates. This does not invalidate the ACV pilots in these two locations—the sampling process was probabilistic and representative. But it instead suggests the limitations of producing such estimates from small-scale surveys. These calculations are data intensive and require significant social and logistic effort that was beyond the scope, budget, and timeframe of the ACV pilots.

The reality of data collection is messy. Social and logistic challenges are faced by everyone doing social science research and the challenges are unique to every pilot location. Being transparent about potential error in survey processes could jeopardize publication and overall validity of survey results. But it is the only way that readers of academic literature can communicate areas for growth. Sampling designs do not have to be complicated, but simple designs need to be executed properly. It is difficult to use statistics to adjust away errors that occurred during sloppy data collection.

[45] The small size of the rural Tanzania village did allow for the calculation of sample weights and population estimates that had smaller variation than in the other pilots. But we are still limited by assumptions and outdated data provided by local government officials.

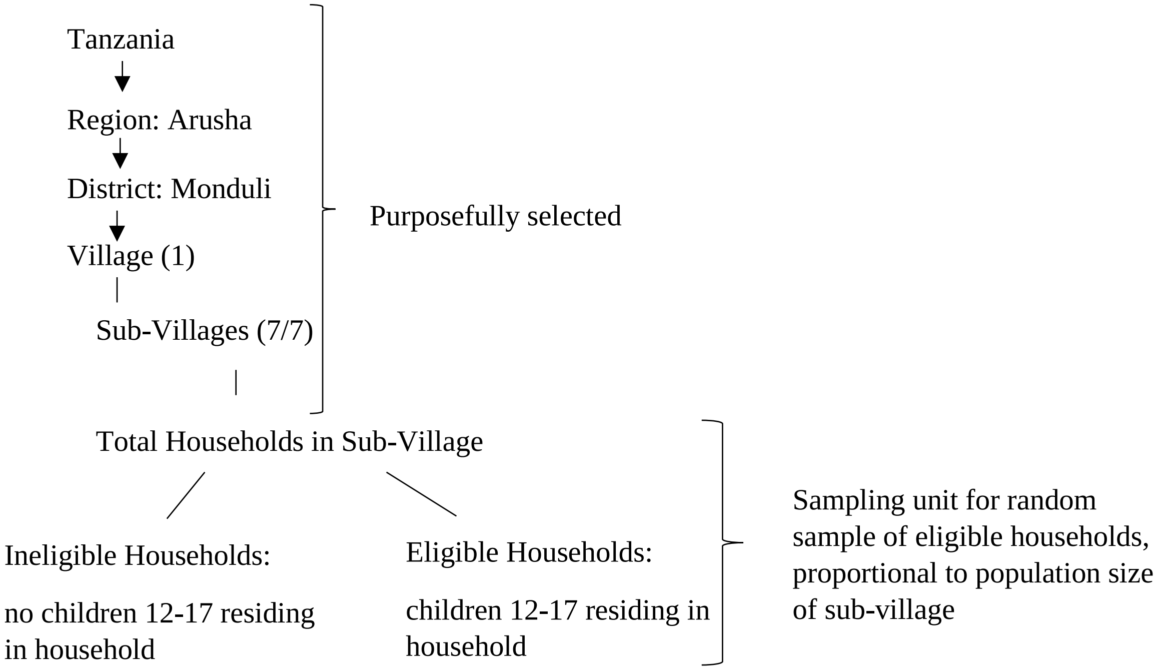
10 Figures

Figure 1: Total Survey Error Components Linked to Steps in the Measurement and Representational Inference Process



Adaptation of Figure 3 in Groves & Lyberg (2010)

Figure 2: Flow Chart of Sampling Process in Rural Tanzania



Tanzania

Region: Arusha

District: Monduli

Village (1)

Sub-Villages (7/7)

Purposefully selected

Total Households in Sub-Village

Ineligible Households:

no children 12-17 residing in household

Eligible Households:

children 12-17 residing in household

Sampling unit for random sample of eligible households, proportional to population size of sub-village

