



## Journal of Official Statistics vol. 30, i. 2 (2014)

- Overview of the special issue on surveying the hard-to-reach** ..... p. 171-176  
*Gordon B. Willis, Tom W. Smith, Salma Shariff-Marco, Ned English,*
- Potential uses of administrative records for triple system modeling for estimation of census coverage error in 2020**..... p. 177-190  
*Richard A. Griffin*
- Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations**..... p. 191-214  
*Kristen Himelein, Stephanie Eckman, Siobhan Murray*
- Enumerating the hidden homeless: strategies to estimate the homeless gone missing from a point-in-time count** ..... p. 215-230  
*Robert P. Agans, Malcolm T. Jefferson, James M. Bowling, Donglin Zeng, Jenny Yang, Mark Silverbush*
- A study of assimilation bias in name-based sampling of migrants**..... p. 230-250  
*Rainer Schnell, Mark Trappmann, Tobias Gramlich*
- Comparing survey and sampling methods for reaching sexual minority individuals in flanders**..... p. 251-275  
*Alexis Dewaele, Maya Caen, Ann Buysse*
- A city-based design that attempts to improve national representativeness of asians**..... p. 277-290  
*Steven Pedlow*
- Recruiting an Internet Panel Using Respondent-Driven Sampling**.....p. 291-310  
*Matthias Schonlau, Beverly Weidmer, Arie Kapteyn*
- Locating Longitudinal Respondents After a 50-Year Hiatus** ..... p. 311-334  
*Celeste Stone, Leslie Scott, Danielle Battle, Patricia Maher*
- Evaluating the Efficiency of Methods to Recruit Asian Research Participants**..... p. 335-354  
*Hyunjoo Park, M. Mandy Sha*
- Reaching Hard-to-Survey Populations: Mode Choice and Mode Preference**..... p. 355-379  
*Marieke Haan, Yfke P. Ongena, Kees Aarts*

## Overview of the Special Issue on Surveying the Hard-to-Reach

*Gordon B. Willis<sup>1</sup>, Tom W. Smith<sup>2</sup>, Salma Shariff-Marco<sup>3</sup>, and Ned English<sup>4</sup>*

### 1. Introduction

Throughout the course of the development of survey methods, critical concerns have arisen in tandem with changes in society that impact the nature and composition of the populations that researchers endeavor to understand. In recent years, a concern that has attracted considerable methodological interest involves the conducting of surveys that include members of so-called Hard-to-Reach (H2R) groups. H2R groups have become increasingly important to include within a range of population surveys, given both a burgeoning emphasis on representation of demographic subgroups (e.g., Asians within the U.S. population), and on groups that are of interest due to their potentially unique characteristics or sociocultural location (e.g., transgender individuals).

To further methodological work in this area, the International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations was held from October 31 to November 3, 2012 in New Orleans. This gathering aimed to connect researchers across a range of fields – including survey methodology, statistics, demography, sociology, anthropology, ethnography, and psychology – and engage them in discussions devoted to advancing our methodology for surveying groups that have proven difficult to include in population surveys. However, from the start this seemingly targeted and even niche-like area of interest proved somewhat hard to encapsulate in a simple definition, and finding solutions turned out to be even more challenging. In particular, the science of ‘reaching the H2R’ requires that we address two vital issues: (a) who in particular are we talking about when we use the label ‘hard-to-reach’? and (b) what do we mean by ‘reaching’ them? Considering both of these challenges, the conference advertisement (currently available at <http://www.amstat.org/meetings/h2r/2012/pdfs/H2R2012Flyer.pdf>) lists a range of topics and subpopulations of interest, and includes the following as examples of H2R groups:

- (a) Racial minorities
- (b) Immigrant populations
- (c) Indigenous populations

<sup>1</sup> Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institute of Health, 9609 Medical Center Drive, Bethesda, Maryland, 20892-9762, U.S.A. Email: [willisg@mail.nih.gov](mailto:willisg@mail.nih.gov)

<sup>2</sup> NORC, University of Chicago, 1155 East 60th street, Chicago, IL 60637, U.S.A. Email: [smith-tom@norc.org](mailto:smith-tom@norc.org)

<sup>3</sup> Cancer Prevention Institute of California, 2201 Walnut Ave, Suite 300, Fremont, CA 94538, U.S.A. Email: [salma.shariff-marco@cpic.org](mailto:salma.shariff-marco@cpic.org)

<sup>4</sup> NORC at the University of Chicago, Statistics and Methodology, 55 E. Monroe St., Suite 2000, Chicago, IL 60647, U.S.A. Email: [ENGLISH-NED@norc.org](mailto:ENGLISH-NED@norc.org)

- (d) Highly mobile and migrant populations
- (e) Homeless and refugee populations
- (f) Sexual minorities
- (g) Populations affected by natural disasters
- (h) Populations in zones of armed conflict
- (i) Stigmatized populations
- (j) Cross-cultural similarities and differences in H2R populations
- (k) Linguistic and cultural minorities

Concerning the second element defining H2R groups, or what we mean when we refer to ‘reaching’ them, [Tourangeau \(2014\)](#) has provided a helpful model within the opening chapter of a book deriving from the conference ([Tourangeau et al. 2014](#)). Tourangeau’s model delineates various points in the survey process that present difficulty, as categorized into the following bins:

- (1) Sampling/Coverage: Those who are difficult to include in a statistical sample;
- (2) Identification: Those who are difficult to identify as eligible survey respondents;
- (3) Location/Contact: Those who are difficult to find and to make contact with for purposes of engaging in a survey;
- (4) Persuasion: Those who are difficult to convince to take part in a survey, once located;
- (5) Interviewing: Those who are willing to be interviewed, but who are difficult to successfully interview.

Survey methodologists will recognize the close association between these concepts and the related statistical error subtypes of coverage, sampling, nonresponse, and response error.

## 2. The JOS Special Issue

The articles in this volume derive from contributed papers delivered at the conference, and supplement those commissioned for a book that includes chapters representing conference invited papers ([Tourangeau et al. 2014](#)). These special issue contributions address varying facets of the hard-to-reach continuum described by Tourangeau. For purposes of simplification – and perhaps based on the notion that “good things come in threes” – we have further aggregated these five elements into three general factors, each representing a basic challenge to successfully interviewing someone we think of as hard-to-reach:

- (1) *Selection*: Choosing *who* we are attempting to interview (or to ‘enumerate,’ in the case that we are counting as opposed to collecting self-report information). Sampling and coverage issues dominate, along with challenges of identification.
- (2) *Recruitment*: Deciding *how to locate* potential respondents, and, in cases where self-report is required, *how to persuade* them to consent to participating in the survey, once they have been sampled and identified.
- (3) *Interviewing*: Determining *how to conduct* the interview. Beyond sampling, identifying, locating, and persuading a potential respondent, we must also consider the survey administration mode to be used (e.g., internet, telephone, mail), as well as

Table 1. Journal of Official Statistics, Special Issue on Surveying the Hard to Reach: Attention devoted by each of the ten articles to three basics challenges in surveying the H2R.

<b>Special Issue article:</b>	<b>(1) Selection</b> <i>Who to interview:</i> Sampling and Identifying	<b>(2) Recruitment</b> <i>How to 'get' the respondent:</i> Locating, Contacting, Persuading	<b>(3) Interviewing</b> <i>How to complete the interview:</i> Administration mode, interviewer, etc.
1) <i>Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020:</i> Griffin	X		
2) <i>Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations:</i> Himelein, Eckman, and Murray	X		
3) <i>Enumerating the Hidden Homeless: Strategies to Estimate the Homeless Gone Missing From a Point-in-Time Count.</i> Agans, Jefferson, Bowling, Zeng, Yang, and Silverbush	X		
4) <i>A Study of Assimilation Bias in Name-Based Sampling of Migrants:</i> Schnell, Trappmann, and Gramlich	X	X	
5) <i>Comparing Survey and Sampling Methods for Reaching Sexual Minority Individuals in Flanders:</i> Dewaele, Caen, and Buysse	X	X	
6) <i>A City-Based Design That Attempts to Improve National Representativeness of Asians:</i> Pedlow	X	X	
7) <i>Recruiting an Internet Panel Using Respondent-Driven Sampling:</i> Schonlau, Weidmer, and Kapteyn	X	X	
8) <i>Locating Longitudinal Respondents After a 50-Year Hiatus:</i> Stone, Scott, Battle, and Maher	X	X	
9) <i>Evaluating the Efficiency of Methods to Recruit Asian Research Participants:</i> Park and Sha			X
10) <i>Reaching Hard-to-Survey Populations: Mode Choice and Mode Preference:</i> Haan, Ongena, and Aarts			X

who should conduct the interview, and what other procedural parameters may be optimal for obtaining truthful and accurate responses.

In the interest of organizational clarity, we categorize the articles in the current volume according to these three factors, as illustrated in [Table 1](#). The table depicts what we consider to be the main elements addressed by each author's contribution to the special issue (with the caveat that each research effort may span multiple areas, and our assignment to category may be arguable due to the lack of firm boundaries between them).

**Articles involving selection.** It is clear that across these articles, significant attention is paid to sampling and coverage issues – that is, with respect to *constructing* the sample frame; seven of the ten articles attend to this issue. Sample frame construction is vital for addressing the companion issue of coverage – ensuring that members of the desired population are adequately represented in the sampling frame. For study of H2R populations, a major challenge to coverage is the needle-in-a-haystack phenomenon, especially where the required population 'units' (i.e., people) are well-hidden among a larger, general population – for example, the homeless, or migrants. First, the article by Griffin focuses on the measurement of coverage error in the U.S. Census in a manner relevant to H2R subpopulations, through reference to administrative records. Himelein, Eckman, and Murray tackle the vexing challenge of surveying a special type of group – nomads – who have no set residential location, and who therefore depart from our usual notions of 'place of residence.' Similarly, Agans, Jefferson, Bowling, Zeng, Yang, and Silverbush consider the way in which members of household units can be relied upon to identify individuals – the homeless – who also have no set place of residence but who may have existing relationships with those who do (i.e., individuals who provide temporary shelter).

Several other articles focus on populations whose members do have a set place of residence, but who are difficult for the survey takers to identify and enumerate as members of an H2R group because they may be 'hidden in plain sight.' Schnell, Trappman, and Gramlich describe a study that involves the use of name-based sampling to create a frame of immigrants contained within society at large. Dewaele, Caen, and Buysse also focus on an H2R population – sexual minorities – who are integrated within the larger population, but who are not readily identifiable with respect to H2R status through any means other than self-identification. On the other hand, research by Pedlow attempts to leverage the fact that some subpopulations (in this case, Asians) do tend to be physically clustered, and can be sampled via a geographically oriented (city-based) sampling approach. Finally, Schonlau, Weidmer, and Kapteyn investigate the development of a sample frame through the use of *respondent-driven sampling*, in which no suitable sampling frame exists or can reasonably be constructed by the researchers, and instead relies upon respondents from a particular H2R group to themselves produce contact information for additional eligible individuals.

**Articles involving recruitment.** The second major factor that we have defined involves recruitment, which can be viewed as literally 'reaching' the respondents, and two of the manuscripts focus mainly on this step. Stone, Scott, Battle, and Mahar regard H2R status as imposed by the calendar. Although the sought-after respondents within this survey were not demographically unique, a fifty-year follow-up interval rendered them literally hard to

reach, and put a premium on methodology related to respondent tracing. Once the identified individuals are located through appropriate detective work, the surveyor then must begin the process of selling the survey, persuading the located individual to participate. Park and Sha consider recruitment from a somewhat different vantage point, concerning how to locate and persuade monolingual Asians for purposes of pretesting survey questionnaires. The obvious solution is to use the language that respondents (literally) use; what is less obvious, and is intensively investigated by the authors by relying on multiple efficiency metrics, is the medium to be employed: flyers, newspaper advertisements, word-of-mouth, or something else that may be specific to the H2R population.

**Articles involving interviewing.** A final article, by Haan, Ongena, and Aarts, mainly addresses the conduct of the survey interview, or how to reach respondents in terms of presenting and then obtaining information in a way that makes sense to them. The key consideration is choice of administration mode, which involves factors related both to access (e.g., do potential respondents have internet service) and social dynamics (do they prefer interaction with a human interviewer – that is, interviewer administration, or would they rather answer a computer screen or a piece of paper, under questionnaire self-administration?). The dynamics of interviewing could address a range of other parameters as well, such as interviewer demographic characteristics and behavior, or the physical location of the interview (at home, or elsewhere).

### 3. Looking to the Future

The articles in this volume attempt to provide answers to the issues listed above, in most cases through use of examples and case studies. However, they also raise fundamental issues that challenge the fledgling science of surveying the hard-to-reach. Reflection on the difficulties and barriers to the enumeration process, or to the conduct of self-report surveys, may even suggest the need for a subtle but important shift in investigator viewpoint – and perhaps in nomenclature. We must recognize that, from the respondent's point of view, he or she may not be at all 'hard to reach.' There are certainly subtypes of potential respondents who truly are challenging to select, recruit, and interview because those individuals take steps to make each of these steps difficult (e.g., undocumented individuals who hide from the administrators of a Federal survey). There are others, however, who present difficulties mainly from the perspective of the researcher. We may fail to reach monolingual Vietnamese speakers largely because we fail to enlist interviewers who can communicate in that language. By way of analogy, one can state that Timbuktu is hard to reach – but this is true only from the vantage point of Western locations, and not if one begins the trip from a nearby town in Mali. As such, the application of the label 'H2R' to a particular group may mainly reflect the separation between researcher and respondent, rather than some immutable characteristic of the latter.

An alternative to 'Hard to Reach' is suggested by [Tourangeau \(2014\)](#), who has selected the general term 'Hard to Survey,' and who regards 'Hard-to-Reach' as one subcategory of specific difficulties encountered (literally, those who are hard to locate and to contact, once they have been identified). This solution may not placate those who object that the general use of the term 'Hard' carries the implication that such members of some groups are 'resistant' or 'uncooperative.' As an alternative, a judgment-neutral approach might be

to state simply that some groups are ‘historically under-represented’ to convey the notion that certain populations have tended to be left out of survey (and other) research.

A final, opposing perspective is that survey methodologists need not be overly concerned with labeling, but rather with the ultimate outcome of their survey practices. If researchers increasingly are committed to enhancing our capacities for including those who are left out, and are committed to expending the appropriate resources and/or effort (as are those represented in the current volume), then it may not matter whether they regard their respondents as hard-to-reach, as under-represented, or – from the point of view of staff on the ground – as ‘difficult completions.’ It is our hope and intent that the directions defined within the current set of manuscripts provide a path towards the continued development of imaginative and effective methodologies for ensuring that our surveys and Censuses are equally inclusive of all.

#### **4. References**

- Tourangeau, R. (2014). Defining Hard-to-Survey Populations. In *Hard-to-Survey Populations*, R. Tourangeau, B. Edwards, T.P. Johnson, K.M. Wolter, and N. Bates (eds). Cambridge: Cambridge University Press, (In Press).
- Tourangeau, R., Edwards, B., Johnson, T.P., Wolter, K.M., and Bates, N. (2014). *Hard-to-Survey Populations*. Cambridge: Cambridge University Press, (In Press).

## Potential Uses of Administrative Records for Triple System Modeling for Estimation of Census Coverage Error in 2020

*Richard A. Griffin*<sup>1</sup>

Heterogeneity in capture probabilities is known to produce bias in the dual system estimates that have been used to estimate census coverage in U.S. Censuses since 1980. Triple system estimation using an administrative records list as a third source along with the census and coverage measurement survey has the potential to produce estimates with less bias. This is particularly important for hard-to-reach populations.

The article presents potential statistical methods for the estimation of net census undercount using three systems for obtaining population information: (1) a decennial census; (2) an independent enumeration of the population in a sample of block clusters; and (3) administrative records. The 2010 Census Match Study will create census-like files for the entire nation using federal and commercial sources of administrative records. The 2010 Census Coverage Measurement Survey is an enumeration in a sample of block clusters that is independent of the 2010 decennial Census.

*Key words:* Heterogeneity; independence; log-linear model.

### 1. Introduction

Heterogeneity in capture probabilities is known to produce bias in the dual system estimates (DSE) which have been used to estimate census coverage in U.S. Censuses since 1980. Triple system estimation using an administrative records list as a third source along with the census and postenumeration survey (PES) has the potential to produce estimates with less bias. This is particularly important for hard-to-reach populations. Based on theory in [Bell \(1993\)](#), the bias in DSE due to causal dependence or heterogeneity in capture probabilities may be greater for hard-to-reach populations. Some of the many references for the theory and practice of Dual System Estimation are [Chandrasekar and Deming \(1949\)](#), [Wolter \(1986\)](#), [Alho \(1990\)](#), and [Mulry and Spencer \(1991\)](#).

For the 2020 Census postenumeration survey, we are carrying out a preliminary investigation on using Triple System Estimation (TSE). The three systems for obtaining population information for TSE are: (1) a decennial census; (2) an independent enumeration of the population in a sample of block clusters; and (3) administrative records.

<sup>1</sup> U.S. Census Bureau, 4210 Southwinds Place Unit 109, White Plains, Maryland 20695, U.S.A. Email: [richard.a.griffin@census.gov](mailto:richard.a.griffin@census.gov)

**Acknowledgments:** This report is released to inform interested parties of research and to encourage discussion. Any views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the U.S. Census Bureau.



For this article, all data are simulated and there is no sampling. When administrative records are mentioned the reader should bear in mind that any real application would use census-like files for the entire nation, using federal and commercial sources of administrative records similar to those created for the 2010 Census Match Study (Rastogi and O'Hara 2012). Similarly, in practice a PES would be an enumeration in a sample of block clusters independent of the census, like the 2010 Census Coverage Measurement Survey.

For this simulation study, it is assumed that all  $N$  individuals in the population are exposed to possible inclusion in all three sources. In practice, sampling is necessary for the postenumeration survey and possibly the administrative list (due to the necessity of follow-up for unresolved match status). Table 1 illustrates the eight cells indicating the possible combinations of captured or not captured by each of the three attempts at enumeration. The count of the population total in each cell is defined as  $N_{jkl}$  where the subscripts  $j$ ,  $k$ , and  $l$  are 1 or 0 to indicate captured or not captured in the Census list, the postenumeration survey, and the administrative list respectively. For example,  $N_{110}$  is the count of persons captured by the Census and PES but not captured by the administrative list. All cells are conceptually observable except  $N_{000}$ .

Creation of the simulated populations assumes autonomous independence, which means that the Census list, the postenumeration survey list and the administrative list are created as a result of  $N$  mutually independent trials from one person to the next (all persons are captured independently of all other persons, even persons in the same household). The counts in Table 1 and all the estimators studied in this article could be constructed even if autonomous independence did not hold. Autonomous dependence could create additional bias in estimates.

The Census Bureau has used dual system estimation for census net error estimation starting with the 1980 Census. The incomplete  $2^3$  table of counts for triple system estimation can be divided into one complete  $2 \times 2$  subtable and one incomplete  $2 \times 2$  subtable. The additional source from administrative records provides data with which to evaluate the previously untestable assumption of independence between the census and the postenumeration survey. Evidence is available in the triple-system tables for odds ratios in  $2 \times 2$  subtables formed by restricting consideration to cases observed in the administrative records source. In this case, complete information is available for all four cells defined by capture or noncapture in the census and postenumeration survey. This additional information is used to formulate the triple system estimates using any of an assortment of model assumptions.

For populations of size 1,000, this article presents simulations for ten estimators of persons missed on all lists, each of which can be combined with observed counts to

Table 1. Population counts by capture status

	In AL		Out of AL	
	In PES 1	Out of PES 0	In PES 1	Out of PES 0
In Census 1	$N_{111}$	$N_{101}$	$N_{110}$	$N_{100}$
Out of Census 0	$N_{011}$	$N_{001}$	$N_{010}$	$N_{000}$

produce estimates of the total population. Each estimate is compared with the corresponding true population value. The ability to estimate the dependence between the census and postenumeration survey for persons *not on* the administrative list (using persons *on* the administrative list) may reduce bias in the estimation of census coverage error. With dual system estimation, we cannot achieve this reduction in correlation bias in the presence of dependence and heterogeneity because we have no data available to estimate the dependence. Log-linear model theory can be useful in formulating and understanding some triple system estimators. These models are supported by empirical evidence that capture in the census or coverage measurement list is only weakly associated with capture on the administrative list. This is plausible since the administrative list is created in a radically different way than the census list and postenumeration survey list, which are independent surveys using similar fieldwork. The ten estimators are described in Section 2.

Seven of these estimators are motivated by hierarchical log-linear models based on Fienberg (1972). Additional references for log-linear models are Bishop et al. (1975), Fienberg (2000), and Agresti (2002). Two of the estimators are based on suggestions from Zaslavsky and Wolfgang (1990 and 1993). For comparison, the traditional dual system estimate (DSE) using only the decennial census and postenumeration survey will be computed.

Other triple system estimators using alternative models, not simulated for this article, are suggested by Darroch et al. (1993). They built an equivalence for the generalized Rasch model and the quasi-symmetric log-linear model. They compared estimates from a partial quasi-symmetry model and a full quasi-symmetry model with the no second order interaction estimator (see Subsection 2.3) as well as with the Zaslavsky and Wolfgang estimators (see Subsections 2.4 and 2.5). Chao and Tsay (1998) developed an estimator that is a function of an expected sample coverage (based on an average over the three lists of the proportions of persons observed as missed on the other two lists) and measures of dependence between lists. They compared their estimator with those of Darroch et al. Fienberg and Manrique-Vallier (2009) looked at a methodology for integrating these multiple system estimation methods with record linkage and missing data issues. Madigan and York (1997) developed a Bayesian methodology that allows for a variety of dependence structures between lists, uses covariates, and explicitly accounts for model uncertainty.

The following assumptions apply to all estimators: (1) Erroneous inclusions have been removed from all lists and (2) Processing and matching procedures have been developed so that there is no matching error as well as no error in the determination that a person is enumerated at the correct address. Section 3 describes the creation of a simulated population of 1,000 persons and Section 4 discusses the replication of this process and the creation of evaluation statistics. Section 5 presents the results and Section 6 provides a discussion.

## 2. Estimators to Be Simulated

All these estimates are motivated based on an assumption of homogeneity in capture probabilities across individual persons. If the particular log-linear model assumptions hold

and capture probabilities are homogeneous, then these estimators are nearly unbiased. Since individual capture probabilities are heterogeneous in the real world, the simulated populations are created using heterogeneous capture probabilities. Estimates using these models are biased given this heterogeneity and we can compare these estimates with the known population total.

### 2.1. Conditionally Independent Models

In order to use common log-linear model notation, let C denote Census, P denote the postenumeration survey, and A denote the administrative list. For example, consider the log linear model {CP, PA}. This log-linear model notation puts sources together if there is an assumed relationship (dependence) between them. This is a conditional independence model where at each level of P, C and A are independent, a unconditional relationship between C and P and between P and A is allowed but not between C and A. Since there is some empirical evidence (Zaslavsky and Wolfgang 1990, 1993, and Darroch et al. 1993) that the C and P lists are dependent conditional on capture on the A list, this model may be reasonable. The same is true for the model {CP, CA}. The third model with exactly two two-factor terms, {CA, PA}, may not be accurate since it assumes at each level of A that C and P are independent and this is not supported by the empirical evidence. Note that the empirical evidence from Zaslavsky, Wolfgang and Darroch is from the 1988 Census Dress Rehearsal and is based on administrative data limited to a few specific geographic areas and based on sources likely to be very different from any sources that might be used for the 2020 Census postenumeration study.

For model {CP, PA} the estimate is  $\hat{N}_{000}^1 = \frac{N_{001}N_{100}}{N_{101}}$ . This is the usual dual system estimate for the unobserved cell in the  $2 \times 2$  table conditional on  $P = 0$ , using the A list and the C list as sources after removing all individuals captured on the P list and assuming A and C are independent.

Models {CP, CA} leading to the estimate  $\hat{N}_{000}^2 = \frac{N_{001}N_{010}}{N_{011}}$  and {CA, PA} leading to the estimate  $\hat{N}_{000}^3 = \frac{N_{010}N_{100}}{N_{110}}$  follow from the appropriate permutations of the capture status indices.

### 2.2. Jointly Independent Models

For example, consider the log-linear model {A, CP} where there is a relationship between C and P, but neither C nor P has a relationship with A. This is a jointly independent model where A is jointly independent of C and P. This is ordinary two-way independence between A and a categorical variable composed of all four combinations of C and P. Given the empirical evidence cited above, this model might be reasonable. The other two jointly independent models, {P, CA} and {C, PA}, assume C and P are independent, but this is not supported by the empirical evidence.

For model {A, CP} the estimate is  $\hat{N}_{000}^4 = \frac{N_{001}(N_{110}+N_{100}+N_{010})}{N_{111}+N_{101}+N_{011}}$ . This is equivalent to a DSE where one list is the administrative list and the other is a list formed by combining the census and PES list (un-duplication required). The combined list is assumed to be independent from the administrative list.

Model {P, CA} leading to the estimate is  $\hat{N}_{000}^5 = \frac{N_{010}(N_{101}+N_{001}+N_{100})}{N_{111}+N_{011}+N_{110}}$  and model {C, AP} leading to the estimate is  $\hat{N}_{000}^6 = \frac{N_{100}(N_{011}+N_{001}+N_{010})}{N_{111}+N_{101}+N_{110}}$ ; both follow from the appropriate permutations of the capture status indices.

