2019

Within-Household Selection Methods: A Critical Review and Experimental Examination

Jolene Smyth
University of Nebraska-Lincoln, jsmyth2@unl.edu

Kristen M. Olson
University of Nebraska - Lincoln, kolson5@unl.edu

Mathew Stange
Mathematica Policy Research, Ann Arbor, MI, mstange@mathematica-mpr.com

Within-Household Selection Methods: A Critical Review and Experimental Examination

Jolene D. Smyth,[1] Kristen Olson,[1] and Mathew Stange[2]

1 Department of Sociology, University of Nebraska-Lincoln, Lincoln, NE, USA
2 Mathematica Policy Research, Ann Arbor, MI, USA

1 Introduction

Probability samples are necessary for making statistical inferences to the general population (Baker et al. 2013). Some countries (e.g. Sweden) have population registers from which to randomly select samples of adults. The U.S. and many other countries, however, do not have population registers. Instead, researchers (i) select a probability sample of households from lists of areas, addresses, or telephone numbers and (ii) select an adult within these sampled households. The process by which individuals are selected from sampled households to obtain a probability-based sample of individuals is called within-household (or within-unit) selection (Gaziano 2005).Within-household selection aims to provide each member of a sampled household with a known, nonzero chance of being selected for the survey (Gaziano 2005; Lavrakas 2008). Thus, it helps to ensure that the sample represents the target population rather than only those most willing and available to participate and, as such, reduces total survey error (TSE).

Published (as Chapter 2) in *Experimental Methods in Survey Research: Techniques that Combine Random Sampling with Random Assignment,* First Edition. Edited by Paul J. Lavrakas, Michael W. Traugott, Courtney Kennedy, Allyson L. Holbrook ,Edith D. de Leeuw, and Brady T. West. John Wiley & Sons, Inc. 2019.

In interviewer-administered surveys, trained interviewers can implement a prespecified within-household selection procedure, making the selection process relatively straightforward. In self-administered surveys, within-household selection is more challenging because households must carry out the selection task themselves. This can lead to errors in the selection process or nonresponse, resulting in too many or too few of certain types of people in the data (e.g. typically too many female, highly educated, older, and white respondents), and may also lead to biased estimates for other items. We expect the smallest biases in estimates for items that do not differ across household members (e.g. political views, household income) and the largest biases for items that do differ across household members (e.g. household division of labor).

In this chapter, we review recent literature on within-household selection across survey modes, identify the methodological requirements of studying within-household selection methods experimentally, provide an example of an experiment designed to improve the quality of selecting an adult within a household in mail surveys, and summarize current implications for survey practice regarding within-household selection. We focus on selection of one adult out of all possible adults in a household; screening households for members who have particular characteristics has additional complications (e.g. Tourangeau et al. 2012; Brick et al. 2016; Brick et al. 2011), although designing experimental studies for screening follows the same principles.

2 Within-Household Selection and Total Survey Error

Inaccurate within-household selection can contribute to TSE in multiple ways. First, every eligible member of the household has to be considered by the household informant during the within-household selection process. The household informant needs to identify a "list" (written down or not) of eligible household members. If eligible members are excluded from the list, undercoverage occurs. If certain people tend to be systematically excluded from household lists (e.g. young men) and their characteristics are related to constructs measured in the survey (e.g. health-care expenditures), their exclusion will

result in increased coverage bias of survey estimates. Second, assuming that the list of eligible household members is complete, the interviewer, if there is one, has to accurately administer, and the informant has to correctly follow, the selection instructions. Mistakenly or intentionally selecting the wrong household member from the (conceptual) list of household members can affect sampling error, especially sampling bias (e.g. if there is similarity across households in the characteristics of those erroneously selected and these characteristics are related to measured survey constructs). Finally, nonresponse error can result if the within-household selection procedure dissuades certain types of households or certain types of selected household members from completing the survey. The joint effects of the within-household selection procedure on any one of these three error sources (coverage, sampling, and nonresponse) may bias survey estimates. As a result, a number of different selection procedures have been developed, some of which prioritize obtaining true probability samples and some of which relax this criteria to potentially reduce coverage, sampling, and nonresponse errors.

3 Types of Within-Household Selection Techniques

Researchers can sample individuals within households using various probability, quasiprobability, and nonprobability, and convenience methods (Gaziano 2005). The Kish (1949), age-order (Denk and Hall 2000; Forsman 1993), and full enumeration (and variations of these) techniques obtain a *probability* sample of individuals from within households by ensuring that each eligible member of a sampled household has a known, nonzero chance of becoming the selected survey respondent. Probability methods of within-household selection require the most information about household members, including the number of people living in the household, and often more intrusive information such as household members' sex and age. The interviewer asks the household informant for the requisite information about the household and then follows systematic procedures to select (or the interviewer's computer selects) a respondent from the household. To our knowledge, full probability sample procedures are rarely used in self-administered surveys because they are so complex; even

in interviewer-administered modes, they pose some challenges from a TSE framework. While full enumeration procedures are intended to reduce coverage error, by requesting sensitive information upfront, they may increase nonresponse error.

The last birthday and next birthday within-household selection techniques are quasiprobability methods because household members' birthdates identify who should be the respondent rather than a truly random selection mechanism. In the birthday techniques, the researcher uses an interviewer (in interviewer-administered modes) or the cover letter (in self-administered modes) to ask the household member who has the birthday that will occur next (next birthday) or who most recently had their birthday (last birthday) relative to a reference date to respond to the survey. Birthday techniques assume that birthdates are functionally random for the purposes of identifying a member of the household to respond. For many topics, this assumption seems warranted; however, for topics where the variables of interest are related to birthdays (i.e. voting at age 18), this method may not be appropriate. These techniques are popular in both interviewer- and self-administered questionnaires because of their ease of implementation, although the selection process is often inaccurately completed in any mode.

Variations that aim to reduce the intrusiveness of the probability methods by combining probability and quasi-probability methods and accounting for household size also exist (e.g. Rizzo method – Rizzo et al. 2004; Le et al. 2013).These methods first obtain information about the number of people in a household and then use different methods for households with two adults (unobtrusive random selection of the informant or the other adult) and households with three or more adults (more obtrusive requests for enumeration, using a birthday method, asking by age position and possibly sex of the adults in the household).These methods reduce the proportion of households subjected to more intrusive methods; for example, most U.S. households have only one or two adults (Rizzo et al. 2004).

Quota or targeted techniques identify a respondent based on demographic criteria, such as the youngest male or oldest female from the selected household, or simply select any adult from the household. These methods are nonprobability methods, meaning the researcher loses the statistical theory linking the sample to the target population,

thus undermining the representation side of the TSE framework. However, they are less costly, less intrusive, and easier to implement accurately. Nonprobability methods can be used in any data collection mode. In a telephone survey, the interviewer may ask for a knowledgeable respondent or take the phone answerer as the respondent; in a mail survey, the instructions will appear in a cover letter, if at all.

4 Within-Household Selection in Telephone Surveys

In a telephone survey, the interviewer (typically assisted by a computer) selects and encourages the sampled household member to participate in the survey using one of the methods described above. In telephone surveys conducted up to the early 2000s, less invasive techniques (i.e. birthday and nonprobability techniques) demonstrated the tradeoff across error sources. They tended to have higher response rates and lower cost but less representative demographic compositions than more invasive probability techniques such as the Kish method (Gaziano 2005; Yan 2009).

More recently, probability-based within-household selection methods continue to result in lower *response rates* than quasi-probability birthday techniques and nonprobability techniques (Marlar et al. 2014; Longstreth and Shields 2005; Beebe et al. 2007). For example, in a comparison of probability, quasi-probability, and nonprobability methods, Marlar et al. (2014) found that the probability-based Rizzo et al. (2004) method garnered response rates that were roughly 2.5 percentage points lower than selecting a respondent based on age/sex criteria and being at home, with quasi-probability and nonprobability methods selecting among all people in the household (not just those at home) in the middle. Longstreth and Shields (2005) had a similar magnitude difference in response rates comparing the last birthday method to the Rizzo method. Beebe et al. (2007) compared the Rizzo method with the next birthday method, finding response rates for the next birthday method about 4 percentage points higher than the Rizzo method.

