Conducting Energy Access Surveys in Rural India
Lessons from a Collaborative Project

Michaël Aklin	Chao-yo Cheng	Karthik Ganesan	Abhishek Jain	Johannes Urpelainen[1]
University of Pittsburgh, USA	University of California, Los Angeles, USA	Council on Energy, Environment and Water, INDIA	Council on Energy, Environment and Water, INDIA	Columbia University, USA

1. Summary

This technical report summarizes lessons from a large energy access survey in rural India. Our goal is to use our experience – a collaboration between three major American universities and a leading Indian policy research institute – to offer concrete suggestions and guidelines for researchers and practitioners, working in the area of energy access. The report covers all stages of the survey project, beginning with questionnaire design and ending with data cleaning and analysis. In between, we also consider the challenges posed by questionnaire review, sampling, and the fieldwork itself.

The survey, which we conducted, was targeted on six major Indian states: Uttar Pradesh, Bihar, Jharkhand, Odisha, Madhya Pradesh, and West Bengal. We interviewed a total of 8,568 households in 714 villages, resulting in one of the largest energy access survey conducted in India, so far. The length of the survey was 45 minutes and it captured both cooking energy and electrification related issues. We reached beyond describing the situation and also captured people's preferences, aspirations, and willingness to pay for various services. We also collected data on the villages themselves.

The report is structured as follows. First, we provide an overview of the stages of a survey project. Next, we go through the different stages: questionnaire design, review, sampling, fieldwork, data cleaning and analysis. Each section contains a summary of our approach and the most important lessons.

2. Project Phases: From Questionnaire Design to Data Analysis

The conduct of a large survey on energy access can be divided into five major phases. The first of them is *questionnaire design*. This stage requires defining the substantive goals of the survey, identifying the information that has to be collected, and drafting a questionnaire that enables the survey team to collect the data. The design is constrained by the budget, as longer surveys are more expensive than shorter surveys. Furthermore, long surveys cause respondent fatigue and may thus compromise the quality of the data.

The next stage is *questionnaire review and revision*. After an initial draft has been prepared, it should be reviewed by external experts who can provide comments and advice on the content. Based on the comments, the research team revises the questionnaire. This stage may require several rounds of

[1] **Corresponding author**. Associate Professor, Department of Political Science, Columbia University. 420 W 118 St, 712 IAB. New York, NY 10027, USA. +1-734-757-0161, ju2178@columbia.edu

revision, as incorporating various comments within the limited time, presents a challenge. For the purposes of revising the questionnaire, it is also important to conduct an actual field pilot. Such a pilot allows the research team to test whether the subjects understand the questions, whether the design of the questionnaire generates useful and reliable knowledge, and whether the conduct of the data collection itself presents important challenges that should be considered in the fieldwork.

The third stage of a survey project is the *sampling* of respondents. To gain a good understanding of the reality in the field, it is important to design a sampling strategy that allows one to draw valid inferences about the study population. A good understanding of the statistics of survey research is essential, especially for surveys in large areas. This stage presents notable challenges. Sampling also has clear implications for the budget.

In the fourth stage, the *fieldwork* itself is conducted. Based on instructions and training provided by the survey team, the field researchers locate the respondents and interview them. This stage is challenging for logistics and communications.

The final stage of the survey project is *data cleaning and analysis*. Once the data has been collected, it has to be carefully validated and prepared for statistical analysis. The analysis of the data itself begins after the dataset has been cleaned and verified for quality.

3. Questionnaire Design

3.1 Our Approach

Our goal was to design a 45-minute survey questionnaire on household energy access in rural India. The length of the survey was determined by the available budget and our expectations about the data collection needs.

From the beginning, we decided to focus on household electricity access and cooking fuels, as incorporating the productive use of energy would have required a longer survey. To ensure that we had an original contribution to offer, we also deliberately chose to reach beyond quantifying the current energy access situation; our survey included modules on satisfaction, preferences, and aspirations from the beginning. To complement the household survey, we also designed a shorter village survey to enable the collection of contextual information.