### 2.3. No-Second-Order-Interaction Log-Linear Model

There is only one no-second-order-interaction log-linear model. Model {CP, CA, PA} assumes that the Census and postenumeration survey have dependence but there is no CPA term (three-way interaction). This is the least restrictive log-linear model for which data is available for estimation. All log-linear models from Subsections 2.1 and 2.2 are special cases of the no-second-order-interaction model (i.e., they all assume no second-order interaction along with additional restrictions).

The incomplete  $2^3$  table of counts in Table 1 is divided into one complete  $2 \times 2$  subtable and one incomplete subtable. Assume the cross-product ratio is the same in both subtables. Then the estimate of the missing cell in the incomplete  $2 \times 2$  table can be estimated using the known cross-product ratio from the complete  $2 \times 2$  table. The assumption is that the dependence in the  $2 \times 2$  table for  $C \times P$  using only those individuals in A is the same as the dependence in the  $2 \times 2$  table for  $C \times P$  using only those individuals not in A. This model is in some sense analogous to the assumption of independence for the  $2 \times 2$  table used for DSE but is one layer deeper. All pairs of sources can exhibit dependence, but the amount of dependence in each pair is assumed to be unaffected by conditioning on the third source. The estimator for this model is

$$\hat{N}_{000}^7 = \frac{(N_{111})(N_{001})(N_{100})(N_{010})}{(N_{011})(N_{101})(N_{110})}.$$

Note that in order to estimate  $N_{000}$ , it is necessary to make an assumption about second-order interaction. This assumption does not have to be that the interaction term in the log-linear model is zero; any other fixed value for the interaction coefficient could be used, although some assumptions might be more plausible than others.

### 2.4. Zaslavsky and Wolfgang 1

This is a DSE, suggested in Zaslavsky and Wolfgang (1990 and 1993), where one source is the administrative list and the other is the combined census and census coverage measurement list. However, persons captured in both the census and postenumeration survey are removed from the administrative list and the combined list.

$$\hat{N}_{000}^8 = N_{001} \frac{N_{100} + N_{010}}{N_{101} + N_{011}}$$

The assumption underlying the use of this estimator is that the probability of capture in the administrative list of persons omitted from the census and postenumeration survey is the same as the average probability of capture for those included in either the census or postenumeration survey, but not both. In other words, persons captured by neither C nor P are more like those captured by only the C or P than those captured by both. This estimator was included by Zaslavsky and Wolfgang based on evidence for four poststrata studied taken from the 1988 Census Dress Rehearsal.

### 2.5. Zaslavsky and Wolfgang 2

For this estimator, also suggested by Zaslavsky and Wolfgang (1990 and 1993), the odds ratio in the  $2 \times 2$  table fixing on capture in the administrative list is calculated. Then, assuming this odds ratio holds, the DSE for the 00+ cell of the marginal  $C \times P$  table is multiplied by this odds ratio.

$$\hat{N}_{000}^9 = \left( \frac{N_{001}N_{111}}{N_{011}N_{101}} \right) \left( \frac{N_{01+}N_{10+}}{N_{11+}} \right) - N_{001}$$

The count in the 001 cell is subtracted to obtain an estimate of the 000 cell. The assumption is that the degree of dependence between the C and P sources is similar in the overall population to that in the subpopulation captured by the administrative list. Zaslavsky and Wolfgang note that this assumption may be conservative for the population as a whole, because the administrative list captures are likely to be more homogeneous than the general population and the odds ratio would be closer to 1 (independence would more nearly hold).

### 2.6. Traditional DSE

For comparison, this is the DSE estimate using only the Census list and postenumeration survey list. The assumption is that C and P are unconditionally independent {C, P}.

$$\hat{N}_{000}^{10} = \frac{N_{10+}N_{01+}}{N_{11+}} - N_{001}$$

### 2.7. Population Total Estimates

For each of the  $t = 1$  to 10  $\hat{N}_{000}^t$  estimates calculated, the total population estimate is

$$\hat{N}^t = \hat{N}_{000}^t + N_{1++} + N_{011} + N_{010} + N_{001}.$$

## 3. Creating the Simulated Populations

Populations of  $N = 1,000$  persons will be simulated, allowing for heterogeneous capture probabilities and homogeneous conditional odds ratios. One conditional odds ratio is the odds ratio for the  $2 \times 2$  table of  $C \times P$  conditional on capture on A and the other is the odds ratio for the  $2 \times 2$  table of  $C \times P$  conditional on not captured (missed) on A.

### 3.1. Creating a Specified Conditional Odds Ratio

Omitting any subscript for an individual member of the population, the  $2 \times 2$  table of conditional capture probabilities for census capture and postenumeration survey capture given capture on the administrative list is given in Table 2.

In order to create a simulated population with a given set of conditional odds ratios, the odds ratio formula for a  $2 \times 2$  subtable is written as a function of an unknown proportion in the 11 cell and the known marginal proportions the 1+ and +1 margins.

Table 2. Capture probabilities for Census and PES given capture on administrative list

	In PES 1	Out of PES 0	
In Census 1	$P_{11}$	$P_{10}$	$P_{1+}$
Out of Census 0	$P_{01}$	$P_{00}$	
	$P_{+1}$		

Accordingly, given  $P_{1+}$ ,  $P_{+1}$ , and odds ratio

$$\theta = \frac{P_{11}P_{00}}{P_{10}P_{01}} = \frac{P_{11}(1 - P_{1+} - P_{+1} + P_{11})}{(P_{1+} - P_{11})(P_{+1} - P_{11})},$$

the equation can be rewritten as

$$(1 - \theta)P_{11}^2 + [1 - P_{1+} - P_{+1} + \theta(P_{1+} + P_{+1})]P_{11} - \theta P_{1+}P_{+1} = 0. \tag{1}$$

This equation can be solved for  $P_{11}$  using the quadratic formula producing two roots, one of which is between 0 and 1 and is the one we want.

This value of  $P_{11}$  and given  $P_{1+}$  and  $P_{+1}$  provides the desired odds ratio  $\theta$ .

The process described starting with Table 2 is repeated for Capture Probabilities for census and PES given not captured (missed) on the administrative list, allowing in some simulations for a different conditional odds ratio  $\theta$ .

### 3.2. Generating a 1,000 Person Population Allowing for Heterogeneity in Capture Probabilities

We want to generate several populations of size  $N=1,000$  persons to have particular capture properties. This is accomplished by specifying two conditional odds ratios.

Let  $\theta_1$  be the odds ratio for census and PES given capture on the administrative list and  $\theta_2$  the odds ratio for census and PES given *not* captured on the administrative list.

Given  $\theta_1$  and  $\theta_2$  (assumed constant over persons) and five beta parameters in the following conditional capture probabilities

$$P_k\langle A \rangle = \frac{\exp(\beta_{10} + \beta_{11}X_k)}{1 + \exp(\beta_{10} + \beta_{11}X_k)}, \quad P_k\langle C|A \rangle = \frac{\exp(\beta_{20} + \beta_{21}X_k)}{1 + \exp(\beta_{20} + \beta_{21}X_k)},$$

$$P_k\langle P|A \rangle = \frac{\exp(\beta_{30} + \beta_{31}X_k)}{1 + \exp(\beta_{30} + \beta_{31}X_k)}, \quad P_k\langle C|notA \rangle = \frac{\exp(\beta_{40} + \beta_{41}X_k)}{1 + \exp(\beta_{40} + \beta_{41}X_k)},$$

$$P_k\langle P|notA \rangle = \frac{\exp(\beta_{50} + \beta_{51}X_k)}{1 + \exp(\beta_{50} + \beta_{51}X_k)},$$

for  $k=1$  to 1,000 independently generate  $X_k \sim N(0,1)$  and calculate

$$P_k\langle A \rangle, P_k\langle C|A \rangle, P_k\langle P|A \rangle, P_k\langle C|notA \rangle, P_k\langle P|notA \rangle.$$

Note that although the conditional odds ratios are assumed constant over persons, the capture probabilities are heterogeneous since variation in the independent variables is created.

Using  $\theta_1$  and  $P_k\{C|A\}, P_k\{P|A\}$ , we use the methodology from Subsection 3.1 and Equation (1) to solve for the probability of capture in both the census and postenumeration survey given capture on the administrative list. Then complete the  $2 \times 2$  table of capture probabilities given capture on the administrative list. Multiplying each of these conditional probabilities by  $P_k\{A\}$  provides  $p_{k,111}, p_{k,101}, p_{k,011}, p_{k,001}$ .

Then, using  $\theta_2$  and  $P_k\{C|notA\}, P_k\{P|notA\}$ , use the methodology from Subsection 3.1 and Equation (1) to solve for the probability of capture in both the census and postenumeration survey given *not* captured on the administrative list. Then complete the  $2 \times 2$  table of capture probabilities given *not* captured on the administrative list. Multiplying each of these conditional probabilities by  $(1 - P_k\{A\})$  provides  $p_{k,110}, p_{k,100}, p_{k,010}, p_{k,000}$ .

Next, generate a number  $u$  from 0 to 1 from the  $U(0,1)$  distribution and use the cumulative distribution of the eight cell probabilities to determine which of the eight cells of Table 1 person  $k$  falls into.

After completing the above for each of the 1,000 population persons, tabulate the seven observed counts from Table 1 and using these compute  $\hat{R}^t = \frac{N^t}{1000}$  for  $t = 1$  to 10. This is the ratio of the estimated population count to the true population count and provides a measure of the accuracy of the estimate.

#### 4. Replication

This article presents results for 1,000 independent replications of the population generation as specified in 3.2 for a given  $\theta_1$  and  $\theta_2$  (assumed constant over persons) and one set of beta parameters (shown in Table 3). This set of beta parameters was selected as they produce average capture probabilities, described in Section 5, that are small (.227), and thus represent what may be considered a hard-to-reach population.

Table 3. Accuracy of alternative estimates of missing count

<b>“Average Capture Probability” = 0.227</b>				
Average $R = \text{Estimated Count/True Count (se)}$				
$\beta_{10} = -0.700$	$\beta_{11} = 0.800$	$\beta_{20} = -1.200$	$\beta_{21} = .500$	
$\beta_{30} = -1.000$	$\beta_{31} = 0.600$	$\beta_{40} = -2.000$	$\beta_{41} = -.300$	
$\beta_{50} = -1.500$	$\beta_{51} = -0.400$			
Estimator	$\theta_1 = 1.5$ $\theta_2 = 1.2$	$\theta_1 = .75$ $\theta_2 = .85$	$\theta_1 = .75$ $\theta_2 = .75$	$\theta_1 = 1.5$ $\theta_2 = 1.5$
1	.965 (.002)	.958 (.002)	<b>.972 (.002)</b>	<b>.946 (.002)</b>
2	1.535 (.007)	1.455 (.007)	1.428 (.006)	1.488 (.007)
3	1.021 (.003)	1.178 (.004)	1.240 (.005)	<b>.947 (.003)</b>
4	1.100 (.002)	1.093 (.002)	1.091 (.002)	1.086 (.002)
5	1.242 (.003)	1.369 (.004)	1.390 (.004)	1.174 (.003)
6	.968 (.002)	<b>1.013 (.002)</b>	<b>1.032 (.002)</b>	.937 (.002)
7	1.177 (.008)	1.022 (.007)	1.056 (.007)	1.060 (.006)
8	1.087 (.002)	1.075 (.002)	1.077 (.002)	1.063 (.002)
9	1.197 (.008)	<b>1.018 (.008)</b>	1.053 (.008)	1.074 (.007)
10	<b>.993 (.004)</b>	1.295 (.004)	1.377 (.007)	.915 (.003)

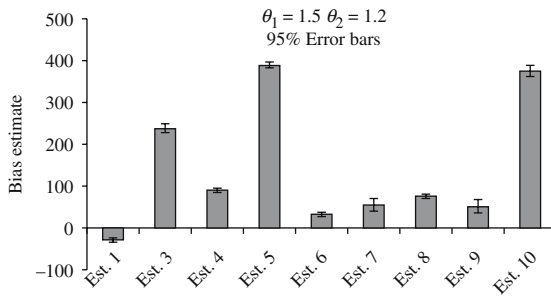


Fig. 1. 95% error bars

For each of the ten estimates, use these 1,000 replicates to compute the empirical mean ratio  $R^t$  denoted as  $\bar{R}^t$ , and its variance,  $Var(\bar{R}^t)$ .

Note that none of the precise assumptions, particularly homogeneity in capture probability, needed for validity of any of these ten estimators is satisfied by any of these simulated populations. Darroch et al. (1993) provide some arguments that no three-way interaction model may be a fair approximation except for heterogeneity. The kind of person-to-person heterogeneity introduced by these simulations might be expected to be a reasonable representation of the reality of list formation. This heterogeneity produces bias in these estimates even if the model assumptions about the relationship between the capture attempts hold.

### 5. Results

Table 3 shows results for each of the ten estimator alternatives for one set of  $\beta$  parameters and four sets of odds ratios  $\theta_1$  and  $\theta_2$  (1.5 and 1.2; .75 and .85; .75 and .75; 1.5 and 1.5). When  $\theta_1 = \theta_2$ , the odds ratio for census capture or not by postenumeration survey capture status is independent of capture status on the administrative list (no second order interaction). When  $\theta_1 \neq \theta_2$  the odds ratio for census capture or not by postenumeration survey capture status is dependent on capture status on the administrative list. The ‘‘Average Capture Probability’’ (ACP) is the average of the five probabilities defined in Section 3 for  $X_k = 0$  (the mean of the random variable  $X$ ). It is used as a measure of

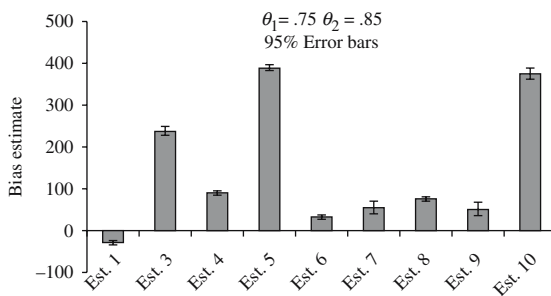


Fig. 2. 95% error bars



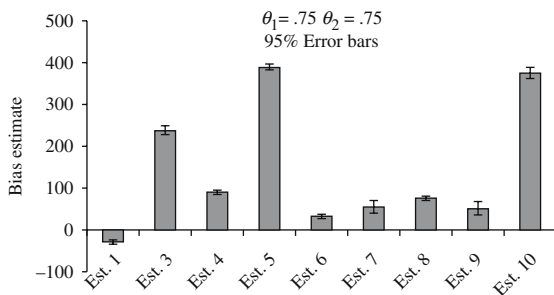


Fig. 3. 95% error bars

“hard to reach” since lower values indicate lower capture probabilities (i.e., harder to reach).

$$ACP = \frac{\sum_{i=1}^5 \frac{e^{\beta_{i0}}}{1 + e^{\beta_{i0}}}}{5}$$

There are ten rows of average ratios  $R$  defined in Section 3 with the standard error of the average  $R$  in parenthesis, one row for each of the ten estimators of total population. There are four columns, one for each  $\theta_1$  and  $\theta_2$  combination. For each  $\theta_1$  and  $\theta_2$  combination, the average  $R$ -value that is closest to 1 is in bold. If the second-best average  $R$  is not statistically different (single pair comparison) than the best, it is shown in bold italics. The standard errors are small (all coefficients of variation less than 0.01). Thus the results are similar for many of the estimators, except for Estimator 2 which produced a large overestimate (close to 50%) for all four columns. To illustrate this, for each of the four sets of odds ratios, a 95% confidence interval error bar chart for the bias estimate is also provided (as Figures 1 through 4) excluding Estimator 2.

For Table 3, the average capture probability was .227. For  $\theta_1 = 1.5$  and  $\theta_2 = 1.2$ , Estimator 10 was the best with an average  $R$  of .993 (se = .004). For  $\theta_1 = .75$  and  $\theta_2 = .85$ , Estimator 6 was the best with an average  $R$  of 1.013 (se = .002). For  $\theta_1 = .75$  and  $\theta_2 = .75$ , Estimator 1 was the best with an average  $R$  of .972 (se = .002). For  $\theta_1 = 1.5$  and  $\theta_2 = 1.5$ , Estimator 1 was the best with an average  $R$  of .947 (se = .003).

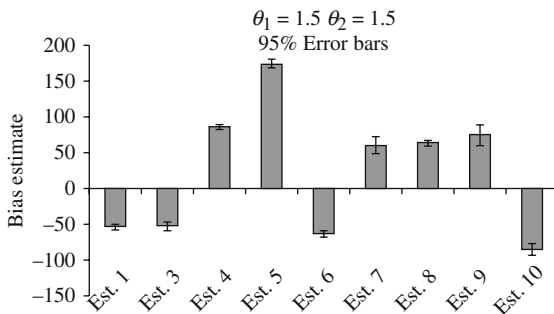


Fig. 4. 95% error bars

### 6. Discussion

Different conditional odds ratios and beta parameters, as well as a new generation of the independent random  $X$  variables for each iteration, would produce different results; thus the simulations shown here serve as an example and only as an indication of what may be expected with varying parameters. Even with the same odds ratios and beta parameters, generating a new 1,000 person population, as described in Section 3, produces different results.

Although it is clear from these results that three sets of capture attempts can produce more accurate estimates than two capture attempts, there are additional things worth considering. First, the cost of three enumeration attempts is considerably greater than for two enumeration attempts. Second, there is likely to be greatly increased matching error going from two attempts to three attempts. For two attempts at capture, there are only four cells in a  $2 \times 2$  table, and given the marginal counts of the total count for each of the attempts, matching is only necessary to obtain the 11 cell (captured in both attempts). For three attempts, there are eight cells. For Estimate 7, no second-order interaction, counts are required for all the other seven cells in order to estimate the 000 cell. Estimate 7 makes a less restrictive assumption (no second-order interaction) than the estimators from Subsection 1.1 (conditionally independent models) and Subsection 1.2 (jointly independent models). In theory, Estimate 7 should be the better than the estimates in Subsections 1.1 and 1.2 as well as 1.6 (traditional DSE), if in reality there is a second-order interaction and if there are no errors in obtaining the counts. Second-order interaction and heterogeneity in capture probabilities are likely in the real world for most populations. For example, both the 111 cell and the 110 cell are required so that both the count of captured in the first two attempts and in the third attempt *and* captured in the first two attempts but missed in the third are necessary. Obtaining all these counts from a complex matching operation may be error prone. Further research using some reasonable matching-error models is planned to investigate whether it may be more effective to use less optimal estimators that require less matching but may be more robust to matching error.

For the simulations in Table 3, Table 4 shows the average ratios  $R$  for the total population estimate for each of the four sets of  $\theta_1$  and  $\theta_2$  for the DSE and the best (lowest  $ABS(R-1)$ ) of all ten estimators of total population. Note that although the standard errors of average  $R$  values are small, for some simulations the second-best estimator was not significantly different than the best estimator. The absolute value ( $ABS$ ) of  $R-1$  is shown for DSE and the best of the ten estimators. This is the absolute relative error.

Table 4. Accuracy of total population estimate:  $R = \text{average estimated total population}/1,000$

$\theta_1$	$\theta_2$	Average $R$ for DSE	$ABS(R-1)$ for DSE	Estimator with best average $R$	Best average $R$	$ABS(S-1)$ for Best	Difference in absolute error DSE – Best
1.5	1.2	0.993	0.007	10 (DSE)	0.993	0.007	0.000
.75	.85	1.295	0.295	6 {C,AP}	1.013	0.013	0.282
.75	.75	1.377	0.377	1{CP,PA}	0.972	0.028	0.349
1.5	1.5	0.915	0.085	1{CP,PA}	0.946	0.054	0.031

For example, the maximum difference is found in Table 3 for  $\theta_1 = .75$  and  $\theta_2 = .75$  where the absolute relative error for the best estimator, Estimator 1, is 2.8% and the absolute relative error for DSE is 37.7%, a 34.9 percentage point difference.

When considering the accuracy of DSE, which requires only two sets of enumeration attempts and is less subject to matching error, it is important to compare two enumeration attempts with one enumeration attempt. For example for  $\theta_1 = .75$  and  $\theta_2 = .75$  the difference in absolute error between the best triple system estimator and the DSE was 34.9 percentage points. The average capture probability is .227. If the capture probability was a constant .227, one capture attempt for the population of 1,000 would have an expected capture of 227 persons and absolute error of 77.3%. The DSE absolute error of 37.7% is much less. Thus two capture attempts followed by DSE (with a 37.7% absolute error) may produce a substantial gain over one capture attempt (with a 77.3% absolute error) even if the absolute relative error of DSE is still rather high. In practice, while likely not sufficient for a Decennial Census, the two independent capture attempts, (1) an attempted 100% enumeration of a hard-to-reach population and (2) the creation of a list using administrative records, followed by dual system estimation may produce a much more accurate population estimate than relying on only one capture attempt.

## 7. References

- Agresti, A. (2002). *Categorical Data Analysis* (2nd edition). New York: John Wiley and Sons.
- Alho, J.M. (1990). Logistic Regression in Capture-Recapture Models. *Biometrics*, 46, 623–635.
- Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.
- Bell, W.R. (1993). Using Information from Demographic Analysis in Post-Enumeration Survey Estimation. *Journal of the American Statistical Association*, 88, 1106–1118. DOI: <http://www.dx.doi.org/10.1080/01621459.1993.10476381>
- Chandrasekar, C. and Deming, W.E. (1949). On a Method of Estimating Birth & Death Rates and the Extent of Registration. *Journal of the American Statistical Association*, 44, 101–115.
- Chao, A. and Tsay, P.K. (1998). A Sample Coverage Approach to Multiple-System Estimation with Aoolication to Census Undercount. *Journal of the American Statistical Association*, 93, 283–293.
- Darroch, J., Fienberg, S., Glonek, G., and Junker, B. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation with Heterogeneous Catchability. *Journal of the American Statistical Association*, 88, 1137–1148. DOI: <http://www.dx.doi.org/10.1080/01621459.1993.10476387>
- Fienberg, S. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2k Contingency Tables. *Biometrika*, 59, 591–603. DOI: <http://www.dx.doi.org/10.1093/biomet/59.3.591>
- Fienberg, S.E. (2000). Contingency Tables and Log-Linear Models: Basic Results and New Developments. *Journal of the American Statistical Association*, 95, 643–647. DOI: <http://www.dx.doi.org/10.1080/01621459.2000.10474242>

- Fienberg, S.E. and Manrique-Vallier, D. (2009). Integrated Methodology for Multiple Systems Estimation and Record Linkage using a Missing Data Formulation. *AStA Advances in Statistical Analysis*, 93, 49–60. DOI: <http://www.dx.doi.org/10.1007/s10182-008-0084-z>
- Madigan, D. and York, J.C. (1997). Bayesian Methods for Estimation of the Size of a Closed Population. *Biometrika*, 84, 19–31. DOI: <http://www.dx.doi.org/10.1093/biomet/84.1.19>
- Mulry, M.H. and Spencer, B.D. (1991). Total Error in PES Estimates. *Journal of the American Statistical Association*, 86, 839–855. DOI: <http://www.dx.doi.org/10.1080/01621459.1991.10475122>
- Rastogi, S. and O'Hara, A. (2012). 2010 Census Match Study. 2010 Census Program for Evaluations and Experiments, Center for Administrative Records Research and Applications. November 19, 2012.
- Wolter, K.M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, 81, 338–346. DOI: <http://www.dx.doi.org/10.1080/01621459.1986.10478277>
- Zaslavsky, A.M. and Wolfgang, G.S. (1990). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (Anaheim, CA, August 1990).
- Zaslavsky, A.M. and Wolfgang, G.S. (1993). Triple-System Modeling of Census, Post-Enumeration Survey, and Administrative List Data. *Journal of Business & Economic Statistics*, 11, 279–288. DOI: <http://www.dx.doi.org/10.1080/07350015.1993.10509955>

Received February 2013

Revised January 2014

Accepted February 2014

## Sampling Nomads: A New Technique for Remote, Hard-to-Reach, and Mobile Populations

*Kristen Himelein<sup>1</sup>, Stephanie Eckman<sup>2</sup>, and Siobhan Murray<sup>3</sup>*

Livestock are an important component of rural livelihoods in developing countries, but data about this source of income and wealth are difficult to collect due to the nomadic and seminomadic nature of many pastoralist populations. Most household surveys exclude those without permanent dwellings, leading to undercoverage. In this study, we explore the use of a random geographic cluster sample (RGCS) as an alternative to the household-based sample. In this design, points are randomly selected and all eligible respondents found inside circles drawn around the selected points are interviewed. This approach should eliminate undercoverage of mobile populations. We present results of an RGCS survey with a total sample size of 784 households to measure livestock ownership in the Afar region of Ethiopia in 2012. We explore the RGCS data quality relative to a recent household survey, and discuss the implementation challenges.

*Key words:* GIS; cluster sampling; pastoralists; livestock surveys.