In these studies, the *composition* of completed samples did not differ unless demographic characteristics were part of the selection method. For example, Beebe et al. (2007) and Longstreth and Shields

(2005) both found no differences in demographic characteristics such as sex, age, race, education, income, and number of people in the household in the completed samples produced by the Rizzo selection procedure and either birthday selection technique (last or next birthday). On the other hand, Marlar et al. (2014) found that the youngest male/oldest female technique resulted in more males in the sample, while selecting the youngest person in the household produces a sample that contained more females. In an international context, Le et al. (2013) found no difference in composition of the respondent pool across sex or age characteristics comparing the Kish method to a new household-size dependent procedure. Moreover, none of the studies found differences in substantive estimates by within-household selection procedure.

Other outcomes used to assess within-household selection methods include the *accuracy of selection* and *cost information*. Among the existing telephone research, only Marlar et al. (2014) examined the accuracy of selection, finding that roughly 20–30% of respondents were inaccurately selected in the quasi-probability birthday methods, which is similar to earlier research (O'Rourke and Blair 1983; Troldahl and Carter 1964; Lavrakas et al. 2000; Lind et al. 2000). By comparison, the nonprobability methods they tested had considerably lower inaccuracy rates (youngest person – 20.1%, youngest male/youngest female – 1.8%, multiquestion youngest male/youngest female – 0.5%) (Marlar et al. 2014). For cost, little information is available. Longstreth and Shields (2005) found that the completion time for interviewers to implement the Rizzo and last birthday methods did not significantly differ, and Beebe et al. (2007) found that the mean number of call attempts to interview was the same across the Rizzo and next birthday methods.

5 Within-Household Selection in Self-Administered Surveys

Unlike telephone surveys, self-administered surveys cannot rely on trained interviewers to administer the selection procedures. In mail surveys, the household informant opens the mail, reads the selection technique typically described in a survey's cover letter or on the questionnaire, and determines which member of the household should complete the survey. The household informant must then complete

the survey if they are selected or must convince the selected person to complete the survey. Problems can arise at any of these steps.

One fundamental difference between mail and telephone surveys is that true probability methods such as the Kish selection method are considered too complex for households to implement in mail surveys (Battaglia et al. 2008; Reich et al. 1986). As such, researchers have most often employed the quasi-probability birthday methods or non-probability techniques to try to reduce coverage, sampling, and non-response errors at the expense of true probability methods.

Unlike telephone surveys, there are few significant differences in *responses rates* by type of within-household selection method in mail surveys (Battaglia et al. 2008; Olson et al. 2014). For example, despite the any-adult technique being minimally burdensome, it yields response rates similar to the next birthday method (Battaglia et al. 2008). Across two studies of Nebraskans, the next and last birthday selection procedures had statistically identical response rates as the oldest adult procedure, but the youngest adult method had a significantly lower response rate, likely driven by lower response rates among younger adults in general (Olson et al. 2014).

Similarly, the demographic *composition* of respondent pools do not significantly differ across the within-household selection techniques in mail surveys and all the methods result in samples that significantly differ from demographic benchmarks in similar ways (Battaglia et al. 2008; Hicks and Cantor 2012; Olson et al. 2014).The any adult, all adult, and next and last birthday techniques all tend to underrepresent younger people and overrepresent non-Hispanic whites, adults with higher education, and married people. Studies also find no significant differences in substantive survey estimates across within-household selection techniques (Battaglia et al. 2008; Hicks and Cantor 2012; Olson et al. 2014).

Selection accuracy is the primary focus of within-household selection evaluations in self-administered surveys. Across studies, up to 30% of within-household selections are inaccurate, with (substantially) higher rates when excluding one-adult households that have accurate selections by default (Stange et al. 2016; Olson and Smyth 2014; Olson et al. 2014; Battaglia et al. 2008; Schnell et al. 2007; Gallagher et al. 1999). Moreover, inaccuracy rates do not significantly differ by selection technique (Olson et al. 2014).

Few studies have examined within-household selection methods in web surveys. In part, this is because web surveys are often used to survey named people, using individualized emails to deliver the survey invitation. When researchers want to send a web survey to a household and administer a within-household selection procedure, they typically do so using a mixed-mode design in which the invitation letter is delivered by postal mail (e.g. Smyth et al. 2010). In this case, researchers have largely adopted mail within-household selection procedures, most often including a quasi-probability selection instruction in the invitation letter. In the only study of which we are familiar that assessed selection accuracy in web surveys (using postal mail invitations), the inaccuracy rate in web was about 20% and did not significantly differ from the inaccuracy rate in mail-only, or in conditions mixing mail and web data collection modes (Olson and Smyth 2014).

6 Methodological Requirements of Experimentally Studying Within-Household Selection Methods

The goal of within-household selection of a single adult is to produce a (quasi-)probability-based sample that mirrors the target population (i.e. minimizes coverage and nonresponse error from a TSE perspective) on characteristics being measured in the survey. As such, there are three general methods for assessing the quality of the within-household selection:

1. comparing the characteristics of the completed sample to benchmark measures for the target population,
2. comparing survey estimates across the experimental treatments, and
3. evaluating how well the completed samples followed the within-household selection instructions by measuring the accuracy of selection.

For example, for state, regional, or national surveys, one can compare the demographic makeup of the completed sample to official statistics for the same geographic region, such as from the American Community Survey (ACS). Of course, this requires that benchmark

outcomes be available for the target population. Comparison of estimates in the survey across experimental treatments should be guided by the mechanisms for what might differ across experimental treatments (e.g. age in the youngest adult method). Accuracy requires obtaining external information about household composition from a rich sampling frame or incorporating methods to assess the accuracy of selection (at least among the respondents) in the survey itself. With these three methods in mind, it is possible to identify the appropriate experimental design for studying within-household selection methods.

Comparing characteristics of the completed sample to benchmark data requires minimizing other sources of survey error that might affect the composition of the final sample. Thus, an experimental study of within-household selection should start with **a sample frame with good coverage** so that coverage error in the sample frame is excluded as an explanation for differences between the completed sample characteristics and benchmark measures. It also means that within-household selection experiments need to **start with a probability sample of housing units** from the sample frame. A probability sample of housing units will produce a sample that mirrors the target population so that any differences between the final completed sample of individuals and the benchmark measures can be attributed to measurable sampling error and the within-household selection techniques after accounting for probabilities of selection. The sample does not have to be a simple random sample as long as information about strata, clusters, and unequal probabilities of selection are maintained and incorporated in the analyses. A sample frame with poor coverage or a nonprobability sample of housing units will make it impossible to tell how much of the difference between the respondent pool and benchmark outcomes is due to coverage and sampling from the household frame versus coverage and sampling of individuals within households. Statistically, it is also necessary to ensure that the **sample size is sufficiently large** to allow for enough power to detect significant differences across treatments. This decision will be driven by a power analysis that accounts for the number of experimental treatments, the outcome of interest, the type of analysis used for evaluating the experiment, and the expected effect size that will result from the experiment for that outcome. Thus, the first requirement of such

a study under a TSE framework is a sufficiently large probability sample of a known population from a good sample frame.

Of course, any experiment needs experimental treatments or factors. The **selection of the experimental factors and items included in the questionnaire should be informed by theory** to anticipate possible effects of the experimental factors. For example, Olson and Smyth (2014) theorized that there were three reasons for inaccurate within-household selections: confusion, concealment, and commitment. They were able to test these theories using a limited set of questions included in their questionnaire as proxies for each reason: size of household, education, and presence of children in the household were proxies for confusion; gender, age, race, income, concern with identity theft, and fear of crime for concealment; and previously reported mode preference (a variable on the sample frame) for commitment. Thus, using theory to guide the selection of experimental factors and including measures that allow researchers to test theoretical reasons for the success or failure of the selection methods can help advance knowledge of why certain methods work or fail and ways to improve them.

The next requirement of an experimental study of within-household selection methods is that **the sampled housing units be randomly assigned to the alternative experimental treatments**. Randomly assigning housing units to the within-household selection method treatments ensures that each treatment is assigned a representative subset of the sample of housing units. Thus, the composition of the respondent pool in each treatment can be attributed to the within-household selection method used in the treatment, not to differences in the composition of housing units assigned to each treatment. Using both a probability sample of households from a known population and then randomly assigning sampled households to treatments ensures that the sampling design and experimental assignments are not confounded with the experimental treatments.