Figure 3: Flow Chart of Sampling Process in Urban Tanzania

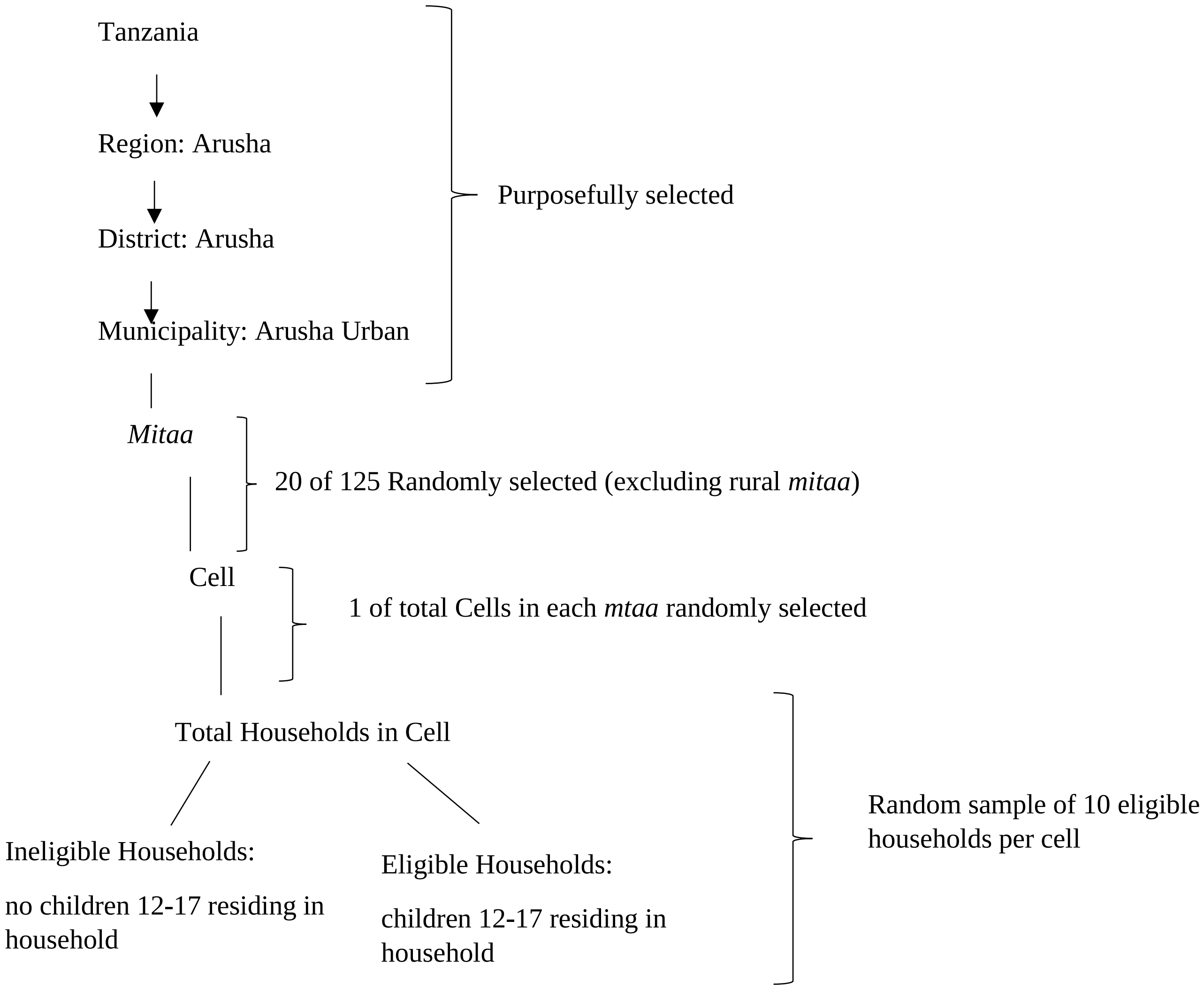

Tanzania
Region: Arusha
District: Arusha
Municipality: Arusha Urban
Purposefully selected
Mitaa
20 of 125 Randomly selected (excluding rural *mitaa*)
Cell
1 of total Cells in each *mtaa* randomly selected
Total Households in Cell
Ineligible Households:
no children 12-17 residing in household
Eligible Households:
children 12-17 residing in household
Random sample of 10 eligible households per cell

Figure 4: Flow Chart of Sampling Process in Nepal

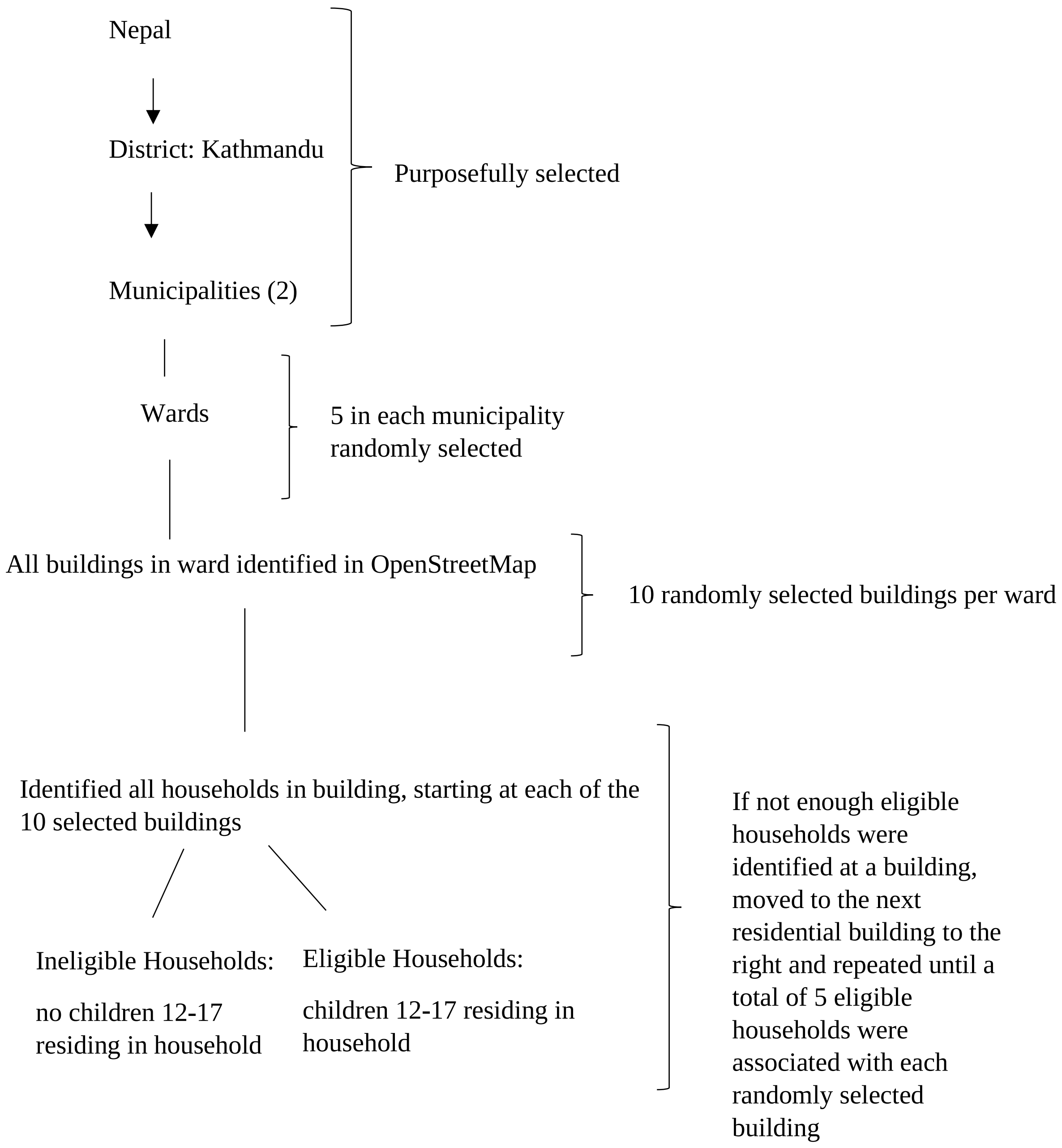
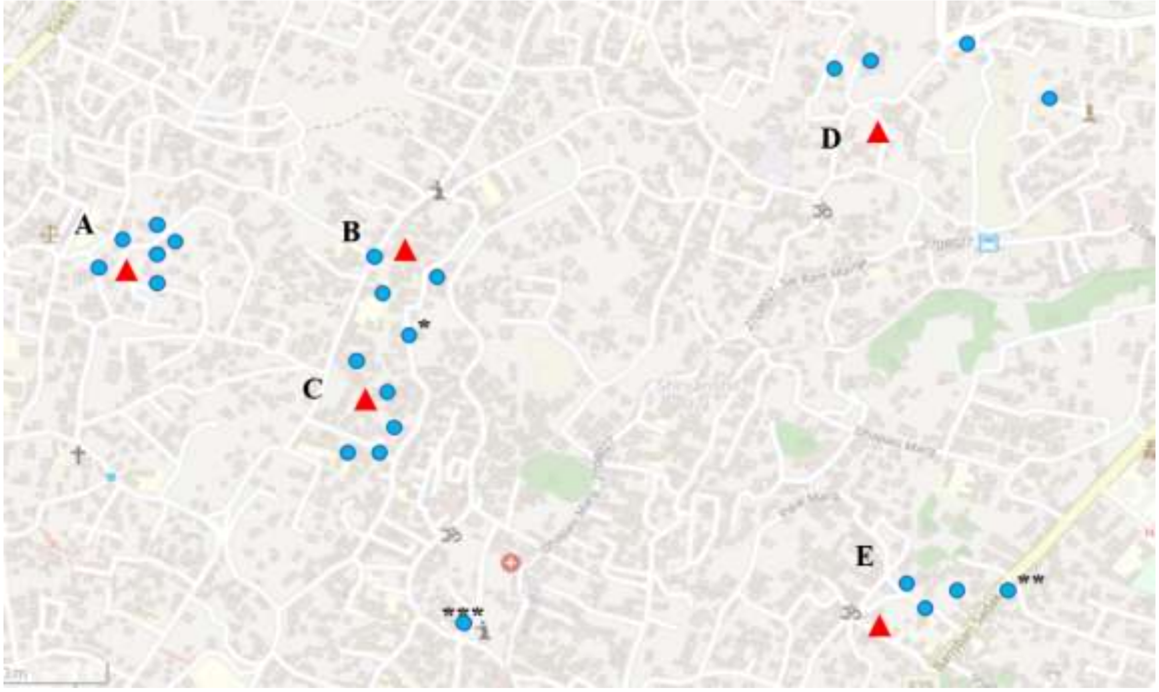