### 1. Introduction

Livestock ownership comprises a large part of rural wealth and well-being in the developing world, serving diverse functions from food source to savings and investment vehicle. The sector, however, has recently come under increasing pressure from a number of sources, including increased demand for meat and dairy products from the expanding middle class, climate change, and loss of traditional pasture land to development. Efforts to understand these evolving dynamics, and their impact on the welfare of livestock-owning households, are hampered by a lack of high-quality data on which to base analyses. Beyond the general data collection issues of definition and quantification, livestock

<sup>1</sup> World Bank – Development Economics Research Group, 1818 H St. NW Washington District of Columbia 20433, U.S.A. Email: khimelein@worldbank.org

<sup>2</sup> Institute for Employment Research, Nuremberg, Germany. Email: stephanie.eckman@iab.de

<sup>3</sup> World Bank – Development Economics Research Group, Washington, District of Columbia, U.S.A. Email: smurray@worldbank.org

**Acknowledgments:** The authors would like to thank their partners in the Ethiopia Central Statistics Agency, in particular Samia Zekaria, Biratu Yigezu, Habekiristos Beyene, Abate Sidelel, Jemal Ali, Abdulaziz Shifa, and the other CSA staff that supported this project. We would also like to thank Alemayehu Ambel and Jon Kastelic of the World Bank for their facilitation and technical assistance, as well as Sarah Walker, Svenja Wippich, Ruben Bach, Angus Cameron, Mike Brick, Keith Rust, the participants at the 2012 International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations conference, and four anonymous reviewers for their comments on earlier concept notes and drafts. Finally, we would like to thank Asmelash Haile Tsegay for his critical work on all levels of the project. Funding for this project was provided by the Bill and Melinda Gates Foundation Trust Fund for Improving the Quality and Policy Relevance of Household-Level Data on Agriculture in Sub-Saharan Africa and the Knowledge for Change programs at the World Bank. All views are those of the authors and do not reflect the views of the World Bank or its member countries.

statistics present particular challenges due to the nomadic and seminomadic nature of many pastoralists.

The most common sample selection methodology for household surveys in the developing world is a multistage stratified sample (Grosh and Munoz 1996). In the first stage, primary sampling units are selected from census enumeration areas. In the second stage, dwellings are selected from a housing unit frame, usually compiled through costly in-field listing. However, seminomadic households that are temporarily absent, as well as fully nomadic households without fixed dwellings, are undercovered by this approach. In areas where a large portion of the poor and vulnerable population engages in pastoralist activities, this undercoverage could lead to substantial bias in livestock and welfare estimates.

This article considers the use of an alternative approach to collecting data from livestock-owning households, Random Geographic Cluster Sampling (RGCS). Similar methods are commonly used by developed world agricultural statistics agencies, such as the United States Department of Agriculture, to measure agricultural production and livestock (USDA 2010), and have also been used by researchers to study farms in Scotland and livestock in Somalia, South Africa, Thailand, and Laos (Emerson and MacFarlane 1995; Cameron 1997; Soumare et al. 2007; von Hagen 2002). They are also common in forestry surveys (Husch et al. 1982; Roesch et al. 1993). This article describes a pilot project to test the RGCS methodology in the Afar region of Ethiopia, carried out collaboratively by the World Bank Development Economics Research Group and the Ethiopian Central Statistical Agency (CSA).

In an RGCS design, the study area is stratified using data from Geographic Information Systems (GIS) sources. Within each stratum, points (latitude and longitude) are randomly selected, and then a circular cluster of a given radius is created around the point. All eligible respondents found within this cluster are selected for the survey. The main advantage of this design is that it captures everyone who resides in the selected circles at the time of the interview, including those who do not have a permanent dwelling or who are temporarily away from their dwelling. Properly implemented, this design eliminates the undercoverage resulting from mobile populations.

There are other alternative methodologies for measuring livestock ownership that we do not use directly in this study. The CSA used a flyover survey in 2004 to estimate the total number of livestock for areas in the Ethiopian Somali region not covered in the agricultural census due to security concerns. In addition to high costs and difficulties in implementation, flyover surveys do not allow researchers to link livestock to households, which severely constrains the use of the data for socioeconomic analysis. Water point surveys are also common, but are biased as they exclude all livestock not found at a known watering point. Adaptive sampling is an approach often used for wildlife studies that are rare and unevenly distributed. However, such a design requires ongoing and close supervision by a sampling statistician and often multiple trips to the same area, neither of which was possible in this project (Thompson 1990; Thompson 1991; Thompson and Seber 1996). Other geographic sampling methods use a grid or hexagon design to eliminate overlap, but are more difficult to implement in the field or would require more expensive GPS hardware (Reams et al. 2005).

We developed the RGCS approach to address the shortcomings of the other available data-collection methodologies while taking into account the limited technical capacity of

the implementing partner. We note some advantages of the RGCS over a traditional household-based survey, but also report the many challenges encountered. Unfortunately, some of the difficulties in implementing the design seem to be due to interviewers' failure to implement the procedures. Though unforeseen challenges, such as natural disasters and ethnic violence, also played a role, it is also possible that the design, which at times required interviewers to cross long distances on foot in very harsh conditions, is not feasible in terms of what it is realistic to require of an interviewer. We conclude with thoughts on the limitations of RGCS specifically in the drylands context but also discuss its potential use in surveys of persons more generally.

## 2. Background on the Afar Region

To test the RGCS approach in the field, we carried out a survey in July and August of 2012 in the Afar region of Ethiopia. This region was selected for the pilot project for a number of reasons. First, the CSA had conducted an agricultural and livestock household survey, the Ethiopia Rural Socioeconomic Survey (ERSS), six months prior to the implementation of the RGCS field work. In Afar the ERSS included a module on pastoralist issues. We had therefore expected to be able to use the ERSS data as a point of comparison for our RGCS results. Unfortunately, we have concerns about the ERSS data as a benchmark, as discussed below.

The second factor in our choice of Afar for this project was the high-quality existing GIS infrastructure at the CSA compared to other potential study areas. The CSA has compiled GIS data layers for the entire country and has several trained staff members. The agency also maintains a stock of GPS devices suitable for the specialized fieldwork. We expected that the CSA's previous experience with the technology used during planning, sample selection, and data collection would be beneficial to the project outcomes.

Third, the Afar region also offered geographic advantages over other pastoralist areas in the region. Afar covers a land area of approximately 72,000 square kilometers located in the north of the country, and is relatively isolated. Well-guarded national boundaries, geographic features, and traditional ethnic hostilities limit the migration of the Afar people outside the boundaries of the region, which simplifies comparability between the RGCS and ERSS data sources.

The Afar region is divided into five administrative zones. The companion ERSS survey covered only Zones 1 and 3. As the RGCS survey was designed to make comparisons to this survey, these two zones were taken as a basis for the new approach. However, since seasonal migration patterns take regular residents of Zones 1 and 3 into Zones 4 and 5, these two zones were also included. Zone 2 in the far north of the region is excluded from both surveys due to extreme weather conditions, recent violence against Western nationals, and its self-contained migration patterns. According to the ERSS, 55 percent of respondent households in the Afar region that own livestock indicated that they had taken their livestock outside of the village to graze for at least one night during the previous season, and 41 percent indicated that they had similar plans for the upcoming dry season. Most respondents (56 percent) made only one trip in the previous year, with an additional 24 percent making two trips. Of those making trips with their livestock, less than one percent travelled outside of Afar and no one reported migrating to Zone 2. These results also support our choice of Afar for this pilot project.



### 3. Study Design

#### 3.1. Stratification

We divided Afar into five strata before selecting points to improve the statistical and cost efficiency of the project. The five strata were defined by the expected likelihood of finding herders and livestock, based on an assumption that herds congregate around limited water sources and available pasture in the driest part of the year. Spatial datasets describing land cover, land use, and other geographic features were used as input to delineate five discrete, mutually exclusive strata.

The first stratum consisted of land in or near towns, defined by population density measures in the AfriPop dataset (Tatem 2010). The second stratum consisted of permanent agriculture, under the assumption that livestock would be largely excluded from these areas. Boundaries were defined based on the interpretation of five meter resolution SPOT Imagery from 2006 from the CSA's Land Cover Mapping project, and included commercial agriculture as well as some small individual farms. Area placed in the first two strata was then excluded from remaining strata definitions.

The third stratum consisted of land within two kilometers of a major water source, including the Awash River and its permanent tributaries, and which also met criteria for pasture based on the average annual mean and range of the long-term normalized difference vegetation index (USGS Earth Resources Observation and Science Center 2012a,b). This stratum was considered to be the most likely to contain livestock. The fourth stratum consisted of land between two and ten kilometers from a major water source which met criteria of pasture land. The remainder of the land was placed into the lowest probability stratum. See Figure 1 for a map of the five strata.

A total of 125 points were selected from these five strata for the survey. The total number of points selected and the allocation between strata was based on sample size calculations from the previously collected data from the 2008/2009 Agricultural Sample Survey, the expected number of households to be found and interviewed in each stratum based on the results of the pretests, and the available budget for the pilot project. The number of selected points was higher in the strata where we expected the highest concentrations of potentially nomadic households and livestock (Stratum 3), and lower in areas of lower expected density (Stratum 5). Points were selected in areas with low likelihood of finding pastoralists, towns and settled agricultural areas, because excluding these areas would bias the total livestock populations. The radii for the circles also varied across the strata. In areas where we expected higher densities, we drew smaller circles to keep the workload reasonable. In areas where we expected few or no livestock, we expanded the circle radius to the largest feasible dimensions to maximize the probability of finding animals. See Table 1 for the definition, sample size, and radius used in each of the five strata.

#### 3.2. Survey Implementation

To develop the framework protocols for the RGCS approach in Afar, two pretests were conducted, the first in December 2011 and the second in June 2012. The first focused on equipment and field practices and on qualitative research into seasonal migration patterns. The second finalized the protocols and tested the survey instrument.



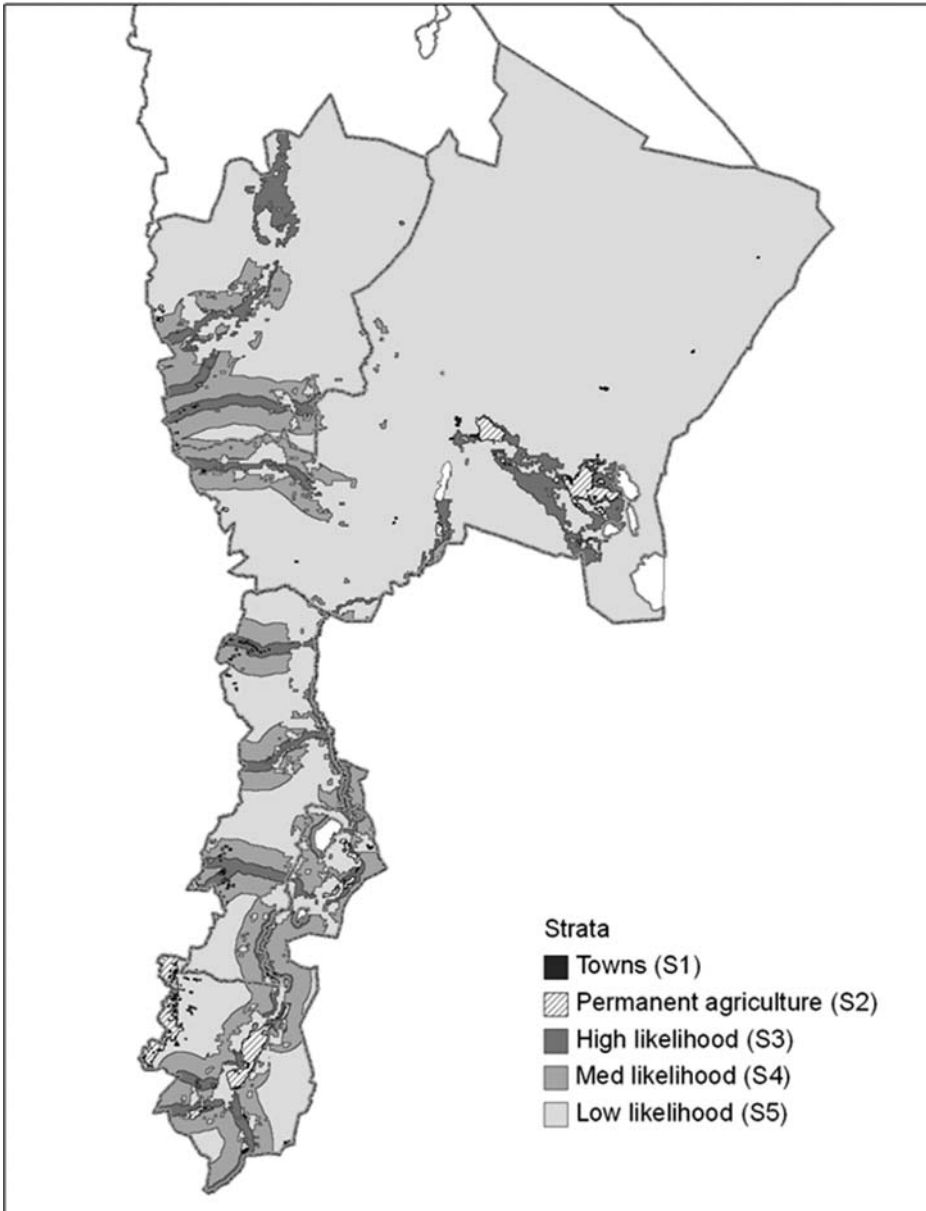


Fig. 1. Stratification Map

The resulting methodology was designed to be relatively straightforward to implement in a low-capacity field environment. Each interviewer was given a GPS device to which the selected points, and the circles around them, had been preloaded. In addition to the usual zoom and pan features, the device always displayed where the interviewer was in relation to selected area and was set to sound an alarm when the interviewer entered the circle. The interviewer teams were to drive as close as possible to the circle and then travel the rest of the way on foot, if necessary. Figure 2 shows an example of a point and circle.

Table 1. Stratification of Afar region

Stratum	Description	Radius (km)	Points Selected	Total area (km <sup>2</sup> )	Percent of total landscape
1	High likelihood: towns	0.1	10	33	< 1
2	Almost no possibility: settled agricultural areas/commercial farms	0.5	15	930	2
3	High likelihood: within 2 km of major river or swamps	1	60	3,538	6
4	Medium likelihood: within 10 km of major river or swamps	2	30	6,921	12
5	Low likelihood: all land not in another stratum	5	10	45,152	80
Total			125	56,574 <sup>a</sup>	100

<sup>a</sup> The total area in the table does not match the total area of Afar due to exclusion of Zone 2 from our study.

The selected circle has a radius of one kilometer and includes both land and water (on the eastern edge). To assist in locating the area, each interviewer was also provided with printed maps such as that shown in Figure 2.

Once inside the circle, the team was assigned to canvas the area and interview all livestock-holding households. The device recorded the interviewer’s path of travel within the circle so that he could navigate back to the starting point.

When a team member encountered a household (or a group of people travelling together) inside the circle, they attempted to complete three questionnaires. The first was a household roster, completed with a household informant, which captured basic

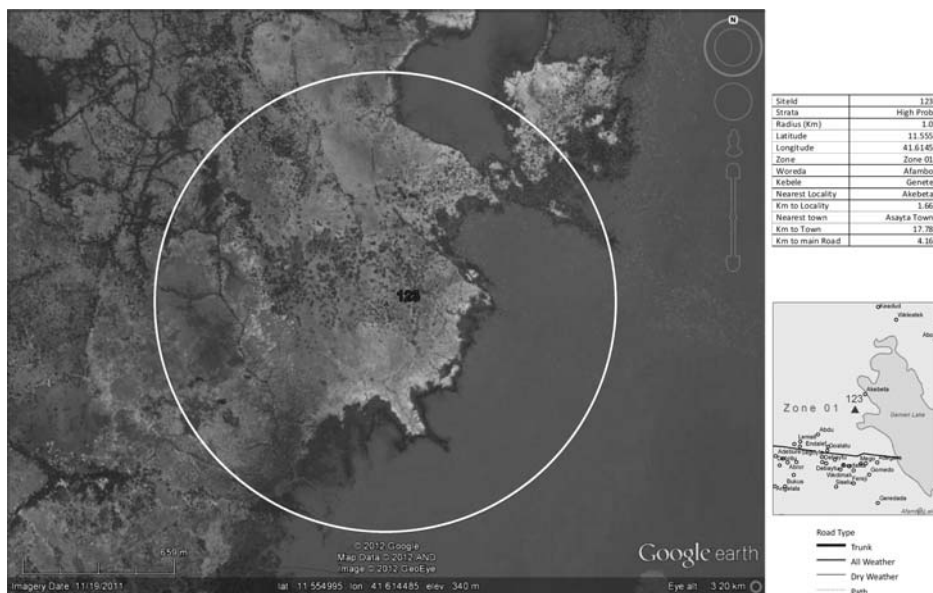


Fig. 2. Example of Selected Point and Circle

demographic information about each member of the household, such as name, age, schooling, and health information. The second questionnaire, also for the household informant, gathered data about the goats, cattle, and camels currently travelling with the household (those away for a day to graze were included in this roster). This questionnaire also asked who owned the livestock and whether the owner was currently travelling or staying with the group. The third questionnaire was administered to each individual livestock holder in the household and contained more information about the animals in his or her possession. Following the completion of each selected circle, the supervisor filled out a cover sheet indicating how many persons or households were found in the circle.

A one-week training for supervisors and enumerators was conducted in the city of Awash in southern Afar in early July 2012. A total of 22 field workers, five supervisors, one field coordinator, and one CSA branch head participated. All participants were recruited by the CSA and some had prior survey experience. The training stressed questionnaire administration, sampling protocols, safety, instruction with the handheld GPS devices, and the use of field guides. Data collection took place from July 10 to August 9, 2012. Interviewers worked in teams consisting of four interviewers and one supervisor.

The use of local field guides was strongly encouraged in this study. During pretesting, we found the most helpful available guides were young men from the local area with extensive knowledge of the terrain and the people living there. The data collection budget included funds to hire such guides whenever necessary. The guides played essential roles in determining the best route from road and river access points to the circle boundary, and acting as intermediaries between the government data collection teams and a suspicious and occasionally hostile local population.

All questionnaires were administered on paper and were provided in Amharic. The interviewers used local translators and the local guides to translate the questionnaire into Afar when necessary. The interview lasted on average 20 minutes per household, though there was substantial variation based on the household size and livestock holdings. At the end of fieldwork, all of the household and holder questionnaires, as well as the supervisor questionnaires, were returned to the CSA headquarters in Addis Ababa where data entry took place.

The fieldwork was facilitated by a survey coordinator who participated in the second pilot, conducted the training, and performed selected field visits. The survey coordinator was contracted independently of CSA, and had extensive experience with primary data collection projects in Ethiopia. The survey coordinator visited the teams throughout the course of fieldwork, accompanying each team to between three and five circles. His visits were not randomized, though he attempted to cover the distribution of teams, zones, and strata.

As the study area encompasses some of the harshest terrain in the region, and the methodology was novel both for the research and implementation teams, a number of unexpected difficulties were encountered. First, the timing of the fieldwork, which was originally designed to coincide with the dry period, unfortunately fell during the annual Ramadan fasting period. As most field guides and respondents were observant Muslims, they were reluctant to participate in activities during daylight hours. Second, the seasonal rains started earlier than had been expected, which created access problems such as flooding of roads and land bordering the rivers. The access issues necessitated longer walks for enumerators, including one incident where a team had to walk 15 km to reach

the selected site. Other obstacles, such as national park boundaries, active volcanoes, and militarized areas further restricted access. Third, ongoing strained relations between local communities and the national government led to a few isolated security incidents, including minor assaults on drivers and fieldworkers, and the (brief) kidnapping of the survey coordinator. Team supervisors repeatedly cited these challenges to explain their lack of progress in completing assigned field tasks.

### 3.3. Weighting

The probabilities of selection for such a design are in principle rather straightforward. Setting aside the issue of stratification for a moment, say we select  $c$  points with replacement and draw an  $r$ -kilometer radius around each one, selecting all households that fall within the circles. To get the probabilities of selection of a given household  $i$ , we invert our reasoning and consider the set of all points such that, if any of those points were selected, household  $i$  would be interviewed (see [Roesch et al. 1993](#) and [Thompson and Seber 1996, p. 108 for a similar approach](#)). Call this set  $A_i$ . For most households,  $A_i$  is simply a circle with radius  $r$  centered at household  $i$ . (For households near the boundary of the study region, the circle may be cut off a bit, but we ignore this issue for the moment.) Then the probability of selection of household  $i$  is one minus the probability that no point in the area surrounding that household is ever selected, across all  $c$  selections (based on [Särndal et al. 1992, p. 50](#)).

$$\pi_i = 1 - \left(1 - \frac{\pi r^2}{\text{total area}}\right)^c$$

However, due to the stratification used in this study, the probabilities of selection of the interviewed households are more complex. Because the strata are quite commingled (see [Figure 1](#)), a circle drawn around a point selected in one stratum could extend outside of the boundaries of that stratum and include land in another stratum. For example, consider a household that lies in Stratum 2 near the boundary of Strata 1 and 2, as shown in [Figure 3](#). Household  $x$  can be selected if points inside Stratum 2 are selected but also if points inside Stratum 1 are selected. In terms of the notation developed above, the selection region for household  $x$ ,  $A_x$ , contains land in both Stratum 1 and Stratum 2.

This issue with stratum boundaries is not trivial. In our study, all land in the town stratum (Stratum 1) is within five kilometers of the low probability stratum (Stratum 5) and thus was also selectable from that stratum. In fact, more than 90 percent of all land area in Strata 1, 2, 3, and 4 falls within the selectable range of points in Stratum 5, due to the very large radius of Stratum 5. Thus, even though each household itself lies in only one stratum, many households were selectable from more than one stratum. The probability of selection of a household  $i$ ,  $\pi_i$ , is equal to the probability that the points in  $A_i$  that lie in Stratum 1 were selected, plus the probability that points in  $A_i$  that lie in Stratum 2 were selected, and so on for the  $H$  strata. Define  $\pi_{i,h}$  as the probability that household  $i$  is selected from stratum  $h$ . The overall probability of selection of household  $i$  is then:

$$\pi_i = \sum_{h=1}^H \pi_{i,h} + \sum_{j=2}^H (-1)^{j+1} \left[ \sum_{h_1 < h_2 < \dots < h_j} \prod_{h_1}^{h_j} \pi_{i,h} \right] + (-1)^{H+1} \prod_{h=1}^H \pi_{i,h} \quad (1)$$

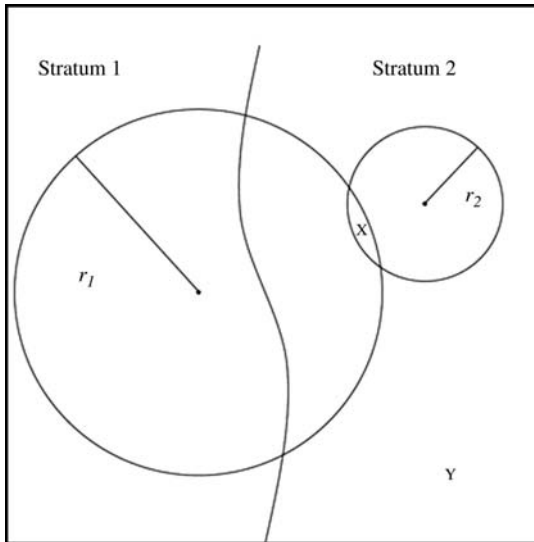


Fig. 3. Overlap between circles in different strata. Household X, in Stratum 2, can be selected by points selected from Stratum 1 or 2.  $r_1$  is the selection radius used in Stratum 1;  $r_2$  is the selection radius in Stratum 2.

where the terms after the first adjust for overlapping probabilities. However, because most households are selectable from only one or two strata, many of these terms are zero and the probabilities simplify a good deal.

Let  $S_h$  be the land within stratum  $h$ , and let  $A_i \cap S_h$  be the land in the selection region of household  $i$  that lies within stratum  $h$ . Let  $|S_h|$  and  $|A_i \cap S_h|$  be the areas of these two sets of land. Then the constituent terms in Equation 1, the probability that household  $i$  was selected from within stratum  $h$ , are each:

$$\pi_{i,h} = 1 - \left( 1 - \frac{|A_i \cap S_h|}{|S_h|} \right)^{c_h},$$

that is, one minus the probability that none of the land in stratum  $h$  that is within the selectable range of household  $i$  is selected, across all  $c_h$  selections in stratum  $h$ . The GIS tools allow us to calculate the areas of  $A_i$  and  $S_h$  precisely. Using the actual areas, rather than the areas of the circles with radius  $r_h$ , addresses the issue of lower probabilities of selection for households near the boundary of the study area (see Barrett 1964 for a discussion of “edge effect bias”). (It is also possible to conceptualize the sampling technique used in this study as a form of indirect sampling. Such an approach would also lead to appropriate probabilities of selection and weights (Lavallée 2007). We have chosen not to take such an approach here because of the difficulty of dividing the study area into slices of land which lead to selection of unique sets of households. See Roesch et al. (1993) for such an approach in the context of RGCS.)

Although 125 circles were selected, only 102 were visited by interviewing teams, as discussed below. For the  $c_h$  values, we use the number of visited circles rather than the number of selected circles, under the assumption that the circles within a stratum that the teams did not visit are missing completely at random. The initial weight for each household is then the reciprocal of its overall probability of selection:  $w_i = \pi_i^{-1}$ .