In addition, **design differences other than the factors being tested between the experimental treatment versions should be eliminated** (i.e. eliminate confounding factors). For example, if incentives are to be used, they should be used in exactly the same way (type, amount, timing, etc.) in all treatments. Likewise, the response device type (e.g. cell phone versus landline phone; computer versus

mobile web, etc.) should not differ across treatments nor should the number, type, and timing of contacts or the information communicated in those contacts, other than changes needed for the factor being tested. If testing methods to try to improve the quality of a single within-household selection method, then all other features of the within-household selection method should be held constant. For example, to compare the effects of the selection instruction wording for the next birthday method on selection accuracy, all treatments should use the same selection method (next birthday) and the same wording for all other aspects of the cover letter other than the relevant part being manipulated. Essentially, the only thing that should differ across the treatments is the within-household selection method or elements that are modifying a single within-household selection method.

Another possible confounding factor for within-household selection method experiments in interviewer-administered modes is the interviewer themselves. Interviewers pose two types of threats to the integrity of these experiments. First, they may be differentially skilled at administering within-household selection methods and/or obtaining cooperation from selected household members. If more skilled interviewers are disproportionately assigned to a particular experimental treatment, that treatment may end up performing better because of the interviewers, not because it is the better method. To solve this problem, **interviewers should be randomly assigned to experimental treatments** so that interviewer characteristics (both observable and unobservable) are equally distributed across the treatments. If interviewers cannot be randomly assigned to treatments, then observable interviewer characteristics such as demographic characteristics, interviewer experience or tenure, and even measures of interviewer skill, such as cooperation rates on previous studies, should be collected so that they can be used to statistically control for potential differences in interviewers across the experimental treatments. Analyses of experiments in interviewer-administered surveys should use multilevel models that can account for the nesting of selection procedures and sample cases within interviewers (e.g. Raudenbush and Bryk 2002; Hox et al. 1991). Likewise, interviewer assessments of the characteristics of each method (e.g. ease, sensitivity) should be evaluated. In addition to randomly assigning interviewers to treatments, **the way that cases are assigned to interviewers should be the same across**

all treatments to ensure that the assignment of cases to interviewers does not confound results across the treatments.

The second type of challenge that interviewers pose is that their knowledge of the experiment itself may lead them to change their behaviors, either intentionally or unintentionally, in ways that undermine the integrity of the experiment. For example, an interviewer may prefer the ease of the next or last birthday method to a full probability method such as the Kish method. Interviewer expectations can affect response rates (Durrant et al. 2010), and thus these preferences or expectations confound the experiment itself with interviewer preferences. This suggests that **each interviewer should only be assigned to one treatment**; the same interviewer(s) should not work on multiple treatments. This topic is covered in more depth in Chapter 12 of this volume by Lavrakas, Kelly, and McClain.

In addition to considering how the sample of households is drawn and assigned to treatments and interviewers, considerable thought should be given to whether there are **variables on the frame or that can be measured in the survey that can help assess the quality of each selection treatment**. For example, to assess the accuracy of the birthday and oldest/youngest adult methods, Olson et al. (2014) included a household roster in their questionnaire that collected relationship to the respondent, age, date of birth, and sex of each person living or staying in the household. They could then check whether or not the respondent actually had the next or last birthday or were the youngest or oldest adult in the household, depending on the assigned selection method. They were also able to examine whether selection accuracy differed by factors such as the size of the household or whether a household member had a birthday during the field period, both of which positively predicted selection inaccuracies.

In sum, the methodological requirements of experiments for studying within-household selection methods under a TSE framework are:

- Identifying analytic outcomes that will be used to evaluate the methods (e.g. benchmarks and/or ways to assess accuracy);
- A sample frame with good coverage from which a probability sample of housing units will be selected;
- Theoretically driven experimental treatments;
- Random assignment of the selected housing units to experimental treatments;

- Elimination of design differences across the treatments that are not the focus of the comparison;
- In interviewer-administered surveys, random assignment of interviewers to experimental treatments, separate interviewer corps for each treatment, and consistent assignment of cases to interviewers across treatments;
- Inclusion of covariates from the frame or measured in the survey to better understand why differences occur between the treatments.

7 Empirical Example

The process of implementing within-household instructions in mail surveys can break down if the informant does not read the instructions, understand them, enumerate a full list of eligible household members, believe in the importance of the selection process, feel motivated to follow the instructions, and/or have the ability to recruit the sampled household member. In our early studies, we found that proxies for confusion such as complexity of the instruction, number of adults in the household, children in the household (Olson and Smyth 2014), and a member of the household having a birthday during the field period (Olson et al. 2014) were associated with higher inaccuracy rates. However, our research designed to reduce confusion (i.e. providing a calendar to help informants place household birthdays in time and providing explanatory instructions to help informants understand why the selection instructions should be followed) failed to improve the quality of sample pools (Stange et al. 2016).The motivation of the informant to implement the instructions and of the selected household member to complete the survey, a factor also discussed by Battaglia et al. (2008), had not been tested. As a result, we designed a new experiment to target motivation. That is, this experiment focuses on the commitment part of the confusion, concealment, and commitment framework theorized by Olson and Smyth (2014). We discuss the theoretical motivation for the experimental treatments, the design of the experiment, and its results here.

One technique previously shown to be effective at encouraging survey participation among unmotivated sample members is providing

prepaid (i.e. noncontingent) cash incentives. Numerous studies show that incentives significantly increase response rates (e.g. Church 1993; James and Bolstein 1992; Singer 2002; Singer and Ye 2013; Trussell and Lavrakas 2004). Importantly, Baumgartner and Rathbun (1996) and Groves et al. (2006) found that incentives also encourage participation among sample members who are less interested in the survey topic. These findings suggest that incentives might improve within-household selection in several ways. First, incentives may increase the likelihood that the letter opener will read the cover letter in the first place and the importance they attribute to the survey (Dillman et al. 2014), thus increasing the likelihood that they see and subsequently follow the within-household selection instruction rather than simply doing the survey themselves (i.e. reducing the potential for sampling error). Second, incentives may increase the likelihood that otherwise reluctant household members are included in the household list (i.e. reducing undercoverage), either because they themselves are the selected respondent and want to receive the incentive or because another informant believes that the reluctant household member would want it. Third, the incentive may increase the otherwise reluctant household members' willingness to respond if selected. Thus, we examine whether providing a prepaid, noncontingent incentive improves the performance of the next birthday within-household selection method.

For the incentive to have the largest impact within the context of within-household selection, the selected respondent should receive the incentive. Their receiving the incentive should reduce resistance to being included in the selection process and increase the likelihood that they respond if selected, thereby improving coverage and response rather than simply increasing response rates from the household more generally. As such, in addition to examining the effects of providing an incentive versus no incentive, we also experimentally varied whether or not wording about the incentive in the cover letter was targeted to the selected respondent.

The experiment had three treatments:

1. No incentive
2. $1 incentive with standard letter wording
3. $1 incentive with targeted letter wording.

The standard letter wording was, "We have enclosed a small token of appreciation to thank you for your help," and the targeted letter wording was, "We have enclosed a small token of appreciation to thank the adult with the next birthday for their help. "The no incentive condition necessarily omitted all mention of incentives. To eliminate confounds in the experiment and ensure that we could attribute all differences across treatments to either the provision of the incentive or the standard versus targeted wording, the remaining content of the letter in all three treatments was identical.

We expected that the incentive would encourage households to notice and follow the selection instruction, include all household members in the list of eligible household members, and encourage participation when the selected household member was uninterested in the survey, thus targeting coverage, sampling, and nonresponse errors. Thus, we hypothesized that the incentive would lead to:

1. a higher response rate,
2. a completed sample that more closely matched ACS benchmarks for the area under study, and
3. a higher rate of accurate selections, determined through the use of information from a household roster.