Our research team comprised three academic researchers from American universities and two policy researchers from an Indian research institute. Because the academic researchers had more prior experience with primary data collection in rural India, they prepared the first draft of the survey. Once the first draft of the survey was ready, the two Indian researchers reviewed it carefully and proposed numerous changes throughout the draft. The entire research team organized multiple conference calls to discuss the individual modules of the survey and ensure the draft meet the overall objectives of the project: compilation of a detailed database of energy access among the study population, the design of measurements and indicators of energy access, and relevance for policy formulation in India. Overall, the preparation of the first draft of the survey required approximately two months of collaborative work.

3.2 Key Lessons

The initial design of the questionnaire yielded four primary lessons. First, it is essential to specify clearly the goals of the survey in advance. The design of any major survey inevitably involves multiple experts, and any unresolved misunderstandings about the goals of the survey could lead to redundant work and delay in the development.

Second, the design of a comprehensive energy survey is not possible without both substantive and technical expertise. We quickly learned that we cannot simply copy and paste modules from generic energy access surveys conducted in other countries. Because the cultural and contextual differences are far too large for this approach, the survey team must comprise people with concrete experience in the study of energy access in the relevant context. At the same time, technical experience with survey design and data analysis is equally necessary. The survey questionnaire must be designed to generate data that lends itself to descriptive and explanatory statistical analysis, and such design requires considerable academic training and experience.

Third, the priorities of the energy access survey must be agreed upon among the collaborators and a procedure should be designed to reconcile any inconsistencies. Even though 45 minutes may initially seem sufficient for a comprehensive survey, in practice, difficult choices about priorities are inevitable. In our case, we found ourselves debating the relative importance of more detailed measurements of the current situation and additional information about people's preferences and aspirations. Even seemingly simple questions, such as the education of the household roster, may require large amounts of time. At the same time, a longer survey would have been significantly more expensive and aggravated the problem of respondent fatigue.

Finally, a careful review of available survey questionnaires, especially from the same context, can be very helpful. Reinvention of the wheel is not necessary because scholars and practitioners have designed surveys for rural India for decades. Although many of these surveys suffer from limitations, they are nonetheless excellent material to inform survey design.

4. Questionnaire Review and Revision

4.1 Our Approach

Once we had prepared a first draft of the survey, we began the review and revision of the questionnaire. We began by soliciting comments from the company that we had commissioned to conduct the fieldwork. Because their management team has extensive experience with fieldwork in rural India, they were able to provide many useful insights into the clarity and contextual relevance of the questions.

Next, we conducted a small pilot study to test the survey questionnaire performs in the field. In the small pilot, we first reviewed a Hindi translation of the questionnaire with our enumerators in the office setting. We then interviewed twenty households in the district of Lucknow to evaluate the clarity, suitability, and length of the questions. Two members of the research team oversaw the pilot. Following the pilot, the survey was once again revised after a conference call between the researchers. The researchers who oversaw the pilot prepared a detailed report on the experience, and this report was used in the revision of the questionnaire.

After the pilot, we requested comments from various academic and practitioners both within and outside India. We also showed the questionnaire to our funders and the advisory board of CEEW. These comments allowed us to verify that the questionnaire was up to the standard of researchers and practitioners active in the field. What is more, we were able to secure an interim stamp of approval from the funders and the advisory board.

4.2 Key Lessons

The first lesson from the questionnaire review and revision process was the importance of consulting various types of stakeholders. We quickly learned that the comments from the field researchers of the survey company, from our practitioner partners, and from academic partners were very different. The field researchers commented on the understandability and contextual suitability of questions, along with questionnaire length. The practitioner partners questioned some of our substantive foci and recommended alternative approaches to question formulation based on experience. Academic researchers cited results from earlier methodological studies to propose improvements for questions on willingness to pay, awareness about policies, and preferences.

Next, we want to emphasize the importance of multiple pilots. When we conducted our initial pilot, we quickly learned that our survey was far too long and that there were many issues with the Hindi-English translation. Moreover, the logistics of some questions, such as those that required the respondent's ranking of alternatives, turned out to be challenging. After the initial pilot, we simplified the questionnaire considerably.

As we revised our questionnaire, we noticed that controlling its length was difficult. As members of the research team improved on the quality of the questions, the length of the questionnaire increased rapidly. Whenever someone proposed cuts, others argued against the cuts. We would have saved a lot of time if we had adopted a more rigorous procedure for preventing increases in length from the beginning.