We make one adjustment to this weight, for the unobserved portions of the selected circles. The fieldwork protocol stipulated that interviewer teams should systematically observe the entire circle, however, this was not always possible due to the challenges discussed above (and possibly also due to low effort by the interviewing teams, which is discussed in more detail below). We calculate an alternative set of weights that adjusts for the portion of each circle that was not observed. The GIS technique of Viewshed analysis uses the tracks recorded by the GPS devices as the interviewers traveled within the circle, along with an altitude map derived from the ASTER Global DEM V2 dataset ([NASA Land Processes Distributed Active Archive Center 2011](#)), to determine what the interviewers were able to observe, that is, the area that was in their line of sight as they travelled around the circle. [Figure 4](#) shows an example of a map produced by the Viewshed analysis. The white tracks are the paths taken by the interviewing team members in circle 134 and the land within the circle that they could observe from those paths. We see that although the interviewers walked only a small portion of the circle, they were able to observe the majority of the area, 72.4 percent in this case. Across all of the visited circles, the observed coverage percentages range from 14.1 to 99.0, with a mean of 84.0.

The multiplicative weight adjustment is the reciprocal of the percent observed in each circle. If we believe that there are households within the unobserved portions of the selected circles, and that these households are similar to those interviewed in the observed portion, then the adjusted weight is appropriate and improves estimates. If, alternatively, we believe that the areas that were not observed were missed because they could not possibly contain any livestock, due, for example, to flood water or vegetation too thick to traverse, the adjustment to the weights is not necessary. We use both the unadjusted and the adjusted weights in the results section. No further adjustment to the weights for household nonresponse was made as the field teams did not report any issues with

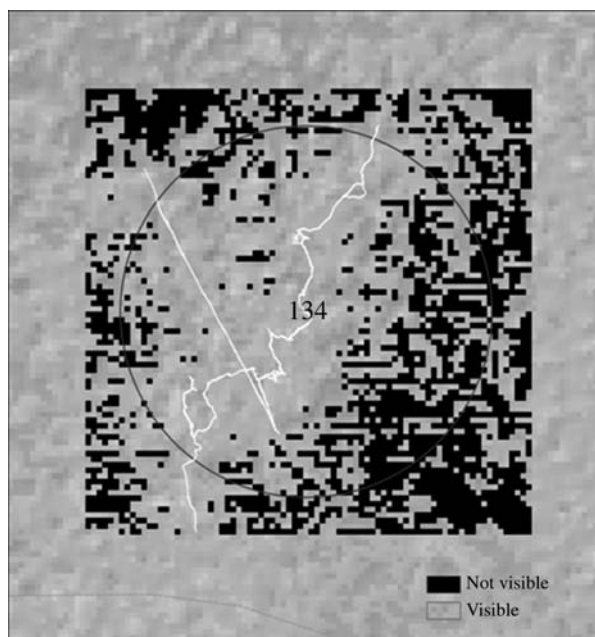


Fig. 4. Viewshed Analysis

participation. Household surveys in rural areas of the developing world, and in particular Ethiopia, have historically had high response rates.

The weights require one further caveat. The probabilities of selection on which the weights are based are accurate only if the people and livestock that the survey aims to capture do not move during the study period. If a man and his camels are selectable in more than one circle over the data collection period, then they have more than one probability of selection, which greatly complicates the weighting. To minimize this complication, we constrained the data collection period to one month. We also asked three questions in the survey regarding the past and future movements of the respondents. The first item asks whether the respondent had traveled with his livestock outside of the area where the interview took place during the dry season, the second asks if the livestock had traveled separately outside the area where they were currently, and finally if the respondent planned to travel outside of the current area during the dry season. In approximately six percent of the cases, the respondent had travelled with their livestock to a different area during the current dry season prior to the survey. In about ten percent of cases, the respondent's livestock had travelled separately to another area previously. Additionally, eleven percent of respondents indicated that they were planning to move with their livestock during the current dry season. Therefore, while mobility remains an issue with this method, in this particular context it is unlikely that it led to substantial bias. We note this issue of case mobility also affects similar area designs such as adaptive sampling.

Using these weights, we apply the Horvitz-Thompson estimator of the mean (Särndal et al. 1992, p. 111). To estimate variances, we use the bootstrap method with 1,000 replications. In each replication, we select a sample of  $c_h$  circles with replacement from the  $c_h$  selected circles within each stratum and recalculate the mean. The estimated variance of the mean estimate is the variance of the replicated means around the full-sample mean (Kolenikov 2010).

## 4. Results

### 4.1. Field Work Results

As mentioned above, of the 125 points selected, 102 were visited. Of those visited, 59 circles (58 percent) contained at least one livestock. In total, the interviewers collected information from 793 households which owned livestock, though nine of these households were shown by their GPS points to be outside of the circle boundaries and are therefore excluded from the analysis, leaving a total sample size of 784. The number of interviewed households per circle with livestock-owning households ranged from one to 65, with a mean of approximately 15, Table 2 shows the full results.

It was also necessary to replace four circles during the course of the fieldwork. These replacements were made at the discretion of the survey coordinator, with input from the CSA and World Bank teams, for locations that fell within restricted areas.

In total, 3,698 individuals living in households owning livestock were identified as part of the survey. Of these, 127 reported having no permanent dwelling, which weights up to an estimate of 4,701 (95% CI: 94, 9,307), or two percent of the livestock-holding population in Zones 1, 3, 4, and 5. All but five of the individuals without a permanent



Table 2. Field Work Results

Stratum	Description	Selected Points	Visited Circles	Households in Circles	Circles without Livestock
1	High likelihood: towns	10	10	69	4
2	Almost no possibility: settled agricultural areas / commercial farms	15	14	113	8
3	High likelihood: within 2 km of major river or swamps	60	49	229	24
4	Medium likelihood: within 10 km of major river or swamps	30	22	182	6
5	Low likelihood: all land not in another stratum	10	7	191	1
Total		125	102	784	43

dwelling lived in households in which all members are completely nomadic. The inclusion of households without permanent addresses in the survey was a main objective of the original research agenda, as this group is traditionally undercovered in dwelling-based surveys. There are, however, very few of them in the study, not enough to perform independent analyses.

#### 4.2. Means and Totals

To assess the RGCS approach, we compare weighted estimates of means and totals from the RGCS survey to estimates from the ERSS, a household survey carried out by CSA during the 2011/2012 agricultural season. The ERSS used the traditional stratified two-stage cluster design to select households and completed interviews with more than 4,500 households throughout Ethiopia. The Afar portion of the survey included 144 households in twelve clusters, ten of which were rural and two of which represented small towns. As discussed above, only two zones were covered in the ERSS survey, Zones 1 and 3, and therefore we limit our comparisons to these two areas. Among the interviewed households, 83 percent (weighted) reported owning livestock, and were administered an additional livestock questionnaire in November and December 2011 ([Central Statistical Agency and World Bank 2012](#)). Thus there is a short time gap between the ERSS livestock survey and our RGCS study, for which we compensate in the analysis. Though each survey has its shortcomings, our expectation was that the surveys should agree in the aggregate.

To construct comparable measures, we use retrospective questions about livestock in the RGCS to derive the number of livestock the household owned at the time of the ERSS survey. These questions account for slaughter, loss, death, purchase, and birth over the six months prior to the survey. We calculate two weighted estimates, one using the base weights and the other using the weight which adjusts for the proportion of the selected circle actually observed, as discussed above. As with the RGCS estimates, the standard errors we calculate for the ERSS means and totals reflect the clustered design.

Compared to the ERSS survey, there are no statistically significant differences in the mean number of animals found per household across the three groups, the RGCS



unadjusted and adjusted, and the ERSS. These numbers are also in the range of secondary source estimates of herd size and composition, though the available estimates are dated (see Sabates-Wheeler et al. 2013, Getachew 2001, and Said 1994 for further discussion). There are, however, large differences with regard to the totals. While the ERSS and RGCS estimates are not statistically different for camels due to the wide confidence intervals on the estimates, the ERSS estimate is more than 70 percent higher than even the higher of the two RGCS estimates. The gap is even wider for goats, where the ERSS estimate is more than 1.5 times higher than the adjusted RGCS, and for cattle, where the ERSS estimate is nearly 5 times higher (Table 3).

We hypothesize two issues that could have led to these discrepancies. The first possible explanation is the *interviewer effort hypothesis*: RGCS interviewers did not make efforts to reach all portions of the circles that they could have and/or did not interview all households in the circles and all holders at those households, and thus systematically excluded many livestock from the survey. The lower levels of effort could be attributed to the weather, which was extremely hot during this period, flooding, which would have made access more difficult by requiring interviewers to take long detours on foot or ford swollen rivers, and also the Ramadan period, which would have limited access to local guides to assist the teams. Low effort by the interviewers could have led to undercoverage of livestock and thus to underestimates of totals in Table 3. The second possible explanation is the *ERSS overestimation hypothesis*: Implementation issues with the ERSS upwardly biased the livestock totals. The following two sections explore these two hypotheses in more detail.

#### 4.3. Test of Interviewer Effort Hypothesis

To further explore the interviewer effort hypothesis, we estimate three regression models in which measures of effort are the dependent variables. The covariates in each of the models are similar and are of two types: measures about the area and the land, and measures about the workload and the interviewers. The models are all run at the level of the circle, rather than on the household or holder level.

The first two models use a logistic regression in which the dependent variable is whether a selected circle was visited (1) or not (0) by a field team, regardless of whether any livestock households were found. Recall that there were 23 circles that were selected but never visited, and this failure to complete assigned workload is one measure of interviewer effort. Since the unit of observation is the circle, it is not necessary to account for weighting or stratification in this analysis. In the model,

$$\Pr(Y = 1) = \frac{1}{1 + e^{-\eta}} \text{ where } \eta = \alpha + \beta X + \varepsilon,$$

$\alpha$  is a constant term,  $X$  is a vector of relevant household and team characteristics, and  $\varepsilon$  is the error term. In the first model, the variables included in vector  $X$  are the distance of the center point of the circle to the nearest paved road, the distance from the center to the nearest locality, the distance from the center to the nearest large body of water, the relief roughness of the terrain (the maximum elevation minus the minimum elevation divided by site radius, based on Meybeck et al. 2001 using the SRTMV4 Digital Elevation database Jarvis et al. 2008), the radius of the circle, a historical mean vegetation index

Table 3. Weighted estimates of total livestock in study area and average livestock held by household (conditional on ownership), by animal type

	Mean (SE)			Total (SE)		
	RGCS (unadjusted weights)	RGCS (adjusted weights)	ERSS	RGCS (unadjusted weights)	RGCS (adjusted weights)	ERSS
Cattle	10.4 (1.5)	10.8 (1.8)	15.3 (3.3)	153,505 (34,384)	186,164 (51,283)	1,092,752 (367,307)
Camels	8.1 (1.4)	7.7 (1.4)	6.2 (1.9)	92,009 (25,893)	139,608 (37,186)	237,568 (116,430)
Goats	20.2 (3.1)	19.7 (3.0)	20.7 (3.1)	566,139 (146,182)	815,310 (222,853)	2,095,876 (488,027)

Standard errors in parentheses

(NDVI, a measure of 'greenness') value, and supervisor-level (or team-level) fixed effects. The distance measures are included to capture how difficult it was for interviewers to access the selected circle; interviewers may have been less likely to visit circles which were further from the road or from a town. Similarly, if the circle was situated in rough terrain, it may have been more difficult to access. Because it was not possible to calculate the NDVI value at the time of the attempt for those sites that were not visited, the 10-year historical average NDVI value for that area is used. Also, in lieu of strata-level fixed effects, we include the circle radius, which, along with the distance to a major water source and the long-term NDVI mean values, constitutes the strata definitions. These results are presented in Column 1 of [Table 4](#).

In addition to the information included in the first regression, we also know that in the 20 cases where the survey coordinator was present, the sites were always successfully visited. The model is re-run to exclude those 20 sites in which the survey coordinator was present. Those results are presented in Column 2 of [Table 4](#).

The third model is a standard OLS model in which the dependent variable is the proportion of the circle observed, measured between 0 and 1 according to the Viewshed calculations discussed above. Again, walking more of the circle and observing the area is a sign of greater effort by the interviewers. This model is conditioned on the interviewers having visited the circle, and thus includes only 102 data points. Here the covariates included are the distance to a main road, distance to nearest locality, distance to a major water source, relief roughness, the radius of the circle, historical mean NDVI values, total rainfall in the week prior to the survey ([NOAA Climate Prediction Center RFE 2.0](#)), current mean NDVI values, the supervisor fixed effects, and the indicator of the coordinator's oversight. There are two additional variables included in this model, total rainfall in the past week, which is added to further explore the teams' assertion that flooding was the main obstacle to coverage, and current NDVI values, to test if perhaps dense vegetation hampered observation rates.

[Table 4](#) presents the results of the three models of interviewer effort. The supervisors had reported that flooding and rough terrain were the main reasons they could not access or fully observe the selected areas, but the models reveal limited support for these claims. The first column in [Table 4](#) shows that the closer a circle is to the main road, the more likely interviewers are to visit it ( $\hat{\beta} = -0.140, p = .014$ ). The interpretation of this result is a bit ambiguous, as it could be due to the need to travel long distances off-road to reach the circle, leaving the teams vulnerable to flooding or other terrain hazards, or it could be interpreted as a lack of willingness by the interviewers to attempt to access these sites. The negative and significant estimated coefficient on the radius size ( $\hat{\beta} = -0.634, p = .090$ ) suggests that teams preferentially worked the circles that were smaller and thus easier, which supports the lack of effort hypothesis. If flooding or rough terrain at the sites themselves were the problem, we would have expected to see negative coefficients on the distance to river variable or relief roughness, but none were found.

Repeating the model excluding those overseen by the survey coordinator, we see similar effects, though their magnitude is larger. Teams are less likely to visit larger and more remote circles in both the full and restricted models (as noted above for the full model, and  $\hat{\beta} = -0.174, p = .007$  and  $\hat{\beta} = -0.933, p = .030$ , respectively, in the second model). They are also less likely to visit circles with historically higher rainfall totals

Table 4. Regression Results

	(1)		(2)		(3)	
	$\hat{\beta}$	Std. Error	$\hat{\beta}$	Std. Error	$\hat{\beta}$	Std. Error
Kilometers to main road	-0.140**	0.057	-0.174***	0.064	-0.005	0.004
Kilometers to nearest locality	-0.070	0.136	-0.116	0.159	-0.002	0.009
Kilometers to river	0.010	0.033	0.000	0.032	-0.000	0.001
Relief roughness	0.006	0.005	0.007	0.005	-0.000	0.000
Circle radius	-0.634*	0.374	-0.933**	0.430	-0.057***	0.016
Historical mean NDVI value	-4.575	3.076	-6.154*	3.354	-0.441*	0.231
Total rainfall week prior to survey					-0.000	0.000
Current mean NDVI value					0.146	0.160
<i>Reference: Supervisor 1</i>						
Supervisor 2	-2.416**	1.207	-3.197**	1.413	-0.085**	0.036
Supervisor 3	-0.249	1.646	-0.086	1.898	-0.215**	0.094
Supervisor 4	-3.211***	1.148	-4.021***	1.383	-0.011	0.045
Supervisor 5	-1.771	1.215	-2.740**	1.395	-0.051	0.037
Overseen by survey coordinator					-0.056	0.046
Constant	6.549***	1.912	8.357***	2.371	1.131***	0.076
Number of observations	125		105		102	
Pseudo R <sup>2</sup> /R <sup>2</sup>	0.264		0.314		0.515	

Note: \*\*\*  $p < .01$ , \*\*  $p < .05$ , \*  $p < .1$

( $\hat{\beta} = -6.154, p = .067$ ) when not accompanied by the survey coordinator. This is in contrast to the assertion from the field teams that those circles closest to the river were the most difficult to access due to flooding. Unsupervised, teams were less likely to visit historically drier circles, which would most likely be located in the harshest terrain.

In the final model, of the percent of the circle directly observed, the only measure related to geography that is significantly related to the percent of the circle observed is the circle radius ( $\hat{\beta} = -0.057, p = .001$ ), indicating that larger circles have lower coverage percentages. In addition, the historical NDVI value is also weakly significant, perhaps again indicating interviewer unwillingness (or inability) to spend long periods of time in harsh climates. The relief roughness, distance to river variables, total rainfall in the past seven days, and current NDVI value, which correspond to the reasons cited by the interviewers as explanations for not observing the whole circle, are not significant.

Perhaps the most striking finding across the three regressions is the consistent significance of the supervisor effects. In terms of the number of sites visited, the teams led by Supervisors 2, 4, and 5 are consistently lower compared to Supervisors 1 and 3. This is particularly true for the cases in which they were not accompanied by the survey coordinator. In terms of the percent of the circle observed, the results are harder to interpret. Supervisors 2 and 3 observed smaller proportions of their assigned circles, which would seem contradictory to the findings in the previous two regressions, which identify Supervisor 2 as low effort and Supervisor 3 as high effort. It may be difficult in this case to separate what is a lack of effort and what is the inability to completely observe a relatively inaccessible site that other teams would not have extended the extra effort to visit. Finally, the coefficient on the “Overseen by survey coordinator” variable in [Table 4](#) indicates that the presence of the survey coordinator was not significant in terms of the area of the circle observed.

Taken together, the three regressions present a picture of what occurred during field implementation and why some areas were not thoroughly worked. Though some evidence on a lack of interviewer effort is confounded by actual obstacles to task completion, such as the distance the team had to travel from a paved road, a general lack of significant findings related to flooding and terrain, the two main difficulties cited by supervisors, point toward a low-effort interpretation. The substantial findings of supervisor-level effects and the survey coordinator effect further support the low-effort hypothesis but also demonstrate that effort level varied across teams. Low field effort can in turn explain why our collected data seems to capture too few livestock, relative to the ERSS household survey, as shown in [Table 3](#).

#### 4.4. Tests of ERSS Quality Hypothesis

In addition to undercoverage by the RGCS, another possible explanation for the discrepant totals in [Table 3](#) is some degree of overestimation in the numbers produced by the ERSS. While we were not directly involved with the data collection for the ERSS, we did observe some cause for concern when working with CSA staff in both the Addis Ababa headquarters and the Afar regional field office. In addition, there have been quality issues in data generated by CSA in the past (see [Dercon and Hill 2009](#) for more detail).

The Afar field office is particularly vulnerable to data quality issues as it is a remote region of the country, has fewer staff members, and generally lower levels of skills and

training compared to headquarters or other regional field offices. Communication between headquarters and the Afar office is difficult, and communication between the field office and teams is even more complicated due to frequent power outages and unreliable cell phone networks. These issues are exacerbated by long distances and a limited road network which make field supervision challenging. In the review of the ERSS data quality, Afar was one of the regions with the highest incidence of problems, including incorrect listing forms, missing questionnaires, and incomplete information in administered surveys. In particular, the release of the livestock data was delayed for almost a year following the end of fieldwork while data cleaning was completed.

In addition, there have concerns raised with some of the procedures in the CSA headquarters. The weight calculations for the ERSS had to undergo a major revision due to incorrect calculations. In addition, serious concerns were raised by outside survey coordinators about the methods used to deal with missing values by the data entrants. In cases where sections were blank or incomplete, entrants would fill in the information from other households in the same EA. If data was missing because respondents did not participate in a given activity, this could introduce substantial overestimation bias into the ERSS data.

We find support for the hypothesis that the ERSS over-reports livestock in the limited secondary source material available. The Global Livestock and Production Health Index (GLiPHA) is produced annually by the [Food and Agriculture Organization \(2010\)](#). This database only offers disaggregation down to the regional level, which would be an underestimate of the densities in Zones 1 and 3 because it would also include the low-population high-area Zone 2, but can offer approximate estimates. In addition, in 2003, as part of the Agricultural Census, USAID contracted a consulting firm based in London and Nairobi to conduct an aerial surveillance estimation of seven of the nine zones in the neighboring Somali region that could not be covered due to remoteness and security considerations. This methodology is limited in that it cannot provide any information at the household or holder level, but it can produce high quality data on livestock totals for a given area (see [CSA 2004](#) for details on estimation techniques). As both Afar and Somali have a largely pastoralist population base and similar climates, we would expect the density of animals to be broadly similar in the two areas.

[Table 5](#) compares estimates of livestock per square kilometer from four different sources. The first three columns give estimates from the RGCS (using both the unadjusted and adjusted weights) and the ERSS survey. The fourth column contains information from the GLiPHA for all of Afar in 2010. The last seven columns give estimates for the seven zones in Somali in 2004. We see that the RGCS estimates are within the range of those from the GLiPHA and the aerial surveillance, while the ERSS estimates are substantially higher. While it should be stressed that these estimates are not directly comparable, as they are for different areas in different time periods, we would expect the ranges to be similar for the reasons stated above. This increases our confidence in the accuracy of the RGCS estimates over those produced by the ERSS, at least with regard to livestock totals.

## 5. Discussion and Conclusion

This pilot project of the RGCS technique to collect livestock data in the Afar region of Ethiopia demonstrated that the implementation of such a design is feasible; however,

Table 5. Livestock Density ( $n/km^2$ )

	Afar Region (Zones 1 & 3)				Somali Region (Aerial Survey 2004)						
	RGCS (unadjusted)	RGCS (adjusted)	ERSS	All Afar 2010 (GLiPHA)	Afdir	Degehabur	Fik	Gode	Korahe	Shinile	Warder
Camels	2.5	3.8	6.4	2.2	2.6	3.6	0.8	2.8	4.2	2.6	8.4
Cattle	4.2	5.1	29.7	5.6	3.0	1.4	0.5	4.1	0.7	5.3	0.8
Goats	15.5	22.3	56.9	9.9	13.2	19.6	4.2	24.3	19.5	21.7	31.6

questions remain as to whether it is the best available method. The project showed that sufficient GIS information is available, often through the public domain, to create strata for the probability of finding livestock, and to select points within those strata. With maps and relatively inexpensive GPS devices, teams can navigate to points and identify eligible respondents within these clusters. These respondents can then be interviewed regarding their households socioeconomic conditions and livestock holdings, creating the linkages necessary to perform poverty analysis on these populations. In addition, using standard statistical methods, it is possible to calculate weights that take into account the varying probabilities of selection and sufficiently address overlap probabilities. Moreover, information generated as part of the GPS field implementation, such as the Viewshed results, can be used to estimate the area observed by individual interview teams, and account for undercoverage, if necessary. And finally, the methodology was able to do what it was designed to do – capture households without permanent dwellings that would have been missed by a traditional dwelling-based sample design. The location and interviewing of these persons is a major benefit to the RGCS technique over the traditional household-based approach to survey sampling.

A number of questions remain as to whether this method should be considered the best practice for collecting this type of information. The RGCS has demonstrated some advantages over the traditional household-based survey methodology, such as eliminating the need to conduct a cluster-listing exercise and allowing data collection to be completed in a single step. The methodology was also successfully implemented in a low-capacity environment and avoided overly technical issues, such as those one would face with an adaptive sampling method. However, problems were still noted in the implementation. Interviewers did not visit all of their assigned areas and did not observe the entire area when they did visit. Some supervisors required supervision themselves by the survey coordinator. Overall, monitoring in this study was difficult compared to the standard household survey as, without a household-listing operation, the paper trail on total cluster size was limited, and the populations are mobile, which limits the usefulness of repeat visits to verify the data collected. The project was also highly dependent on the cooperation of local guides, which are outside of the management structure and may be unreliable in some areas. Perhaps most importantly, the terrain in which the survey was implemented is difficult. The weather was extremely hot and numerous natural obstacles to the successful completion of the survey tasks occurred. Although these factors affect any survey in Afar, they were particularly troublesome for the RGCS, which required a good deal of driving and walking to reach the selected areas. ‘High effort’ was required from supervisors and interviewers throughout the project to implement the design as developed. Unfortunately we are not able to thoroughly evaluate how much these factors impacted the quality of the data collected. We have concerns that weighted estimates from the data do not accurately capture the number of livestock held in Afar, but we have no reliable comparable standard against which to compare our numbers.

Based on our experience in Afar with the RGCS, we have a number of suggestions that would improve the implementation. In particular, we recommend more careful planning to avoid conducting the survey during the Ramadan period and the completion of data collection before the onset of the seasonal rains. In addition, training should better explain to supervisors and interviewers the goals of the survey in order to elicit more ‘high-effort’



fieldwork. Providing incentives to teams that complete more circles and observe greater percentages of assigned circles may further increase effort. Though it is possible that the RGCS approach may work better in a different country context, the limited capacity of statistics bureaus and the potentially dangerous terrain of the study area are common to nearly all pastoralist areas in the developing world.

Despite the limitations noted above, drylands areas remain difficult to survey and the RGCS offers a viable alternative to traditional approaches. The Ethiopia CSA has decided to extend the method and include it as part of the data-collection method in pastoralist areas nationwide as part of the upcoming Agricultural Census. Beyond the specific livestock in drylands context, we wonder if this technique might have applications to other contexts, such as the measuring of homeless persons. We believe that the RGCS approach deserves more study in both the developed and the developing world.

