Comparing Two Common Approaches to Within-Household Sampling: A Field Experiment in Costa Rica

Noam Lupu[1], J. Daniel Montalvo[1], Mitchell A. Seligson[1], Elizabeth J. Zechmeister[1], and Kirill Zhirkov[2]

[1] Vanderbilt University

[2] University of Virginia

[Draft. Please don't cite or circulate]

Abstract

We report on an original field experiment that directly compares a quasi-probabilistic method of selecting individuals within households (*last birthday*) to a nonprobabilistic approach that selects respondents according to known parameters on age and gender (*frequency matching*). We evaluate three outcomes: fieldwork efficiency (time) and economy (cost), success in recruiting a representative sample, and differences across a set of attitudinal and behavioral measures. We find that the frequency matching approach performs better on the first dimension and no different than the last birthday method on the second and third. We conclude that researchers' choice of selection methods should be guided by both theoretical benefits and practical tradeoffs.

Keywords: developing countries, frequency matching, last birthday, respondent selection, sampling, survey design

In general population surveys of individuals, researchers follow a probabilistic approach down to the selection of households, whether via a physical address (face-to-face, mail) or telephone (e.g., RDD).[1] At that point, researchers need a method to select survey respondents from within the household.[2] The decision is important as it shapes subsequent estimates (Ziniel, 2008). *In theory*, a pure probability-based approach should be carried through to this final stage.[3] Yet, especially in developing country contexts, two other approaches are more common *in practice*: quasi-probability and nonprobabilistic approaches. According to one assessment, the majority of surveys in developing countries use nonprobabilistic methods for respondent selection within the household, and the second most common approach is a quasi-probability "last birthday" approach in which interviewers select the eligible household member with the most recent birthday (Lupu & Michelitch, 2018). Major comparative opinion projects in developing contexts apply these approaches: the Afrobarometer alternates selecting men versus women within households, a nonprobabilistic method designed to match to known binary gender frequencies; the ArabBarometer's approach is similar, though sometimes uses a last birthday selection; and the AmericasBarometer uses a nonprobabilistic approach in which interviewers select respondents within the household based on known age and gender frequencies.[4]

Regardless of country context, pure probabilistic sampling in the household is often the exception. This approach entails enumerating the household and then selecting an individual at random from that list (Kish, 1949). Many researchers choose, instead, a second option: a quasi-probability approach, such as selecting the person with the last, or next, birthday (Gaziano, 2005). Quasi-probability approaches reduce the burden that pure probability approaches place on interviewees and interviewers, which can lead to suboptimal outcomes such as lower participation rates (see, e.g., Binson et al., 2000; Battagila et al., 2008; Gaziano, 2005;

Jabkowski, 2017). In developing countries, one estimate is that 22% of researchers fielding face-to-face surveys use a last-birthday approach (Lupu & Michelitch, 2018).

What is even more common in developing country contexts is a third approach: nonprobabilistic selection of individuals within households according to known population parameters, typically gender and/or age. By one count, this method is used by 59% of developing country survey practitioners (Lupu & Michelitch, 2018). In such an approach, in each cluster, interviewers typically work off a grid that describes potential participants according to age ranges and/or gender, and recruit one (or more, depending on the design) of each type.

Why is there such frequent departure from approaches that retain some probabilistic element? The answer lies in the challenge of executing (quasi-)probability-based selection in the face of practical constraints that are common in face-to-face surveys fielded in developing country contexts. In addition to the standard challenge of high nonresponse rates, these projects face hurdles such as a lack of pre-existing registers of names and addresses, large family sizes that can complicate designs using a probabilistic mode of selection, and the cost and security risks of maintaining teams in rural hamlets to permit sufficient recontact attempts.

In short, there are two predominant approaches to within-household selection in developing countries: a quasi-probability approach such as a "birthday" method and a nonprobablistic frequency matching approach. Scholars have given a good deal of consideration to variations in outcomes when using a pure probability approach versus a quasi-probability approach, with many concluding that the benefits of the latter can outweigh the drawbacks.[5] However, scholars have yet to compare quasi-probabilistic and nonprobabilistic methods within face-to-face research using random assignment of the method.[6]

To address this gap, we designed a large-scale field experiment to directly compare quasi-probabilistic within-household selection of individuals to a nonprobabilistic within-household approach that selected individuals according to known population frequencies of gender and age. Our decision to focus on these two methods deliberately prioritizes the study of approaches that are frequently used in practice in developing contexts. We carried out the experiment in the Greater Metropolitan Area of San José, the capital of Costa Rica. Within pairs of segments in the same survey, interviewers were instructed to apply in one segment a quasi-probabilistic, last birthday method of within-household selection, with up to twenty recontacts. In the other segment, recruitment followed a gender and age frequency match approach without recontact.

We assess the results along three core dimensions: fieldwork efficiency (time) and economy (cost), success in recruiting a representative sample, and differences in selected attitudinal and behavior estimates derived from each approach. We focus on two types of opinion variables for comparison. The first is a set of personality measures, relevant across a range of disciplines. The second is a set of outcome measures relevant to comparative public opinion research: attitudes and behaviors that are associated with stable and efficient democratic governance. After assessing the three dimensions, we conclude that – on the outcome measures that we are able to assess – the frequency matching method can offer practical advantages that outweigh the theoretical advantages of a quasi-probability approach.

Our results highlight the interplay between the theoretical foundations of survey methods and the distortions that challenges in the real world can introduce. We demonstrate that these distortions can be highly consequential for fieldwork efficiency and data quality, and methods that appear suboptimal in theory can sometimes yield better outcomes in practice.

Selection of Individuals within Households

Probabilistic methods are widely considered the gold standard for within-household selection of respondents in household surveys. Kish (1949) pioneered an approach that involves enumerating all the individuals living in a given household and then randomly selecting a person from that list. The drawbacks of the Kish method are its length and intrusiveness: it increases the time necessary to complete an interview and the probability that individuals refuse to cooperate because of its complexity or concerns about sharing household details (Battagila et al., 2008; Gaziano, 2005; Jabkowski, 2017). In addition, the approach is burdensome to interviewers, which can compound suboptimal outcomes when the method is applied (Binson et al., 2000).