Nepal
District: Kathmandu
Purposefully selected
Municipalities (2)
Wards
5 in each municipality randomly selected
All buildings in ward identified in OpenStreetMap
10 randomly selected buildings per ward
Identified all households in building, starting at each of the 10 selected buildings
If not enough eligible households were identified at a building, moved to the next residential building to the right and repeated until a total of 5 eligible households were associated with each randomly selected building
Ineligible Households:
no children 12-17 residing in household
Eligible Households:
children 12-17 residing in household

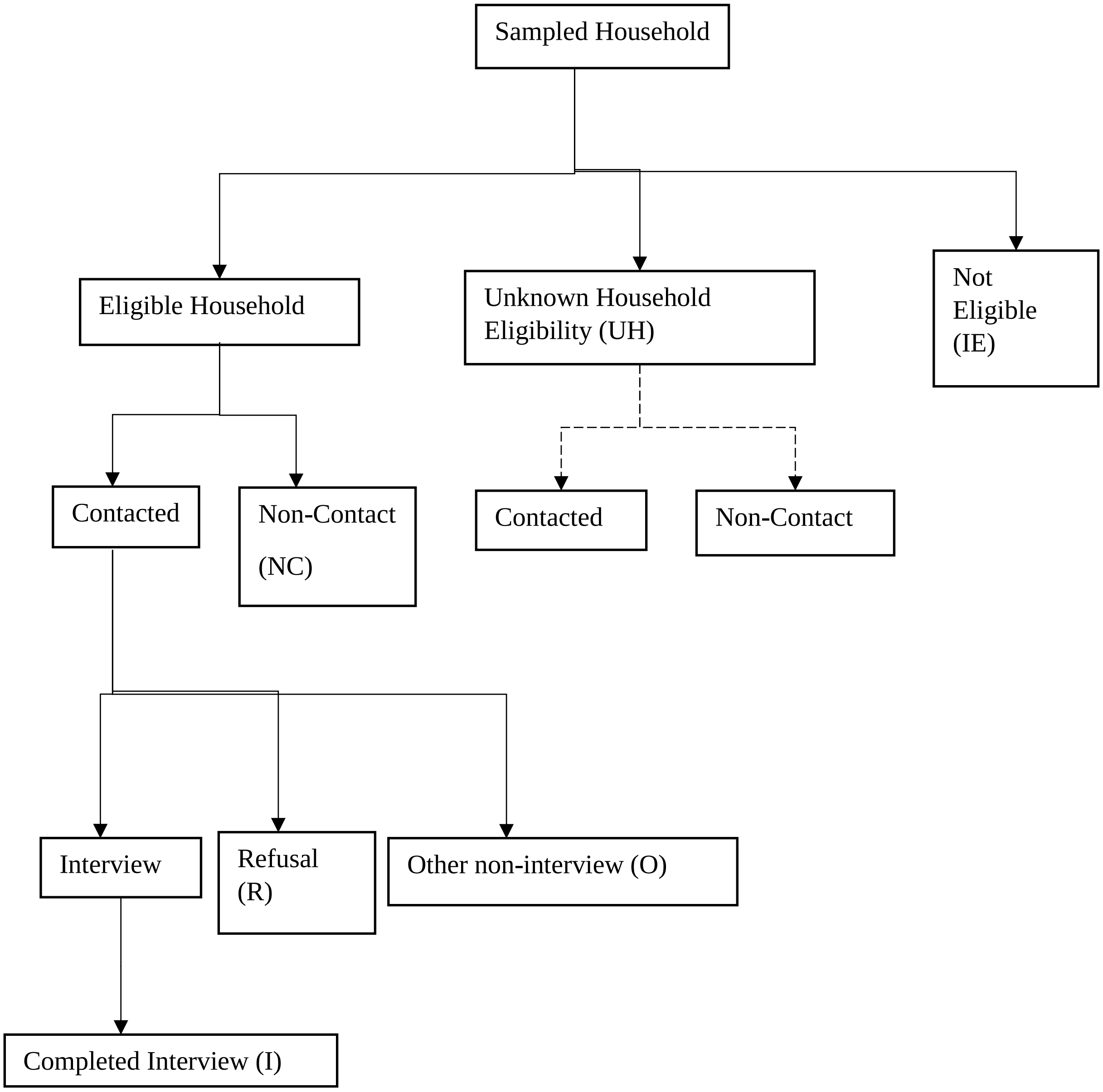
Figure 5: Stylized Map to Show Geospatial Sampling in Peri-Urban Nepal

Red triangle represents sampled buildings
Blue circles identify Eligible households
*Represents a household that falls between 2 sampled buildings
**Represents a household outside of the ward boundary (on the other side of the yellow road on the map)
***Represents a household interviewed in a location far from any originally sampled household, such a community center or café.



Note: Figure does not represent actual geographic locations of sampled buildings or interviewed respondents

Figure 6: Flow Chart of AAPOR Categorizations of Household Eligibility



Adapted from Beerteen, Lynn, Laiho, and Martin (2015)

Figure 7: Rural Tanzania Final Weights (W) by Sub-village (1-7)