## 6. References

- Barrett, J.P. (1964). Correction for Edge Effect Bias in Point-Sampling. *Forest Science*, 10, 52–55.
- Cameron, A.R. (1997). Active Surveillance and GIS as Components of an Animal Health Information System for Developing Countries – Thailand and Laos as Examples. Queensland: University of Queensland.
- Central Statistical Agency (2004). Livestock Aerial Survey in the Somali Region. November 2003. Available at: [www.dppc.gov.et/Livelihoods/Somali/Downloadable/Livestock%20Aerial%20Survey%20in%20the%20Somali%20Region%20November%202003.pdf](http://www.dppc.gov.et/Livelihoods/Somali/Downloadable/Livestock%20Aerial%20Survey%20in%20the%20Somali%20Region%20November%202003.pdf) (Accessed July 3, 2013).
- Central Statistical Agency & World Bank (2012). Living Standards Measurement Study-Integrated Surveys on Agriculture: Ethiopia Rural Socioeconomic Survey Basic Information Document. (December 2012).
- Dercon, S. and Hill, R.V. (2009). Growth from Agriculture in Ethiopia: Identifying Key Constraints. IFPRI's ESSP-II policy conference 'Accelerating agricultural development, economic growth and poverty reduction in Ethiopia', Hilton Hotel, Addis Ababa, October 22–24, 2009, (p. 22–24).
- Emerson, H. and MacFarlane, R. (1995). Comparative Bias Between Sampling Frames for Farm Surveys. *Journal of Agricultural Economics*, 46, 241–251. DOI: <http://www.doi.org/10.1111/j.1477-9552.1995.tb00770.x>
- Food and Agricultural Organization (2010). Global Livestock Production and Health Atlas (GLiPHA). Available at: <http://kids.fao.org/glipha> (accessed July 26, 2013).
- Getachew, K.N. (2001). Among the Pastoral Afar in Ethiopia: Tradition, Continuity and Socio-Economic Change. Utrecht: International Books.
- Grosh, M.E. and Munoz, J. (1996). A Manual for Planning and Implementing the Living Standards Measurement Study Survey. Living Standards Measurement Study (LSMS) Working Paper No. LSM 126. Washington, DC: The World Bank. Available at: <http://documents.worldbank.org/curated/en/1996/05/438573/manual-planning-implementing-living-standards-measurement-study-survey> (accessed January 4, 2013).
- Husch, B., Miller, C.I., and Beers, T.W. (1982). *Forest Mensuration*. New York: Wiley.

- Jarvis, A., Reuter, H.I., Nelson, A., and Guevara, E. (2008). Hole-Filled Seamless SRTM data V4, International Centre for Tropical Agriculture (CIAT). Available at: <http://srtm.csi.cgiar.org>.
- Kolenikov, S. (2010). Resampling Variance Estimation for Complex Survey Data. *Stata Journal*, 10, 165–199.
- Lavallée, P. (2007). *Indirect Sampling*. New York: Springer-Verlag.
- Meybeck, M., Green, P., and Vörösmarty, C. (2001). A New Typology for Mountains and Other Relief Classes. *Mountain Research and Development*, 21, 34–45. DOI: [http://www.dx.doi.org/10.1659/0276-4741\(2001\)021\[0034:ANTFMA\]2.0.CO;2](http://www.dx.doi.org/10.1659/0276-4741(2001)021[0034:ANTFMA]2.0.CO;2)
- NASA Land Processes Distributed Active Archive Center (2011). ASTER Global DEM V2 data. Sioux Falls, South Dakota: USGS/Earth Resources Observation and Science (EROS) Center. Available at: [https://lpdaac.usgs.gov/get\\_data](https://lpdaac.usgs.gov/get_data) (accessed January 14, 2013).
- NOAA Climate Prediction Center Famine Early Warning System African Rainfall Estimation Algorithm Version 2 (RFE 2.0), daily estimates. Available at <http://www.cpc.ncep.noaa.gov/products/fews/data.shtml> (accessed September 4, 2013).
- Reams, G.A., Smith, W.D., Hansen, M.H., Bechtold, W.A., Roesch, F.A., and Moisen, G.G. (2005). The Forest Inventory and Analysis Sampling Frame. In *The Enhanced Forest Inventory and Analysis Program – National Sampling Design and Estimation Procedures*, W.A. Bechtold and P.L. Patterson (eds). Asheville, NC: USDA Forest Service, Southern Research Station, 11–26.
- Roesch, F.A., Green, Jr. E.J., and Scott, C.T. (1993). An Alternative View of Forest Sampling. *Survey Methodology*, 19, 199–204.
- Sabates-Wheeler, R., Lind, J., and Hodidinott, J. (2013). Implementing Social Protection in Agro-Pastoralist and Pastoralist Areas: How Local Distribution Structures Moderate PSNP Outcomes in Ethiopia. *World Development*, 50, 1–12. DOI: <http://www.dx.doi.org/10.1016/j.worlddev.2013.04.005>
- Said, A. (1994). *Pastoralism and the State Policies in Mid-Awash Valley: The Case of the Afar, Ethiopia*. Uppsala, Sweden: Scandinavian Institute of African Studies.
- Särndal, C.-E., Swensson, B., and Wretman, J. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
- Soumare, B., Tempiab, S., Cagnolatic, V., Mohamoudb, A., van Huylenbroeckd, G., and Berkvensa, D. (2007). Screening for Rift Valley Fever Infection in Northern Somalia: A GIS Based Survey Method to Overcome the Lack of Sampling Frame. *Veterinary Microbiology*, 121, 249–256. DOI: <http://www.dx.doi.org/10.1016/j.vetmic.2006.12.017>
- Tatem, A.J. (2010). *Ethiopia AfriPop Data 2010 (alpha version)*. Gainesville, Florida: Emerging Pathogens Institute, University of Florida. Available at: [http://www.clas.ufl.edu/users/atatem/index\\_files/Ethiopia.htm](http://www.clas.ufl.edu/users/atatem/index_files/Ethiopia.htm) (accessed June 15, 2012).
- Thompson, S.K. (1990). Adaptive Cluster Sampling. *Journal of the American Statistical Association*, 85, 1050–1059. DOI: <http://www.dx.doi.org/10.1080/01621459.1990.10474975>
- Thompson, S.K. (1991). Stratified Adaptive Cluster Sampling. *Biometrika*, 78, 389–397. DOI: <http://www.dx.doi.org/10.1093/biomet/78.2.389>
- Thompson, S.K. and Seber, G.A.F. (1996). *Adaptive Sampling*. New York: Wiley.

United States Department of Agriculture Area Frame Section. Available at: <http://www.nass.usda.gov/research/AFS.htm> (accessed November 8, 2010).

USGS Earth Resources Observation and Science Center (2012a). eMODIS NDVI Africa (monthly means). Available at: <http://earlywarning.usgs.gov/fews/africa/index.php> (accessed June 20, 2012).

USGS Earth Resources Observation and Science Center (2012b). eMODIS NDVI Africa (pentadal). Available at: <http://earlywarning.usgs.gov/fews/africa/index.php> (accessed October 30, 2012).

von Hagen, C. (2002). Using an Area Sampling Frame to Calculate Livestock Statistics in the Gauteng Province, South Africa, within a GIS. *Directions Magazine*. (August 20, 2002).

Received February 2013

Revised October 2013

Accepted November 2013

## Enumerating the Hidden Homeless: Strategies to Estimate the Homeless Gone Missing From a Point-in-Time Count

*Robert P. Agans<sup>1</sup>, Malcolm T. Jefferson<sup>1</sup>, James M. Bowling<sup>1</sup>, Donglin Zeng<sup>1</sup>,  
Jenny Yang<sup>1</sup>, and Mark Silverbush<sup>2</sup>*

To receive federal homeless funds, communities are required to produce statistically reliable, unduplicated counts or estimates of homeless persons in sheltered and unsheltered locations during a one-night period (within the last ten days of January) called a point-in-time (PIT) count. In Los Angeles, a general population telephone survey was implemented to estimate the number of unsheltered homeless adults who are hidden from view during the PIT count. Two estimation approaches were investigated: i) the number of homeless persons identified as living on private property, which employed a conventional household weight for the estimated total (Horvitz-Thompson approach); and ii) the number of homeless persons identified as living on a neighbor's property, which employed an additional adjustment derived from the size of the neighborhood network to estimate the total (multiplicity-based approach). This article compares the results of these two methods and discusses the implications therein.

*Key words:* Homeless count; hidden homeless; unsheltered homeless population; Horvitz-Thompson estimator; multiplicity-based estimator.

### 1. Introduction

How to best estimate homelessness has historically been a difficult and costly venture (Link et al. 1994; Tompsett and Toro 2004; Toro and Janisse 2004; Toro 2005, 2006). As a highly mobile population, it is difficult to contact and track the homeless due to their unstable living situations. Furthermore, when performing homeless counts of the unsheltered population, enumerators are typically required to count late at night when the shelters are closed for the evening and the homeless on the streets are more easily identifiable. Night-time counts, however, can be problematic because visibility is reduced and vulnerable populations, such as children, hide from public view. In addition, homelessness is not a permanent state. A person's housing situation can change rapidly and homeless people can relocate quite easily. Thus the true value of homelessness is constantly in flux, which creates inherent variability between estimates taken at different points in time. Estimates therefore can vary depending on the assumptions and

<sup>1</sup> University of North Carolina, Carolina Survey Research Laboratory, 730 Martin Luther King Jr. Blvd., Bolin Creek Center, Chapel Hill, NC, U.S.A. Corresponding author Email: [agans@unc.edu](mailto:agans@unc.edu)

<sup>2</sup> Los Angeles Homeless Services Authority, 811 Wilshire Blvd., Los Angeles, CA 90017, U.S.A

**Acknowledgments:** Special thanks is extended to Dr. William Kalsbeek for having conceived the original idea of using a multiplicity-based estimation approach to improve the precision of estimates around hidden homeless estimate in greater Los Angeles Homeless Count.

methodology applied in the count. Nevertheless, homelessness needs to be measured, especially in the U.S., which has one of the highest rates of homelessness among developed nations (Tompsett et al. 2003; Toro et al. 2007). In addition to the important social issues that surround homelessness, there are practical reasons to obtain the best possible estimates. One such agency that depends on reliable measures of the homeless population is the Los Angeles Homeless Services Authority (LAHSA).

LAHSA, a joint powers authority of the County and City of Los Angeles, coordinates and manages government funds for programs that provide shelter, housing, and other services to the homeless in 85 of the 88 cities of Los Angeles County (and all of the unincorporated areas). In order to receive federal dollars, LAHSA is required to conduct a homeless count every two years. Given that Los Angeles is the largest urban county in the U.S. with more than ten million residents and a geographic area of 4,083 square miles that encompasses 88 cities, this task is challenging. Los Angeles County also has one of the largest disparities between wealthy and low-income people in the nation. It manages one of the largest welfare systems in the country and contends with one of the nation's largest homeless populations (Bring Los Angeles Home 2006).

In 2009, the Carolina Survey Research Laboratory (CSRL) at the University of North Carolina at Chapel Hill collaborated with LAHSA for the 2009 Los Angeles Homeless Count (HC09). Homelessness was measured at the Continuum of Care (CoC) level and included all of Los Angeles County except the cities of Glendale, Pasadena, and Long Beach, which produce their own independent estimates. HC09 included a street count, a shelter count, and a youth count as well as a hidden homeless estimate, all in an attempt to measure the county's homeless population. The HC09 findings estimated that 42,694 people were homeless and that over 20 percent of the homeless population was hidden (2009 Greater Los Angeles Homeless Count Report). The hidden homeless estimate was derived from interviewing 4,288 households in the CoC and asking them if a hidden homeless person lived on their property. Though a relatively large number of sample respondents provided interviews, only 16 hidden homeless persons were identified as meeting the HUD criteria for hidden homelessness. The rarity of this event produced a rather imprecise estimate with a relatively large variance where the total was 9,451 and the standard error 2,339.

The focus of this article is on the hidden homeless estimate, which made up a significant portion of the homeless population in HC09. To redress the issue of large sampling errors commonly associated with survey estimates of rarely occurring events, frame coverage was increased in the 2011 Los Angeles Homeless Count (HC11) by asking survey participants to report not only on homeless individuals currently residing on private property, but also on homeless individuals in their immediate neighborhood. Though this method increases the number of hidden homeless persons that get reported, it also introduces the potential for multiple reports of the same hidden homeless persons by members of the same neighborhood network. Consequently, a multiplicity-based estimator must be derived which adjusts for the possibility of multiple reports of hidden homeless. It is through this neighborhood network approach that we hope to improve the statistical precision of the hidden homeless estimate. This article examines the utility of the multiplicity-based approach for estimating the hidden homeless population in Los Angeles.

## 2. Methods

To minimize the difficulties that arise in sample surveys designed to estimate a rare event, [Birnbaum and Sirken \(1965\)](#) proposed a stratified random sample design that included obtaining information about the “multiplicity” of individuals with a rare condition as reported by health care providers (i.e., the primary sampling units). To derive an estimate of the number of diagnosed cases of a rare disease, these investigators designed the survey to increase sample coverage while accounting for patients being reported by several providers. Following this idea, an alternate approach was proposed in HC11 that expands beyond reporting for hidden homeless individuals located on a private property, as was the sole approach for HC09. In the present study, respondents were asked to not only report the number of hidden homeless on their residential property but also any hidden homeless on their neighbor’s property. Consequently, coverage was expanded and the possibility of discovering hidden homelessness was broadened, thus hopefully making a rare event less rare. This approach has been called network or multiplicity sampling in the literature ([Sudman and Freeman 1988](#); [Flores-Cervantes and Kalton 2008](#); [Kalton 2009](#)), because the typical one-to-one counting rule in conventional sampling takes into account the inclusion of larger observational units or networks. Instead of being self-weighting (conventional approach), the multiplicity approach must take into account the network size. The success of this approach, however, rests largely on the validity and accuracy of the information provided by the respondent, not only for the critical measure of hidden homelessness, but also for a new, additional measure of neighborhood network size. In order for the multiplicity estimate to be valid, we must accurately gather the size of the neighborhood and the probability that more than one person (viz., neighbors) could report on the same hidden homeless person or persons. In this article, comparisons were made between the statistical precision of a more conventional Horvitz-Thompson approach (as applied in HC09) and the multiplicity-based approach proposed here.

### 2.1. Sample Design

Households for the telephone interview were identified from a disproportionately stratified dual-frame sample of landline telephone numbers obtained from Marketing Systems Group in Horsham, Pennsylvania. Stratification was defined by frame source (list-assisted RDD and electronic white page “EWP” listings), median household income of the exchange area (EA) in which the telephone number was located (delineated into *high* and *low* designations at \$50,680), the percent of single-family households in the EA (delineated into “high” and “low” designations at 60%), as well as other local area information thought to be predictive of hidden homelessness ([Table 1](#)). The latter was used to form an *item predictor score (IPS)* based on the distribution of such scores for all listed telephone numbers on the frame. IPS was considered *low* in designated areas if it resulted in a summative score of 0 or 1, and high if the IPS ranged between 2 and 6. IPS scoring criteria consisted of the following items:

- Single family dwelling: Yes = 1; No = 0;
- High African American concentration: Yes = 1; No = 0;

Table 1. Strata used to sample telephone numbers for the hidden homeless telephone survey

		Median Household Income in Exchange Area:					
		Low		High			
		Listing Status of Phone Number:		Listing Status of Phone Number:			
		Directory Listed	NOT Directory Listed	Directory Listed	NOT Directory Listed		
% Single Family Dwelling Units in Exchange Area:	High	Item Predictor Sum:	High	1	5	2	7
		Low	9		11		
	Low	Item Predictor Sum:	High	3	6	4	8
		Low	10		12		

- Below the 20th percentile for length of time in current residence (measure of mobility): Yes = 1; No = 0;
- Below the 20th percentile on household income: Yes = 1; No = 0;
- In a block group (or census tract, if only available at this level) above the 80th percentile on percent vacancy rate: Yes = 1; No = 0; and
- In a Census Tract above the 80th percentile on rate of street homeless count per 100,000 population members as of the 2000 Census: Yes = 1; No = 0.

2.2. Questionnaire Design and Pilot Testing

In general, *hidden homeless persons* are defined as those who live among, but not directly with, the residential population of a community. In this study, people were classified as hidden homeless if they were sleeping on private property in such places as an unconverted garage, carport, back porch, tool shed, tent, camper, car, encampment, and so on. These people were likely to go missing during a point-in-time count because they were not on the streets nor readily visible to enumerators, but nevertheless were considered unsheltered homeless according to the U.S. Department of Housing and Urban Development (HUD) definition. This definition is in contrast to the *precariously housed* or *at risk of literal homelessness* definitions, which include individuals temporarily staying within a household because they have no regular or adequate place to stay and lack the means or money to provide it.

Neighborhood was operationally defined by street block neighborhood (SBN). The SBN for any household was defined as the set of households whose front entrance faces the street – bounded by one linear segment of street on which the referent household is located. In other words, a respondent reported on the number of hidden homeless for the set of households on both sides of a one-block long street that also includes the respondent’s household. Residents of apartment, condominium, or single room housing



complexes were asked to provide estimates of hidden homeless on complex property. An additional requirement was that respondents must also give a measure of neighborhood network size. For single family dwellings, respondents need to provide a count of homes contained within the SBN. For complexes, respondents need to report the number of housing units in their building and the number of buildings in their housing development in order to calculate their respective neighborhood network size.

For the pilot study, 2,500 random digit dial telephone numbers within the Los Angeles CoC were placed into calling between December 14 and December 29, 2010. Of those telephone numbers, 1,662 were finalized as ineligible, 55 as refusals, 747 were given an unknown status because eligibility was never determined, and 36 cases resulted in completed interviews. Interviews were completed with an adult who either owned or rented the property. After reviewing the pilot data, minor changes were made to the instrument to improve the flow of the telephone interview. The question about any hidden homeless on personal property read: *Not including dependents or adult children, is there anyone living with you or staying on your property because they do not have a regular or adequate place to stay due to a lack of money or other means of support?*

The questions on neighborhood hidden homeless piloted well, but went through minor reconstruction to simplify telephone administration. The previous version read: *Now I'd like to ask you some questions about your neighborhood. Have you seen anyone staying [on your neighbors' property (IF X10 = single dwelling)] [in your complex (IF X10 = housing complex)] who you believe does not have a regular or adequate place to stay due to a lack of money or other means of support? For this question, your [neighborhood includes houses on both sides of the street confined by two intersection streets] OR [complex includes all the housing units in your complex or development]. INTERVIEWER NOTE: THIS DOES NOT INCLUDE DEPENDENTS OR ADULT CHILDREN.*

The new version read: *Now I'd like to ask you some questions about the houses on your block. Have you seen anyone who appears to be homeless staying [on your neighbors' property (IF X10 = single dwelling)] [on the complex grounds (IF X10 = housing complex)] [WHO YOU BELIEVE DOES NOT HAVE A REGULAR OR ADEQUATE PLACE TO STAY DUE TO A LACK OF MONEY OF OTHER MEANS OF SUPPORT]? For this question, your [neighborhood block includes houses on both sides of the street confined by the closest intersection in either direction] OR [complex includes all the housing units in your complex or development].*

### 2.3. Main Study

All totaled, 32,826 randomly selected telephone numbers were worked to complete 3,390 hidden homeless interviews. The overall selected sample of telephone numbers was disproportionately allocated among the twelve strata, following a similar allocation pattern of disproportionality to that observed in the HC09 survey. Data collection took place from January 16, 2011 to April 10, 2011. The CSRL has an advanced CATI operation consisting of 42 interviewer workstations and three monitoring stations. Supervisors and clients can silently monitor interviewers' audio and keyed responses from computers in the monitoring room. This monitoring capability helps ensure that data

collection for the study meets the highest quality standards. During data collection, interviewing took place Saturday through Thursday (EST). Monday through Thursday shifts typically were conducted from 12 noon to 12 midnight. Saturday sessions occurred between 1:30 pm until 5:30 pm. Sunday shifts were typically held from 5:30 pm to 12 midnight.

In addition to questionnaire programming, the CSRL also utilizes Blaise’s (Version 4.8, 2008) call-scheduling capabilities to maximize the probability of contacting potential respondents. A central file server takes sample telephone numbers and arranges automatic call scheduling for interviewer administration. The system enables calls to be scheduled so that different times of the day and week are represented. In this study, no cases were withdrawn from calling until a minimum of eight unsuccessful call attempts were made and had been at least one weekend call, one evening call, and one daytime call. Calls could also be scheduled at times specified by the respondent, thus ensuring that calls were made at optimum times.

CSRL supervisors closely monitor data collection to ensure that data are being collected and entered correctly according to guidelines and policies reviewed in training. In addition, several steps were taken to both reduce the occurrence of refusals and to improve refusal conversion. First, we attempted to minimize refusals by introducing techniques for dealing with reluctance and refusal during general interviewer training. This was often accomplished through role-playing sessions that enable trainees to become familiar with and to rehearse a variety of refusal situations. Upon encountering a refusal, interviewers documented the following information for each refusal: reason for the refusal, the point in the interview at which the refusal occurred, and the gender and approximate age of the respondent. Refusal documentation is standard procedure at the CSRL because it enables the next interviewer, the refusal converter, to tailor her approach in eliciting participation from the potential respondent, thereby optimizing the likelihood of conversion. Finally, as part of interviewer monitoring, interviewers’ individual refusal rates were closely watched. Only experienced refusal converters recontact respondents who initially refuse.

2.4. Final Outcomes and Response Rates

The final outcomes from calling may be grouped into four broad categories (see Table 2) that were used to calculate the overall response rate: (i) a complete interview (I = 3,390); (ii) not eligible (NE = 13,503) because the telephone numbers were found to be nonworking, dedicated fax or computer lines, or a business/cell line; (iii) no interview or response from an eligible household (NR = 2,593); or (iv) unknown or indeterminate

Table 2. Final dispositions for hidden homeless survey by strata

OUTCOME	STRATA												Totals
	1	2	3	4	5	6	7	8	9	10	11	12	
I	590	379	608	209	151	39	85	88	440	254	316	231	3,390
NR	518	357	467	94	120	32	50	48	378	190	192	147	2,593
NE	1,147	859	838	289	3,150	1,056	1,670	2,795	662	441	271	324	13,503
U	1,983	1,563	2,047	685	1,085	322	532	735	1,497	916	1,027	948	13,340

( $U = 13,340$ ) because we never had the opportunity to talk to a person or someone in the household refused participation before we could verify eligibility.

The response rate was based on the American Association for Public Opinion Research (AAPOR) Standard Definitions (2011). Weighted (42.6%) and unweighted (33.6%) response rates were determined by AAPOR’s Response Rate 3 (RR3).

### 2.5. Final Sample Weights

A standard three-step sample weighting procedure was followed in producing sample weights (Kalsbeek and Agans 2008). The base weight was computed using the sampling rate for telephone numbers in each stratum, adjusted for the portion of the stratum samples that were placed in calling and the number of phone lines that reached the household (Step 1). The base weight was then adjusted for differential household-level nonresponse among sampling strata using the inverse of the stratum-specific household-level RR3 response rate as the adjustment factor (Step 2). The nonresponse-adjusted household sample weight was then calibrated to population counts as estimated from the American Community Survey sample by cross-tabulating on: (i) the race-ethnicity of the reference person/knowledgeable adult (white non-Hispanic/Hispanic/Other), (ii) the type of dwelling (single-family/all other types), and (iii) the education of the reference person/knowledgeable adult ( $<$  college bachelor’s degree/ $\geq$  college bachelor’s degree) (Step 3). The multiplicative effect of variable sample weights or  $Meff_w$  (Kish 1965) was somewhat large ( $Meff = 2.44$ ), so weights were trimmed at the nonresponse adjustment stage as recommended by Potter (1988).

### 2.6. Horvitz-Thompson Estimator

The Horvitz-Thompson (HT) estimator of a population total was used to estimate the number of hidden homeless on private property for the HC11 survey. This conventional method was also employed to estimate the total number of hidden homeless individuals in the LA for the HC09 survey. First, we define

- $N$                     Number of closed street block neighborhood (SBN) networks in the target population
- $Y_i$                     The actual number of hidden homeless persons in the  $i$ -th SBN
- $M_i$                     Number of household residences in the  $i$ -th SBN
- $I_{ij}$                     Sample selection indicator for the  $j$ -th household in the  $i$ -th SBN  
 $\Rightarrow 1$  if household is selected, 0 otherwise
- $\pi_{ij} = E(I_{ij})$         The selection probability for the  $j$ -th survey household in the  $i$ -th SBN

Now let  $t_{HH}$  denote the total (relevant) hidden homeless count that is to be estimated such that

$$t_{HH} = \sum_{i=1}^N Y_i = \sum_{i=1}^N \sum_{j=1}^{M_i} H_{ij}^{(Res)},$$

where  $H_{ij}^{(Res)}$  denotes the number of hidden homeless persons located on their private property at the  $j$ -th survey household in the  $i$ -th SBN. Then the HT estimator of a

population total for the number of hidden homeless people identified in the HC11 survey is given as:

$$\hat{t}_{HT} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{I_{ij} H_{ij}^{(Res)}}{\pi_{ij}}.$$

2.7. Multiplicity-Based Estimator

To apply the multiplicity-based alternative approach, the telephone survey interview required all respondents (located in the  $j$ -th household in the  $i$ -th SBN) to provide the number of hidden homeless persons located on the property of all other surrounding households in their SBN, also denoted as  $H_{ij}^{(SBN)}$ . Respondents who did not provide SBN size data were called back for data retrieval. Hot deck imputation was used to handle missingness for households still missing SBN size data after unsuccessful attempts to retrieve the information through a callback. Specifically, records for which measure of size data was missing were imputed using the mean SBN value of other record that shared that particular record's household type: For instance, in the case where a participant informed the interviewer that they lived in an apartment but refused or were unable to provide SBN size data, missing information was imputed using completed SBN data collected from the other participants who also classified their residence as an apartment.