We hypothesized that the effect of the incentive on response rates would be attenuated somewhat in the targeted letter wording as this wording reinforces the idea that the incentive and therefore the survey is for the specifically selected person in the household. Thus, we thought it was more likely in this condition that if the selected person refused, the survey would be discarded rather than returned by another adult. Because of this, however, we expected the incentive and targeted wording treatment to most closely match ACS benchmarks and to have the highest rate of accurate selections (i.e. we expected a tradeoff between response rates and selection accuracy in this treatment).

8 Data and Methods

We embedded the incentive and cover letter wording experiment in the 2014 Nebraska Annual Social Indicators Survey (NASIS), which is

an annual, omnibus mail survey of Nebraska adults aged 19 and older (Bureau of Sociological Research 2014). NASIS 2014 included 93 questions (some with multiple prompts) across 11 pages about natural resources, underage drinking, vaccinations, the Affordable Care Act, invasive plant species, household characteristics, finances, and demographics. The surveys were administered in English only. After obtaining institutional review board approval for the study within our university, the questionnaires were mailed on 20 August 2014. A postcard reminder was sent one week later, and a replacement survey packet was sent to nonrespondents on 18 September 2014. The survey cover letter instructed (with bolded text) that the household member with the next birthday after August 1, 2014, should complete the survey.

The sample consisted of a simple random sample of $n = 3500$ addresses from across Nebraska drawn by Survey Sampling International (SSI) from the USPS computerized delivery sequence (CDS) file. NASIS 2014 was an ideal survey for this experiment because of its use of an address-based sample frame with excellent coverage of US households (Iannacchione 2011) and a probability sampling method. Thus, with frame-based confounds minimized, differences between characteristics of our completed sample and ACS estimates for the state of Nebraska (i.e. a key outcome) can be attributed to coverage, sampling, and response *within* households rather than coverage and sampling *of* households.

The NASIS 2014 sample size was also sufficiently powered to allow us to test our hypotheses. For example, previous years' NASIS surveys, which did not use incentives, yielded response rates around 25% (Bureau of Sociological Research 2013). At the planning stage, we assumed a similar response rate for NASIS 2014 would yield 875 completes or roughly 291 completes per treatment. Based on these assumptions, **Table 1** shows the effect sizes we anticipated being able to detect with a given level of power across treatments with an alpha of 0.05. If our assumptions held, we would be able to detect effect sizes of 10.7 percentage points with power 0.8 (a typical minimum power level). Table 1 also shows the effect sizes we would be able to detect by power level if we compared one treatment to two others combined, to determine the overall effects of the incentive (i.e. the no incentive treatment compared to the two incentive treatments). Ultimately, a total of $n = 1018$ sampled households completed NASIS 2014 for a

Table 1. Detectable effect sizes (proportions) by power level for anticipated and actual response rates (α = 0.05).

	Anticipated response rate = 25%; n = 875		*Actual response rate = 29.1%; n = 1018*	
Power	*Effect size comparing any two treatments*	*Effect size comparing one treatment to the other two combined*	*Effect size comparing any two treatments*	*Effect size comparing no incentive treatment to both incentive treatments combined*
0.4	0.064	0.056	0.048	0.056
0.5	0.074	0.064	0.055	0.064
0.6	0.083	0.072	0.062	0.073
0.7	0.094	0.082	0.070	0.082
0.8	0.107	0.092	0.080	0.093
0.9	0.124	0.107	0.093	0.107

29.1% response rate (AAPOR RR1) (for effect sizes by post hoc power level with actual response rates, see the right half of Table 1).

To ensure that we could attribute differences across the three experimental treatments to the features of the treatments themselves and not other factors (i.e. different types of households assigned to different treatments), each sampled household was randomly assigned to one of the three experimental treatments. This resulted in 1166 households assigned to the no incentive treatment and 1167 households assigned to each of the $1 incentive with standard letter wording and $1 incentive with targeted letter wording treatments. Comparisons of household characteristics provided with the sample (e.g. FIPS [Federal Information Processing Standard] code – a geographic code identifying counties and county equivalents, Census tract, delivery type, race of population in Census tract, age, children, homeowner versus renter, length of residence, and gender) revealed that the randomization worked; there were virtually no significant differences in the types of households assigned to each treatment. The exception is that the no incentive treatment was assigned to slightly more black households and slightly fewer white households than the other treatments. Both of these differences were small in magnitude – less than 2 percentage points and likely attributable to Type I error (Type I error refers to a statistical test being significant by chance alone – that is, a false positive; results available from the authors).

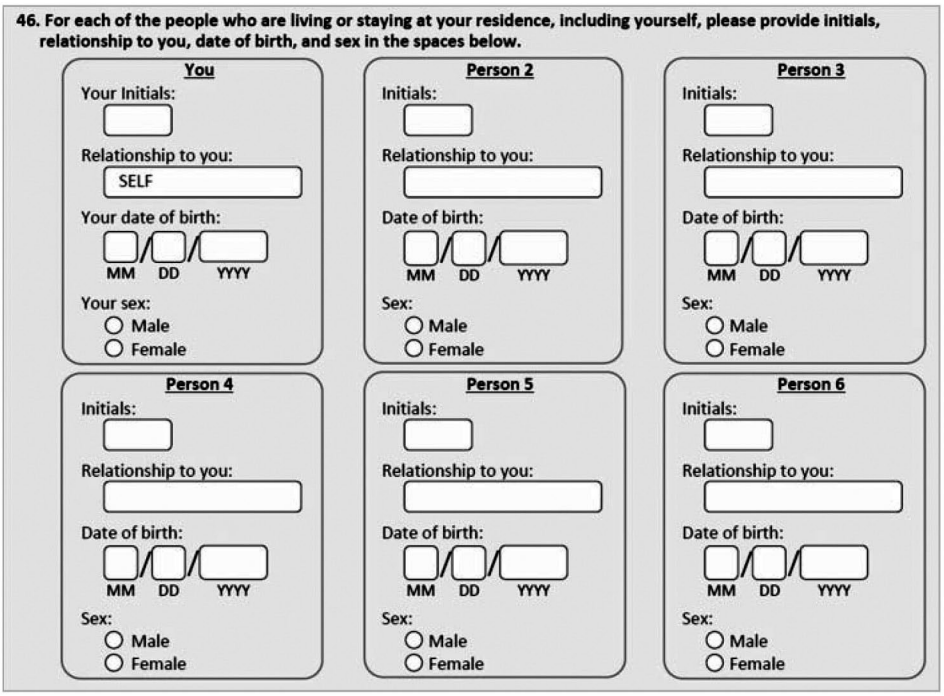

Figure 1. Household roster from 2014 Nebraska Annual Social Indicators Survey.

In addition to using a sample frame, probability sampling method, and random assignment to treatments to allow for comparison to the ACS benchmark, NASIS 2014 included a household roster (see **Figure 1**) that we used to determine whether the person answering was the adult in the household with the next birthday (i.e. accuracy/inaccuracy of selection). Following Olson and Smyth (2014), the questionnaire also included a set of covariates designed to reflect theoretically guided correlates of confusion, concealment, or commitment in within-household selection. However, because we had more control over questionnaire content in this experiment, a more extensive set of proxies were included. The confusion proxies included respondent education, children in the household, respondent's marital status, number of adults in the household, and whether the respondent lived in the same household as they did two years ago. These variables capture aspects of cognitive ability (Krosnick 1991; Narayan and Krosnick 1996) or complexity of the household makeup (Martin 1999, 2007; Martin and Dillman 2008; Olson and Smyth 2014), both of which are expected to increase confusion and thus increase inaccurate

selections. The concealment proxies included sex, age, income, and race because previous research has shown that young black men are underrepresented in surveys, with a hypothesis that the household is concealing household members (Tourangeau et al. 1997; Valentine and Valentine 1971).They also include a measure of how often respondents are concerned with identity theft (never to always) and measures of whether the respondent believes most people cannot be trusted, is suspicious of others, is concerned about personal privacy and the number of days the respondent felt sad or hopeful in the past seven days. These measures all reflect respondents' openness to the outside world; those who worry about intrusions from others, feel sad, or lack hopefulness are expected to be more hesitant to engage with the outside world and thus more likely to conceal themselves or family members (Caplan 2003; Kim et al. 2011; Malhotra et al. 2004; McKenna et al. 2002; Olson and Smyth 2014; Phelps et al. 2000; Segrin 2000), leading to inaccurate within-household selections. Finally, the commitment proxies include a set of items measuring who controls entrance into the household (the household gatekeeper) and thus would be the one to initially handle an incoming mail survey. We hypothesized that this person would be more likely to erroneously complete the survey because they are the household member who introduces it to the household, but that this effect would be diminished in the incentive treatments, especially with the targeted letter wording. The gatekeeper covariates included measures of who in the household opens the mail, answers the landline telephone (if available), opens the door for friends and relatives, and opens the door for strangers. These were recoded into dichotomous variables indicating whether the respondent was the person most likely to do each task (0 = no, 1 = yes). Under commitment, we also included an item measuring how likely the respondent is to answer surveys "like this one."

