Sampling to Balance Community Engagement and Representativeness

Brett Close, TRC, Portland, OR
Donna Whitsett, TRC, Chattanooga, TN
Christine Zook, TRC, Chicago, IL

ABSTRACT

Across the industry, we see declining response rates for surveys and interviews, which are driving up costs and putting the reliability of results at risk. At the same time, the need to incorporate insights from respondents who are less likely to respond to traditional research recruitment methods, including people in disadvantaged communities (DACs), has taken on increased relevance. One way to address these needs is by partnering with community-based organizations (CBOs) that have existing relationships and can help contact respondents. But while the relationships of these organizations are often deep—leading to improved response rates in specific communities—they are often narrow—not covering significant portions of the population, risking the overall representativeness of samples. To address these challenges, we developed a hybrid geographic sampling approach to leverage the support of CBOs while still gathering responses from a wide range of the population. The paper describes this sampling approach in detail, discussing the approach, the challenges, and findings from a market characterization project in California. At a high level, this sampling approach involves identifying and defining geographic regions (on the level of a local community, town, or city) based on US Census designations, identifying the locations where partner CBOs have relationships, and developing a hybrid clustered hierarchical sample design that combines the regions with CBO relationships with randomly sampled regions.

Introduction

Primary research into households' or businesses' perspectives and experiences is one of the foundations of evaluation and related research, but researchers' ability to successfully use surveys and interviews to represent the relevant populations is increasingly at risk due to non-representative samples. With very low response rates becoming increasingly common, respondents may generally be different from non-respondents in important ways, and in particular may be more likely to be white, relatively well-off financially, and either be retired or have a flexible job that allows them to respond to surveys or participate in interviews. Views, perspectives, and experiences from these groups are not likely to reflect the true distribution of preferences, or the distribution researchers could expect from a true random sample. In a setting where response rates are often below 5%, we as an industry are very far from random sampling—even if we are randomly pulling names from contact lists—because random sampling requires that all members of the population be equally as likely to be in the sample, not just equally as likely to be contacted. As research budgets are squeezed along with program budgets, more intensive methods of follow-up to increase response rates become less feasible for researchers, exacerbating a challenging problem.

This paper describes a hybrid sampling approach designed to improve the representativeness of survey results while mitigating recruitment costs. In particular, the approach combines random sampling from general population contact lists with convenience sampling by partnering with community-based organizations (CBOs) in order to gather a broad and representative sample while also increasing sample response rates in disadvantaged communities (DACs). While we believe this approach had substantial benefits in terms of representativeness, it also presented substantial challenges in terms of feasible design of the sample and analysis of the results. We do not claim that this approach is a perfect approach, but hope it can add an additional tool for researchers and be a starting point for future advancement in methodology.

The typical approach to sampling in energy program evaluation involves developing a solid sample frame from a fixed source (such as a utility program participation list, utility customer list, or other market source); developing a stratification (such as the method of Dalenius and Hodges (1959)); and then an optimal allocation (such as a Neyman allocation (1934)). Our approach deviates substantially from that the typical one, not because of problems with that approach per se, but because of the logistical infeasibility of using it in a setting where specific groups of potential respondents—in particular members of disadvantaged communities—are not on available lists and may not respond to recruitment from unknown recruiters.

We begin with a description of the challenge the industry faces in relying on surveys and interviews based on outreach to recruit respondents. We then describe our overall approach and the project where we implemented it, discussing design, implementation, and analysis challenges. Finally, we provide an overview of the results of the study and provide concluding thoughts.

The Challenge of Survey Sample Representativeness

The insight we can gain from survey research can only be as good as the representativeness of the sample, and declining responses rates for surveys present a fundamental challenge for primary data collection. As response rates decline overall, that presents an additional risk that people who do respond to surveys are increasingly unrepresentative of the overall population (Economist 2023). This issue is particularly acute for people in extremely marginalized populations, such as low-income migrant workers. For example, according to the World Bank, the percentage of the US population without access to electricity is 0.0% (2025). But within our sample we found multiple respondents who credibly reported having no access to electricity, presumably because they live in temporary housing that has not been connected to the electric grid. It is likely that the typical data collection techniques to estimate the population without access to electric service is failing to appropriately account for the difficulty of reaching the population without access to electricity. At the same time, there is increasing pressure on many research projects to limit budgets. These two trends of declining response rates and declining budgets have the potential to substantially impact the quality of the survey research used in evaluation and market research.