Following the [Birnbaum and Sirken \(1965\)](#) approach to multiplicity estimation, an estimator of the overall number of hidden homeless persons in the target population is given as

$$\hat{t}_{HH} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{I_{ij} Y_{ij}}{\pi_{ij} M_i}, \quad \text{where } Y_{ij} = \left[ H_{ij}^{(Res)} + H_{ij}^{(SBN)} \right].$$

Ignoring the biasing effects of nonsampling error (i.e., due to frame, nonresponse, and measurement),  $\hat{t}_{HH}$  can be shown to be an unbiased estimator of  $t_{HH}$ :

$$E(\hat{t}_{HH}) = \sum_{i=1}^N \sum_{j=1}^{M_i} \left\{ \frac{Y_{ij}}{\pi_{ij} M_i} \right\} E(I_{ij}) = \sum_{i=1}^N \sum_{j=1}^{M_i} \left\{ \frac{Y_{ij}}{\pi_{ij} M_i} \right\} \pi_{ij} = \sum_{i=1}^N \sum_{j=1}^{M_i} \left\{ \frac{Y_{ij}}{M_i} \right\}$$

Noting that  $Y_{ij} = Y_i$  for all  $M_i$  households in the  $i$ -th SBN

$$E(\hat{t}_{HH}) = \sum_{i=1}^N \sum_{j=1}^{M_i} \left\{ \frac{Y_{ij}}{M_i} \right\} = \sum_{j=1}^{M_i} \left\{ \frac{M_i Y_i}{M_i} \right\} = \sum_{i=1}^N Y_i = t_{HH}.$$

The variance of  $\hat{t}_{HH}$  can be obtained using the standard formula

$$\sum_{i=1}^N \sum_{j=1}^{M_i} \sum_{i'=1}^N \sum_{j'=1}^{M_{i'}} \frac{\pi_{ij} \pi_{i'j'} - \pi_{ijj'}}{\pi_{ij} \pi_{i'j'}} \frac{Y_{ij} Y_{i'j'}}{M_i M_{i'}},$$

where  $\pi_{ijj'}$  is the second inclusion probability of the  $j$ -th household in the  $i$ -th block and the  $j'$ -th household in the  $i'$ -th block. As compared to the variance of the Horvitz-Thompson's estimator  $\hat{t}_{HT} = \sum_i^n \frac{y_i}{\pi_i} = \sum_{i=1}^N \sum_{j=1}^{M_i} \frac{I_{ij}}{\pi_{ij}} H_{ij}^{(Res)}$ , whose variance is  $\sum_{i=1}^N \sum_{j=1}^{M_i} \sum_{i'=1}^N \sum_{j'=1}^{M_{i'}} \frac{\pi_{ij} \pi_{i'j'} - \pi_{ijj'}}{\pi_{ij} \pi_{i'j'}} H_{ij}^{(Res)} H_{i'j'}^{(Res)}$ , we can see that the only difference is

that we replace  $H_{ij}^{(Res)}$  by  $\frac{Y_{ij}}{M_i}$ . Since the latter tends to be less variable for the households in the same block (ideally, zero variability) as compared to the homeless count from each individual household, we expect that in our stratified simple random sampling design, the multiplicity estimator should have a smaller variance as compared to the Horvitz-Thompson estimator.

2.8. *Producing the Estimates*

Hidden homeless estimates were produced using SUDAAN (Version 10), a statistical software package developed by RTI International that specializes in providing efficient and accurate analysis of data from complex studies. The estimated total number of hidden homeless persons in the Los Angeles CoC was produced using the DESCRIPT procedure in SUDAAN. A stratified with replacement (STRWR) design nested by stratum was specified in the procedure. A weight statement was also used in the procedure to account for varying selection probabilities in the telephone sample. The final calibrated household sample weight (without adjustment for neighborhood reporting) was used to produce the total number of hidden homeless individuals in a fashion similar to the way it was obtained for HC09. To obtain the total number of hidden homeless individuals using the multiplicity-based approach, a sample weight was used that applied an adjustment for the size of each respondent’s SBN – specifically, the conventional calibrated HH weight for each respondent was divided by the respondent’s estimated SBN count of HHs (or its multiplicity). Taylor series linearization methods are employed for robust variance estimation of descriptive statistics in the DESCRIPT procedure.

3. Results

The estimated total of hidden homeless persons in the Los Angeles CoC using the Horvitz-Thompson method in 2011 was 10,800 (SE = 3,421) and was based on the entire sample of 3,390 completed interviews in which only 13 households responded that a hidden homeless person was present (see Table 3). This estimate was comparable to the HC09 estimate of 9,451 (SE = 2,339). Again, a household could only qualify as having a hidden homeless person if they had someone living on their property in an unconverted garage, a back porch, or in an encampment, camper or car. These individuals were considered homeless by HUD and estimated counts were added to the total homeless estimate.

The hidden homeless estimate based on the multiplicity-based approach (see Table 4) led to a substantial increase in the hit rate of hidden homelessness ( $n = 322$ ), which

Table 3. *Hidden homeless estimate using conventional (Horvitz-Thompson) method*

Survey year	Overall Hidden Homeless Estimate (Personal Residence)			
	Raw count	Weighted estimate	Weighted standard error	Relative standard error
2011	13	10,800	3,421	32%
2009	16	9,451	2,339	25%

Table 4. Hidden homeless estimate using multiplicity-based method

Survey year	Overall Hidden Homeless Estimate (Neighborhood Network)			
	Raw count	Weighted estimate	Weighted standard error	Relative standard error
2011	322	18,622	2,889	15%

improved the precision around the weighted estimate as demonstrated by a major reduction in the relative standard error (32% vs. 15%). The estimated total, however, was substantially higher (10,800 vs. 18,622) but was not significant based on a normal distribution test ( $p = 0.06$ ).

### 3.1. Quality of Reported Data

The success of the neighborhood reporting approach, however, hinges on the accuracy of the data reported by respondents. Respondents were asked to identify homeless persons living on their personal property and persons living on their neighbor's property (immediate neighborhood). When asked about homeless persons living on a neighbor's property, respondents were also required to remark whether they were *very sure*, *somewhat sure*, or *not very sure at all*. To ensure that the count was computed using quality responses, only hidden homeless individuals reported by respondents who were *very sure* and *somewhat sure* of the status and number of homeless individuals residing on their neighbors' properties were counted. Of the 118 households who provided a response to the question and reported homeless persons living on their neighbor's property, 101 (86%) respondents were *very sure* of the status and number of homeless individuals residing in their immediate neighborhood, while just 17 (14%) were *somewhat sure*. Confidence in the quality of the counts reported by respondents is gained based on this knowledge. An inconsistency may however be present in the respondents' reporting of the size of their street block neighborhood (SBN).

### 3.2. Simulations

We have conducted extensive simulation studies to compare the performance of the multiplicity-based estimator versus the Horvitz-Thompson estimator. To imitate the actual LA homeless study, we generated a population of  $N$  households randomly assigned to  $d$  blocks. A proportion  $\pi$  of these households were assumed to have hidden homeless people on property. We used simple random sampling to select  $n$  of these households to collect whether there was a homeless person in the selected property. For comparison, we estimated the total count of homeless using the Horvitz-Thompson estimator and the multiplicity estimator. In the simulation studies, we varied the number of blocks ( $N$ ), the hidden homeless proportion ( $\pi$ ) and the selected household number ( $n$ ). The results from 1,000 replicates are summarized in the table in the [Appendix](#).

It is clear from the table that both the Horvitz-Thompson estimator and the multiplicity-based estimator produced a precise estimate of the total count of the homeless. However, the multiplicity-based estimator has a greater efficiency gain: The ratio of the mean

squared errors of the Horvitz-Thompson estimator versus the multiplicity-based estimator ranges from 15 to about 20, while the ratio of the relative standard error is between 4 and 5. The efficiency gain of the multiplicity-based estimator seems to increase with increasing sample size and the prevalence of the homeless. All simulation work was done in SAS (Version 9.3).

#### 4. Discussion

Our estimation approach produced hidden homeless estimates within the methodological framework utilized in the 2009 Los Angeles Homeless Count (HC09) and successfully employed an alternative, multiplicity-based method in the 2011 Los Angeles Homeless Count (HC11). The Horvitz-Thompson (HT) estimate in HC09 (9,451 hidden homeless) is comparable to the HT estimate in HC11 (10,800 hidden homeless). As similar as these two estimates are, they have large standard errors thereby producing wide intervals at the 95 percent confidence level (HC09  $\pm$  4,584; HC11  $\pm$  6705). In terms of relative standard errors, both estimates are not very reliable (HC09 = 25%; HC11 = 32%), thus providing the motivation for multiplicity sampling.

Our approach takes the simplest form of multiplicity sampling in that we only have to adjust for the size of the network, which here was operationalized as a respondent's street-block neighborhood. Unlike other network approaches (viz., snowball or respondent-driven sampling), we did not have to interview the subjects of our investigation (i.e., the hidden homeless), only count them. To produce unbiased estimates using a multiplicity-based approach, eligible respondents need only to report the size of their network ( $n$ ) which is then weighted by the reciprocal  $1/n$ . The main benefit of such an approach is the reduction of sampling error due to the increase in sample size. When costs are fixed, the multiplicity-based approach, we argue, will produce a more precise estimate of a rare event, as we found in the present study (RSE 32% in HT versus RSE 15% in MB). While significantly decreasing the amount of variance around our estimate, we can also see that the multiplicity approach produced a more sensitive measure that detected an additional 7,822 cases of hidden homeless (versus 10,800). Consequently, the multiplicity-based approach produced a more sensitive estimate of the hidden homeless population in Los Angeles, which was more precise than the conventional approach and should be implemented in the next homeless enumeration. [Sudman and Freeman \(1988\)](#), however, suggest that network sampling will lose its attractiveness as the proportion of an event in the population grows and that more conventional methods (such as the Horvitz-Thompson estimator) will remain superior. Future research should explore what that cut-off in the population is likely to be and under what conditions it best applies.

Future work should also look for ways to improve the measure of network size needed to adjust for multiplicity. Recall that in our case, to adjust for the wider reporting framework developed through the multiplicity approach all sample weights were reduced by a factor of the size of each respondent's SBN. Initial framing of the neighborhood network size question required respondents to provide an estimate of how many houses/units were on their block or in their development. Because the selection of random telephone numbers allowed the chance of a house, apartment, or mobile home to be included in the sample, neighborhood network size questions were tailored to household



type. Respondents living in homes or townhomes were asked to provide the number of homes on their block, which consisted of both sides of a street between the two nearest intersections. Respondents living in condominiums or an apartment complex were asked to approximate how many units were in their building and how many buildings were in their complex. The purpose of these questions was to quantify the size of each SBN for which the respondents were reporting hidden homeless. However, the analysis of the data showed that the framing of the question may have been confusing to the complex respondents. In fact, some of these respondents were initially unable to estimate the size of their SBN. As a consequence, imputation techniques and callbacks were used to retrieve missing and erroneous neighborhood size values for nearly a quarter of the households reporting hidden homeless persons. Callbacks were issued for cases where abnormally high values of neighborhood network size were observed. Distinguishing between which measures were valid and which were erroneous, an inherently subjective task, required all neighborhood size data to be re-evaluated. In general, any respondent reporting their SBN size as having more than 500 households/housing units was flagged for further investigation. Nearly all the cases where a respondent reported their SBN size as being more than 500 households/housing units also identified their housing type as a condo or apartment. A potential limitation to this approach might be the difficulties some respondents have estimating SBN. Further research should examine ways to reduce this error, perhaps by shrinking large multi-unit complexes into smaller, manageable units, such as the respondents' street block as contained within the complex, thus reducing the area to be covered within large multiunit complexes. Successfully addressing these limitations holds the promise of making multiplicity-based estimation a more stable, statistically precise and preferred approach for estimating hidden homeless individuals in future homeless counts.



Appendix

Appendix. Summary of Simulation Study Results for Comparing the Horvitz-Thompson Estimator and Multiplicity-based Estimator

Total Number of Blocks (d)	Total Household Population (N)	Proportion of Households with Hidden Homeless (p)	Sample Size (n)	Horvitz Thompson Estimation				Multiplicity Estimation						
				Mean Estimate	Standard Error (SE)	Mean Squared Error (MSE)	RSE	Mean Estimate	Standard Error (SE)	Mean Squared Error (MSE)	RSE			
1407	28147	0.24%	15	63.80	63.80	115,665.54	533.29%	68.73	56.00	7,665.01	127.44%	15.09	4.18	
			30	73.18	71.36	70,371.81	362.61%	81.18	53.74	3,707.87	74.94%	18.98	4.84	
			60	60.99	58.97	28,435.95	276.53%	57.21	33.71	1,372.58	20.72	68.75%	17.88	4.27
			15	277.72	265.53	521,698.41	260.15%	289.68	153.51	29,173.82	58.98%	17.88	4.41	
			30	247.69	228.80	229,999.74	193.35%	279.14	108.06	12,270.08	39.69%	18.74	4.87	
			60	288.98	246.82	130,401.42	125.02%	286.98	79.87	6,332.24	27.73%	20.59	4.51	
		5.00%	15	1,313.53	1,090.37	2,204,319.09	112.90%	1,410.52	346.04	129,631.61	25.52%	17.00	4.42	
			30	1,481.47	1,005.06	1,275,880.84	76.13%	1,420.05	251.33	63,005.29	17.67%	20.25	4.31	
			60	1,431.74	759.84	603,227.92	54.26%	1,410.05	174.70	32,186.86	12.73%	18.74	4.26	
			15	5,644.41	2,815.31	8,399,557.01	51.36%	5,698.52	641.42	429,791.04	11.51%	19.54	4.46	
			30	5,518.69	2,017.83	3,945,789.56	36.01%	5,548.82	447.20	194,043.87	7.94%	20.33	4.54	
			60	5,381.55	1,440.42	2,203,058.07	26.60%	5,365.22	327.95	116,631.98	6.14%	18.89	4.33	
14073	281474	0.24%	150	656.77	590.55	1,291,528.07	172.82%	716.97	254.94	67,062.01	36.12%	19.26	4.78	
			300	647.39	533.92	614,409.81	121.13%	651.85	171.34	30,214.91	26.68%	20.33	4.54	
			600	643.64	461.04	287,098.63	83.20%	668.87	123.39	14,705.35	18.14%	19.52	4.59	
			150	2,951.72	2,024.15	5,414,402.82	78.69%	2,810.92	504.40	262,338.27	18.22%	20.64	4.32	
			300	2,737.80	1,497.02	2,647,774.81	59.46%	2,765.29	358.28	127,539.36	12.92%	20.76	4.60	
			600	2,796.44	1,112.99	1,313,745.34	41.01%	2,789.27	254.44	63,046.37	9.00%	20.84	4.55	
140737	2814740	0.24%	1500	6,867.97	3,437.44	12,299,052.30	51.09%	6,818.99	796.89	655,087.14	11.87%	18.77	4.30	
			3000	6,817.30	2,481.81	6,100,243.77	36.24%	6,765.28	561.85	324,008.61	8.42%	18.83	4.31	
			6000	6,683.13	1,753.55	2,954,527.12	25.70%	6,768.77	396.87	156,173.25	5.84%	18.92	4.40	
			1500	28,046.07	7,160.85	50,203,412.83	25.27%	28,230.96	1,616.93	2,657,546.94	5.78%	18.89	4.37	
			3000	27,917.53	5,070.78	26,880,793.49	18.57%	28,050.11	1,141.13	1,422,574.90	4.25%	18.90	4.37	
			6000	28,073.42	3,603.42	13,275,478.46	12.94%	28,000.56	805.17	639,259.30	2.86%	20.77	4.55	
140737	2814740	5.00%	1500	141,600.19	15,863.65	244,333,568.30	11.04%	141,821.14	3,551.83	13,458,268.96	2.59%	18.15	4.27	
			3000	140,355.13	11,177.59	127,111,006.08	8.04%	140,338.58	2,492.92	6,670,671.46	1.89%	19.06	4.36	
			6000	141,754.06	7,943.89	65,035,869.35	5.67%	141,007.95	1,770.73	3,330,679.41	1.24%	19.53	4.38	
			1500	565,072.19	29,106.27	843,594,702.23	5.13%	563,609.38	6,501.05	41,463,829.65	1.14%	20.35	4.49	
			3000	560,960.79	20,526.31	456,460,140.76	3.81%	561,647.27	4,594.29	21,382,841.57	0.82%	21.35	4.62	
			6000	564,178.04	14,546.30	231,048,663.33	2.69%	563,461.16	3,250.42	10,061,108.72	0.56%	22.96	4.78	

Note: RSE is the relative standard error defined as the ratio between the standard error and the total count estimate.

## 5. References

- The American Association for Public Opinion Research (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, (7th edition): AAPOR.
- Birnbaum, Z.W. and Sirken, M.G. (1965). Design of Sample Surveys to Estimate the Prevalence of Rare Diseases: Three Unbiased Estimates. Vital and Health Statistics, Series 2, No. 11. DHEW publication no (PHS) 1000. Washington, DC: U.S. Government Printing Office, 1–8.
- Blaise 4.8 [computer software] (2007). Voorburg/Heerlen: Statistics Netherlands.
- Bring Los Angeles Home: The Campaign to End Homelessness (2006). Los Angeles Housing Services Authority, Web. Available at: [http://www.bringlahome.org/docs/BRINGLAHOME\\_book\\_final.pdf](http://www.bringlahome.org/docs/BRINGLAHOME_book_final.pdf) (accessed January 6, 2010).
- Flores-Cervantes, I. and Kalton, G. (2008). Methods for Sampling Rare Populations in Telephone Surveys. In *Advances in Telephone Survey Methodology*, J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japac, P.J. Lavrakas, M.W. Link, and R.L. Sangster (eds). New York: Wiley and Sons.
- Kalsbeek, W.D. and Agans, R.P. (2008). Sampling and Weighting in Household Telephone Surveys. In *Advances in Telephone Survey Methodology*, J.M. Lepkowski, C. Tucker, J.M. Brick, E.D. de Leeuw, L. Japac, P.J. Lavrakas, M.W. Link, and R.L. Sangster (eds). New York: Wiley and Sons.
- Kalton, G. (2009). Methods for Oversampling Rare Subpopulations in Social Surveys. *Survey Methodology*, 35, 125–141.
- Kish, L. (1965). *Survey Sampling*. New York: Wiley and Sons.
- Link, B.G., Susser, E., Stueve, A., Phelan, J., Moore, R.E., and Struening, E. (1994). Lifetime and Five-Year Prevalence of Homelessness in the United States. *American Journal of Public Health*, 84, 1907–1912, DOI: <http://www.dx.doi.org/10.2105/AJPH.84.12.1907>.
- Los Angeles Homeless Services Authority (2009). Greater Los Angeles Homeless Count Report. Available at: <http://www.lahsa.org/docs/2011-Homeless-Count/HC11-Detailed-Geography-Report-FINAL.PDF> (accessed October 13, 2012).
- Potter, F. (1988). Survey of Procedures to Control Extreme Sampling Weights. *Proceedings of the American Statistical Association Section on Survey Research Methods*, 453–458.
- SAS Statistical Software [computer program]. (Version 9.3), Cary, North Carolina.
- SUDAAN [computer program] (Version 10). Research Triangle Park, North Carolina.
- Sudman, S. and Freeman, H.E. (1988). The Use of Network Sampling for Locating the Seriously Ill. *Medical Care*, 26, 992–999.
- Tompsett, C.J., Toro, P.A., Guzicki, M., Schlien, N., Blume, M., and Lombardo, S. (2003). Homelessness in the US and Germany: A Cross-National Analysis. *Journal of Community and Applied Social Psychology*, 13, 240–257, DOI: <http://www.dx.doi.org/10.1002/casp.724>.
- Tompsett, C.J. and Toro, P.A. (2004). Public Opinion. *Encyclopedia of Homelessness*: SAGE Publications. Available at: [http://sageereference.com/homelessness/Article\\_n133.html](http://sageereference.com/homelessness/Article_n133.html) (accessed January 18, 2010).

- Toro, P.A. and Janisse, H.C. (2004). Homelessness, patterns of. In D. Levinson (ed.). Encyclopedia of homelessness (pp 244–250) Thousand Oaks, CA: Sage.
- Toro, P.A., Tompsett, C.J., Lombardo, S., Phillippot, P., Nachtergaeel, H., Galand, B., Schlienz, N., Stammel, N., Yabar, Y., Blume, M., MacKay, L., and Harvey, K. (2007). Homelessness in Europe and the United States: A Comparison of Prevalence and Public Opinion. *Journal of Social Issues*, 63, 505–524, DOI: <http://www.dx.doi.org/10.1111/j.1540-4560.2007.00521.x>.
- Toro, P.A. (2005). Community Psychology: Where Do We Go from Here? *American Journal of Community Psychology*, 35, 9–16, DOI: <http://www.dx.doi.org/10.1007/s10464-005-1883-y>.
- Toro, P.A. (2006). Trials, Tribulations, and Occasional Jubilations While Conducting Research With Homeless Children, Youth, and Families. *Merrill-Palmer Quarterly*, 52, Academic OneFile. Available at: [http://find.galegroup.com/gtx/start.do?prodId=AONE&userGroupName=unc\\_main](http://find.galegroup.com/gtx/start.do?prodId=AONE&userGroupName=unc_main). (accessed January 11, 2010).

Received February 2013

Revised October 2013

Accepted November 2013

# A Study of Assimilation Bias in Name-Based Sampling of Migrants

Rainer Schnell<sup>1</sup>, Mark Trappmann<sup>2</sup>, and Tobias Gramlich<sup>3</sup>

The use of personal names for screening is an increasingly popular sampling technique for migrant populations. Although this is often an effective sampling procedure, very little is known about the properties of this method. Based on a large German survey, this article compares characteristics of respondents whose names have been correctly classified as belonging to a migrant population with respondents who are migrants and whose names have not been classified as belonging to a migrant population. Although significant differences were found for some variables even with some large effect sizes, the overall bias introduced by name-based sampling (NBS) is small as long as procedures with small false-negative rates are employed.

*Key words:* Hard-to-Reach populations; sampling; undercoverage; onomastic sampling.

## 1. Sampling Migrants

Migrants are of particular interest in the social sciences. However, in many countries research on migrants is hampered by the lack of appropriate sampling frames for migrant populations. Census or register data or other lists of the population of interest may exist, but these sampling frames are usually not available for any purpose other than official statistics. Since the proportion of migrants is often small, and registers unavailable, special sampling procedures for rare populations have to be used (Sudman and Kalton 1986; Kalton 2009).

### 1.1. Common Sampling Procedures for Migrants

In some situations, lists of subgroups of migrant populations are available for sampling. Examples include membership lists of migrant organisations or training seminars for naturalisation interviews (Kosmidis et al. 1980; Rutishauser and Wahlquist 1983). If the migrant population of interest tends to segregate, cluster sampling of areas with a high concentration of members from the target population could be used (Blane 1977; Ecob and Williams 1991). Occasionally, quota sampling and snowball sampling (Bertelsmann Stiftung 2009; Sulaiman-Hill and Thompson 2011) are used. All these methods have serious methodological problems.

<sup>1</sup> University of Duisburg-Essen, Methodology Research Unit, Lotharstr. 65, 47057 Duisburg, Germany. Email: rainer.schnell@uni-due.de

<sup>2</sup> Institute for Employment Research, Regensburger Str. 104, 90478 Nuremberg, Germany and University of Bamberg, Germany. Email: mark.trappmann@iab.de

<sup>3</sup> University of Duisburg-Essen, Methodology Research Unit, Lotharstr. 65, 47057 Duisburg, Germany. Email: tobias.gramlich@uni-due.de

**Acknowledgment:** Schnell had the idea for this study and the classification program, supervised data analysis and wrote the final version. Schnell and Trappmann designed the study, Trappmann suggested the mechanisms, contributed to the text, improved the data analysis and provided data access. Gramlich wrote the first draft, computed the tests and performed the classification for the PASS data.

### 1.2. Name-Based Screening

Even though separate lists of migrants may not be available, very often sampling frames for a general population contain names of individuals. Thus these general population frames can be screened for names likely to belong to members of migrant populations. Name-based sampling methods (NBS) have been used in different countries and for a variety of purposes (Mateos 2007). Most often, the lists used for NBS consist of names considered to be typical for migrants or are constructed ad hoc by members of the target population; but of course, more sophisticated methods have also been used. For example, Braun and Santacreu (2009) used names which featured more frequently than a specific threshold in telephone directories of different countries to identify likely members of migrant populations.

There are also examples of the use of carefully compiled dictionaries of names which have a high positive predictive value for classification as migrant (“onomastic sampling”, Humpert and Schneiderheinze 2000). All of these methods rely on more or less error-free records of migrants’ names exactly as listed in the dictionary. However, names or spelling variations not listed in the dictionary cannot be classified. This also applies to names which contain typographical errors. It seems safe to assume that typographical errors are more common for names which might be unusual for database maintainers.