9 Analysis Plan

For the analyses, we first use unweighted chi-square tests to examine response rate differences across the experimental treatments. We then examine whether the demographic makeup of the completed samples differ by the incentive treatments. To account for item nonresponse

in the demographic and other predictor variables, we use a sequential regression imputation approach (the user-written ice command in Stata) to multiply impute missing values (Raghunathan et al. 2001). We created five imputed datasets.[1] We also created probability of selection weights; households were selected as a simple random sample, and one adult was selected out of all adults in the households (1/# adults).Thus, the probability of selection weight is proportionate to the number of adults in the household. We cap this weight at 3 to minimize increases in variance due to weighting (Kish 1992). All analyses of demographics and substantive variables account for this multiple imputation and are weighted by the inverse of the probability of selection (unweighted estimates available on request).We did not use poststratified weights because our analyses are focused on comparisons to benchmark data; the fully weighted estimates would artificially make the experimental treatments match the benchmark data.

We test whether the demographic variables differ across experimental treatments by predicting each demographic variable using ANOVA and regression approaches, accounting for multiple imputation and probability weights using the mi estimate procedures in Stata13. Using *t*-tests, we then compare the characteristics of the completed samples in each treatment to ACS 2014 five-year estimates benchmarks for Nebraska obtained from American Fact Finder (factfinder.census.gov). For these analyses, we look at respondent's sex, education, whether there are children in the household, age, family income, and race.

We then use birthdate information from the household roster to examine if the household member who completed the survey was the household member with the next birthday following August 1 (i.e. accurate versus inaccurate selection).We examine this for all households and those households with two or more adults because one-adult households automatically have accurate within-household selections. We then test for differences in accuracy by the incentive treatments and examine associations between our proxy measures for confusion, concealment, and commitment and accuracy of selections using

1 Creating five imputed data sets is consistent with established convention for data sets with low missing data rates and small fractions of missing information (Rubin and Schenker 1987; Raghunathan et al. 2001).More data sets are needed when the fraction of missing information is high, but our overall low item nonresponse rate (maximum <10%) and low fraction of missing information (maximum <0.18) suggest that five is adequate.

logistic regression. In these analyses, we also include control variables for whether the household was located on a farm, open country (not a farm), or a town or city; whether the home was owned, and whether it was a single family dwelling to account for any potential household composition differences across these characteristics. We look at these predictors overall, and whether there are any differences across the experimental treatments using interaction terms between the treatment indicators and proxies.

For all analyses, consistent with our power analysis, we adopt a $p<0.050$ cutoff for determining statistical significance. However, consistent with the American Statistical Associations statement on *p*-values (Wasserstein and Lazar 2016), we recognize more than a *p*-value has to be considered in assessing the importance of statistical results. Therefore, we also discuss results with *p*-values ranging from 0.050 to 0.100 where effect sizes are also large enough to be meaningful.

10 Results

10.1 Response Rates

As hypothesized, the incentive increased response rates. The response rate (AAPOR RR1) for the no incentive condition was 22.3% compared to 32.5% for the two incentive conditions combined ($\chi^2(1) = 39.05$, $p<0.001$). Also consistent with expectations, among the two incentive conditions, the response rates were 34.3% with the standard letter wording and 30.7% with the targeted wording. Both incentive conditions significantly differed from the no incentive condition (standard $\chi^2(1) = 41.25$, $p<0.001$; targeted $\chi^2(1) = 21.03$, $p<0.001$), and the 3.6 percentage point difference (a 10.4% reduction) between the standard and targeted incentive conditions approached significance ($\chi^2(1) = 3.45$, $p = 0.060$).

10.2 Sample Composition

As **Table 2** shows, the sample composition only differed significantly across the three treatments on sex ($F = 8.38$, $p<0.001$). The incentive with the standard letter wording treatment yielded a sample that was

Table 2. Demographic composition and comparison to 2014 five-year ACS estimates by treatment.

	Composition (%)					*Significance tests across treatments*				*Significance tests NASIS vs. ACS estimates*			
	All	*No $ n = 260*	*$+stand. n = 400*	*$+target n = 358*	*ACS*	*Overall F*	*t stand. vs no $*	*t target vs no $*	*t stand. vs target*	*All vs ACS*	*No $ vs ACS*	*Stand. vs ACS*	*Target vs ACS*
Sex													
Female	54.6	51.7	62.9	47.6	50.9	8.38***	−2.68**	0.93	3.97***	2.21*	0.23	4.67***	−1.17
Male	45.4	48.3	37.1	52.4	49.1					−2.21*	−0.23	−4.67***	1.17
Education													
HS or less	24.6	24.1	22.9	26.7	37.3	1.34	−0.33	0.71	1.15	−8.93***	−4.68***	−6.55***	−4.21***
Some college	35.5	37.0	32.7	37.5	36.2		−1.05	0.12	1.30	−0.44	0.24	−1.39	0.47
BA+	40.0	39.0	44.4	35.8	26.4		1.30	−0.76	−2.25*	8.17***	3.91***	6.72***	3.44***
Children in HH													
No kids	66.4	68.4	66.6	64.6	68.1	0.43	0.44	0.92	0.54	−1.06	0.11	−0.55	−1.27
Has kids	33.6	31.6	33.4	35.4	31.9					1.06	−0.11	0.55	1.27
Age													
19–34	13.3	12.0	13.1	14.3	20.7	1.89+	0.39	0.77	0.43	−6.60***	−4.02***	−4.17***	−3.30***
35–54	30.2	28.4	30.6	31.0	25.4		0.54	0.63	0.11	3.01**	1.01	2.02*	2.09*
55–64	25.1	21.3	29.6	22.9	12.2		2.14*	0.43	−1.89+	8.71***	3.30***	6.95***	4.39***
65+	31.5	38.2	26.7	31.8	13.8		−2.95**	−1.57	1.46	11.48***	7.73***	5.49***	6.96***
Family income													
<$50k	38.0	39.8	36.3	38.5	35.1	0.42	−0.82	−0.31	0.57	1.78	1.45	0.47	1.22
$50–99k	40.3	41.4	40.6	39.1	38.4		−0.17	−0.51	−0.38	1.11	0.90	0.81	0.25
$100k+	21.7	18.8	23.0	22.4	26.4		1.15	0.97	−0.18	−3.22**	−2.77**	−1.41	−1.66
Race													
Nonwhite	9.5	9.2	10.7	8.3	18.8	0.59	−0.63	0.33	1.06	−9.69***	−5.09***	−4.99***	−6.82***
Non-Hispanic white	90.5	90.8	89.3	91.7	81.2					9.69***	5.09***	4.99***	6.82***

+ $p<0.100$; * $p<0.050$; ** $p<0.010$; *** $p<0.0001$

Data are weighted for probability of selection and adjusted for multiple imputations.

62.9% female, which was about 11 percentage points higher than the no incentive treatment (t = 2.68, p = 0.007) and 15 percentage points higher than the incentive with targeted wording treatment (t = 3.97, p<0.001).This is also about 12 percentage points higher than the ACS estimate (t = 4.67, 0<0.001). Thus, the incentive on its own resulted in an overrepresentation of women but using the targeted wording with the incentive appears to have corrected for this overrepresentation.

The distribution of age was moderately significantly different across the three treatments (design-adjusted F = 1.89, p = 0.079). The addition of the incentive reduced the percent of respondents in the oldest age group (65+) by 6.4 percentage points in the targeted wording treatment (t =−1.57, p<0.117) and 11.5 percentage points in the standard wording treatment (t = −2.95, p<0.010). In the standard wording treatment, this reduction was accomplished primarily through an 8.3 percentage point increase in the percent of respondents in the next highest age category (55–64), but in the targeted wording treatment, the increase was spread among all the younger age categories.