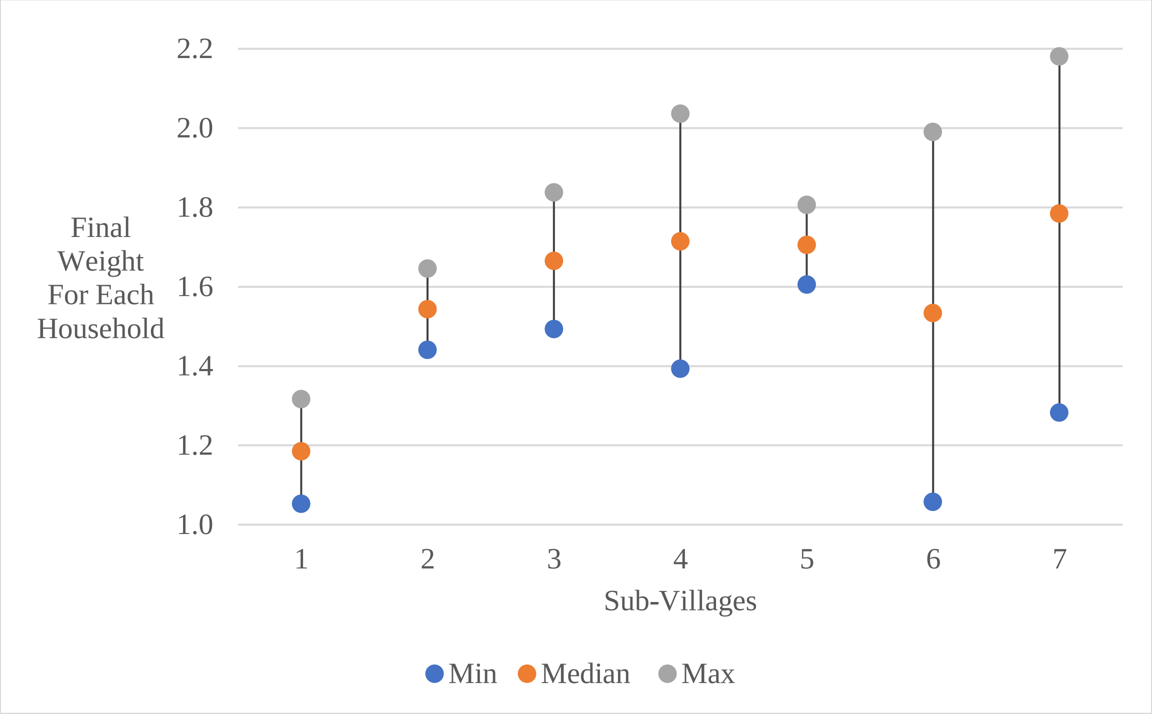


Figure 8: Urban Tanzania Final Weight (W) by Mtaa

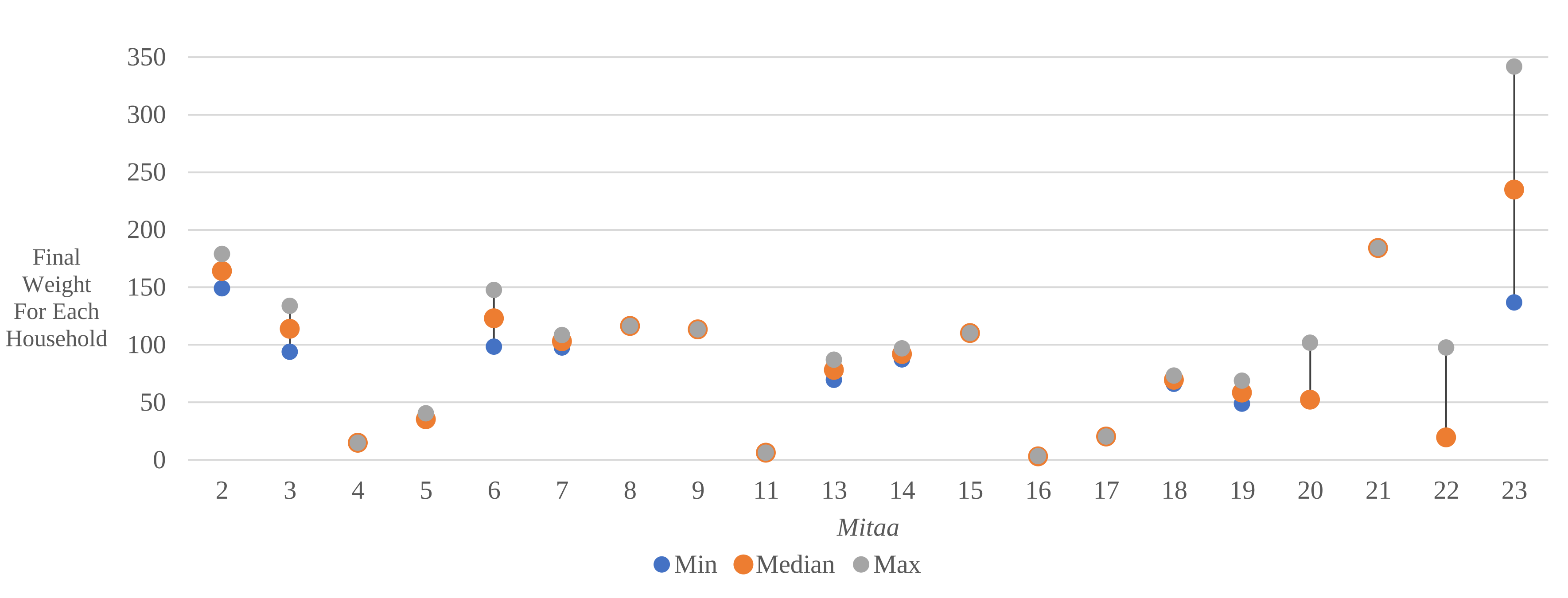


Figure 9 A & B: Nepal Final Weights by Ward (10 wards)
Note the difference in magnitude along the y-axis between Figure A and Figure B.