To overcome these problems, methodologists have developed approaches that maintain the arbitrary nature of respondent selection without enumerating all the individuals within the household.[7] Examples of such "quasi-probability" approaches are the next- and last-birthday methods: the first person contacted is asked to name the individual with the upcoming or the most recent (last) birthday, who is automatically selected (Salmon & Nichols, 1983). Technically, these are quasi-probabilistic rather than true probabilistic methods since births are distributed non-randomly and non-uniformly throughout the year (Gaziano, 2008). Survey methodology research shows that, compared to the Kish enumeration approach, next- and last-birthday methods yield little to no differences in substantive responses (Oldendick et al., 1988), while they simultaneously decrease survey length and intrusiveness (O'Rourke & Blair, 1983), and result in fewer dropouts (Binson et al., 2000). There is also evidence that these quasi-probabilistic methods produce response rates that are higher than more intrusive probabilistic approaches and similar to nonprobabilistic selection of any adult member of the household (Battaglia et al., 2008; Salmon & Nichols, 1983). Of the practitioners fielding face-to-face surveys in developing contexts, only 8% reported using a full Kish table, whereas 22% reported using the last birthday method (Lupu & Michelitch, 2018).

Another large class of methods for within-household selection is not based on probabilistic or quasi-probabilistic techniques. Instead, individuals are chosen on the basis of target demographic characteristics, usually age and gender population frequencies derived from the census. There are many different selection methods within this class (e.g., Bryant, 1975; Hagan & Collier, 1983; Paisley & Parker, 1965), often designed to compensate for underrepresentation of young males in most probabilistic sampling approaches. These methods trade off the theorized advantages of random selection within households for lower fieldwork times,[8] higher cooperation rates, and less need for calibration weighting. In one estimate, 59% of developing country practitioners use a frequency match approach (Lupu & Michelitch, 2018).

In theory, the use of a nonprobabilistic approach for the selection of individuals can generate samples that are biased toward respondents with certain personality traits. Researchers speculate, for instance, that nonprobabilistic approaches end up selecting more cooperative individuals who are more interested in the survey topic (Brick, 2011; Clark & Steel, 2007; Koch, 2018). Yet, we lack studies that test whether these theoretical intuitions obtain empirically. If anything, existing research shows that probabilistic methods for selecting respondents within households do not always perform better, and suggests that nonprobabilistic methods can offer substantial cost savings (Cumming, 1990; Czaja et al., 1982; Olson et al., 2014).

Moreover, field research reveals that the theoretically superior probabilistic and quasi-probabilistic selection methods have their own implementation challenges. One of these is related to enumeration of household members or, in a predominant quasi-probabilistic approach, identifying the individual with the most proximate birthday: respondents often have trouble completing this task accurately, which can cause up to a third of respondents to be selected in error (Battaglia et al., 2008; Martin, 1999; Tourangeau et al., 1997). This issue is even more acute in developing contexts where large extended families living in a single household are still common (e.g., McBurney, 1988).[9] Cultural differences can also complicate implementation of probabilistic methods. For instance, in the Middle East, researchers have to match the genders of interviewer and respondent, making random selection impractical (Le et al., 2014).

Further, many factors that Kish (1967) himself acknowledged, such as larger households or multiple families per dwelling, systematically reduce the probability of selecting a poor versus a rich person when the norm of one interview per household is applied (see also Nemeth, 2004).[10] In addition, even with multiple recontacts, samples drawn in this way can end up imbalanced on demographic traits such as gender (Jabkowski, 2017). In brief, all approaches are imperfect when it comes to how they perform in practice, leaving open the question: do the theoretical advantages of probabilistic or quasi-probabilistic approaches for within-household selection translate into better outcomes than nonprobabilistic methods in the field?

Hypotheses

Drawing on the prior discussion and the scholarship that informs it, we posit a set of hypotheses regarding outcomes from nonprobabilistic and quasi-probabilistic selection methods. We specifically focus on two approaches that are commonly used in general population surveys in developing contexts: frequency matching and last birthday. Our hypotheses are as follows:

1. Compared to a last birthday method, the frequency match method yields lower fieldwork time, lower per-interview cost, and higher cooperation rates.
2. Compared to a last birthday method, the frequency match method produces a sample closer to the census-based population on various demographic measures.
3. Compared to individuals selected using a last birthday method, respondents selected using frequency matching will differ in personality traits (e.g., cooperativeness).

4. There are few differences between attitudinal and behavioral estimates when using frequency match versus last birthday methods for within-household selection.

Study Design

Our objective is to test whether, in practice, a face-to-face survey yields different results when the within-household selection of respondents uses a quasi-probabilistic approach (specifically, last birthday) versus a frequency matched nonprobabilistic approach. Our focus is on research in developing contexts, where nonprobabilistic within-household approaches are common (Lupu & Michelitch, 2018). While in theory it could be beneficial to include a pure probability approach within the study design, we opted to focus our study on two commonly-applied approaches in general population survey research in developing contexts: quasi-probability (last birthday) and nonprobability (frequency matching). We implemented a large-scale field experiment as part of a survey study in Costa Rica, a middle-income country in Central America.[11] Average household size in Costa Rica in 2011 (the most recent census) was 3.46, which is typical for a middle-income country. We selected Costa Rica because it has detailed census information and a strong capacity for high-quality survey research. In addition, we have years of experience and an extensive network of research partners in the country.