Overall Approach

In order to address the challenge of survey sample representativeness, TRC implemented a hybrid sampling approach for recruiting survey respondents that combined random sampling from contact lists with convenience sampling recruitment by community-based organizations (CBOs) that have high-quality contacts and strong name recognition in key communities. This section describes the approach at a high level.

TRC implemented this approach as part of a study on propane utilization among residential homes and businesses for the California Air Resources Board (CARB) in California.[1] The purpose of the study was to inform the development of the zero-emission space and water heater standards and related building decarbonization policies that CARB is developing. To achieve this, the study had two primary objectives:

- Examine and characterize utilization of propane in nonresidential and residential buildings across California and characterize propane and woodburning users in California.
- Evaluate potential solutions to the adoption of zero-emission space and water heater standards with the communities that currently rely on propane.

[1] This paper focuses on the residential respondents, not the nonresidential portion.

We addressed these research objectives by analyzing existing data sources—such as the American Community Survey (ACS), conducted by the US Census Bureau (2022)—and conducting a survey of residential and nonresidential users of propane and wood for space heating. We designed the approach to be able to report on statewide results, as well as results for four regions: Central Valley, Northern Coastal and Sierra, Southern Coastal, and Southern Inland.

The population of interest for the study was all homes and businesses in California that use propane or wood for space heating. Unlike a study of homes and businesses that use electricity or natural gas funded by utilities or regulators, we did not have a list of relevant homes and businesses to target. Instead, we focused on geographic-based sampling as propane and wood usage is highly spatially correlated: most wood and propane users are in areas where natural gas service is unavailable and are in close proximity to other wood or propane users. But even with a geographic-based approach, we still needed a way to contact respondents and decided to work with CBOs and purchase contact lists. In order to limit the costs of the contact lists, which often have costs based on the number of contacts, we needed to select areas with relatively high rates of propane and wood usage. That is, it did not make sense for us to buy contact information for homes or businesses that were very unlikely to use propane or wood.

Our overall sampling approach for this project created a hybrid sample that combined random sampling from a set of randomly selected communities along with convenience sampling by CBOs from specific identifiable communities that could be treated as approximately random for analysis. The random portion of the sample design was based on US Postal Service (USPS) city names with high proportions of propane and wood home heating based on the ACS. USPS cities correspond to the areas where an address uses that city name and contains outlying areas outside the legal boundaries of the municipality, and thus does not exclude rural areas that are in unincorporated areas. For example, Dunlap, CA has a population of 131 based on the US Census, but the Census block groups with the USPS city name of “Dunlap” have a total population of 1170 as they encompass surrounding unincorporated areas.

To select the areas to target, we aggregated all of the block groups with the same USPS city name and calculated the percentage of homes using propane or wood for space heating. We then selected cities with at least 25% propane or wood home heating and assigned a random ordering. We then selected enough cities within each region, based on the random ordering, to have a sufficiently large pool to draw from to achieve our sample targets. The random ordering also provided us with a pre-set list of back-up communities to add to the sample pool if our recruitment had a lower response rate than assumed. For the random sampling portion of the sample, we sent letters to residential addresses in sampled communities and purchased a list of contacts from a survey panel company.

For the convenience portion of the sample, we worked with our partner CBOs to identify the areas where they have strong connections for outreach in specific communities that were designated Disadvantaged Communities (DACs) based on California State guidelines[2] and recruited directly in those communities. For example, one partner had particularly strong contacts in specific communities, such as Cantua Creek, Monterey Park Tract, and Fairmead, CA. Most of these were small communities that did not correspond to distinct towns or cities. In some cases, the communities targeted by the CBO partners were in USPS cities that were randomly selected to be included in our random recruitment. In those cases, we treated any responses from those areas as coming from the random sample.