Finally, extensive review of the questionnaire inevitably resulted in contradictory recommendations. While some recommendations were clearly about best practice, others were subjective preferences or based on individual experiences. To deal with contradictory recommendations without compromising the integrity and consistency of the questionnaire, it is essential that everyone on the team has a clear vision of the goals of the survey. This issue should be discussed already before the circulation of the questionnaire begins.

5. Sampling

5.1 Our Approach

One of the exciting but challenging aspects of our project was sampling. We decided to study energy poverty in a very large geographic area comprising the states of Madhya Pradesh, Uttar Pradesh, Bihar, Jharkhand, West Bengal, and Odisha. The population of the study area is approximately 500 million and the distances between villages are significant, at times. Thus, surveying a simple random sample was out of the question for budgetary reasons.

Instead, we decided to follow a stratified approach:

- Each of the six Indian states is divided into administrative divisions. To ensure that entire state would be adequately covered, we conducted surveys in all of the administrative divisions for each state.
- Within each division, we randomly chose one district for the survey (in West Bengal, where divisions are large, we chose two districts within each of the three divisions).

Through this procedure, we had reduced our survey to a random sample of 51 districts. The next challenge was to select villages within each district. To ensure sufficient variation, we adopted the following approach:

- Using data from the 2011 Census of India, we first split each district into two groups: small and large villages. Each group consisted of 50% of all rural households in the district. We split the villages into two groups because we wanted to make sure we would survey both small and large villages. A simple random sampling of villages could have resulted in a skewed sample.
- Within each district, we selected a random sample of 7 small and 7 large villages. The villages were chosen in proportion to the number of households within them. This allowed us to ensure that the resulting 14 villages were statistically representative of the rural population in the district.

In total, we now had 714 villages. Within each village, we created a list of habitations and their populations. We then chose a random sample of 12 households such that their allocation across habitations was approximately proportional to their population sizes. The randomization was conducted by sending the enumerator team to a random spot in the different habitations and then selecting a random direction for the surveys. Based on prior experience in India and elsewhere, this simple approach provides a genuinely random sample of the village population. In total, we interviewed 8,568 households, making our survey, one of the largest one on energy access, to date.

5.2 Key Lessons

From the beginning, we realized that the design of the sampling frame was of critical importance. Our review of existing surveys revealed that, with the exception of census and national sample surveys, most relevant surveys do not provide sufficient information about sampling. The randomization procedure and the sampling frame are often not properly defined. Unfortunately, these omissions mean that the survey results do not have a clear interpretation with reference to the study population. We realized from the beginning that, if we were to offer an original contribution and data that would be useful for a large number of researchers and practitioners, we would have to avoid this pitfall.

In conducting the sampling and randomization, we were able to achieve a good outcome by relying on latest census and administrative data. We began by putting together all relevant data from the 2011 Census of India for the six states under consideration. Specifically, the Primary Census Abstract (PCA) was used to first select a random sample of districts and then, within each district, a random sample of villages for the surveys. Because the PCA also contained the number of households within each village, it was used to ensure that villages were sampled with a probability proportional to their household number. By writing clear and carefully documented scripts for randomization, we were able to ensure a rigorous and replicable procedure. Investing in gaining access to such data is definitely a good idea for survey projects on energy access.

Finally, we quickly realized that sampling households in an area comprising six major states would present an important budgetary challenge. We worked closely with the survey company to avoid cost overruns without sacrificing the representativeness of the survey. In particular, we agreed in advance on the budget based on detailed calculations provided by the director of the company. Moreover, the budget was monitored throughout the project and any challenges, such as unexpectedly high travel expenses, were discussed regularly. Our multi-level sampling strategy was helpful in achieving the goal of cost-effectiveness.

6. Fieldwork

6.1 Our Approach

After we had finalized the survey and conducted the sampling, we were ready to begin the fieldwork. Because we had to do the survey in six states and in three languages (Hindi, Bangla, Oriya), finalizing and reviewing the translations was the first step. Then, we began the training of the survey team. We conducted the following training sessions:

- Uttar Pradesh
- Madhya Pradesh
- Bihar and Jharkhand
- West Bengal
- Odisha

Members of the research team, including at least one person fluent in Hindi, attended the three first trainings. The same survey company representative was also present at all of the trainings to ensure consistency of instructions across the teams.