The study was as an area probability sample of 900 adults within the voting age population of the Greater Metropolitan Area of San José (GMASJ), administered in September–October 2018.[12] The core of the experiment is random assignment of census segments to one of two *within-household* sampling methods: (1) a frequency match approach in which individuals are selected to match known distributions on age and gender, and (2) a quasi-probabilistic approach in which the individual with the most recent birthday is selected. The first approach – the Frequency Match (FM) approach – is a quota-based approach, but must be distinguished from quota-based approaches that sample from “flow points” (e.g., shopping malls); rather, our sample is probabilistic down to the household, at which point individuals are selected to match gender and age frequencies in the population. The second method – the Last Birthday (LB) approach – selects individuals using a quasi-probabilistic approach, is considered by many to be a reasonable substitute for the gold-standard Kish method (e.g., Gaziano, 2005).

The sampling frame was the list of municipalities (*cantones*), districts (*distritos*), census segments, and maps for the GMASJ from the 2011 census provided by the Centro Centroamericano de Población. Within the GMASJ, a municipality was a primary sampling unit (PSU). We selected PSUs using two criteria. First, municipalities with more than 100,000 inhabitants were self-selected; that is, the selection probability for those municipalities equaled one. We selected the remaining municipalities applying probability proportional to estimated size (PPeS). Following this design, we selected eight PSUs within the GMASJ: Alajuelita, Aserrí, Desamparados, Escazú, Goicochea, Montes de Oca, San José, and Tibás.[13]

In the next stage, we selected census segments with PPeS, after urban/rural stratification. These segments constituted the secondary sampling unit (SSU). To avoid contamination of the experiment at the household level (i.e., houses on the same block falling into both conditions), we assigned the treatment at the SSU level. We first determined the FM segments, by selecting at random 75 census segments from the 7,703 segments in the selected municipalities. We fixed the number of selected households within each segment to six.[14] Households were chosen using systematic selection, with enumerators beginning at the northeast corner of the segment, moving clockwise, skipping one house after each completed interview and selecting one respondent per household. To implement the frequency match, enumerators used a matrix of three age cohorts (18–29, 30–46, 47+) and gender (male/female) categories to select individuals so that the sample matched known frequencies on these characteristics as established by census data.[15] Enumerators were free to complete the matrix in any order and selected households were not revisited. If the enumerator could not secure an interview at a given household (due to refusals, ineligibility, or no one at home), the enumerator moved on to the next household.

We next selected the LB segments, using a probabilistic approach designed to effectively match these segments to the FM segments. To do this, we identified all the segments that were contiguous to the selected FM segments. We sorted these neighboring segments by size and discarded segments with insufficient numbers of households.[16] We then randomly selected one segment from the pool of neighbors for each FM segment, yielding a total of 75 LB segments. Table S1 in Supplementary Material shows the number of neighboring segments by district and the final number of segments selected for the LB condition.

In the LB condition, enumerators were also instructed to complete six interviews in each segment and to designate the northeast corner as the starting point. In these cases, the enumerator approached the first house and asked to interview the adult with the most recent (last) birthday, thus establishing the conditions for quasi-probabilistic selection within the household. As in the FM approach, if successful, the enumerator was instructed to skip one residential unit before approaching the next one. If no one was home or the selected individual was not available, the enumerator was instructed to revisit the household up to nine more times (for a total of ten attempts). If the selected individual declined or remained unreachable after all the attempts were exhausted, the enumerator moved to the adjacent dwelling unit and repeated the same protocol. If still unsuccessful, no additional substitutions were permitted, meaning that segments with fewer than six completed interviews were possible so long as either both attempts resulted in refusals or the maximum number of attempts was exhausted.[17]

A highly regarded and experienced local survey firm carried out the interviews. To minimize interviewer effects, the same enumerators were assigned to each segment within any given FM–LB pair.[18] A total of 22 interviewers were deployed in the LB condition and 27 in the FM condition, but 4 FM interviewers completed only one interview each. All interviewers were trained by the research team and their office personnel were also trained in a rigorous quality control program. Computer-assisted personal interviewing permitted capturing dozens of quality control markers including location, voice recordings, and time stamps. Every interview was audited by the local team; approximately one quarter were re-audited by the research team.[19]

Comparing Methods of Selection

We begin our analysis by documenting the key indicators of survey fieldwork efficiency for the FM and LB selection methods: numbers of complete interviews, response rates, and fieldwork time and cost. The comparisons are presented in Table 1. Importantly, the FM method resulted in more than twice as many complete interviews as the LB condition: 451 vs. 220. However, the response rates (AAPOR RR1) for the two methods are similar. The higher number of completes for the FM method is achieved by making many more interview attempts.[20]

[Table 1 about here]

Consistent with what prior studies suggest will occur, implementing the LB selection method substantially increases fieldwork time. In our study, fieldwork was complete in 29 days for the FM condition and 49 days with the LB method. These differences are even more stark when we calculate time per completed interview: the FM approach (0.06 days per interview) was far more time-efficient than the LB method (0.22 days per interview). The difference is most likely due to the time interviewers in the LB condition spent on repeated contact attempts.[21]

The result of this additional fieldwork time is that survey costs also increase roughly in proportion. We do not calculate the precise difference in cost between the two methods because we do not have detailed information on the local firm's pay scale; however, an estimate provided by the local survey firm that implemented the fieldwork informs us that each interview in the LB condition was over four times more expensive than an interview using the FM approach.

How do these two approaches compare in terms of the representativeness of the samples they produce? We compare the LB and FM conditions by contrasting both unweighted and weighted sample means to the available census data. We developed the calibration weights by raking over known population distributions of gender and age categories (18–30, 31–45, 46+),[22] and we account for the clustered sample design. In assessing the samples, we examine self-reported years of formal education. We also examine wealth using a measure based on a battery of items on whether the household possesses certain essentials, a common measure of socioeconomic status in developing contexts (see Filmer & Pritchett 2001). According to a factor analysis, five items form a wealth index: internet access, computer (either a desktop or a laptop), at least one cell phone, car, and LCD TV.[23] We calculated the additive total of these indicators, so that wealth ranges from zero (household has none of the items) to 5 (all items).