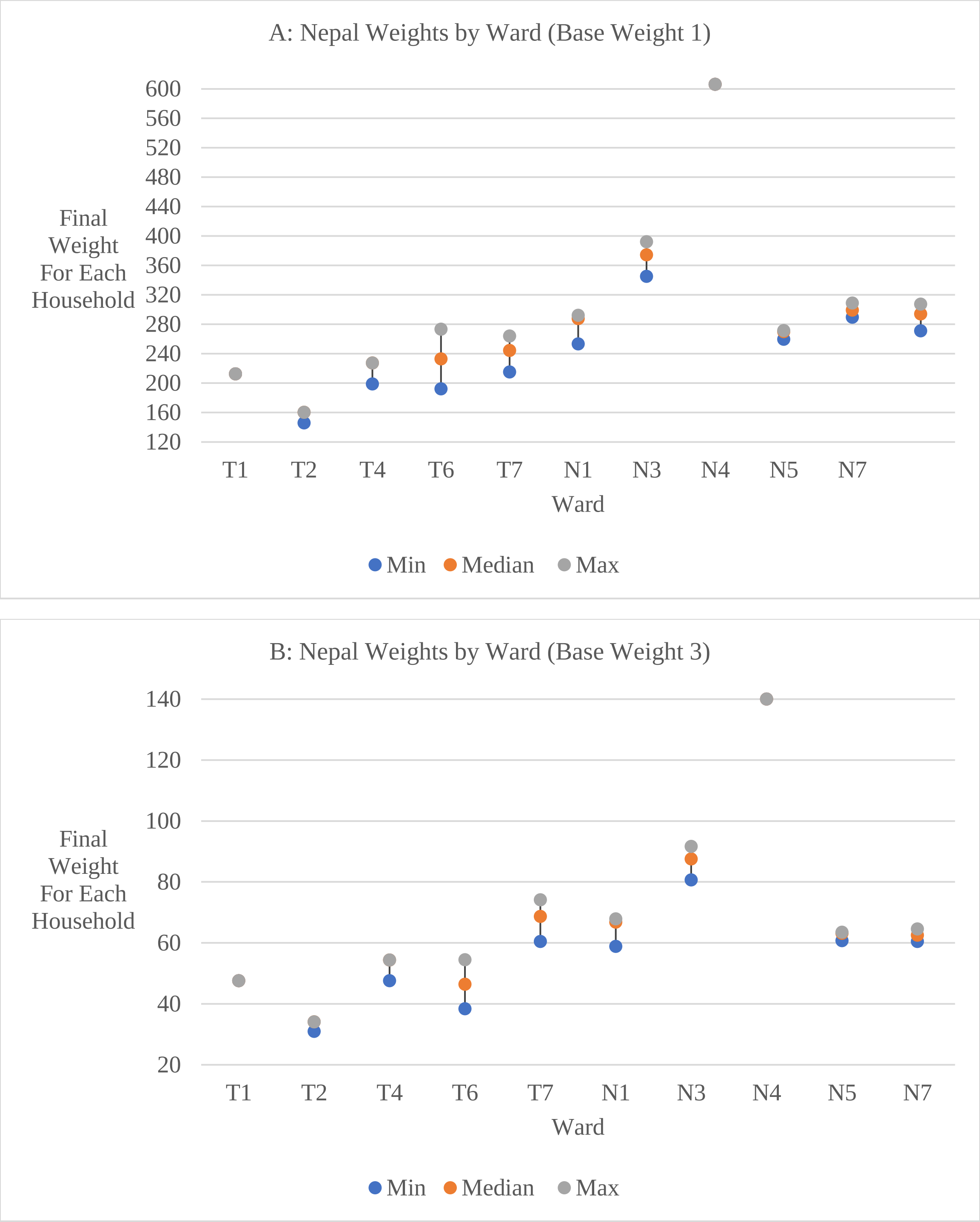
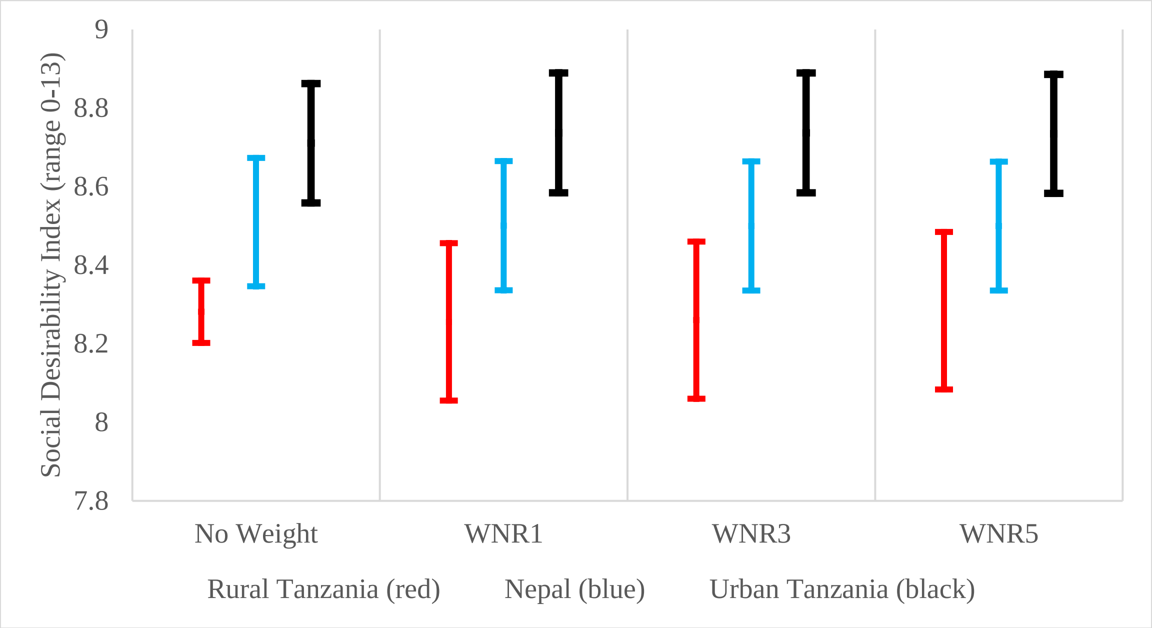


Figure 10: Comparison of Nonresponse Weights on the Social Desirability Index (SDI)



11. Tables

Table 1: Summary of sampling methods as proposed in Fottrel and Byass (2008) and the application of the technique to the context of the 3 pilot studies

Pilot	**Sampling Method**	**Technique**	**Translation to Context**
Rural Tanzania	Proportional Stratified Sampling	Step 1: Determine the proportion of sampling units needed in each strata	Among 7 sub-villages of village, determine number of households needed based on posted population sizes
		Step 2: Assign a random number to each sampling unit	Assign households a number based on order
		Step 3: Select sampling units from each strata using simple random methods until the desired sample size and ratio between strata is obtained.	Using a random number generator, select households with children 12-17 from each of the sub-villages until the determined sample size for each sub-village is reached.
Urban Tanzania	Multi-Stage Sampling	Step 1: Randomly select geographical area for sampling	Randomly selected 20 of 125 *mitaa* in predetermined urban city
		Step 2: Assign a random number to each sampling unit in the select area	Within selected *mitaa*, identified all cells and randomly identify one (or more)
		Step 3: Sort sampling units by their random number	Within the selected cell, identify all households with children 12-17
		Step 4: Select sample units in ascending order of random number until desired sample size is reached	Randomize order of households with 12-17 and sample until desired sample size within each *mtaa*
Urban Nepal	Geographically Dispersed	Step 1: Randomly select # geographic areas	Purposefully selected 2 municipalities within Kathmandu District. ACV did not do this randomly to maximize variation within the municipalities.
		Step 2: Assign a random number to each sampling unit in each of the selected areas	Identify wards within municipalities
		Step 3: Sort sampling units by their random number	Randomly select 50% of the wards in each municipality
		Step 4: Select sampling units in ascending order of random number until 50% of the desired sample in selected from each geographic area	Identify all buildings (sampling units) within selected wards. Randomize order of sampling units. Establish 10 as the target number of units per ward. Select first 10 random sampling units within each ward. Further steps identify which households contain children 12-17 using a random walk method.

Table 2: AAPOR Outcome Rate Formulas (AAPOR 2016)

Response Rates	
RR1	$\frac{I}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + (\mathrm{UH} + \mathrm{UO})}$
RR3	$\frac{I}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + \mathrm{e}(\mathrm{UH} + \mathrm{UO})}$
RR5	$\frac{I}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O})}$
Cooperation Rate	
COOP1	$\frac{I}{(\mathrm{I} + \mathrm{P}) + \mathrm{R} + \mathrm{O}}$
COOP3	$\frac{I}{(\mathrm{I} + \mathrm{P}) + \mathrm{R}}$
Refusal Rates	
REF1	$\frac{R}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + (\mathrm{UH} + \mathrm{UO})}$
REF2	$\frac{R}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + \mathrm{e}(\mathrm{UH} + \mathrm{UO})}$
REF3	$\frac{R}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O})}$
Contact Rates	
CON1	$\frac{(I + P) + R + O}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + (\mathrm{UH} + \mathrm{UO})}$
CON2	$\frac{(I + P) + R + O}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O}) + \mathrm{e}(\mathrm{UH} + \mathrm{UO})}$
CON3	$\frac{(I + P) + R + O}{(\mathrm{I} + \mathrm{P}) + (\mathrm{R} + \mathrm{NC} + \mathrm{O})}$

Key:

- Eligible Households that were interviewed
    - Completed Interview (I)
    - Partial Interview (P)
- Eligible households that were not interviewed
    - Refusals and break-off (R)
    - Non-contact (NC)
    - Not interviewed for other reasons (O)
- Households not interviewed and unknown if the household would be eligible (UH)
    - Unknown for any other reasons (UO)
- Ineligible households (IE)
- e = estimate of the probability of UH being eligible

Table 3: Specific Sub-Categorizations of Unknown Households (UH)

"Not attempted"	In these households, the field research team did not attempt to contact the household at all via phone or visit. In Nepal, the geographic layout of the sampling meant that some wards were very remote and if the team ran out of time to visit all households, the ward was not visited again.
"Household unsafe or unable to reach"	The team often encountered households that were located in areas that the jeep or transport was unable to travel to. Additionally, we encouraged field researchers not to put themselves in danger if approaching a house with a guard dog, if they were unable to call or shout to the inhabitants of the house.
"Unable to locate"	This category includes notes in the paradata like "gave up", "didn't find", or "no one home". In other words, there was an attempt to find household members, but it was unsuccessful.
"Unable to make contact via phone"	In a special category for Nepal, problems with the phone connections were recorded. In Nepal, there was a pre-screening process that collected the phone numbers of sampled households about 2-6 weeks before the field team went to the locations[46]. Because the primary contacting of the household was done via phone, the Nepal team encountered barriers when the service was shut off or the line was blocked.

[46] This is in contrast to Tanzania, where a community leader was showing the field team to the home in person.

Table 4: AAPOR Reporting Outcomes Measures for ACV Pilots in Rural Tanzania, Urban Tanzania, and Nepal

	Rural Tanzania	Rural Tanzania (BS)	Urban Tanzania	Urban Tanzania (BS)	Nepal	Nepal (BS)	Total	Total (BS)
Response Rates								
RR1	76%	64%	77%	73%	79%	76%	78%	72%
RR3	80%	67%	77%	73%	80%	77%	79%	73%
RR5	91%	75%	84%	80%	86%	83%	87%	79%
Cooperation Rates								
COOP1	99%	99%	94%	94%	88%	88%	93%	93%
COOP3	100%	100%	95%	95%	90%	90%	94%	94%
Refusal Rates								
REF1	0%	0%	4%	4%	8%	8%	5%	4%
REF2	0%	0%	4%	4%	9%	9%	5%	5%
REF3	0%	0%	4%	4%	9%	9%	5%	5%
Contact Rates								
CON1	77%	81%	82%	83%	90%	90%	83%	85%
CON2	80%	83%	82%	83%	91%	91%	85%	86%
CON3	90%	75%	89%	85%	96%	93%	92%	85%

BS = inclusion of households with eligible children away at boarding school students as non-contact (NC). Definitions of AAPOR outcome measures found in Table 2

References

Abelsæth, Anne. 2012. “Tutorial: Development of Data Entry- and CAPI Applications in CSPro.”

Bauer, Johannes J. 2016. “Biases in Random Route Surveys.” *Journal of Survey Statistics and Methodology* 4 (2): 263–87. https://doi.org/10.1093/jssam/smw012.

Beerten, Roeland, Peter Lynn, Johanna Laiho, and Jean Martin. 2014. “Response Rates as a Measure of Survey Quality.”

Bennett, Carol, Sara Khangura, Jamie C. Brehaut, Ian D. Graham, David Moher, Beth K. Potter, and Jeremy Grimshaw. 2011. “Reporting Guidelines for Survey Research: An Analysis of Published Guidance and Reporting Practices.” *PLoS Medicine* 8 (8): 1–11. https://doi.org/10.1371/journal.pmed.1001069.

Bennett, Steve, Tony Woods, Winitha M Iiyanagec, and Duane L. Smith. 1991. “A Simplified General Method for Cluster-Sample Surveys of Health in Developing Countries.” *World Health Stat Q.* 44 (3): 98–106.

Blom, Annelies G. 2009. “Nonresponse Bias Adjustments : What Can Process Data Contribute ?” *ISER Working Paper Series*.

Carley-Baxter, Lisa R, Craig A Hill, David J Roe, Susan E Twiddy, Rodney K Baxter, and Jill Ruppenkamp. 2009. “Does Response Rate Matter? Journal Editors Use of Survey Quality Measures in Manuscript Publication Decisions.” *Survey Practice* 2 (7). https://doi.org/10.29115/SP-2009-0033.

Caviglia-Harris, Jill, Simon Hall, Katrina Mulllan, Charlie Macintyre, Simone Carolina Bauch, Daniel Harris, Erin Sills, Dar Roberts, Michael Toomey, and Hoon Cha. 2012. “Improving Household Surveys through Computer-Assisted Data Collection: Use of Touch-Screen Laptops in Challenging Environments.” *Field Methods* 24 (1): 74–94. https://doi.org/10.1177/1525822X11399704.

Chen, Xinguang, Hui Hu, Xiaohui Xu, Jie Gong, Yaqiong Yan, and Fang Li. 2018. “Probability Sampling by Connecting Space with Households Using GIS/GPS Technologies.” *Journal of Survey Statistics and Methodology* 6 (2): 149–68. https://doi.org/10.1093/jssam/smx032.

Couper, Mick P. 2005. “Technology Trends in Survey Data Collection.” *Social Science Computer Review* 23 (4): 486–501. https://doi.org/10.1177/0894439305278972.

Crowne, D. P., and D. A. Marlowe. 1960. “A New Scale of Social Desirability Independent of Pathology.” *Journal of Consulting Psychology* 24: 351. https://doi.org/10.1097/DER.0000000000000036.

Dooblo. n.d. “SurveyToGo.” https://www.dooblo.net/.

Escamilla, Veronica, Michael Emch, Leonard Dandalo, William C Miller, and Irving Hoffman. 2014. “Sampling at Community Level by Using Satellite Imagery and Geographical Analysis.” *Bull World Health Organ* 92 (June): 690–94.

Esri. 2019. “World Street Map.” https://www.arcgis.com/home/item.html?id=3b93337983e9436f8db950e38a8629af.

Fincham, Jack E. 2008. “Response Rates and Responsiveness for Surveys, Standards, and the Journal.” *American Journal of Pharmaceutical Education* 72 (2): 43. https://doi.org/10.5688/aj720243.

Galdo, Jose, Ana C Dammert, and Degnet Abebaw. 2019. "Child Labor Measurement in Agricultural Households: Seasonality, Proxy Respondent and Gender Information Gaps in Ethiopia." Bonn, Germany. https://glm-lic.iza.org/wp-content/uploads/2018/06/glmlic-wp043.pdf.

Galea, Sandro, and Melissa Tracy. 2007. "Participation Rates in Epidemiologic Studies." *Annals of Epidemiology* 17 (9): 643–53. https://doi.org/10.1016/j.annepidem.2007.03.013.