Under each training session, we began by going through the questionnaire in great detail with the entire enumerator team. The survey instrument went another round of revision after the Uttar Pradesh training session but remained largely unchanged after the others.

Once the trainings were completed, the fieldwork began. The survey company provided us with a data collection plan. Approximately once every two weeks we requested an update on the status of the survey. We also requested immediate contact upon any difficulties.

After the first district was completed in Uttar Pradesh, the survey team sent us some sample data to review. We provided extensive comments on the sample data and the survey company incorporated our recommendations to their working plan. We also requested another batch of sample data from West Bengal to verify that contextual differences across states and languages would not cause any problems. Again, the survey company incorporated our recommendations to their working plan.

6.2 Key Lessons

The fieldwork is typically the most challenging part of any survey project, and our project was no exception. The first important lesson from the experience is the importance of a self-contained questionnaire. Because we had five large teams of people in the field during the fieldwork, the consistency of data collection was a constant challenge for us. When the questionnaire itself clearly specified the instructions to the enumerator, the risk of misunderstanding or inconsistent practice across the teams minimizes. When the enumerators relied on extra instructions given at the training sessions, this risk was higher. Therefore, we recommend making the questionnaires as self-contained as possible, with all instructions to enumerators explicitly written on the survey instrument.

While a self-contained questionnaire goes a long way toward avoiding inconsistencies, the enumerators must nonetheless be carefully trained. We were able to avoid problems by ensuring that one research manager was responsible for all trainings in the different states. While this practice is obviously demanding for the research manager, it is nonetheless essential to avoid any inconsistencies between the instructions given to the enumerators. For example, when we began cleaning and validating the data, we were able to quickly and conclusively reconcile any differences because there was one person who was able to answer our questions.

One part of the training, which needs special mention, is the importance of role-plays and mock interviews during the training sessions. These role-plays significantly improve the enumerators' understanding of the questionnaire and the context of the study. It also makes them confident about the questionnaire content and well-conversant with the flow of the questionnaire.

In conducting the surveys, we also learned that investing into finding experienced enumerators and team leaders was important for the quality of data collection. In contexts such as India, finding qualified enumerators to conduct surveys on specialized topics such as energy access is always a challenge. We had to disqualify several enumerators because they were not able to pose the questions properly to the respondents.

The next important lesson pertains to the translations. Besides the English draft of the questionnaire, we had to prepare translations in Hindi, Bangla, and Oriya. The survey company recruited the translators. Constant revisions made it difficult to ensure that the questionnaires were fully consistent. We adopted the approach of paying particular attention to the Hindi translation because our team included two researchers fluent in Hindi and because it was used in four of the six states. In retrospect, it would have been ideal to approach qualified translators for Bangla and Oriya to ensure all of the translations were consistent when the fieldwork began.

As to the actual data collection, we found it essential to request frequent updates from the survey company. Every 1-2 weeks, we had a research supervisor from the main office send us a spreadsheet with the details of progress in all of the states, both in terms of districts and villages. We also discussed any issues with the fieldwork at the time of progress updates, so as to react to any problems promptly and effectively.

The review of sample data also presented an important challenge. Because we had to review the sample data while the data collection was in process, we had to provide a prompt review. At the same time, it was essential to provide a comprehensive and careful review, as any missed problems would have found their way to the final dataset. Based on this experience, we recommend preparing a rigorous review procedure and clear deadlines for review, well in advance of the receipt of the sample data.

We recommend incorporating the checks for common problems, typically encountered in survey based research, into the review process. These include verifying that individual enumerators are not biasing answers to certain questions; comparing values of variables across states; cross-checking the survey responses with other surveys, especially in the case of objective questions such as energy use. One practice that we found very useful was to call a set of respondents afterwards to discuss their responses and verify that they had understood the questions.

Any large survey inevitably faces challenges and delays. We soon learned that the distances between villages in some districts where very large. This forced enumerators to work very long days, resulting in fatigue and low morale. We solved this issue by giving the enumerators more time off between workdays and by reserving more time for transportation between locations. In retrospect, it would have been better to compute the distances between villages well in advance and account for them in the planning.