Table 2 presents descriptive data on the samples that the FM and LB methods produced. Note that the census data used for this analysis include only those GMASJ municipalities that were also included in the survey sampling frame. By design, the FM sample is extremely close to the census benchmarks on age and gender without weights. As anticipated by prior research, the LB sample underrepresents men and younger people, requiring calibration weights to make the quasi-probabilistic sample data demographically representative of the population. Interestingly, the samples are close to the census benchmark on education, with or without weights.

[Table 2 about here]

It is also the case that households surveyed in both the LB and FM samples are more affluent than the corresponding census data would suggest. We cannot say for sure what may have generated that outcome, but there are two probable causes. One is that it results from less affluent individuals being less willing or available to respond to surveys. This tendency to over-sample individuals higher in socioeconomic status has been documented in the United Kingdom (Sturgis et al., 2018), and in the United States (AAPOR, 2017), among others. Another possibility is an overall increase in wealth in the region from 2011 (the census year) to 2018 (when our data were collected). Evidence from LAPOP's AmericasBarometer shows that, on the same 0–5 index, average household wealth in San José, Costa Rica increased from 2.38 in 2010 and 2012 to 3.73 in 2018 ($\Delta = 1.35$, $p < .001$). Of course, these are survey data, so we cannot be sure that true increases in wealth are driving the difference between our samples and the 2011 census. Still, when it comes to basic demographics, both the LB (although only with weighting) and FM methods are comparable in their capacity to produce representative samples.

Even if the two approaches both yield samples that are demographically similar, the sampled respondents may have different personality dispositions, biasing estimates on attitudinal measures. To test this, we assess differences in key psychological indicators across the two sampling methods. In measuring personality traits, we draw on the Big Five model that has been shown to impact both survey response style and substantive preferences (Hibbing et al., 2019; Valentino et al., 2020). We also compare the samples on political attitudes and behaviors that, according to the literature, are particularly important to the quality and endurance of democratic governance. This decision reflects our own area of expertise, in the field of comparative public opinion, which dictated the contents of the questionnaire used in the experiment.

The first of the political measures is political participation (e.g., Verba et al., 1995): self-reported voting and a composite index for other forms of civic engagement.[24] Second, symbolic ideology helps citizens organize their political preferences (Knutsen, 1997). Ideological labels in Latin America are known to depend on context (Zechmeister, 2006), making it interesting to see whether they differ across the two samples. Third, we estimate differences in political orientations that many believe are conducive to liberal-democratic regimes (Almond & Verba 1963; Inglehart & Welzel 2005; Lipset 1959; Putnam 1993). Specifically, we compare support for democracy (Mattes, 2018), political trust (Levi & Stoker 2000),[25] and political tolerance (Sullivan & Transue, 1999). See Table 3 for the short descriptions of the survey items.

[Table 3 and Table 4 about here]

Table 4 presents the results, with all the variables normalized to the same 0–100 scale for ease of comparison. These estimates use calibration weights for both FM and LB samples to account for the demographic skew in the latter. Overall, we find no significant differences between the two samples on *any* of the analyzed indicators. Note that the differences are also extremely small in terms of magnitudes: the largest absolute difference found, the one on the Big Five trait conscientiousness, is only −2.06 units within the possible range from −100 to 100.

Conclusion

Area probability samples using probabilistic or quasi-probabilistic approaches to selecting individuals within households are difficult to draw in developing contexts. In addition to the omnipresent challenge of high nonresponse rates, such studies run up against the lack of registers with which to select and pre-contact households. They also confront elevated expenses and security risk associated with maintaining teams in the same location long enough to make numerous in-person recontacts. It is critical that survey practitioners and data users take these constraints seriously. When researchers insist on practices that are optimal in theory, but excessively challenging in practice, the actual methods applied in the field may be inconsistent and irreproducible if enumeration teams deviate from protocols and, worse, obscure their work.

Most survey researchers working on major social science projects in developing contexts recognize this challenge, and use some form of frequency matching as a practical recourse. One estimate is that 59% of investigators on such projects use frequency matching, very few use quasi-probabilistic approach, and even fewer use a pure probability method for within-household selection (Lupu & Michelitch, 2018). Yet, this choice has been something of a leap of faith, with no prior study systematically putting the question to a test: does this deviation from the theoretical ideal affect outcomes? Our study was designed to address this question directly.

Practically speaking, no survey sample is perfect. For instance, both samples have relatively low response rates (just under 20%). In that sense, samples drawn in Costa Rica seem to be subject to the same types of challenges that are present in surveys around the world (e.g., among many, Curtin et al., 2005; Singer, 2006; Smith, 1995). And yet, our results show that a frequency match approach is not only more efficient in time and money, but also produces a sample that is comparable to what one can gather when applying a quasi-probabilistic approach to within-household selection in combination with calibration weights.

We cannot identify with certainty the specific reasons behind the drastic differences we observe in efficiency between the FM and LB methods. The most likely explanation is that interviewers in the LB condition spend more time recontacting selected respondents, whereas the interviewers in the FM condition have more discretion. In theory, such discretion can introduce biases to estimates, if interviewers select individuals within households who are systematically different from the general population. However, our results reveal no such biases, at least on the variables we observe, across the two selection methods.

Our design is not without limitations. We did not include a true probabilistic selection method, such as the Kish grid. Such methods are rarely used in developing contexts because they are known to increase fieldwork time and reduce cooperation (Battaglia et al., 2008; Binson et al., 2000; O'Rourke & Blair, 1983; Salmon & Nichols, 1983). Yet, this means we cannot know whether a probabilistic selection method might have yielded different results.

Second, our selection mechanism was not perfectly random at the household level. Substitution of non-responding households was allowed, meaning that the size of the gross sample was not defined in advance. Practically, the final gross sample in the FM condition turned out to be twice as large as the gross sample in the LB condition. Therefore, differences between the two conditions cannot unambiguously be attributed to within-household selection methods alone. Implementing our experiment more perfectly would have meant, first, constructing a sample of households and, then, randomly assigning each household in this predefined sample to either the FM or LB condition. However, implementing this design in a country like Costa Rica would have been prohibitively costly. Such an experiment would also have deviated substantially from the practices actually employed by field researchers.