Galway, Lindsay P., Nathaniel Bell, Sahar A.E. Al Shatari, Amy Hagopian, Gilbert Burnham, Abraham Flaxman, Wiliam M. Weiss, Julie Rajaratnam, and Tim K. Takaro. 2012. "A Two-Stage Cluster Sampling Method Using Gridded Population Data, a GIS, and Google Earth TM Imagery in a Population-Based Mortality Survey in Iraq." *International Journal of Health Geographics* 11: 1–9. https://doi.org/10.1186/1476-072X-11-12.

Gelman, Andrew. 2007. "Struggles with Survey Weighting and Regression Modeling." *Statistical Science* 22 (2): 153–64. https://doi.org/10.1214/088342306000000691.

Gosen, Steganie. 2014. "Social Desirability in Survey Research : Can the List Experiment Provide the Truth ?"

Grahe, Jon. 2018. "Another Step towards Scientific Transparency: Requiring Research Materials for Publication." *Journal of Social Psychology* 158 (1): 1–6. https://doi.org/10.1080/00224545.2018.1416272.

Grais, Rebecca F, Angela M C Rose, and Jean-paul Guthmann. 2007. "Don't Spin the Pen : Two Alternative Methods for Second-Stage Sampling in Urban Cluster Surveys." *Emerging Themes in Epidemiology* 4 (8). https://doi.org/10.1186/1742-7622-4-8.

Graner, Elvira. 2001. "Labor Markets and Migration in Nepal." *Mountain Research and Development* 21 (3): 253–59. https://doi.org/10.1659/0276-4741(2001)021[0253:lmamin]2.0.co;2.

Groves, Robert M. 2011. "Three Eras of Survey Research." *Public Opinion Quarterly* 75 (5 SPEC. ISSUE): 861–71. https://doi.org/10.1093/poq/nfr057.

Groves, Robert M., Robert B. Cialdini, and Mick P. Couper. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly* 56: 475–95.

Groves, Robert M., and Lars Lyberg. 2010. "Total Survey Error: Past, Present, and Future." *Public Opinion Quarterly* 74 (5): 849–79. https://doi.org/10.1093/poq/nfq065.

Groves, Robert M., and Emilia Peytcheva. 2008. "The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis." *Public Opinion Quarterly* 72 (2): 167–89. https://doi.org/10.1093/poq/nfn011.

Haenssgen, Marco J. 2015. "Satellite-Aided Survey Sampling and Implementation in Low- and Middle-Income Contexts: A Low-Cost/Low-Tech Alternative." *Emerging Themes in Epidemiology* 12 (1): 1–10. https://doi.org/10.1186/s12982-015-0041-8.

Harling, Guy, Dumile Gumede, Tinofa Mutevedzi, Nuala McGrath, Janet Seeley, Deenan Pillay, Till W. Bärnighausen, and Abraham J. Herbst. 2017. "The Impact of Self-Interviews on Response Patterns for Sensitive Topics: A Randomized Trial of Electronic Delivery Methods for a Sexual Behaviour Questionnaire in Rural South Africa." *BMC Medical Research Methodology* 17 (1): 1–14. https://doi.org/10.1186/s12874-017-0403-8.

Helmes, Edward, and Ronald R Holden. 2003. "The Construct of Social Desirability: One or Two

Dimensions?" *Personality and Individual Differences* 34: 1015–23. https://doi.org/10.1016/S0191-8869(02)00086-7.

Henderson, R. H., H. Davis, D. L. Eddins, and W. H. Foege. 1973. "Assessment of Vaccination Coverage, Vaccination Scar Rates, and Smallpox Scarring in Five Areas of West Africa." *Bulletin of the World Health Organization* 48 (2): 183–94.

Hubbard, Frost, Yu-chieh (Jay) Lin, Dan Zahs, and Mangyao Hu. 2016. "Sample Design." In *Cross- Cultural Survey Guidelines*, 99–149. http://unstats.un.org/unsd/hhsurveys/pdf/Household_surveys.pdf.

Hughes, Sarah M, Samuel Haddaway, and Hanzhi Zhou. 2016. "Comparing Smartphones to Tablets for Face-to-Face Interviewing in Kenya," no. 2010: 1–12. https://doi.org/10.13094/SMIF-2016-00001.

JAMA. n.d. "JAMA Instructions to Authors." Accessed November 28, 2019. https://jamanetwork.com/journals/jama/pages/instructions-for-authors.

Johnson, T.P., D.O. O'Rourke, J. Burris, and Linda Owen. 2002. "Culture and Survey Nonresponse." In *Survey Nonresponse*, edited by Robert M. Groves, 55–69. Wiley.

Johnson, Timothy, and Linda Owens. 2002. "Survey Response Rate Reporting in the Professional Liturature." *American Association for Public Opinion Research - Section on Survey Research Methods*, 127–33.

Johnson, Timothy P., Beth-Ellen Pennell, Ineke A.L. Stoop, and Brita Dorer, eds. 2019. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Context (3MC)*. John Wiley & Sons, Inc.

Johnson, Timothy P, and Fons J R van de Vijver. 2003. "Social Desirability in Cross-Cultural Research." *Cross-Cultural Survey Methods*, 195–204.

Kalton, G. 1983. *Introduction to Suvey Sampling*. Newbury Park, CA: Sage Publications.

Kalton, Graham, and Ismael Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19 (2): 81–97.

Kish, L. 1965. *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kondo, Michelle C, Kent D W Bream, Frances K Barg, and Charles C Branas. 2014. "A Random Spatial Sampling Method in a Rural Developing Nation." *BMC Public Health* 14 (1): 1–8. https://doi.org/10.1186/1471-2458-14-338.

Kumar, Naresh. 2007. "Spatial Sampling Design for a Demographic and Health Survey." *Population Research and Policy Review* 26 (5–6): 581–99. https://doi.org/10.1007/s11113-007-9044-7.

Kviz, Frederick J. 1977. "Toward a Standard Definition of Response Rate." *The Public Opinion Quarterly* 41 (2): 265–67.

Lalwani, Ashok K., Sharon Shavitt, and Timothy Johnson. 2006. "What Is the Relation between Cultural Orientation and Socially Desirable Responding?" *Journal of Personality and Social Psycology* 90 (1): 165–78.

Leeper, Thomas J. 2019. "Where Have The Respondents Gone? Perhaps We Ate Them All." *Public Opinion Quarterly* 83 (Special Issue): 280–88. https://doi.org/10.1093/poq/nfz010.

Levison, Deborah, and Anna Bolgrien. 2020. “Using Cartoon Videos to Survey Children and Adolescents in the Global South: A Tanzanian Example.” *Statistical Journal of the IAOS* 36 (S1): S147–59. https://doi.org/10.3233/SJI-200698.

Lynn, Peter, and Gerry Nicolaas. 2010. “Making Good Use of Survey Paradata.” *Survey Practice* 3 (2): 1–5. https://doi.org/10.29115/sp-2010-0010.

Magnani, Robert, Keith Sabin, Tobi Saidel, and Douglas Heckathorn. 2005. “Review of Sampling Hard-to-Reach and Hidden Populations for HIV Surveillance.” *AIDS, Supplement* 19 (2): 67–72. https://doi.org/10.1097/01.aids.0000172879.20628.e1.