With contacts and sample points coming from multiple sources, we needed a system for categorizing the responses into the appropriate strata for weighting and analysis. The hierarchical relationship of the sample is as follows:

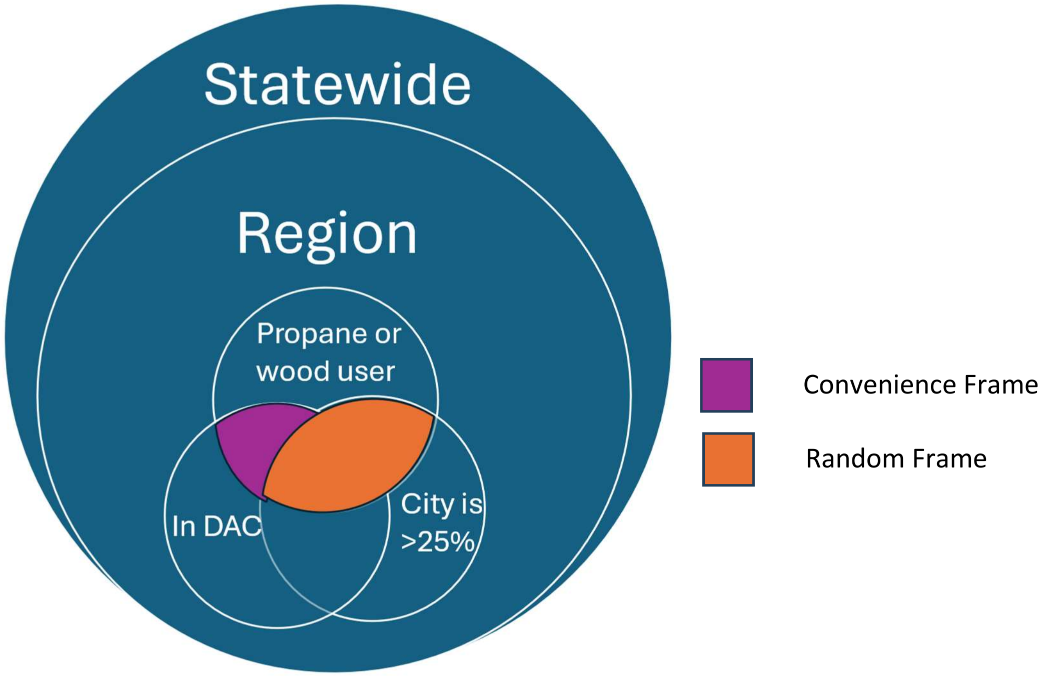
- We assigned all areas to one of the four regions;
- We set the random portion of the sample frame in a region to the set of propane or wood users who were also in communities with at least 25% propane or wood usage;

[2] https://oehha.ca.gov/calenviroscreen/sb535.

- We set the convenience portion of the sample frame in a region to the set of propane or wood users who were also in a disadvantaged community, but not in a community with at least 25% propane or wood usage.

That is, all respondents had to be wood or propane users. If they were in a USPS city with at least 25% wood and propane use they were considered to be in the random sample (regardless of DAC status); if they were not in a USPS with at least 25% wood and propane use and they were in a DAC they were in the convenience sample. No one was included in the sample who was not either in a DAC or a USPS city with at least 25% wood and propane even if they used propane or wood. Figure 1 shows this relationship for a single region with the random frame as the orange area and the convenience frame as the purple area.

Figure 1. Sample Frame Hierarchy



Sample Design and Implementation Challenges

There were a number of practical challenges to designing and implementing this hybrid sampling approach. The primary challenges were

- defining geographic units in a way that it was feasible for partner organizations to recruit from them and for purchasing contact lists, while also having demographic data for analysis;
- defining and implementing recruitment quotas so that no area had too many responses; and,
- maintaining quality control over the survey.