Even with these limitations in mind, we draw two conclusions. First, it is critical that both theory and practice guide survey design. While theory can tell which methods are preferable in the abstract, the tradeoffs presented in practice must also direct choices in the field. Second, in practice, frequency matching approaches to within-household sampling can yield better fieldwork outcomes, especially in developing countries: shorter fieldwork times, lower costs, and samples comparable to those generated via quasi-probabilistic methods. Quasi-probability within-household selection can be costly and time inefficient while, at least in some cases, this approach may yield no detectable gains over other widely-used alternative field methods.

References

AAPOR. (2017). An evaluation of 2016 election polls in the U.S. https://www.aapor.org/Education-Resources/Reports/An-Evaluation-of-2016-Election-Polls-in-the-U-S.aspx

Almond, G. A., & Verba, S. (1963). *The civic culture: Political attitudes and democracy in five nations*. Princeton University Press.

Battaglia, M. P., Link, M. W., Frankel, M. R., Osborn, L., & Mokdad, A. H. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly*, *72*(3), 459–469. https://doi.org/10.1093/poq/nfn026

Binson, D., Canchola, J. A., & Catania, J. A. (2000). Random selection in a national telephone survey: A comparison of the Kish, next-birthday, and last-birthday methods. *Journal of Official Statistics, 16*(1), 53–59.

Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly, 75*(5), 872–888. https://doi.org/10.1093/poq/nfr045

Bryant, B. E. (1975). Respondent selection in a time of changing household composition. *Journal of Marketing Research*, *12*(2), 129–135. https://doi.org/10.1177/002224377501200201

Clark, R. G., & Steel, D. G. (2007). Sampling within households in household surveys. *Journal of the Royal Statistical Society*. *Series A: Statistics in Society*, *170*(1), 63–82. https://doi.org/10.1111/j.1467-985X.2006.00434.x

Cumming, R. G. (1990). Is probabilistic sampling always better? A comparison of results from a quota and a probabilistic sample survey. *Community Health Studies*, *14*(2), 132–137. https://doi.org/10.1111/j.1753-6405.1990.tb00033.x

Curtin, R., Presser, S., & Singer, E. (2005). Changes in telephone survey nonresponses over the past quarter century. *Public Opinion Quarterly, 69*(1), 87–98. https://doi.org/10.1093/poq/nfi002

Czaja, R., Blair, J., & Sebestik, J. P. (1982). Respondent selection in a telephone survey: A comparison of three techniques. *Journal of Marketing Research, 19*(3), 381–385. https://doi.org/10.1177/002224378201900312

Filmer, D., & Pritchett, L. H. (2001). Estimating wealth effect without expenditure data or tears: An application to educational enrollments in states of India. *Demography, 38*(1): 115–132. https://doi.org/10.1353/dem.2001.0003

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly, 69*(1), 124–157. https://doi.org/10.1093/poq/nfi006

Gaziano, C. (2008). Last-birthday selection. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 417–418). Sage. https://doi.org/10.4135/9781412963947.n267

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37*(6), 504–528. https://doi.org/10.1016/S0092-6566(03)00046-1

Guignard, R., Wilquin, J.-L., Richard, J.-B., & Beck, F. (2013). Tobacco smoking surveillance: Is quota sampling an efficient tool for monitoring national trends? A comparison with a random cross-sectional survey. *PLOS One, 8*(10), Article e78372. https://doi.org/10.1371/journal.pone.0078372

Hagan, D. E., & Collier, C. M. (1983). Must respondent selection procedures for telephone surveys be invasive? *Public Opinion Quarterly, 47*(4), 547–556. https://doi.org/10.1086/268811

Hibbing, M. V., Cawvey, M., Deol, R., Bloeser, A. J., & Mondak, J. J. (2019). The relationship between personality and response patterns on public opinion surveys: The Big Five, extreme response style, and acquiescence response style. *International Journal of Public Opinion Research, 31*(1), 161–177. https://doi.org/10.1093/ijpor/edx005

Inglehart, R., & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790881

Jabkowski, P. (2017). A meta-analysis of within-household selection impact on survey outcome rates, demographic representation and sample quality in the European Social Survey. *Research & Methods, 26*(1), 31–60. https://doi.org/10.18061/1811/81932

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association, 44*(247), 380–387. https://doi.org/10.1080/01621459.1949.10483314

Kish, L. (1967). *Survey sampling* (2nd ed.). John Wiley & Sons.

Knutsen, O. (1997). The partisan and the value-based component of left–right self-placement: A comparative study. *International Political Science Review, 18*(2), 191–225. https://doi.org/10.1177/019251297018002005

Koch, A. (2018). Within-household selection of respondents. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methodology: Multinational, multiregional, and multicultural contexts* (pp. 139–157). John Wiley & Sons. https://doi.org/10.1002/9781118884997.ch5

Le, K. T., Brick, J. M., Diop, A., & Alemadi, D. (2013). Within-household sampling conditioning on household size. *International Journal of Public Opinion Research,* 25(1), 108–118. https://doi.org/10.1093/ijpor/eds008

Le, K. T., Diop, A., Wittrock, J., Al-Emadi, D., & Elawad, E. (2014). Gender prespecified sampling for cost control. *International Journal of Public Opinion Research, 26*(4), 441–452. https://doi.org/10.1093/ijpor/edt035

Levi, M., & Stoker, L. (2000). Political trust and trustworthiness. *Annual Review of Political Science, 3*, 475–507. https://doi.org/10.1146/annurev.polisci.3.1.475

Lipset, S. M. (1959). Some social requisites of democracy: Economic development and political legitimacy. *American Political Science Review, 53*(1), 69–105. https://doi.org/10.2307/1951731

Lupu, N., & Michelitch, K. (2018). Advances in survey methods for the developing world. *Annual Review of Political Science, 21*, 195–214. https://doi.org/10.1146/annurev-polisci-052115-021432