We also had to deal with changes in districting and the disappearance of some small villages that were abandoned by the inhabitants. In each case, we had to rapidly find appropriate replacements. Fortunately, we had access to the full data of Census 2011, which allowed us to find replacements immediately, following the randomization process.

Finally, we even faced security issues with Naxalite rebels in Jharkhand. In one district, our survey team was unable to enter villages because the army and the policy had raised roadblocks and prevented all entry and exit by outsiders. In such a situation, we decided to find a different district because we were concerned about any difficulties with the government and, more importantly, the safety of our enumerators.

7. Data Cleaning and Analysis

7.1 Our Approach

The final stage of the research project consists of cleaning the data and analyzing them. Data cleaning and processing is a critical step. The reliability of the analysis highly depends on the quality of the data. Therefore, a rigorous data cleaning process is essential.

By data cleaning, we refer to all manipulations that transform the unprocessed ('raw') original data into datasets that can be used for quantitative and qualitative analyses. Such manipulations include, but are not limited to:

- Identifying possible coding mistakes (e.g. impossible values such as a negative age)
- Flag suspicious observations, either for correction or for removal
- Label variables and observations to facilitate analysis
- Standardize open-ended responses
- Verify compliance to sample design (e.g. the number of respondent per village must match the original design)

For this purpose, we prepared a "Data Cleaning Protocol." This protocol divided the data cleaning process in multiple steps.

To begin with, the data had to be imported in a statistical software and variables had to be named in a systematic manner. Not only does this reduce the risk of errors during the analysis stage, but it also facilitates the cleaning if/when new survey data are collected.

Next, we sought to examine the collected data in greater details. To do so, we listed all responses to all variables by all respondents. By examining one variable after the other, we paved the way for a closer inspection of the data. In addition, whenever the variable was numerical, we also presented descriptive statistics such as the mean response, the lowest value, the highest value, and so on so forth.

The third step was then to carefully search the data for potential issues. Although it was impossible to verify the accuracy of every single observation, we were able to identify dubious values in two ways. First, we looked at the distribution of the observations. For instance, were there any absurdly low or high values? Second, we compared answers to one question with those to another related question. To give an example, a respondent who claimed not to be connected to the grid should not report paying anything for it.

As a last step, all team members compared their notes and listed their queries. Any unresolved questions were then forwarded to the survey team to clarify variables that were flagged in the review process. We organized a teleconference with the head of the survey team to go through all remaining issues.

Each task was assigned to a separate member of the team. We divided the survey in six modules, processing two per week.

7.2 Key Lessons

Data cleaning is both an important and an educative stage. Not only does it allow detecting potential failures, it also ensures that researchers have a good understanding of the data that they have collected. This exercise was certainly valuable and highlighted the following important lessons.

First, it is important that all team members agree on a given protocol. A streamlined process is central to avoid inefficiencies and delays. The validation and verification of data quality is challenging work, and a clear protocol ensures both quality and timely completion.

Second, the process needs to be divided into multiple steps, especially as the number variable increases. With multiple modules in a long survey, it is not possible to verify the quality of data in one session.

Third, it is critical to allow for delays. A degree of flexibility is required to accommodate unexpected events. This is particularly true for larger research teams, as problems with coordination and communication are inevitable.

Fourth, the most critical part of the data cleaning process is the identification of issues and inconsistencies in the data. To some degree, there is no automated way to detect them; we had to rely on a careful evaluation of each variable. Therefore, the effectiveness of the process relies on the efforts of each analyst. Possibly, this could be improved by putting individuals in charge of particular parts of the survey. For instance, somebody could be the lead person, monitoring questions pertaining to cooking energy. Another way could be to identify variables that are particularly critical and that may signal possible problems, before the data cleaning begins

8. Conclusion

This report has summarized and collated our experience with a comprehensive energy access survey in rural India. The report has both described our approach and summarized the most important lessons for both researchers and practitioners. All our questionnaires and data will also be made publicly available within 12 months of the end of the survey project.

While the report has been tailored to the case of India, we believe it is also useful for other practitioners and researchers. While the content of the survey itself must reflect the national context and be adjusted according to the ground realities, the technical challenges of survey design and implementation are largely common across national contexts. It is our hope that this report promotes rigorous research on energy access across different contexts, both within and outside India.