Practical geographic boundaries often differ from those used in statistical data and sample design. For instance, U.S. Census and American Community Survey (ACS) data are organized by Census block groups and tracts. However, CBOs cannot easily target outreach by block group, since their contact lists are rarely geocoded and community members usually do not know which block group they live in. That is, people typically know their street address or ZIP code, but not their Census block group, so recruiters cannot simply ask and match responses to a sampling list. This is less of an issue when working with a defined population, such as a utility's customer list, but it poses a major challenge for studies of broader populations, such as all residents of a state or region. In practice, geographic identifiers that are obvious from an address—such as ZIP code or city—are much more feasible.

To address the challenge of balancing recruitment feasibility and data availability, we used Census block groups as the sampling units to sample from but defined the sampling approach for recruitment based on identifiable address characteristics that correspond to the recruitment approach. That is, we included all block groups within a USPS city for recruitment for the random sampling because the contact lists we could obtain from data agencies were based on cities. This meant that the USPS cities could have areas with low propane and wood space heating along with areas with higher propane and wood usage, but by only selecting USPS cities with at least 25% propane or wood, we limited the number of ineligible survey targets.

For the convenience sample, we selected the block groups that corresponded to the Census Designated Places (CDPs) for the communities targeted by our recruitment partners, rather than the entire USPS city, which would have been much larger due to Census block groups, ZIP codes, and USPS cities not having simple one-to-one relationships.[3] For example, Cantua Creek, CA is contained within a single Census block group. That block group contains a small portion of a ZIP code that extends all the way down to Coalinga, CA, and thus the USPS city for Cantua Creek's block group is Coalinga. By only selecting the specific areas where the CBOs have relationships, we were able to focus more directly on the demographics of those areas to improve our estimates of the sample frame demographics.

Another implementation challenge was balancing the sample sizes within any specific area so that our recruitment partners achieved a balance of sample points across targeted areas. We wanted to avoid a situation where we ended up filling the overall sample primarily from one or two regions and not having sufficient representation of the more hard-to-reach respondents in all regions. Therefore, we needed to set quotas by region.

In order to set quotas for each region for our recruitment partners, we treated the task as a constrained optimization problem: given a fixed number of sample points across regions, how could we minimize the variance (maximize the precision) of the estimate we would calculate once the data were collected? We solved this set of conditions by the method of Lagrange multipliers, an approach that incorporates the objective function (the variance formula) and the constraint (the sum of the sample sizes equal to the total sample size) into a system of equations that can then be solved by traditional optimization methods—in this case, setting the partial derivative of each equation jointly to zero with respect to the regional sample allocations.

That is, our objective function and constraints were

$$\min_{w_i \geq 0} V = \sum w_i^2 V_i = \sum w_i^2 \frac{\hat{\sigma}_i^2}{\sqrt{n_i}} \ subject\ to \sum n_i = N$$

The Lagrangean for this optimization then became,

$$\mathcal{L}(n_1, \ldots, n_k, \lambda) = \sum w_i^2 \frac{\hat{\sigma}_i^2}{\sqrt{n_i}} - \lambda * (N - \sum n_i)$$

where the first term on the right side is the variance function and the second term is a constraint term: here, the constraint that the total sample size be equal to the sum of the sample size for each group multiplied by the Lagrange multiplier, λ.

[3] Census blocks groups are mutually exclusive and exhaustive geographic areas covering the US with similar numbers of residents all contained within a county, whereas ZIP codes are linear groups of streets meant to organize postal routes. A single ZIP code can cover multiple cities and even cross county boundaries. There can be people within the same block group who are in different ZIP codes. There are generally people within the same ZIP code from multiple block groups and even from multiple USPS cities. People within the same USPS city can be in multiple block groups and multiple ZIP codes.