Name-based methods are applicable when names of migrants are different from those of the domestic population. If such initial differences do exist, they are likely to persist for at least a few decades. In countries with ongoing immigration from specific regions or countries, name-based methods may differentiate between recent migrants and descendants of previous cohorts of migrants as long as names differ between different cohorts.

### 1.3. A Screening Procedure Based on Trigrams of Names

Schnell et al. (2013a,b) proposed a new screening procedure for sampling migrants that neither requires reviewing names manually nor relies on exhaustive or error-free dictionaries. This procedure does not classify complete first or given names but splits names into substrings of three consecutive characters (trigrams) and classifies the names according to the relative frequencies of these trigrams within a database of names from specific countries using a naive Bayes classifier (Mitchell 1997). Since no dictionaries are used, this procedure is resilient against spelling varieties and typographical errors. Even names not used for the training of the Bayes classifier can be classified as long as they are similar to names listed. Schnell et al.’s (2013a,b) algorithm was trained using frequencies of given names and surnames by nationality of all the employees liable for social security contributions in Germany. The training database was specifically constructed by the Research Data Center of the Federal Employment Agency to develop this technique. The database covers more than 80 percent of the working population of Germany.

The classification process can be applied separately to a given first name (GN) and a given surname (SN). For the classification of a person, the four possible outcomes of the name classification (GN, SN \* classification result) must be reduced to the classification of the person as migrant or not. Different rules could be used for this reduction process.

Only two simple rules will be considered here:

**Rule 1:** A person is classified as migrant when GN *or* SN are classified as foreign.

**Rule 2:** A person is classified as migrant when GN *and* SN are classified as foreign.

For the evaluation of the trigram-based Bayes classification, Schnell et al. (2013a,b) used names and survey data from 18,795 respondents of a national household panel. In that dataset, Rule 1 (GN or SN) yielded about 40% false positives, but the false negative rate was lower than 5% for most countries of origin. Rule 2 (GN and SN) yielded less than 10% false positive cases and the false negative rate varied between 15–20% for most countries of origin. For the current study, the results of both classification rules will be compared.

## 2. Definitions and Basic Demographics of Migrant Populations in Germany

The term “migrants” denotes heterogeneous groups (among others: foreigners, refugees, asylum seekers, immigrants with or without domestic nationality, domestic born descendants of migrants with/without foreign nationality). We will concentrate on one specific subgroup in this example: migrants with a foreign nationality. While this is a rather narrow definition, it is useful for many applications. Furthermore, even the most basic of sampling methods for migrants should be able to detect this group. For NBS, using nationality as the criterion for a correct classification will increase the number of false negatives compared to wider definitions of migration status. Due to screening after selection by NBS, only false negatives will contribute to bias. Therefore, we consider our results to be the lower limit of bias in name-based sampling.

The discussion of social processes following migration is extensive (as an example, see Alba and Nee 1997). In this study we base our conceptualisation of assimilation on the approach taken by Kalter and Granato (2002, p. 200), that is, assimilation is the similarity of distributions for categories of relevant variables related to the central dimensions of education, work and family. Since each country has different migrant populations, some details on the migration to Germany is important in order to evaluate the results reported here.

The vast majority of foreign names in Germany originate from two waves of immigration: the first wave (1955–1973) was the recruitment of workers for the heavy manufacturing industry. These workers were mainly recruited from Turkey, Greece, Spain, Italy and former Yugoslavia. Most of these immigrants were unskilled workers with low or no educational and vocational qualifications from their home countries (Fassmann and Münz 1994). Although initially these immigrations were to be time limited, many of these migrants remained in Germany. The subsequent migration of family members of these immigrants accounted for a large part of immigration to Germany after 1973 (e.g., see Milewski 2007; Liebig 2007). A second wave of migration began with the demise of the Soviet Union. Many Eastern Europeans (mainly from Poland and former Soviet Republics) who had German ancestors became eligible to emigrate to Germany. Although generally better educated than the earlier wave of migrants, these more recent immigrants still encountered problems in labour market participation, mainly due to insufficient command of the German language (Milewski 2007). Every empirical study has reported that the mentioned migrant populations in Germany have a lower average education and income, higher average number of children and stronger religious attitudes than the domestic population (Babka von Gostomski 2010; Fassmann and Münz 1994).

## 3. Bias in Name-Based Sampling

Every selection procedure will result in biased parameter estimates of variables of interest if the selection probabilities are correlated with the variables of interest (Bethlehem

2009, p. 222), and if the estimation does not correct for the unequal selection probabilities. In name-based sampling, the probability of having a name classified as belonging to a migrant population may correlate with indicators of assimilation: better-assimilated migrants are likely to have a higher probability of having their name classified as domestic. At least four different mechanisms can explain this:

1. In most European countries (outside academic circles), it is common for women to adopt the family name of their husband after marriage. Since immigrants intermarry with domestic partners, the proportion of domestic surnames will be higher for female migrants. It seems very likely that female migrants with domestic partners are more assimilated than female migrants with migrant partners. However, there seems to be no official data on the frequency of name changes after marriages in general or for binational couples. Especially with binational couples in which the husband has no migration background, this will result in misclassification of the wife's migration background in many cases.
2. The given names chosen by migrant parents for their children will reflect their preferences: better assimilated migrants more often prefer domestic names (Becker 2009; Gerhards and Hans 2009).
3. If the languages of the country of origin and the host society are similar (as, for example, in Germany, Austria and large parts of Switzerland), the probability of misclassification increases: again, migrants whose names are misclassified as domestic are more likely to be more assimilated.
4. Naturalised migrants may choose to modify their names to assimilate to the host society more. At the same time, more assimilated migrants may show a greater desire to become naturalised.

Mechanisms 2 and 4 do not have to hold in countries which encourage multiculturalism and ethnic distinctiveness. In some countries (for example, Germany) all mechanisms are plausible. Given all or some of these mechanisms, name-based sampling of migrants—whether dictionary-based or not—is more likely to include less assimilated migrants with higher probability. Since the sampling depends solely on the classification of the name, false negative classifications of migrants will result in biased estimates when false negative classified (*FN*) persons differ from true positive classified (*TP*) persons:

$$E(B(\bar{Y})) = \frac{n_{fn}}{n_{fn} + n_{tp}} (\bar{Y}_{tp} - \bar{Y}_{fn})$$

with

$E(B(\bar{Y}))$  the overall bias in the mean of variable  $Y$

$n_{fn}$  the number of false negative classifications

$n_{tp}$  the number of true positive classifications

$\bar{Y}_{tp}$  the mean in variable  $Y$  for the true positives

$\bar{Y}_{fn}$  the mean in variable  $Y$  for false negatives

The expected bias is the product of the proportion  $(n_{fn})/(n_{fn} + n_{tp})$  of false negatives among all migrants, with the difference in means for a variable of interest between true positives and false negatives,  $\bar{Y}_{tp} - \bar{Y}_{fn}$ . There may be no bias despite a high rate of false negative classifications if there are no differences between false negatives and true positives. However, there may be large bias despite a low rate of false negative classifications in the case of large differences between the two groups. Whilst false positives can be excluded after a screening interview, false negatives are excluded from the sampling frame. Thus false positive classified persons only increase sampling costs, but false negative classified persons may lead to coverage bias.

There is little published literature on the estimated false negative rate of different sampling procedures and even less on differences between falsely negative and truly positive classified persons in migration surveys. This lack of studies is surprising, since most studies of migrants are dedicated to the study of some dimension of assimilation. A systematic bias in a sampling procedure would compromise the results of such a study. Because name-based sampling is considered one of the best techniques for sampling migrants if no other sampling frame is available, we conducted a study to examine the differences in characteristics between false negative and true positive classified migrants.

#### 4. Hypothesis

As detailed in Section 3, this article uses the definition of assimilation as similarity of distributions over categories of relevant variables. Accordingly, more assimilated migrants in Germany are predicted to have fewer children, smaller households, higher rates of intermarriage, higher incomes, a better education, a less traditional religious orientation and a better command of the majority (German) language. Although there are some differences between countries of origin, this process of assimilation can be observed for most migrant populations in Germany (Babka von Gostomski 2010, p. 79–113; Statistisches Bundesamt 2011, p. 193–199). Due to the four mechanisms described above that link the degree of assimilation to names, we predict significant differences with respect to assimilation variables between migrants identified by name-based sampling compared to nonidentified migrants.

#### 5. Data

We examined data from the first wave of one of the largest German panel studies (PASS, Trappmann et al. 2010, Bethmann and Gebhardt 2011, Trappmann et al. 2013). PASS is a general population household panel survey, oversampling low-income households and households receiving welfare benefits. The survey is based on a sequential mixed mode design of CATI and CAPI interviews. To enhance survey participation of migrants in PASS, foreign-language questionnaires in Turkish, Russian, and English were administered by foreign-language CATI interviewers. The survey includes questions on nationality, country of birth, year of immigration, and the social and economic situation of the respondents. Additionally, there is information on nationality, country of birth, and year of immigration for the respondents' parents and grandparents. We compared variables expected to be related to the assimilation process between all migrants in the sample and those migrants



who would be identified by a name-based sampling procedure. For this study, the names of the respondents were classified with the Bayes classifier suggested by Schnell et al. (2013a,b). Names and survey data were held separately within the governmental agency owning the data; all procedures were approved by the responsible data protection agents. Among the 18,795 respondents whose names were classified, 1,610 persons reported a foreign nationality and 3,104 reported being born outside of Germany. Here, we focus on data from the 1,610 foreigners (migrants with foreign nationality). Although there may be response error in the indicators of migration background as well as the variables related to different dimensions of assimilation, the following calculations consider these reports as true since previous studies of PASS showed neither serious measurement error nor nonresponse bias for variables which could be validated with administrative data (Sakshaug and Kreuter 2012; Kreuter et al. 2010; Schnell et al. 2010). With the first rule 1,509 out of 1,610 migrants with foreign nationality were classified as true positive (101 false negative classifications). With the second classification rule 63 percent true positive classifications were observed (1,020 TP, 590 FN).

All analyses are conducted with the unweighted PASS sample. While typically researchers are interested not in the sample per se but in inferences which can be drawn from the sample, in this methodological article the goal is to isolate one source of error, namely coverage error (although due to legal restrictions we are only able to investigate this for respondents). This can best be done by neglecting sample weights, which mainly correct for unequal sampling probabilities and differential nonresponse between demographic groups. By displaying a weighted analysis we would have a mix of coverage error, differential nonresponse rates between migrants and domestic respondents and adjustment error in the weighting procedure. While the effect of the interaction of these error sources on total survey error is an interesting topic for future research, we felt this to be too complex an issue since currently not much is known about coverage bias alone.

## 6. Results

Results of our empirical analysis for all subgroups of migrants can be found in Table 1 for Rule 1 (GN or SN) and in table 2 for Rule 2 (GN and SN). Both tables contain estimates of bias and raw differences. For brevity we will focus on the raw differences between false negatives and true positives.

Both tables are organised in the same way. Column (1) shows the mean of a variable for all migrants in the PASS survey. Column (2) shows the mean for those respondents who would be missed by NBS (false negatives, FN). Column (3) is the mean for those migrants who would have been selected by NBS (true positives, TP). Column (4) is the difference between Columns (3) and (1). This is equivalent to bias as calculated using the formula in Section 3. Column (5) gives the difference between the detected (TP) and nondetected (FN) migrants. Column (6) gives the  $p$ -value for a  $t$ -test on mean difference (or difference in proportion) between TP and FN. Finally, Column (7) shows the effect size for the difference between TP and FN. Cohen's  $d = (M_1 - M_2)/(SD_{\text{pooled}})$  was used as a measure of effect size. Values of  $d$  greater than 0.2 are considered as small effects, values greater than 0.5 as medium effects (Ellis 2010, p. 41).

Table 1. Differences between all migrants, false negative and true positive cases (Rule 1: GN or SN)

	(1) all migrants	(2) false negative	(3) true positive	(4) bias $\Delta$ (3)-(1)	(5) $\Delta$ (3)-(2)	(6) $p(t)$	(7) Cohen's $d$
<b>Household characteristics</b>							
Household size (persons)	3.3	2.8	3.3	0.0	+0.6	<0.01	0.34
Number of children	2.2	1.9	2.3	0.0	+0.3	0.04	0.27
<b>Basic demographics</b>							
Proportion of females (%)	52.2	64.4	51.4	-0.1	-13.0	0.01	0.26
Age (years)	37.6	42.6	37.3	-0.3	-5.3	<0.01	0.39
Born outside Germany (%)	83.1	83.2	83.1	0.0	0.0	0.98	0.00
Years since immigration	20.9	21.3	20.9	0.0	-0.4	0.76	0.04
Binational marriage <sup>a</sup> (%)	31.3	61.4	29.3	-2.0	-32.2	<0.01	0.70
<b>Education</b>							
Without degree (%)	17.0	10.9	17.4	+0.4	+6.5	0.09	0.17
University entrance diploma (%)	23.8	41.6	22.6	-1.2	-19.0	<0.01	0.45
Years of education	11.1	12.8	11.0	-0.1	-1.7	<0.01	0.55
<b>Employment</b>							
Employed (%)	19.7	30.3	19.1	-0.7	-11.3	0.01	0.28
Unemployed (%)	42.4	36.0	42.7	+0.4	+6.8	0.21	0.14
Welfare benefit receipt (%)	59.9	44.6	61.0	+1.1	+16.4	<0.01	0.37
Maternity leave (%)	3.3	9.0	2.9	-0.4	-6.1	<0.01	0.34
<b>Income</b>							
Net household income (Euro)	1475.4	1876.2	1448.1	-27.3	-428.1	<0.01	0.48
<b>Deprivation</b>							
Deprivation index (raw count)	6.7	4.8	6.8	+0.1	+2.0	<0.01	0.49
Deprivation index (weighted)	2.1	1.4	2.1	0.0	+0.7	<0.01	0.48

Table 1. Continued

	(1) all migrants	(2) false negative	(3) true positive	(4) bias $\Delta$ (3)-(1)	(5) $\Delta$ (3)-(2)	(6) $p(t)$	(7) Cohen's $d$
<b>Religion</b>							
Member in rel. community (%)	74.1	65.6	74.6	+0.5	+9.0	0.05	0.21
Self-rating as religious (%)	64.2	53.3	64.9	+0.7	+11.6	0.02	0.24
Proportion of Muslims (%)	37.4	5.9	39.5	+2.1	+33.6	<0.01	0.70
<b>Language not German</b>							
During interview (%)	19.0	9.9	19.6	+0.6	+9.7	0.02	0.25
Within household (%)	71.2	44.6	72.9	+1.7	+28.3	<0.01	0.63
Among friends (%)	44.7	26.3	46.4	+1.7	+20.0	<0.01	0.41
<b>Subjective satisfaction</b> <sup>b</sup>							
Housing conditions	6.6	7.1	6.5	0.0	-0.6	0.03	0.22
Living standard	5.7	6.4	5.6	0.0	-0.8	<0.01	0.30
Life in general	6.2	6.7	6.2	0.0	-0.5	0.06	0.19
Health	6.8	6.7	6.8	0.0	+0.1	0.73	0.04
Social participation	6.4	6.5	6.4	0.0	-0.1	0.77	0.03
Sample size	1,610	101	1,509				

<sup>a</sup>Partner has German citizenship.<sup>b</sup>Mean on 10-point scale.

Table 2. Differences between all migrants, false negative and true positive cases (Rule 2: GN and SN)

	(1) all migrants	(2) false negatives	(3) true positives	(4) bias $\Delta(3)-(1)$	(5) $\Delta(3)-(2)$	(6) $p(t)$	(7) Cohen's $d$
<b>Household characteristics</b>							
Household size (persons)	3.3	3.0	3.5	+0.2	+0.5	<0.01	0.28
Number of children	2.2	2.0	2.4	+0.1	+0.4	<0.01	0.29
<b>Basic demographics</b>							
Proportion of females (%)	52.2	60.5	47.4	-4.1	-13.2	<0.01	0.27
Age (years)	37.6	39.4	36.6	-1.0	-2.7	<0.01	0.20
Born outside Germany (%)	83.1	86.1	81.3	-1.8	-4.8	0.01	0.13
Years since immigration	20.9	18.8	22.2	+1.3	+3.4	<0.01	0.29
Binational marriage <sup>a</sup> (%)	31.3	45.8	22.3	-8.9	-23.4	<0.01	0.52
<b>Education</b>							
Without degree (%)	17.0	10.7	20.6	+3.6	+9.9	<0.01	0.27
University entrance diploma (%)	23.8	30.6	19.8	-4.0	-10.8	<0.01	0.26
Years of education	11.1	12.0	10.6	-0.5	-1.3	<0.01	0.42
<b>Employment</b>							
Employed (%)	19.7	21.2	18.9	-0.9	-2.4	0.26	0.06
Unemployed (%)	42.4	39.8	43.8	+1.4	+4.0	0.13	0.08
Welfare benefit receipt (%)	59.9	55.4	62.6	+2.7	+7.2	<0.01	0.15
Maternity leave (%)	3.3	4.3	2.7	-0.6	-1.7	0.08	0.09
<b>Income</b>							
Net household income (Euro)	1475.4	1562.4	1424.0	-51.4	-138.4	<0.01	0.15
<b>Deprivation</b>							
Deprivation index (raw count)	6.7	6.1	7.0	+0.3	+0.8	<0.01	0.20
Deprivation index (weighted)	2.1	1.9	2.2	+0.1	+0.3	<0.01	0.22
<b>Religion</b>							
Member in rel. community (%)	74.1	67.5	77.8	+3.7	+10.2	<0.01	0.24
Self rating as religious (%)	64.2	60.7	66.3	+2.1	+5.6	0.03	0.18
Proportion of Muslims (%)	37.4	18.5	48.3	+10.9	+29.9	<0.01	0.65

Table 2. Continued

	(1) all migrants	(2) false negatives	(3) true positives	(4) bias $\Delta$ (3)-(1)	(5) $\Delta$ (3)-(2)	(6) $p(t)$	(7) Cohen's $d$
<b>Language not German</b>							
During interview (%)	19.0	19.0	19.0	0.0	0.0	0.99	0.00
Within household (%)	71.2	63.5	75.6	+4.4	+12.1	<0.01	0.27
Among friends (%)	44.7	43.1	45.8	+1.1	+2.7	0.48	0.05
<b>Subjective satisfaction<sup>b</sup></b>							
Housing conditions	6.6	6.8	6.5	-0.1	-0.3	0.02	0.12
Living standard	5.7	5.8	5.6	-0.1	-0.2	0.07	0.10
Life in general	6.2	6.4	6.1	-0.1	-0.3	0.01	0.13
Health	6.8	6.7	6.8	0.0	+0.1	0.38	0.05
Social participation	6.4	6.3	6.4	+0.1	+0.2	0.22	0.06
Sample size	1,610	590	1,020				

<sup>a</sup>Partner has German citizenship.

<sup>b</sup>Mean on 10-point scale.

### 6.1. Overall Comparison of Rule 1 Versus Rule 2

Rule 2 results in quite a substantial false negative rate of 36.7% but generates few false positives (6.8%). By contrast, Rule 1 has a false negative rate of only 6.3%, but produces more false positives (41.1%). Thus screening costs would be lower for Rule 2.

A general pattern is obvious from the tables: On most (20 of 28) variables the absolute difference between false negatives and true positives (Column 5) is higher for Rule 1 (GN or SN) than for Rule 2. This is exactly what we would predict based on our hypotheses: Migrants who differ in given name and surname from the domestic population are less well assimilated than persons who only differ in either first name or last name. However, as bias is equal to the product of this difference with the false negative rate and the false negative rate is more than five times higher for Rule 2 than for Rule 1 the resulting absolute bias (Column 4) is, in almost all cases, larger for Rule 2 than for Rule 1. Thus, Rule 1 should be used if there are concerns about a possible assimilation bias.

### 6.2. Demographic and Assimilation Variables

In accordance with the hypotheses about the assimilation process, correctly identified persons (TP) live in larger households with more children than FN (who would have been missed by NBS). Furthermore, NBS would miss more females. All three effects are significant (with  $p < 0.05$ ). However, NBS would yield small biases for household size and number of children. This holds for both rules. The exception is the proportion of females: NBS with Rule 2 would result in an underestimation of 4.1 percentage points, since the proportion of females is higher among the FN. The smaller proportion among the TP is most probably due to the adoption of the name of the husband after marriage. This could be indicative of more general assimilation bias in NBS, since gender is correlated with many assimilation indicators (for example, employment status and education).

Whereas the individual probability of a change of name due to naturalisation or marriage is expected to decrease with age, the overall probability for a domestic name should increase with duration of stay; furthermore the proportion of domestic names is expected to be higher for children of migrants. Therefore, the age distribution of all migrants should differ from the distribution of those who could be found by NBS. As can be seen from the tables, in fact the migrants from NBS were significantly younger on average (Rule 1: 5.3 years; Rule 2: 2.7 years). From the kernel-density plots in [Figure 1](#), it is also clear that both rules more likely miss the younger migrants (aged 15–22 years). In particular, Rule 2 seems to favour older migrants above the age of 45. In total, this is a small effect, causing a small bias of less than one year in the estimation of the mean age. As age is closely related to years since immigration (for those who were born outside Germany) a similar effect could be expected for this variable. The estimation of years since immigration is unbiased for Rule 1 and slightly biased for Rule 2 (1.3 years difference); giving a small but significant effect for Rule 2. Almost the same pattern can be seen for the percentage of respondents born outside Germany.

#### 6.2.1. Binational Marriage

Binational marriage is defined here as an officially recorded partnership of a migrant with a German partner. This is a rare event in Germany: Based on official statistics,

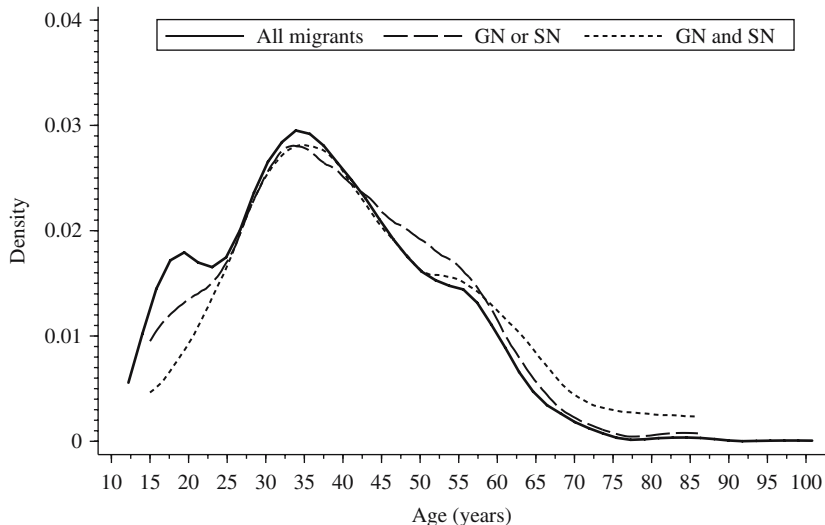


Fig. 1. Distribution of age for all migrants and false negative cases for Rule 1 and 2

Haug (2011) reports about 17% of the first generation male migrants (20% for females) to be married to a German partner. For the second generation of migrants these figures rise to 28% and 21% for males and females respectively. In this dataset, about 31% of the respondents in a recorded partnership reported a binational marriage. Amongst those who would have been missed by NBS, the percentage is almost double (61.4%). Amongst those selected by NBS, the percentage is about 29% (Rule 1) and 22% (Rule 2). Both differences are significant with medium effect sizes. Due to the small proportion of false negatives for Rule 1, however, the overall underestimation (i.e., bias) by NBS would be 2 percentage points for Rule 1 and 8.9 percentage points for Rule 2.

### 6.2.2. Education

Education is considered as a key requirement for successful assimilation. Therefore, higher educational attainment would be expected amongst more assimilated migrants. Hence, differences between population parameters and NBS estimates are most probable. In fact, regardless of which indicator is used to reflect educational attainment in PASS, the predicted differences can be observed. Tables 1 and 2 show the same pattern of results for the proportion of migrants with no school qualifications, the proportion of migrants with a university entrance diploma and years of education. On average, those missed by NBS are better educated than those found by NBS. Five of six comparisons between FN and TP are significant, but only one indicator has a medium effect size. The indicator 'years of education' has one of the largest effects reported here with  $d = 0.55$  for Rule 1 and  $d = 0.42$  for Rule 2. Nonetheless, the resulting underestimation would be small (0.1 and 0.5 years). Figure 2 shows the effect clearly: better-educated persons are more likely to be missed by both rules, but the proportion of highly educated persons amongst the migrants is so small that the effect on overall mean estimation is also small.