With the exception of sex and age, none of the other demographic variables differed significantly across the treatments. Moreover, the overall pattern is that all three treatments significantly differed from the ACS estimates on a number of the demographic characteristics, especially education (overrepresented high education), age (underrepresented the young and overrepresented the old), and race (overrepresented non-Hispanic whites). The no incentive treatment and the incentive with targeted wording treatment did not differ from the ACS on sex or children in the household and the two incentive conditions did not differ from the ACS on family income.

Because the treatments differed in how their estimates compared to the ACS, it is difficult to say that one treatment is better than another from these analyses. One way to assess the overall performance of the treatments is to examine the average absolute differences between the estimates produced by each treatment for each demographic and the corresponding ACS estimate. Looking across all characteristics, the treatment with the incentive and targeted wording had the lowest average absolute difference from the ACS estimates at 6.5 percentage points versus 6.8 percentage points for the no incentive treatment and 8.1 percentage points for the treatment with incentive and standard wording. Taken altogether, the sample composition

results suggest that while the differences are not large in magnitude, the treatment with the incentive and targeted wording produced demographic estimates that most closely matched the ACS estimates.

10.3 Accuracy

Sufficient information about household members had to be provided in the household roster to determine whether or not the within-household selection was done accurately for each responding household. Accuracy could be determined for 92.6% of households; the accuracy analyses thus are limited to the 943 cases where accuracy could be determined. Households with complete versus incomplete roster information did not differ on any characteristic other than the likelihood to answer surveys – those for whom accuracy could not be determined rated their likelihood of answering surveys like this one significantly lower (2.92 on a 4 point scale) than those for whom accuracy could be determined (3.34; $t = -3.15$, $p = 0.002$).

Table 3 shows accuracy rates overall and by treatment for both the full sample ($n = 943$) and the sample limited to households with two or more adults ($n = 660$). In the full sample, 63.2% of respondents were selected accurately with accuracy rates ranging from 59.9% in the no incentive condition to 66.2% in the incentive condition with targeted letter wording (a 6.3 percentage point difference); the overall difference in accuracy by treatment was not significant ($F = 1.07$, $p = 0.343$). The accuracy rate in the sample limited to households with

Table 3. Selection accuracy rates overall and by treatment for the full sample and for households with at least two adults.

	All households	*Two+ adult households*
All sample ($n = 943$/$n = 660$)	63.2%	55.2%
No incentive treatment ($n = 243$/$n = 165$)	59.9%	50.4%
Incentive+standard wording treatment ($n = 371$/$n = 261$)	62.5%	54.3%
Incentive+targeted wording treatment ($n = 329$/$n = 234$)	66.2%	59.5%
Overall *F*	1.07	1.58
t Incentive+standard wording vs. no incentive	0.61	0.76
t Incentive+targeted wording vs. no incentive	1.44	1.75+
t Incentive+standard wording vs. incentive+targeted wording	0.93	1.12

+ $p<0.100$

two or more adults was lower because households with only one adult can only get the selection correct – 55.2% overall and ranging from 50.4% in the no incentive condition to 59.5% in the incentive condition with targeted letter wording. Among this sample in which errors of selection could occur, the overall difference across treatments was not significant ($F = 1.58$, $p = 0.207$), but there was a 9 percentage point difference between the no incentive and incentive with targeted wording treatments ($t = 1.75$, $p = 0.081$). With more statistical power, this sizable difference would likely reach statistical significance.

While there were a few demographic differences across the treatments as discussed above (Table 2), none of the estimates of the theoretically driven concealment or commitment proxies (e.g. concern over identity theft, trust, mail opener, likelihood to answer surveys, etc.) significantly differed across treatments (results available from the authors).

Table 4 shows the results of logistic regression models predicting accuracy by experimental treatment; the proxy measures for confusion, concealment, and commitment; and control variables. Contrary to hypotheses, the full models indicate that the incentive with standard wording treatment was no more effective at producing accurate selections than the no-incentive treatment (full sample: $t = 1.22$, $p = 0.224$; 2+ adults sample: $t = 1.05$, $p = 0.295$), but, consistent with hypotheses, the incentive with targeted wording treatment was 63% more likely to produce accurate selections than the no incentive treatment in the full sample ($t = 2.27$, $p = 0.024$) and 52% more likely in the two or more adult households ($t = 1.92$, $p = 0.055$).

The results for the other predictors of accuracy were fairly consistent across the full and 2+ adult samples, indicating that survey estimates of these predictors differ for accurately and inaccurately selected households. Larger households were 42–64% less likely to make accurate selections (full sample: $t = -5.18$, $p<0.0001$; 2+ adults sample: $t = -2.90$, $p = 0.004$). Households where the respondent was the household member who is most likely to answer the door for friends or family were about 35% less likely to make accurate selections (full sample: $t = -1.86$, $p = 0.063$; 2+ adults sample: $t = -2.06$, $p = 0.040$). Respondents age 35–54 were 76–89% more likely than their younger counterparts to be accurately selected (full sample: $t = 2.26$, $p = 0.024$; 2+ adults sample: $t = 1.94$, $p = 0.053$). In the full sample, those who

Table 4 Logistic regression predicting accurate selection by experimental treatment, confusion, concealment, commitment, and control variables.

	Complete sample (n = 943)						*Two+ adult households (n = 660)*					
	Model 1 F(2, 940) = 1.07			*Model 2 F(31, 932.3) = 4.26****			*Model 1 F(2, 657) = 1.58*			*Model 2 F(31, 653.7) = 1.28*		
	Coeff.	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*
Constant	0.40**	0.141	1.49	2.51***	0.877	12.25	0.02	0.159	1.02	1.71+	0.886	5.51
Experimental treatments												
No incentive (omitted)												
Incentive+standard	0.11	0.184	1.12	0.25	0.203	1.28	0.16	0.204	1.17	0.22	0.213	1.25
Incentive+targeted	0.27	0.190	1.31	0.49*	0.215	1.63	0.37+	0.209	1.44	0.42+	0.218	1.52
Confusion proxies												
Education												
High school or less (omitted)												
Some college				−0.12	0.239	0.89				−0.08	0.239	0.92
BA+				0.02	0.247	1.02				0.01	0.248	1.01
Kids in household				0.18	0.215	1.20				0.13	0.212	1.14
Marital status												
Married (omitted)												
Never married				0.46	0.322	1.58				−0.21	0.382	0.81
Div./sep./widow				1.13***	0.351	3.09				0.01	0.431	1.01
Number of adults in HH				−1.02***	0.197	0.36				−0.55**	0.189	0.58
Same residence 2 yr				0.05	0.253	1.05				0.04	0.259	1.04
Concealment proxies												
Sex (male)				0.27	0.204	1.31				0.22	0.206	1.25
Age												
19–34 (omitted)												
35–54				0.64*	0.282	1.89				0.56+	0.291	1.76
55–64				0.47	0.323	1.60				0.30	0.334	1.35
65+				0.43	0.331	1.54				0.30	0.344	1.36

Continued

Table 4 Continued.