Makela, Susanna, Yajuan Si, and Andrew Gelman. 2014. “Statistical Graphics for Survey Weights.” *Revista Colombiana de Estadística* 37 (2Spe): 285–95. https://doi.org/10.15446/rce.v37n2spe.47937.

Marton, By Krisztina, and Lowndes F Stephens. 2001. “The New York Times’ Conformity to AAPOR Standards of Disclosure for the Reporting of Public Opinion Polls.” *Journalism & Mass Communication Quarterly* 78 (3): 484–502.

Milligan, Paul, Alpha Njie, and Steve Bennett. 2004. “Comparison of Two Cluster Sampling Methods for Health Surveys in Developing Countries.” *International Journal of Epidemiology* 33 (3): 469–76. https://doi.org/10.1093/ije/dyh096.

Montana, Livia, Peter M. Lance, Chris Mankoff, Ilene S. Speizer, and David Guilkey. 2016. “Using Satellite Data to Delineate Slum and Non-Slum Sample Domains for an Urban Population Survey in Uttar Pradesh, India.” *Spatial Demography* 4 (1): 1–16. https://doi.org/10.1007/s40980-015-0007-z.

Onwuegbuzie, Anthony J. 2007. “A Typology of Mixed Methods Sampling Designs in Social Science Research A Typology of Mixed Methods Sampling Designs in Social Science” 12 (2): 281–316.

OpenStreetMap Contributors. 2019. “OpenStreetMap.” https://planet.openstreetmap.org.

Reierson Draugalis, Jolaine, Stephen Joel Coons, and Cecilia M Plaza. 2008. “Best Practices for Survey Research Reports: A Synopsis for Authors and Reviewers.” *American Journal of Pharmaceutical Education* 72 (1): 1–8.

Ritter, Joseph, Monique Borgerhoff Mulder, Kari Hartwig, Susan James, Deborah Levison, Ester Ngadaya, and Craig Packer. 2010. “The Whole Village Project : A Platform for Evaluating Rural Development Projects The Whole Village Project : A Platform for Evaluating Rural Development Projects Department of Applied Economics , University of Minnesota Department of Anthropology , Univer.” Minnesota Population Center Working Paper Series. Minneapolis.

Sadler, Georgia Robins, Hau Chen Lee, Rod Seung Hwan Lim, and Judith Fullerton. 2010. “Recruitment of Hard-to-Reach Population Subgroups via Adaptations of the Snowball Sampling Strategy.” *Nursing and Health Sciences* 12 (3): 369–74. https://doi.org/10.1111/j.1442-2018.2010.00541.x.

Salganik, Matthew J. 2006. “Variance Estimation, Design Effects, and Sample Size Calculations for Respondent-Driven Sampling.” *Journal of Urban Health* 83 (7 SUPPL.): 98–112. https://doi.org/10.1007/s11524-006-9106-x.

Savel, Craig, Stan Mierzwa, Pamina Gorbach, Michelle Lally, Gregory Zimet, Kristin Meyer, and Sameer Souidi. 2014. “Web-Based, Mobile-Device Friendly, Self-Report Survey System

Incorporating Avatars and Gaming Console Techniques." *Online Journal of Public Health Informatics* 6 (2): 1–11. https://doi.org/10.5210/ojphi.v6i2.5347.

Schooler, Jonathan W. 2014. "Metascience Could Rescue the 'Replication Crisis.'" *Nature* 515 (7525): 9. https://doi.org/10.1038/515009a.

Shannon, Harry S., Royce Hutson, Athena Kolbe, Bernadette Stringer, and Ted Haines. 2012. "Choosing a Survey Sample When Data on the Population Are Limited: A Method Using Global Positioning Systems and Aerial and Satellite Photographs." *Emerging Themes in Epidemiology* 9: 1–7. https://doi.org/10.1186/1742-7622-9-5.

Smyth, Jolene D., Kristen Olson, and Mathew Stange. 2019. "Within-Household Selection Methods: A Critical Review and Experimental Examination." *Experimental Methods in Survey Research*, 23–45. https://doi.org/10.1002/9781119083771.ch2.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–16. https://doi.org/10.3368/jhr.50.2.301.

Stedman, Richard C, Nancy A Connelly, Thomas A Heberlein, Danial J. Decker, and Shorna B Allred. 2019. "The End of the (Research) World As We Know It? Understanding and Coping With Declining Response Rates to Mail Surveys." *Society & Natural Resources* 32 (10): 1139–54. https://doi.org/10.1080/08941920.2019.1587127.

Survey Research Center. 2016. "Guidelines for Best Practice in Cross-Cultural Surveys." Ann Arbor, MI: Retrieved June 19, 2020.

The American Association for Public Opinion Research. 2016. "Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 9th Edition."

The DHS Program. 2021. "DHS Methodology." The DHS Program: Demographic and Health Surveys. 2021. https://dhsprogram.com/Methodology/Survey-Types/DHS-Methodology.cfm.

Turner, Anthony G., Robert J. Magnani, and Muhammad Shuaib. 1996. "A Not Quite as Quick but Much Cleaner Alternative to the Expanded Programme on Immunization (EPI) Cluster Survey Design." *International Journal of Epidemiology* 25 (1): 198–203. https://doi.org/10.1093/ije/25.1.198.

UNICEF. 2020. "Choosing the Sample." In *Multiple-Indicator Survey Handbook*. New York. http://mics.unicef.org/files?job=W1siZiIsIjIwMTUvMDQvMDMvMDYvNDIvNDgvMjg2L2NoYXAwNC5wZGYiXV0&sha=d31cdb905d60500d.

Verardi, Sabrina, Donatien Dahourou, Jennifer Ah-Kion, Uma Bhowon, Caroline Ng Tseung, Denis Amoussou-Yeye, Marcel Adjahouisso, et al. 2010. "Psychometric Properties of the Marlowe-Crowne Social Desirability Scale in Eight African Countries and Switzerland." *Journal of Cross-Cultural Psychology* 41 (1): 19–34. https://doi.org/10.1177/0022022109348918.

Vu, Alexander, Nhan Tran, Kiemanh Pham, and Saifuddin Ahmed. 2011. "Reliability of the Marlowe-Crowne Social Desirability Scale in Ethiopia, Kenya, Mozambique, and Uganda." *BMC Medical Research Methodology* 11 (1): 162. https://doi.org/10.1186/1471-2288-11-162.

Wampler, Peter J., Richard R. Rediske, and Azizur R. Molla. 2013. "Using ArcMap, Google Earth, And Global Positioning Systems to Select and Locate Random Households in Rural Haiti." *International Journal of Health Geographics* 12. https://doi.org/10.1186/1476-072X-

12-3.

Watters, John K, and Patrick Biernacki. 1989. “Targeted Sampling : Options for the Study of Hidden Populations.” *Social Problems* 36 (4): 416–30.

Yansaneh, Ibrahim S. 2003. “Construction and Use of Sample Weights.” *Designing of Household Sample Surveys*, no. November: 1–12.

Zimmerman, Linnea. 2017. “PMA2020 General Sampling Strategy.” https://www.pma2020.org/reports/pma2020-sdp-sampling-memo-en.