Martin, E. (1999). Who knows who lives here? Within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly, 63*(2), 220–236. https://doi.org/10.1086/297712

Mattes, R. (2018). Support for democracy. In *Oxford research encyclopedia of politics*. https://doi.org/10.1093/acrefore/9780190228637.013.622

McBurney, P. (1988). On transferring statistical techniques across cultures: The Kish grid. *Current Anthropology, 29*(2), 323–325. https://doi.org/10.1086/203642

Morris, S. D., & Klesner, J L. (2010). Corruption and trust: Theoretical considerations and evidence from Mexico. *Comparative Political Studies, 43*(10), 1258–1285. https://doi.org/10.1177/0010414010369072

Nemeth, R. (2004). Representativeness problems inherent in address-based sampling and a modification of the Leslie Kish grid. *Bulletin of Sociological Methodology, 83,* 43–60. https://doi.org/10.1177/075910630408300105

Oldendick, R. W., Bishop G. F., Sorenson, S. B., & Tuchfarber, A. J. (1988). A comparison of the Kish and last birthday methods of respondent selection in telephone surveys. *Journal of Official Statistics, 4*(4), 307–318.

Olson, K., Stange, M., & Smyth, J. (2014). Assessing within-household selection methods in household mail surveys. *Public Opinion Quarterly, 78*(3), 656–678. https://doi.org/10.1093/poq/nfu022

O'Rourke, D., & Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research, 20*(4), 428–432. https://doi.org/10.1177/002224378302000409

Paisley, W. J., & Parker, E. B. (1965). A computer-generated sampling table for selecting respondents within households. *Public Opinion Quarterly, 29*(3), 431–436. https://doi.org/10.1086/267343

Putnam, R. D. (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton University Press.

Rizzo, L., Brick, J. M., & Park, I. (2004). A minimally intrusive method for sampling persons in random digit dial surveys. *Public Opinion Quarterly, 68*(2), 267–274. https://doi.org/10.1093/poq/nfh014

Salmon, C. T., & Nichols, J. S. (1983). The next-birthday method of respondent selection. *Public Opinion Quarterly, 47*(2), 270–276. https://doi.org/10.1086/268785

Singer, E (2006). Introduction: Nonresponse bias in household surveys. *Public Opinion Quarterly, 70*(5), 637–645. https://doi.org/10.1093/poq/nfl034

Smith, T. W. (1995). Trends in non-response rates. *International Journal of Public Opinion Research,* 7(2), 157–171. https://doi.org/10.1093/ijpor/7.2.157

Sturgis, P., Kuha, J., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Lauderdale, B. E., & Smith, P. (2018). An assessment of the causes of the errors in the 2015 UK general election opinion polls. *Journal of the Royal Statistical Society: Series A, 181*(3), 757–781. https://doi.org/10.1111/rssa.12329

Sullivan, J. L., & Transue, J. E. (1999). The psychological underpinnings of democracy: A selective review of research on political tolerance, interpersonal trust, and social capital. *Annual Review of Psychology,* 50, 625–650. https://doi.org/10.1146/annurev.psych.50.1.625

Tourangeau, R., Shapiro, G., Kearney, A., & Ernst, L. (1997). Who lives here? Survey undercoverage and household roster questions. *Journal of Official Statistics, 13*(1), 1–18.

Valentino, N. A., Zhirkov, K., Hillygus, D. S., & Guay, B. (2020). The consequences of personality biases in online panels for measuring public opinion. *Public Opinion Quarterly, 84*(2), 446–468. https://doi.org/10.1093/poq/nfaa026

Verba, S., Schlozman, K. L., & Brady, H. E. (1995). *Voice and equality: Civic voluntarism in American politics*. Harvard University Press.

Zechmeister, E. (2006). What's left and who's right? A Q-method study of individual and contextual influences on the meaning of ideological labels. *Political Behavior, 28*(2), 151–173. https://doi.org/10.1007/s11109-006-9006-5

Ziniel, S. (2008). Within-unit coverage error. In P. J. Lavrakas (Ed.), *Encyclopedia of survey research methods* (pp. 962–964). Sage. https://doi.org/10.4135/9781412963947.n635

Footnotes

[1]As exceptions, many online and mobile phone studies target individuals without consideration of household.

[2]This is the case at least in our area of research, comparative public opinion, where individuals are key because researchers are more interested in the behaviors and attitudes of individuals (e.g., their vote choice) than in household-level variables (e.g., aggregate income).

[3]This is because meaningful sampling errors can be calculated when each eligible person in each household has a known and nonzero probability of being included in the sample. In addition, arbitrary (nonrandom) selection of respondents within households can lead to significant biases if people who are first or most willing to respond are systematically different (e.g., more cooperative) compared to the rest of the population.

[4]https://afrobarometer.org/surveys-and-methods/sampling-principle; https://www.arabbarometer.org/survey-data/methodology/; https://www.vanderbilt.edu/lapop/ab2016/AmericasBarometer_2016-17_Sample_Design.pdf.

[5]The application of a pure probabilistic approach carries a burden that tends to decrease cooperation rates and increase drop-offs, though it is an open question whether that comparatively less ideal outcome is driven by the interviewees or the interviewers (see, e.g., Binson et al., 2000; Gaziano, 2005).

[6]Guignard et al. (2013) compare random selection to quota selection of respondents across two phone surveys, but the sampling frames differ across the two studies.

[7]There are also within-household selection methods that combine probabilistic and quasi-probabilistic elements (e.g., Le et al., 2013; Rizzo et al., 2004).

[8]And generally, more efficient fieldwork efforts, given that recontacts are eliminated.

[9]Even though this concern was first expressed more than 30 years ago, differences in household size and composition between developed and developing countries still persist. For instance, average household size in 2015 ranged from 2.22 in France to 8.32 in Senegal according to the most recent available data from the UN Population Division (https://www.un.org/development/desa/pd/data/household-size-and-composition).