	Complete sample (n = 943)						*Two+ adult households (n = 660)*					
	Model 1 F(2, 940) = 1.07			*Model 2 F(31, 932.3) = 4.26****			*Model 1 F(2, 657) = 1.58*			*Model 2 F(31, 653.7) = 1.28*		
	Coeff.	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*	*Coeff.*	*SE*	*Odds ratio*
Family income												
<50k (omitted)												
50–99k				−0.02	0.245	0.98				−0.02	0.242	0.98
100k+				−0.02	0.300	0.98				−0.04	0.290	0.96
Race (white)				0.48	0.299	1.61				0.41	0.298	1.51
Worry about identity theft				−0.04	0.104	0.97				−0.06	0.107	0.95
Trust (be careful)				0.01	0.197	1.01				0.00	0.197	1.00
Suspicious of others				0.20	0.203	1.23				0.24	0.200	1.27
Privacy concern				−0.16	0.131	0.86				−0.14	0.130	0.87
Days sad				0.06	0.060	1.06				0.06	0.060	1.06
Days hopeful				0.00	0.036	1.00				−0.01	0.036	0.99
Commitment proxies												
Opens mail				−0.24	0.229	0.79				−0.25	0.216	0.78
Answers phone				0.29	0.192	1.34				0.21	0.190	1.23
Answers door for strangers				−0.01	0.231	0.99				−0.07	0.229	0.93
Answers door for friends/family				−0.44+	0.236	0.64				−0.46*	0.224	0.63
Likelihood to answer surveys				0.10	0.100	1.10				0.10	0.105	1.10
Control variables												
Rural/urban												
Farm (omitted)												
Open country, not farm				0.10	0.376	1.10				0.03	0.363	1.03
Town or city				−0.18	0.291	0.83				−0.20	0.286	0.82
Own home				−0.15	0.328	0.86				−0.18	0.330	0.84
Single family dwelling				−0.59+	0.301	0.56				−0.39	0.321	0.68

+ $p<0.100$; * $p<0.050$; ** $p<0.010$; *** $p<0.001$

were divorced, widowed, or separated were over three times as likely as their married counterparts to be selected accurately (t = 3.22, p = 0.001), and those living in a single family dwelling were 44% less likely than those in other types of housing to be selected accurately (t = −1.94, p = 0.052). These were not significant in the 2+ adult sample. No interactions between the proxies for confusion, concealment, or commitment, and the experimental treatments were statistically significant.

11 Discussion and Conclusions

Within-household selection is an important step for maintaining a probability sample of individuals. Unlike sampling housing units or households, within-household selection requires household members to identify who are members of the household and follow rules to garner a (quasi-)random selection of adults. Thus, within-household selection has implications for coverage, sampling, and nonresponse survey errors. Although this process is fairly straightforward when interviewers are present, it is much more difficult in self-administered surveys when no interviewer is present to assist the household. Previous research and the experimental results presented here suggest that households get this selection wrong at high rates. In fact, in households with more than one adult, the chance that the correct adult is selected is roughly equivalent to a coin flip. Thus, understanding how well different within-household selection methods work, why they may fail, and how to improve them is important. This kind of understanding is facilitated by the use of experimental methods.

Experimental tests of within-household selection methods are the strongest when they have good external validity through a sample frame with good coverage of households and a probability sample of households from that frame and strong internal validity through unconfounded experimental treatments and outcomes identified prior to data collection with requisite information collected in the questionnaire. This requires paying close attention to the design and its implementation. For instance, although implementing a within-household selection technique is easier overall in an interviewer-administered survey, implementing an experiment to test alternative

Table 5. Outcome rates for sampled households (n = 3500) by experimental treatment.

	Nonresponding household	*Responding household: correct selection*	*Responding household: incorrect selection*	*Responding household: unknown correctness*
Percent within each treatment				
No incentive	77.7	14.0	6.9	1.5
Incentive and standard wording	65.7	21.9	9.9	2.5
Incentive and targeted wording	69.3	20.1	8.1	2.5
Differences between treatments				
Incentive and standard wording minus no incentive	−12.0	8.0	3.0	1.0
Incentive and targeted wording minus no incentive	−8.4	6.2	1.2	1.0
Incentive and targeted wording minus incentive and standard wording	3.6	−1.8	−1.8	0.0

within-household selection methods is more difficult with interviewers because they can introduce (unobserved) confounding factors through their attitudes or expectations about a given method.

One challenge in implementing within-household selection field experiments on a probability sample of the general population is that multiple errors of nonobservation can be impacted by the experimental treatments. We do not know exactly what was happening inside sampled households as they processed the survey materials. Thus, we cannot be fully certain whether it was coverage, sampling, or nonresponse, or a combination of these errors that produced the differences we observed across the selection methods. Yet, **Table 5** is suggestive about possible mechanisms. It shows that adding the incentive increased response rates by increasing the percentage of households responding with the correctly selected household member between 6.2 and 8 percentage points but also increasing the percentage responding with an incorrectly selected household member by between 1.2 and 3 percentage points. This finding suggests that the incentive may not only have improved the coverage and response propensity of reluctant household members but also may have slightly increased errors in the sample selection (perhaps due to informants selecting themselves to get the incentive). Likewise, among the two incentive conditions, the targeted wording decreased response rates by about 3.6 percentage points with half of the decrease (1.8 percentage points) coming from responding households with correct selections

and the other half from those with incorrect selections. Thus, the incentive with targeted wording resulted in a lower response rate and a lower percentage of households that responded with correct selection when looking at the entire sample (Table 5) but a higher percentage of households with correct selection when looking at only the respondent pool (Table 3).Multiple error sources were clearly at play in this treatment. Again, while not definitive, we believe that we have a tradeoff between nonresponse and coverage/sampling errors occurring in this treatment. Overall, we believe that the increased accuracy rate outweighs the decrease in response rate because, even though the response rate was lower, this treatment had better alignment with the target population on important demographic characteristics. For experimental design, this example shows the importance of identifying multiple outcomes of interest prior to conducting the field experiment so that these different effects, and their relative importance, can be jointly weighed. For survey practice, if incentives are to be used in self-administered surveys with within-household selection of an adult, targeted wording about who should receive the incentive should be used in the cover letters as such wording improved the composition of the final sample compared to standard letter wording, especially on the characteristic of sex.

In this chapter, we provided an example of an experimental study of the effects of incentives on sample composition, variables theoretically measuring the mechanisms of confusion, concealment, and commitment, and accuracy of selection. Our results suggest that the incentive with the targeted wording yielded slightly better representation relative to official benchmarks and more accurate selection than the other two approaches. Even with these improvements, roughly 40% of respondents in households with two or more adults were not the correct respondent. Thus, there is ample room for improvement. The research here should be replicated with different types of samples and survey topics and additional strategies for improving the accuracy for within-household selection should be tested.

Designing an experiment to evaluate and, potentially, improve within-household selection methods requires careful planning and thoughtful consideration of theory, design, and implementation challenges. With a theoretically guided set of experimental factors, implemented to minimize any other confounding features, and a thorough

set of outcomes examining the multiple possible error sources, within-household selection experiments can yield useful and important insights. These experiments are even more necessary as self-administered surveys continue to grow in use and importance.

References

Baker, R., Brick, J.M., Bates, N.A. et al. (2013). *Report of the AAPOR Task Force on Non-probability Sampling*. American Association for Public Opinion Research. Retrieved 29 January 2016 from http://www.aapor.org/AAPOR_Main/media/MainSiteFiles/NPS_TF_Report_Final_7_revised_FNL_6_22_13.pdf

Battaglia, M.P., Link, M.W., Frankel, M.R. et al. (2008). An evaluation of respondent selection methods for household mail surveys. *Public Opinion Quarterly* 72: 459–469.

Baumgartner, R.M. and Rathbun, P.R. (1996). Prepaid monetary incentives and mail survey response rates. Paper presented at the Annual Conference of the American Association of Public Opinion Research. Norfolk, VA.

Beebe, T.J., Davern, M.E., McAlpine, D.D., and Ziegenfuss, J.K. (2007). Comparison of two within-household selection methods in a telephone survey of substance abuse and dependence. *Annals of Epidemiology* 17 (6): 458–463.

Brick, J.M., Andrews, W.R., and Mathiowetz, N.A. (2016). Single-phase mail survey design for rare population subgroups. *Field Methods* https://doi.org/10.1177/1525822X15616926

Brick, J.M., Williams, D., and Montaquila, J.M. (2011). Address-based sampling for subpopulation surveys. *Public Opinion Quarterly* 75 (3): 409–428.

Bureau of Sociological Research (2013). *NASIS 2012–2013 Methodology Report*. Lincoln, NE: Department of Sociology, University of Nebraska-Lincoln. https://digitalcommons.unl.edu/bosrreports/1

Bureau of Sociological Research (2014). *NASIS 2013-2014 Methodology Report*. Lincoln, NE: Department of Sociology, University of Nebraska-Lincoln. https://digitalcommons.unl.edu/bosrreports/2

Caplan, S.E. (2003). Preference for online social interaction: a theory of problematic Internet use and psychosocial well-being. *Communication Research* 30 (6): 625–648.