Taking first-order conditions for optimality—that is, taking the first-order partial derivatives of the Lagrangean and setting them all equal to zero—gave us a system of *k*+1 equations that supplies the conditions for a constrained optimum[4]:

$$\frac{\partial \mathcal{L}}{\partial n_1} = -\frac{\hat{\sigma}_1^2 w_1^2}{2n_1^{3/2}} + \lambda = 0$$

$$\cdots$$

$$\frac{\partial \mathcal{L}}{\partial n_k} = -\frac{\hat{\sigma}_k^2 w_k^2}{2n_k^{3/2}} + \lambda = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = \sum n_i - N = 0$$

The first *k* equations provide the conditions on choice variables, and the final equation restates the constraint. We made the assumption that the true variances in each group were equal (but the means could still be different), leading to the optimality condition,

$$\frac{w_1^2}{n_1^{3/2}} = \cdots = \frac{w_k^2}{n_k^{3/2}}.$$

That in turn led to the condition that the sample proportion in each region is the sum of the weight in each region to the four-thirds power, divided by the sum across all regions of the weight to the four-thirds power:

$$p_i = \frac{w_i^{4/3}}{\sum_{j\ in\ R} w_j^{4/3}}$$

where p_i is the sample proportion for stratum *i*, w_i is the total weight for stratum *i*, and *R* is the set of strata. In this case, the weight is proportional to the number of eligible homes in target USPS cities within each region for the random portion of the sample and the number of eligible homes in DACs (but not in target USPS cities) in the region. We used this condition to set target sample quotas for the CBOs and target sample sizes for the random sample in each region. Note that this deviates from the outcome in a typical Neyman allocation where different assumptions are made about the structure of the variance.

A final practical challenge we faced was maintaining data quality with the survey results. In particular, because our CBO partners were recruiting participants, including through online platforms, we discovered that automated digital bots were completing the survey to try to get the participation incentive. This led to us needing to spend significant extra time cleaning the data to remove invalid responses.

Analysis Challenges

The primary challenge for analyzing the results of the survey was determining the appropriate weighting scheme for the results; that is, balancing the representativeness of responses for reporting on population-level results. The three primary elements of the weighting were deciding how to operationalize the hybrid sample frame, constructing the weights for each response, and determining the appropriate use of weighted vs unweighted results.

[4] Because the objective function is convex and monotonic, the conditions of the Theorem of Lagrange are necessary, but also sufficient, so there is no need to verify second-order conditions or check for an interior solution—that is, a solution where the sum of the sample sizes in each group is less than the total maximum sample size.

Operationalizing the Hybrid Sample Frame

With our hybrid sampling approach, we had a sample frame that was composed of two types of elements: propane or wood users in USPS cities with at least 25% propane or wood use, and propane or wood users in disadvantaged communities. We then operationalized the sample frame as the set of homes within those two groups. Although our sampling approach focused on convenience sampling within the disadvantaged communities, we treated them as representative of the population in other disadvantaged communities for the purpose of determining weights because they were drawn from a wide range of communities throughout California. That is, we treated those respondents as if they had been randomly sampled from disadvantaged communities, even though their inclusion in the sample was not random.

Determining Response Weights

Based on this sample frame definition, we weighted our sample based on the four regions within the state and whether it came from the random sampling communities or the targeted communities for the CBO partners. In the cases where the CBO target communities happened to be within a USPS city that was selected for random sampling, we treated responses from the CBO recruitment as coming from the random sampling for determining weights. The logic of the weighting scheme was to replicate the two portions of the sample frame: communities with at least 25% propane or wood heating, and disadvantaged communities. We did not weight on any demographic characteristics of the respondents themselves, partially because we did not have good information on the demographic characteristics of propane or wood users within the target communities (separate from the demographics of the overall communities), and partially to simplify the weighting scheme and avoid situations where responses from individuals received excessively large weights due to small sample sizes within individual region-frame-demographic groups. Thus, we had eight values: four regions, each with a value for the random sampling and the convenience sampling responses.

The total weight for each group was proportional to the number of propane or wood users in that grouping based on the ACS data. For the random sampling proportion, that was straightforward as we had a clear definition that we had developed for the sample design: USPS cities with at least 25% propane or wood space heating; therefore, we summed up the number of homes with propane or wood usage across all of these USPS cities. For the convenience sampling portion, we treated our responses as representative of disadvantaged communities under the California state definition.[5] We determined which Census block groups had CalEnviroScreen[6] percentiles greater than 75 but were not in the random sampling portion of the frame, and summed the number of homes with propane or wood usage across all of these block groups. The total regional weight was then determined as the proportion of propane or wood users from the sample frame in each region, and the weight within each sampling group (random or convenience) within each region was the proportion of propane or wood users in that region and also in the sampling group. We then weighted each observation such that its weight was equal to the total group weight divided by the number of responses in the group. We set the sum of the weights to be equal to the total number of responses (225).