[10]Kish (1967) notes that, “the proper fractional representation of each adult is approximated closely with only the eight tables” (p. 399).

[11]In 2019, the average GDP per capita for the 17 Latin American countries for which the World Bank has data was $8,105 USD; in Costa Rica this figure was $12,238.

[12]As is standard in survey research, the sample excludes people in institutionalized settings like boarding schools, hospitals, military barracks, and the like. We focus on the GMASJ to increase control over the design and implementation of the experiment while decreasing the costs associated with the study. The study had IRB approval from the authors’ institution.

[13]San José was the only self-selected municipality, subdivided into two PSUs.

[14]Fixing the number of households within segments allows us to have an evenly distributed and dispersed number of interviews per segment. For the study, a household is considered to be a group of people who eat their meals together. In apartment buildings, enumerators identify a starting point and select households on each floor, skipping one unit after each completed interview. If enumerators do not complete the interviews on one floor, they must continue on the next floor, either upstairs or downstairs.

[15]The use of a binary categorization of gender is common in survey research, reflecting known parameters available in most census statistics, especially in developing country contexts. More inclusive approaches are not yet established as the norm in area probability sampling that

applies frequency matching methods, but we see this as a frontier the discipline should address.

[16]The information on the size of each segment as well as the cartography came from the 2011 census. Segments varied substantially in terms of the number of households: while some segments had as few as four households, others had as many as 286. The decision to remove segments with an insufficient number of households was taken in order to avoid major obstacles during fieldwork. Twenty-one (out of the 214 neighbor segments) had fewer than 10 dwellings and were not included as candidates in the selection of segments for the LB condition.

[17]If additional substitutions were permitted, the method would depart from the intention of the design and move closer in practice to the nonprobabilistic approach.

[18]Our review of location markers confirmed that enumerators were following protocols for the selection of households, but also revealed an unanticipated deviation. This deviation occurred because of incorrect spatial geolocations for the selected FM segments. Due to a miscommunication with the fieldwork team, those spatial geolocations were catalogued using a coding scheme that matched the prior census (conducted in 2000) rather than the 2011 census. The LB segments were selected using the more recent census and, because of updates in the intervening years, not all these segments were contiguous. This deviation is important to note but is orthogonal to the objective of the study. Assignment was still random. Moreover, we find no significant demographic differences between LB segments that are closer to versus farther away from the originally planned locations. See Table S2 in Supplementary Material.

[19]Interviewees were told parts of the interview would be recorded for quality control. Based on this quality control process, a small number of interviews were canceled (5.1%), mostly because of errors in the reading of questions; canceled interviews were replaced in real time by sending the team back to the same segments with the assigned sampling protocol.

[20]"Completes" are the raw numbers of obtained interviews (while response rates reflect the proportion of completes out of all attempts). Interviewers in FM segments made more attempts because they could knock on many doors in order to find a cooperative respondent with matching age and gender. In LB segments, in contrast, interviewers made fewer attempts because they had to try repeatedly to contact the same households.

[21]Analyzing interview duration confirms this conjecture. We compare the amount of time interviewers spent with respondents (net time) to the amount of time that passed from the first contact attempt until the interview was completed (gross time). Net interview times are approximately the same across the two conditions: 0.75 hours in the FM condition and 0.72 hours in the LB condition ($\Delta = 0.03$, $p = 0.073$). However, gross interview times are very different: 0.92 hours in the FM condition and 42.36 hours in the LB condition ($\Delta = 41.44$, $p < 0.001$). Of course, interviewers in the LB condition could complete other interviews between recontact attempts, so these figures may understate their efficiency. Still, these results strongly suggest that the difference in time efficiency between the FM and LB conditions is primarily due to interviewers in the LB condition spending time recontacting selected respondents.

[22]The parameters were taken from the 2011 Costa Rica census.

[23]See Table S3 in Supplementary Material for the results of the factor analysis.

[24]The indices are calculated as the simple average of the constituent items. Reliability analyses for the index variables are presented in Table S4 in Supplementary Material.

[25]Following the common practice in comparative political research, including in Latin America (e.g., Morris & Klesner, 2010), we calculate the trust index using only explicitly political or government institutions. Reported results do not change if a broader index that includes trust in civic institutions like the media and the Catholic Church is used instead.

Table 1

Comparison of Response Rates and Fieldwork Time

	FM	LB
Interviews		
Total initial attempts	2,435	1,213
Refusals and breakoffs	435	406
Other incompletes	1,549	587
Completes	451	220
AAPOR RR1	19.6%	18.2%
Fieldwork time (days)		
Total	29	49
Average per completed interview	0.06	0.22

Note. Fieldwork time is defined as the number of days elapsed from the first to the last day of data collection.

Table 2

Comparison of FM and LB Sample Representativeness

	Unweighted		Weighted		2011
	FM	LB	FM	LB	Census
Female	0.501	0.582	0.522	0.522	0.522
	(0.023)	(0.033)	(0.008)	(0.023)	
Age	40.8	47.6	40.3	41.1	39.8
	(0.82)	(1.28)	(0.40)	(1.73)	
Education	10.7	10.7	10.7	10.7	10.3
	(0.21)	(0.33)	(0.54)	(0.56)	
Wealth index	3.80	3.60	3.81	3.65	2.92
	(0.06)	(0.09)	(0.07)	(0.10)	

Note. Standard errors in parentheses

Table 3

Short Descriptions of the Personality and Political Variables

Variable	Description
Big Five personality traits	Ten-item personality inventory (TIPI; Gosling, Rentfrow, and Swann 2003)
Voted last election	Respondent voted in the last presidential election
Political participation index	During the last campaign, respondent (1) put a party flag, (2) used a sticker, (3) did door-knocking, (4) donated to a candidate, (5) attended a political meeting
Ideology	Scale from 1 = *Left* to 10 = *Right*
Support for democracy	Respondent agrees that democracy is better than other forms of government
Political trust index	Respondent trusts (1) the Congress, (2) the police, (3) political parties, (4) the President, (5) the Supreme Court, (6) the Fourth Chamber [of the Supreme Court], (7) the local government, (8) elections
Political tolerance index	Respondent thinks that political dissidents should be allowed to (1) vote, (2) conduct peaceful demonstrations, (3) run for public office, (4) appear on television

Note. See Table S5 in Supplementary Material for the exact formulations of all questions and responses.