Church, A.H. (1993). Estimating the effect of incentives on mail survey response rates: a meta-analysis. *Public Opinion Quarterly* 57: 62–79.

Denk, C.E. and Hall, J.W. (2000). Respondent selection in RDD surveys: a randomized trial of selection performance. Paper presented at the annual meeting of the American Association for Public Opinion Research, Portland, OR.

Dillman, D.A., Smyth, J.D., and Christian, L.M. (2014). *Internet, Phone, Mail, and Mixed-Mode Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.

Durrant, G.B., Groves, R.M., Staetsky, L., and Steele, F. (2010). Effects of interviewer attitudes and behaviors on refusal in household surveys. *Public Opinion Quarterly* 74 (1): 1–36.

Forsman, G. (1993). Sampling individuals within households in telephone surveys. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Charles, IL, USA.

Gallagher, P.M., Fowler, F.J. Jr., and Stringfellow, V.L. (1999). Respondent selection by mail obtaining probability samples of health plan enrollees. *Medical Care* 37: MS50–MS58.

Gaziano, C. (2005). Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly* 69: 124–157.

Groves, R.M., Couper, M.P., Presser, S. et al. (2006). Experiments in producing nonresponse bias. *Public Opinion Quarterly* 70 (5): 720–736.

Hicks, W. and Cantor, D. (2012). Evaluating methods to select a respondent for a general population mail survey. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Orlando, FL, USA.

Hox, J.J., de Leeuw, E.D., and Kreft, I.G.G. (1991).The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In: *Measurement Errors in Surveys* (ed. P.P. Biemer, R.M. Groves, L.E. Lyberg, et al.). New York: Wiley.

Iannacchione, V.G. (2011).The changing role of address-based sampling in survey research. *Public Opinion Quarterly* 75 (3): 556–575.

James, J.M. and Bolstein, R. (1992). Large monetary incentives and their effect on mail survey response rates. *Public Opinion Quarterly* 56: 442–453.

Kim, J., Gershenson, C., Glaser, P., and Smith, T.W. (2011).The polls-trends: trends in surveys on surveys. *Public Opinion Quarterly* 75: 165–191.

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of American Statistical Association* 44: 380–387.

Kish, L. (1992).Weighting for unequal Pi. *Journal of Official Statistics* 8 (2): 183–200.

Krosnick, J.A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5: 213–236.

Lavrakas, P.J. (2008).Within-household respondent selection: how best to reduce total survey error? Unpublished report prepared for the Media Rating Council, Inc. Retrieved 29 January 2016 from http://www.mediaratingcouncil.org/MRC%20Point%20of%20View%20-%20Within%20HH%20Respondent%20Selection%20Methods.pdf

Lavrakas, P.J., Stasny, E.A., and Harpuder, B. (2000). A further investigation of the last-birthday respondent selection method and within-unit coverage error. In: *JSM Proceedings, Survey Research Methods Section*, 890–895. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/papers/2000_152.pdf

Le, K.T., Brick, J.M., Diop, A., and Al-Emadi, D. (2013).Within-household sampling conditioning on household size. *International Journal of Public Opinion Research* 25: 108–118.

Lind, K., Link, M., and Oldendick, R. (2000). A comparison of the accuracy of the last birthday versus the next birthday methods for random selection of household respondents. In: *JSM Proceedings, Survey Research Methods Section*, 887–889. Alexandria, VA: American Statistical Association. Retrieved from http://www.asasrms.org/Proceedings/papers/2000_151.pdf

Longstreth, M. and Shields, T. (2005). A comparison of within household random selection methods for random digit dial surveys. Paper presented at the annual meeting of the American Association For Public Opinion Association, 12–15 May 2005, Fontainebleau Resort, Miami Beach, FL.

Malhotra, N.K., Kim, S.S., and Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Information Systems Research* 15: 336–355.

Marlar, J., Jones, J., Manas, C., et al. (2014).Within-household selection for telephone surveys: an experiment of eleven selection methods. Paper presented at the Midwest Association for Public Opinion Research Annual Conference. 21–22 November 2014, Chicago, IL.

Martin, E. (1999).Who knows who lives here: within-household disagreements as a source of survey coverage error. *Public Opinion Quarterly* 63: 220–236.

Martin, E. (2007). Strength of attachment: survey coverage of people with tenuous ties to residences. *Demography* 44: 427–440.

Martin, E. and Dillman, D.A. (2008). Does a final coverage check identify and reduce census coverage errors? *Journal of Official Statistics* 24: 571–589.

McKenna, K.Y.A., Green, A.S., and Gleason, M.E.J. (2002). Relationship formation on the Internet: what's the big attraction? *Journal of Social Issues* 58 (1): 9–31.

Narayan, S. and Krosnick, J.A. (1996). Education moderates some response effects in attitude measurement. *Public Opinion Quarterly* 60: 58–88.

O'Rourke, D. and Blair, J. (1983). Improving random respondent selection in telephone surveys. *Journal of Marketing Research* 20: 428–432.

Olson, K. and Smyth, J.D. (2014). Accuracy of within-household selection in web and mail surveys of the general population. *Field Methods* 26 (1): 56–69.

Olson, K., Stange, M., and Smyth, J.D. (2014). Assessing within-household selection methods in household mail surveys. *Public Opinion Quarterly* 78 (3): 656–678.

Phelps, J., Nowak, G., and Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. *Journal of Public Policy & Marketing* 19: 27–41.

Raghunathan, T.E., Lepkowski, J.M., van Hoewyk, J., and Solenberger, P. (2001). A Multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology* 27: 85–95.

Raudenbush, S.W. and Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*, 2e. Newbury Park, CA: Sage.

Reich, J., Yates, W., and Woolson, R. (1986). Kish method for mail survey respondent selection. *American Journal of Public Health* 76: 206.

Rizzo, L., Brick, J.M., and Park, I. (2004). A minimally intrusive method for sampling persons in random digit dial surveys. *Public Opinion Quarterly* 68 (2): 267–274.

Rubin, D.B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: a case study using census agriculture industry codes. *Journal of Official Statistics* 3 (4): 375–387.

Schnell, R., Ziniel, S., and Coutts, E. (2007). Inaccuracy of birthday respondent selection methods in mail and telephone surveys. Presentation at the European Survey Research Association Conference, 29 June, Prague.

Segrin, C. (2000). Social skills deficits associated with depression. *Clinical Psychology Review* 20 (3): 379–403.

Singer, E. (2002).The use of incentives to reduce nonresponse in household surveys. In: *Survey Nonresponse* (ed. R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), 163–178. New York: Wiley-Interscience.

Singer, E. and Ye, C. (2013).The use and effects of incentives in surveys. *Annals of the American Academy of Political and Social Science* 645 (1): 112–141.

Smyth, J.D., Dillman, D.A., Christian, L.M., and O'Neill, A.C. (2010). Using the Internet to survey small towns and communities: limitations and possibilities in the early 21st century. *American Behavioral Scientist* 53: 1423–1448.

Stange, M., Smyth, J.D., and Olson, K. (2016). Using a calendar and explanatory instructions to aid within-household selection in mail surveys. *Field Methods* 28 (1): 64–78.

Tourangeau, R., Kreuter, F., and Eckman, S. (2012). Motivated underreporting in screening interviews. *Public Opinion Quarterly* 76 (3): 453–469.

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

Troldahl, V.C. and Carter, R.E. Jr. (1964). Random selection of respondents within households in phone surveys. *Journal of Marketing Research* 1: 71–76.

Trussel, N. and Lavrakas, P.J. (2004).The influence of incremental increases in token cash incentives on mail survey response: is there an optimal amount? *Public Opinion Quarterly* 68 (3): 349–367.

Wasserstein, R.L. and Lazar, N.A. (2016).The ASA's statement on p-values: context, process, and purpose. *The American Statistician* 70 (2): 129–133.

Valentine, C.A. and Valentine, B.L. (1971). *Missing Men: A Comparative Methodological Study of Underenumeration and Related Problems*. Washington, DC: U.S. Census Bureau. Retrieved 5 February 2016 from https://www.census.gov/srd/papers/pdf/ex2007-01.pdf

Yan, T. (2009). A meta-analysis of within-household respondent selection methods. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Hollywood, FL, USA.