For example, the Southern Coastal region had 22,525 homes that used propane or wood in USPS cities with more than 25% propane or wood usage and 25,667 additional propane or wood users in disadvantaged communities, constituting 11.49% of the propane or wood users in the sample frame across the state, and 11.49% of the total weight, with that spread as 5.37% for the random frame and 6.12% for the convenience frame. We had 20 responses from the random sampling frame and 17 from the convenience sampling frame, so the sample case weights were 0.0537*225/20=0.604 for the random

[5] https://oehha.ca.gov/calenviroscreen/sb535.

[6] https://oehha.ca.gov/calenviroscreen

sampling and 0.0612*225/17=0.810 for the convenience frame. The sample weight breakdowns for all regions are shown in Table 1.

Table 1. Weighting Scheme Breakdown by Region and Frame

Region	Central Valley	Northern Coastal & Sierra	Southern Coastal	Southern Inland
Homes in Random Frame	171,072	110,756	22,525	61,795
Homes in Convenience Frame	18,360	2,369	25,667	6,869
Regional Percent	45.2%	27.0%	11.5%	16.4%
Regional Percent in Random Frame	90.3%	97.9%	46.7%	90.0%
Regional Percent in Convenience Frame	9.7%	2.1%	53.3%	10.0%
Statewide Percent in Random Frame	40.8%	26.4%	5.4%	14.7%
Statewide Percent in Convenience Frame	4.4%	0.6%	6.1%	1.6%
Responses in Random Frame	17	49	20	41
Responses in Convenience Frame	31	28	17	22
Case Weight for Responses in Random Frame	5.40	1.21	0.60	0.81
Case Weight for Responses in Convenience Frame	0.32	0.05	0.81	0.17
Total Weight for Responses in Random Frame	91.77	59.42	12.08	33.15
Total Weight for Responses in Convenience Frame	9.85	1.27	13.77	3.68
Percent of Weight for Responses in Random Frame	40.8%	26.4%	5.4%	14.7%
Percent of Weight for Responses in Convenience Frame	4.4%	0.6%	6.1%	1.6%

Reporting Weighted vs. Unweighted Results

In addition to determining the weighting scheme, we needed to decide how to use it. That is, we needed to decide when we would report weighted results and when we would report unweighted results. The purpose of weighting is to make the reported results reflect the true population characteristics as much as possible. With a true random sample (where every member of the population is equally likely to be in the sample) this is not necessary, but in a stratified sample with imperfect contact lists and non-response, the unweighted results can differ substantially from the true population results. In the case of our study, we were trying to deal with both stratification and imperfect randomness. The stratification came primarily from the regional stratification: we wanted to be able to report results by region and so wanted to have sufficient sample size in each region, and not have proportional sampling that would have had very small samples in the Southern Coastal and Southern Inland regions. While it may seem that the imperfect randomness in our sample came from the use of convenience sampling through partnering with CBOs, this was in fact an attempt to address the more fundamental underlying imperfect randomness that non-response induces. That is, substantial evidence shows that certain groups (particularly people with higher incomes and who are white) have higher likelihood to respond to surveys than other groups (particularly people with lower incomes, people who are non-white, and people for whom English is not their primary language). Our recruitment approach was an imperfect attempt to address that imperfect randomness with an alternate form of imperfect randomness that would end up with a more representative sample.

Based on these factors, we decided to report overall results and results at the regional level with weighted values, but demographic-specific values with unweighted values. That is, when reporting results by income level, ownership status, and primary language, we reported unweighted values. Our rationale was that when reporting results at the regional- or statewide -level, the distortions from non-randomness from stratification and non-response would be substantial and weighted results would help address them. But when reporting results by demographic group, regional differences and recruitment method were less likely to be substantial and weighting with small sample sizes could prove unreliable.