Table 4

Comparison of Key Estimates

	FM	LB	Difference
Big Five personality traits			
Openness to experience	75.0	75.1	0.14
			(1.99)
Conscientiousness	79.4	77.4	−2.06
			(1.57)
Extraversion	73.6	73.6	−0.04
			(1.84)
Agreeableness	67.5	69.1	1.60
			(2.55)
Emotional stability	65.2	64.2	−1.02
			(1.95)
Political variables			
Participation			
Voted last election	72.9	71.2	−1.76
			(5.01)
Political participation index	10.5	9.5	−0.96
			(1.46)
Ideology (right)	48.7	49.6	0.98
			(2.88)
Pro-democratic values			
Support for democracy	74.4	73.6	−0.83
			(2.66)
Political trust index	47.2	47.6	0.33
			(1.55)
Political tolerance index	54.6	54.4	−0.24
			(2.84)

Note. Weighted estimates. Standard errors in parentheses. All variables normalized to 0–100 scale. None of the estimated differences are statistically significant at the 95% level.

Supplementary Material

Table S1

Last Birthday Segment Selection

	FM neighboring segments	
Municipality	Total	Selected
Alajuelita	24	8
Aserrí	21	8
Desamparados	25	9
Escazú	24	8
Goicoechea	25	8
Montes de Oca	24	8
San José	51	18
Tibás	20	8
Total	214	75

Table S2

Comparison of LB Segments by Closeness to Originally Planned Locations

	Closer	Farther	Difference
Female	0.48	0.55	0.07
			(0.08)
Age	40.79	41.77	0.98
			(2.32)
Education	11.42	9.89	−1.53
			(1.13)
Wealth index	3.78	3.45	−0.32
			(0.24)

Note. Distances between centroids of originally planned and actually surveyed LB segments. Closer = distance less than median. Farther = distance greater then median. Weighted estimates. Standard errors (for difference estimates only) in parentheses. None of the estimated differences are significant on the 95% confidence level.

Table S3

Wealth Index: Results of Factor Analysis

Variable	Factor loading
Internet	0.88
	(0.04)
Computer	0.86
	(0.04)
Cellphone	0.71
	(0.09)
Car	0.63
	(0.05)
LCD TV	0.60
	(0.06)

Note. Estimator = weighted least squares adjusted for means and variances (WLSMV). $\chi^2 = 6.45$, df = 5, p = 0.265; RMSEA = 0.021; CFI = 0.998. Geomin rotation. Standard errors in parentheses. All factor loadings significant at the 99.9% confidence level.

Table S4

Reliability Statistics for Multi-item Indices

	No. of items	Cronbach's alpha
Political trust	10	0.87
Political participation	5	0.68
Political tolerance	4	0.80

Table S5

Wording of Survey Items [Translated from Spanish]

Big Five personality traits

"Now I am going to read you a series of personality traits that may or may not apply to you. Please use the 1–7 ladder to indicate the extent to which you agree or disagree that these statements apply to you. You should rate the extent to which the pair of traits applies to you, even if one characteristic applies more strongly than the other."

Openness to experience

- A person [who is] open to new experiences and intellectual
- An uncreative and unimaginative person (reversed)

Conscientiousness

- A dependable and self-disciplined person
- A disorganized and careless person (reversed)

Extraversion

- A sociable and active person
- A quiet and shy person (reversed)

Agreeableness

- A critical and quarrelsome person (reversed)
- A generous and warm person

Emotional stability

- An anxious and easily upset person (reversed)
- A calm and emotionally stable person

Answers coded from 1 = "Strongly disagree" to 7 = "Strongly agree"

Voted last election

"Did you vote in the last presidential elections?"

Answers: 1 = "Yes" and 2 = "No"

Political participation index

"During the last campaign, did you…"

1. Put a flag of a political party on your house
2. Stick a political campaign stick on your car or home
3. Visit houses to take voters to the polls
4. Contribute money to help a candidate
5. Attend a political meeting or demonstration

Answers: 1 = "Yes" and 0 = "No"

Ideology

"On this card there is a 1–10 scale that goes from left to right. The number one means left and 10 means right. Nowadays, when we speak of political leanings, we talk of those on the left and those on the right. In other words, some people sympathize more with the left and others with the right. According to the meaning that the terms 'left' and 'right' have for you, and thinking of your own political leanings, where would you place yourself on this scale? Tell me the number."

Answers coded from 1 = "Left" to 10 = "Right"

Support for democracy

"…Democracy may have problems, but it is better than any other form of government. To what extent do you agree or disagree with this statement?"

Answers coded from 1 = "Strongly disagree" to 7 = "Strongly agree"

Political trust index

1. To what extent do you trust the National Congress?
2. To what extent do you trust the National Police?
3. To what extent do you trust the political parties?
4. To what extent do you trust the President?
5. To what extent do you trust the Supreme Court?
6. To what extent do you trust the Fourth Chamber [of the Supreme Court]?
7. To what extent do you trust the local or municipal government?
8. To what extent do you trust elections in this country?

Answers coded from 1 = “Not at all” to 7 = “A lot”

Political tolerance index

1. There are people who only say bad things about the country form of government, not just the current government but the system of government. How strongly do you approve or disapprove of such people’s right to vote?
2. How strongly do you approve or disapprove that such people be allowed to conduct peaceful demonstrations in order to express their views?
3. Still thinking of those who only say bad things about the country form of government, how strongly do you approve or disapprove of such people being permitted to run for public office?
4. How strongly do you approve or disapprove of such people appearing on television to make speeches?

Answers coded from 1 = “Strongly disapprove” to 7 = “Strongly approve”