Overview of Results

Based on the survey results, we found that respondents typically use propane appliances because that was the existing equipment when they moved in, and both propane and wood were seen as more affordable than running electric equipment. Looking across all survey respondents, most respondents had heard of heat pumps, although disadvantaged groups (low-income, Native American, those who primarily speak Spanish at home, and those living in mobile or manufactured homes) had lower awareness of heat pumps. Also, all respondents had lower awareness of heat pump water heaters than heat pumps for space heating and cooling.

If propane equipment were no longer available, survey respondents generally preferred heat pumps (described in the survey as having a higher first-cost but lower operating cost compared to a propane furnace) as an alternate to their current heating system, and few preferred electric resistance appliances (described in the survey as having a similar first-cost but higher operating cost, compared to a propane furnace). However, low-income respondents preferred plug-in space heaters and those in the Southern Inland region preferred burning wood. Most respondents were not willing to pay $6,000 extra for heat pump equipment, suggesting that large incentives would be necessary to help cover incremental costs compared to propane equipment, especially for low-income residents. While most respondents were favorable towards zero-emissions equipment if purchase and installation costs were taken care of, some were not, citing concerns such as power outages. Respondents indicated that they value backup or dual-fuel equipment. In addition to lower utility bills, non-energy benefits associated with heat pump equipment resonated with respondents, including the addition of air conditioning, improving safety, and reducing greenhouse gas emissions. Respondents had concerns regarding electricity costs for switching to electric equipment, as well as concerns regarding power outages and reliability.

We also assessed the demographic composition of the two portions of the sample. The breakdown for income level, primary language, race category, home ownership status, and age is shown in Table 2. Across each of these categories, we were able to reach higher percentages of respondents from under-represented groups in our convenience sample than in the random sample. This supports the effectiveness of the approach of coordinating with CBOs to improve the representativeness of our sample.

Table 2. Demographic Comparison of Convenience and Random Sample

	Low-Income	Non-English Speaking	Non-White	Renter	Under 55
Convenience	50%	10%	55%	32%	60%
Random	21%	2%	19%	15%	43%

Conclusions

Survey researchers working in program evaluation and market research face a challenge of declining response rates and declining budgets along with increased attention to the representativeness of the results. These challenges are even larger in cases where the target population does not have a well-defined contact list. In order to address these challenges, we developed a hybrid sampling approach that combines random sampling from traditional approaches like address lists, with convenience sampling by partnering with CBOs. We implemented this approach in a market characterization study of propane and wood users in California.

This approach substantially improved the representativeness of our sample by increasing the number of responses we were able to get from households in disadvantaged communities compared to what we would have achieved with traditional random sampling approaches. None the less, there were substantial challenges. Design and implementation challenges included defining the sampling approach in a way that CBOs could interpret for recruitment while still having demographic data for analysis, determining appropriate sampling quotas, and maintaining quality control over survey responses. Additional analysis challenges included determining weights and the appropriate application of weights for analysis.

We believe this approach has promise but that much more work could be done to advance this technique as a method for improving representativeness and mitigating cost. In particular, future research should work on improving the approach for coordinating between the random and convenience elements and setting appropriate weights for analysis. In particular, setting up a data collection mechanism to determine more directly which source (from random recruitment or CBO recruitment) a response came from and then developing more targeted definitions of the two population frames could help make sure the weighting reflects the underlying populations more directly.

References

U.S. Census Bureau, “American Community Survey 5-Year Estimates: Comparison Profiles 5-Year,” 2022, <http://api.census.gov/data/2022/acs/acs5>

Dalenius, T. and J. L. Hodges 1959. “Minimum Variance Stratification.” *Journal of the American Statistical Association* 54, no. 285 (1959): 88–101. https://doi.org/10.2307/2282141.

Neyman, J. 1934. "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection". Journal of the Royal Statistical Society. 97 (4): 558–625.

The Economist. “As response rates decline, the risk of polling errors rises”, June 22, 2023. https://www.economist.com/united-states/2023/06/22/as-response-rates-decline-the-risk-of-polling-errors-rises.

World Bank 2025. “Number of people without access to electricity”. https://ourworldindata.org/grapher/people-without-electricity-country.