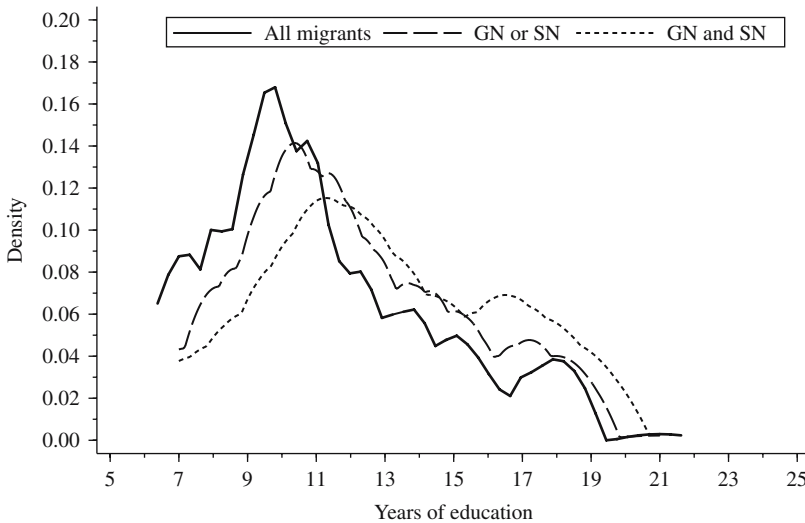


Fig. 2. Distribution of years of education for all migrants and false negative cases for Rule 1 and 2

### 6.2.3. Employment Status and Receipt of Benefits

About 20% of all migrants in the total sample are employed. Regardless of which rule is used, the proportion of employed persons is always higher among the migrants who would have been missed by NBS. For Rule 1, the difference between FN and TP is more than 11 percentage points (2.4 percentage points for Rule 2): a significant, albeit small effect. The resulting bias in the estimation of employed persons would be lower than 1 percentage point. The same pattern of results can be observed for the estimation of unemployment and the proportion of females on maternity leave.

The effect is slightly more prominent in the estimation of households receiving welfare benefits. About 60% of the migrants in the unweighted PASS sample received welfare benefits. Migrants who would have been missed by NBS received welfare benefits less frequently than the population of all migrants and less frequently than those who would have been selected by NBS. The difference of 16.4 percentage points (Rule 1; for Rule 2: 7.2 percentage points) is significant, with only a small effect size. It leads to a bias of +1.1 percentage points for Rule 1 and +2.7 percentage points for Rule 2.

### 6.2.4. Household Income

Income is a central variable in social science and economic research and also a major indicator of successful assimilation. Therefore, if NBS has an assimilation bias, lower income would be expected for migrants identified by NBS compared to migrants who would have been missed by NBS. This hypothesis is clearly supported by the data: the differences of 428 Euros (Rule 1) and 138 Euros (Rule 2) between true positives and false negatives are significant. For Rule 2, the effect size of  $d = 0.48$  is among the largest effects reported here. The absolute bias of 27 Euros (Rule 1; 51 Euros for Rule 2) approximates to nearly 1.9% (3.5%) of total income. A plot (Figure 3) of the estimated



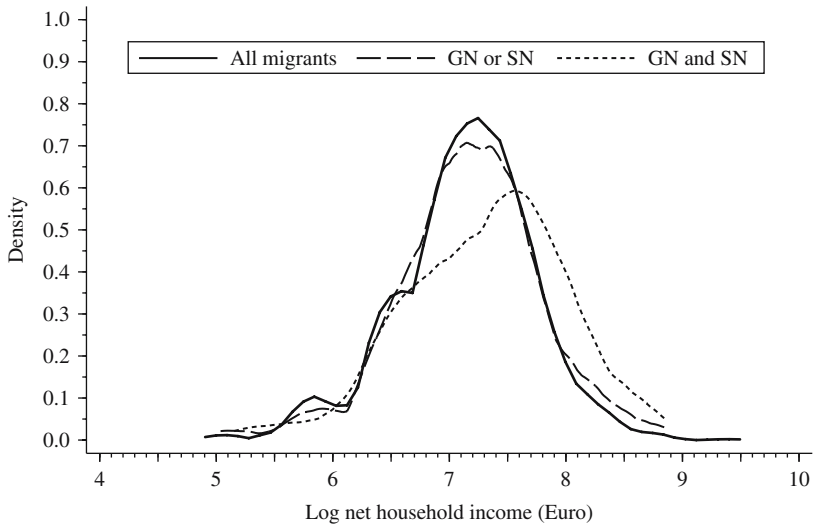


Fig. 3. Distribution of log net household income (Euro) for all migrants and false negative cases for Rule 1 and 2

income distributions for all migrants and the false negative cases of both rules shows the large bias in income which would occur if NBS with Rule 2 were used.

#### 6.2.5. Deprivation

Based on a list of 26 goods, facilities and social activities, [Christoph et al. \(2008, p. 46\)](#) defined an “index of deprivation” as the number of items or facilities which the members of a household do not possess or use due to financial restrictions. The list contained goods and facilities such as central heating, indoor toilet, washing machine, refrigerator, TV and social activities such as inviting friends for dinner or visiting cultural events. The weighted version of this index counts items if the respondents rate them as ‘necessary’. This deprivation index can be seen as a measure of successful assimilation, and therefore differences between better and less well-assimilated migrants are likely. Thus an assimilation bias of NBS would be expected in this instance.

[Figure 4](#) shows the estimated distributions of the deprivation index for all migrants and migrants who would have been missed by NBS according to Rule 1 or Rule 2. For both subgroups of false negatives, the distribution shifted to the left, so that true positives of NBS would overestimate the index. The difference between the subgroups is significant with a small to medium effect size. Again, the overestimation effect is larger for Rule 2 (bias +.3 points compared to +.1 points for Rule 1).

#### 6.2.6. Religion

The traditional assimilation process in modern Western societies usually leads to weakened religious values and decreased religious behaviour. So differences in indicators of religious behaviour due to NBS are to be expected. PASS has three indicators of religious behaviour: Membership in a religious community, the proportion of Muslims, and the degree of self-reported religiousness (see [Tables 1 and 2](#)). The pattern of results is

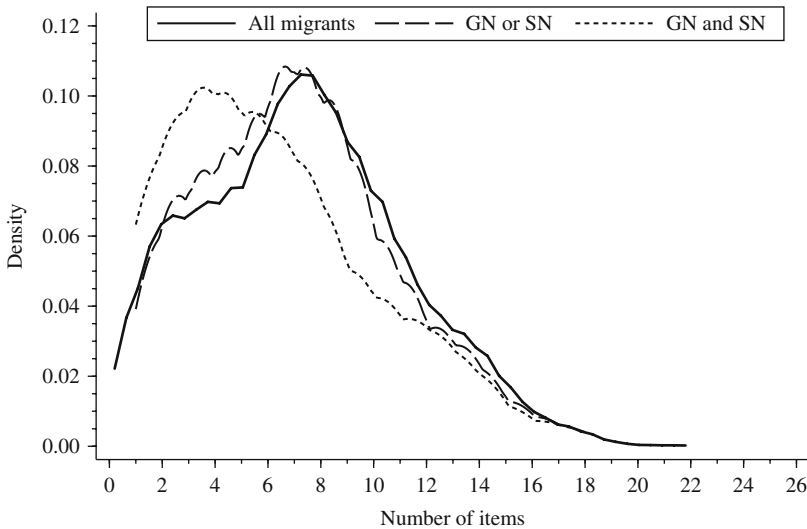


Fig. 4. Deprivation index for all migrants and false negative cases for Rule 1 and 2

the same for all three indicators and both rules: those migrants who would have been missed by NBS are less religious than the sample of all migrants and those who would have been identified by NBS are more religious than the sample of all migrants. The difference in the proportions of Muslims is striking. In the population of all migrants in PASS, the proportion of Muslims is about 37%, but among those missed by NBS, the proportion is only about 6% (Rule 1) and about 19% (Rule 2). All six differences between FN and TP are significant. However, all but the effect sizes regarding the proportion of Muslims ( $d = 0.70$  with Rule 1 and  $d = 0.65$  with Rule 2) should be regarded as small. Bias is much larger for Rule 2 than for Rule 1 for all three variables, peaking 10.9 percentage points for the proportion of Muslims which is the largest bias reported here.

### 6.2.7. Language

The ability to read, speak and write the language of the host society is considered to be of crucial importance for assimilation. Sufficient language skills enable contact with the domestic population and participation in social, political and everyday activities. Success in school and in the labour market also depends on language skills. Therefore differences in indicators relating to the command of the language of the host society due to NBS can be expected. PASS has three direct or indirect indicators of language skills and language use. PASS records the language in which an interview was conducted, there is a question about the language used predominantly in the household and finally the predominant language used amongst friends outside the household is recorded. Table 1 shows the same pattern of results as for most other indicators for Rule 1. Those missed by NBS seem to be more assimilated than those detected by NBS. All three language indicators differ significantly between FN and TP. However, despite some medium-sized effects, the differences between TP and the total sample are small. The effect sizes for Rule 2 are even

smaller. The largest effect for both rules can be seen for the dominant language within the household. Here, Rule 2 produces a bias of +4.4 percent. Surprisingly, results for language use do not seem to be strongly biased by NBS.

#### 6.2.8. *Subjective Satisfaction*

These results are based on demographic variables or language use. When subjective variables such as satisfaction with housing conditions and living standards were considered, there were only minor differences between the overall population and the two groups of false negatives. In general, satisfaction was higher amongst the smaller group of false negatives (with first and last names classified false negative) than in the population. As might be expected, subjective satisfaction with health and social participation, which are less related to the assimilation process, do not show the same pattern: overall bias for subjective measures is never larger than .1 points on the eleven point scale employed.

#### 6.3. *Summary*

Our central hypothesis is clearly supported by the data. For most of the assimilation-related variables considered here, there are significant differences between migrants identified by NBS versus migrants not identified by NBS: in total, 21 of 28 tests for Rule 1 and 20 of 28 tests for Rule 2 showed significant differences with  $p \leq 0.05$ . However, the effect sizes for Rule 1 were small for 17 of 28 tests, four additional effect sizes were medium whereas only three Rule 2 effect sizes were larger than  $d = 0.3$ . In summary: name-based sampling results in biased estimates, but the effect sizes are mostly small.

### 7. Discussion

Name-based sampling procedures differ with respect to details for the generation of sampling frames. However, since all name-based methods share the common problem of undercoverage through assimilation, they are all vulnerable to the same selection effects. Therefore, it is reasonable to expect the same overall pattern of results for all name-based sampling procedures.

In this study on assimilation bias we found significant, but usually small differences in education, employment, frequency of intermarriage, income and religious behaviour between the complete sample and the name-based sample. Although effect sizes were on average larger for Rule 1 than for Rule 2, Rule 1 produced almost no bias due to the low FN rate; no single bias exceeded 2.1 percentage points for this rule. Rule 2 on the other hand showed a substantial FN rate which caused considerable bias for some variables despite small effect sizes. For example, the proportion of Muslims was overestimated by 10.9 percentage points and the proportion of binational marriages was underestimated by -8.9 percentage points. In studies with sample sizes in the order of 2,000 to 5,000 respondents, which are quite common, this systematic bias will exceed the sampling error. If high accuracy for assimilation variables is required by a research project, procedures with lower FN rates than Rule 2 should be used. Although this is the first study to date on bias in name-based samples, it has some limitations. The restrictions imposed by German

law limit name classification to survey respondents only; names of refusals cannot be used. Therefore, the results depend on the assumption that response in the survey is not highly correlated with the classification result. Although it seems possible that false negative migrants in NBS have higher refusal or noncontact rates, we consider our results to be robust with regard to this assimilation effect. However, this issue merits further examination in a separate study.

Furthermore, we have to assume that the survey report on migration background is unbiased and that answers to the questions about dependent variables are either unbiased or bias is the same across the groups that we have compared. If there was bias in the responses to questions on these dependent variables such as income or welfare receipt and this bias differed in size between all migrants, whether false negatives or false positives, this would affect the results of this study. There is, however, no evidence for this artifact.

This study focused on foreign national migrants, which was justified by the theoretical assumption that limiting the study to this subgroup should reduce bias. We tested the effect of this definition further using broader definitions of migrant groups. As expected, bias increased when broader definitions of migration status were used. For example, when migrants were defined as everyone who has foreign nationality, or was born abroad, or has at least one parent born abroad, bias in [Table 1](#) increased for 25 of 28 variables. Using other definitions of migration status also seemed to increase bias; we therefore consider our results indicative of the lower limit of bias.

The results of this study are of course dependent on the current German migrant population. Each country will have different results at different points in time. These results are likely to depend on the proportions and characteristics of migrant groups, likelihood of intermarriages, legal requirements for naturalisation or name changes and so on. However, it is highly plausible that the four mechanisms leading to biased estimates for name-based sampling discussed above will be universal across most cultures: changing names after marriage and naturalisation, giving offspring names that are common in the host culture, and higher probability of migration from countries with a common language. These mechanisms will result in a higher probability of exclusion by name-based sampling procedures for better-assimilated migrants in almost every society. The basic message of this study thus also applies to countries other than Germany: when name-based methods are used, potential bias should be carefully examined since name-based sampling methods have the potential to enforce stereotypes about migrant populations by oversampling less-assimilated individuals.

## 8. References

- Alba, R. and Nee, V. (1997). Rethinking Assimilation Theory for a New Era of Immigration. *International Migration Review*, 31, 826–874.
- Babka von Gostomski, C. (2010). *Fortschritte der Integration*. Nuremberg: Bundesamt für Migration und Flüchtlinge.
- Becker, B. (2009). Immigrants' Emotional Identification With the Host Society: The Example of Turkish Parents' Naming Practices in Germany. *Ethnicities*, 9, 200–225. DOI: <http://www.dx.doi.org/10.1177/1468796809103460>

- Bertelsmann Stiftung (2009). *Zuwanderer in Deutschland. Ergebnisse einer repräsentativen Befragung von Menschen mit Migrationshintergrund*. Gütersloh: Bertelsmann Stiftung.
- Bethlehem, J. (2009). *Applied Survey Methods: A Statistical Perspective*. Hoboken, NJ: Wiley.
- Bethmann, A. and Gebhardt, D. (2011). User Guide “Panel Study Labour Market and Social Security” (PASS). Wave 3. (FDZ-Datenreport No. 04-2011). Nuremberg: Research Data Centre FDZ of the German Employment Agency.
- Blane, H.D. (1977). Acculturation and Drinking in an Italian American Community. *Journal of Studies on Alcohol*, 38, 1324–1346.
- Braun, M. and Santacreu, O. (2009). Methodological Notes. In *Pioneers of European Integration: Citizenship and Mobility in the EU*, E. Recchi and A. Favell (eds). Cheltenham: Edward Elgar, 241–254.
- Christoph, B., Müller, G., Gebhardt, D., Wenzig, C., Trappmann, M., Achatz, J., Tisch, A., and Gayer, C. (2008). Codebook and Documentation of the Panel Study Labour Market and Social Security (PASS): Introduction and Overview, Wave 1 (2006/2007) (FDZ-Datenreport No. 05-2008). Nuremberg: Research Data Centre FDZ of the German Employment Agency.
- Ecob, R. and Williams, R. (1991). Sampling Asian Minorities to Assess Health and Welfare. *Journal of Epidemiology and Community Health*, 45, 93–101. DOI: <http://www.dx.doi.org/10.1136/jech.45.2.93>
- Ellis, P.D. (2010). *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results*. Cambridge: Cambridge University Press.
- Fassmann, H. and Münz, R. (1994). *European Migration in the Late Twentieth Century: Historical Patterns, Actual Trends, and Social Implications*. Frankfurt: Campus.
- Gerhards, J. and Hans, S. (2009). From Hasan to Herbert: Name-Giving Patterns of Immigrant Parents Between Acculturation and Ethnic Maintenance. *American Journal of Sociology*, 114, 1102–1128.
- Haug, S. (2011). Binationale, Interethnische und Interreligiöse Ehen in Deutschland. *Familie, Partnerschaft, Recht*, 10, 417–422.
- Humpert, A. and Schneiderheinze, K. (2000). Stichprobenziehung Für Telefonische Zuwanderumfragen. Einsatzmöglichkeiten der Namensforschung. *ZUMA-Nachrichten*, 24, 36–64.
- Kalter, F. and Granato, N. (2002). Demographic Change, Educational Expansion and Structural Assimilation of Immigrants: The Case of Germany. *European Sociological Review*, 18, 199–216. DOI: <http://www.dx.doi.org/10.1093/esr/18.2.199>
- Kalton, G. (2009). Methods for Oversampling Rare Subpopulations in Social Surveys. *Survey Methodology*, 35, 125–141.
- Kosmidis, G., Rutishauser, I., Wahlquist, M., and McMichael, A. (1980). Food Intake Patterns Amongst Greek Immigrants in Melbourne. *Proceedings of the Nutrition Society of Australia*, 5, 165.
- Kreuter, F., Müller, G., and Trappmann, M. (2010). Nonresponse and Measurement Error in Employment Research: Making Use of Administrative Data. *Public Opinion Quarterly*, 74, 880–906. DOI: <http://www.dx.doi.org/10.1093/poq/nfq060>

- Liebig, T. (2007). The Labour Market Integration of Immigrants in Germany (OECD Social, Employment and Migration Working Paper No. 47). Paris: Organisation for Economic Co-operation and Development OECD.
- Mateos, P. (2007). A Review of Name-Based Ethnicity Classification Methods and Their Potential in Population Studies. *Population, Space and Place*, 13, 243–263. DOI: <http://www.dx.doi.org/10.1002/psp.457>
- Milewski, N. (2007). First Child of Immigrant Workers and Their Descendants in West Germany: Interrelation of Events, Disruption, or Adaptation? *Demographic Research*, 17, 859–895.
- Mitchell, T.M. (1997). *Machine Learning*. Boston, MA: McGraw-Hill.
- Rutishauser, I.H. and Wahlquist, M. (1983). Food Intake Patterns of Greek Migrants to Melbourne in Relation to Duration of Stay. In *Proceedings of the Nutrition Society of Australia*, 8, 49–55.
- Sakshaug, J. and Kreuter, F. (2012). Assessing the Magnitude of Non-Consent Biases in Linked Survey and Administrative Data. *Survey Research Methods*, 6, 113–122.
- Schnell, R., Gramlich, T., Mosthaf, A., and Bender, S. (2010). Using Complete Administration Data for Nonresponse Analysis: The PASS Survey of Low-Income Households in Germany. In *Proceedings of Statistics Canada Symposium 2010. Social Statistics: The Interplay among Censuses, Surveys and Administrative Data*. Ottawa: Statistics Canada, 104–109.
- Schnell, R., Trappmann, M., Gramlich, T., Bachteler, T., Reiher, J., Smid, M., and Becher, I. (2013a). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. *Methoden – Daten – Analysen*, 7(2), 5–33.
- Schnell, R., Trappmann, M., Gramlich, T., Bachteler, T., Reiher, J., Smid, M., and Becher, I. (2013b). A new method for name-based sampling of migrants using n-grams. (Working Paper No. 2013-04). German Record Linkage Center, Nuremberg. Available at: <http://www.record-linkage.de/-download=wp-grlc-2013-04.pdf>
- Statistisches Bundesamt (ed.). (2011). *Datenreport 2011*. Berlin: Bundeszentrale für politische Bildung.
- Sudman, S. and Kalton, G. (1986). New Developments in the Sampling of Special Populations. *Annual Review of Sociology*, 12, 401–429.
- Sulaiman-Hill, C.M. and Thompson, S.C. (2011). Sampling Challenges in a Study Examining Refugee Resettlement. *BMC International Journal of Health and Human Rights*, 11, 2–11. DOI: <http://www.dx.doi.org/10.1186/1472-698X-11-2>
- Trappmann, M., Gundert, S., Wenzig, C., and Gebhardt, D. (2010). PASS: A Household Panel Survey for Research on Unemployment and Poverty. *Schmollers Jahrbuch. Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 130, 609–622.
- Trappmann, M., Beste, J., Bethmann, A., and Müller, G. (2013). The PASS Panel Survey after Six Waves. *Journal for Labour Market Research*, 46, 275–281. DOI: <http://www.dx.doi.org/10.1007/s12651-013-0150-1>

Received February 2013

Revised October 2013

Accepted November 2013

## Comparing Survey and Sampling Methods for Reaching Sexual Minority Individuals in Flanders

Alexis Dewaele<sup>1</sup>, Maya Caen<sup>2</sup>, and Ann Buysse<sup>1</sup>

As part of a large sexual health study, we used two different approaches to target Sexual Minority Individuals (SMIs). Firstly, we drew on a probability sample (1,832 respondents aged 14–80) of the Flemish population in Belgium. Secondly, we set up a targeted sampling design followed by an Internet survey. Our focus was to explore how two different sampling procedures and survey designs could lead to differences in sample characteristics. Results showed that for female SMIs (we excluded male SMIs from the analyses due to their low numbers) the population sample differed from the Internet sample in terms of sociodemographic characteristics (the latter included younger and more highly educated respondents) and scores on sexual orientation dimensions (the population sample included more respondents who didn't identify as lesbian or bisexual but reported same-sex sexual experiences and desire). Respondents' scores on sexual health indicators differed between the samples for two of the seven variables. We discuss implications for improving the quality and validity of nonrandom samples.

*Key words:* Hard-to-reach populations; self-selection bias; nonrandom samples.

### 1. Introduction

Lesbian women, gay men and bisexuals (LGBs) are widely considered a hard-to-reach population. As part of a larger systematic study of sexual health, we used two different approaches to target this population. In our national sample, we drew on a probability sample (1,832 respondents aged 14–80) of the Flemish population (the Dutch-speaking community in Belgium) based on the Belgian National Register. For the other arm of the study we set up a targeted sampling design followed by an Internet survey, using a near-identical questionnaire. By identifying a population of sexual minority individuals (we prefer using the acronym SMI instead of LGB since the latter implies self-identification; SMI also refers to individuals who *do not* identify as LGB but who have or have had same-sex partners) in the national survey, we explored how two different sampling procedures

<sup>1</sup> Faculty of Psychology and Educational Sciences, Department of Experimental Clinical and Health Psychology, Ghent University, B-9000, Ghent, Belgium. Email: Alexis.dewaele@ugent.be

<sup>2</sup> Department of Sociology, Research team CuDOS, Ghent University, B-9000, Ghent, Belgium.

**Acknowledgments:** The Sexpert study group includes Ann Buysse (Ghent University: Department of Experimental, Clinical and Health Psychology), Paul Enzlin (KU Leuven: Department of Development and Regeneration, Institute for Family and Sexuality Studies and UPC KU Leuven, Context – Centre for Couple, Family and Sex Therapy), Guy T'Sjoen (Ghent University Hospital: Department of Endocrinology and Center for Sexology and Gender Problems), John Lievens, Mieke Van Houtte and Hans Vermeersch (Ghent University: Department of Sociology, Research Team Cultural Diversity: Opportunities and Socialisation). The Sexpert study was funded by the Strategic Basic Research program of the Flemish Agency for Innovation by Science and Technology.



and survey designs led to differences in sample characteristics, especially where sexual health and sociodemographic characteristics are concerned. Hence, we provide insight into self-selection processes and explore future strategies for improving the quality and validity of nonrandom samples that rely upon self-selection. This validation, through comparison of data collected by different sampling methods, is especially useful when referring to survey data for program and policy development (Schwarcz et al. 2007). In this study, we took the unique opportunity to compare two datasets from a larger sexual health study in Flanders, Belgium. Due to the low numbers of male SMIs in our population sample, we only focused on female SMIs.

### *1.1. Gathering Data on Sexual Minority Individuals Through Probability and Nonprobability Samples*

Gathering a population-based probability sample while including multidimensional and continuous measures for sexual orientation guarantees that respondents from diverse age groups, levels of education, geographical locations, and, presumably, also with different sexual orientations are covered (see e.g., Laumann et al. 1994; Bajos and Bozon 2008). However, the proportions of men and women who identify as nonheterosexual in population-based probability surveys are often small (Rothblum 2007). The low proportions of SMIs are particularly troublesome as small (absolute) numbers lead to difficulties in estimating reliable parameters. This is especially true for communities such as Flanders (the northern, Dutch-speaking part of Belgium, which has about 6 million inhabitants), where high-quality, population-based representative surveys via face-to-face interviews, typically with samples drawn from the National Register, are not only very expensive in terms of sampling procedures but also require enormous resources (in terms of workforce, time and general effort) in order to organize and carry out data collection. Secondly, researchers have used self-identity, sexual activity and cohabiting status as ways to find nonheterosexual respondents. However, these dimensions are usually not highly correlated. For example, in census data the gender of partners who are cohabiting is sometimes used to capture SMIs. This of course leaves out information about those who are single and those who are not living with their partner (Rothblum 2007). All of these problems can be addressed in Internet surveys.

Making the questionnaire available on the Internet through Computer Assisted Self Interviewing (CASI) has several advantages (see e.g., Wright 2005; Evans and Mathur 2005; Couper 2008). It is a relatively cheap and fast way to gather data (Best and Krueger 2004; Heerwegh 2001). For the SMI target group in particular, it offers a highly accessible and anonymous way of posing delicate questions and gathering information about private matters (Bauermeister et al. 2012). Administration of a questionnaire through the Internet can increase the level of reporting of sensitive information and has a positive effect on accuracy (Kreuter et al. 2008; Tourangeau et al. 2003). Some authors also refer to high rates of Internet use by (young) SMIs because this medium offers them opportunities to find peers (Silenzio et al. 2009).

The most obvious drawback of Internet surveys is that they may not be representative of the population of interest because the subpopulation with access to the Internet may be quite specific (i.e., coverage error) (Couper 2000; Schonlau et al. 2009). Moreover,



Internet surveys typically rely upon self-selection and thus tend to reach respondents with a particular interest in the survey's topic (Couper 2000). On the other hand, intrinsic interest in the survey's topic to some level determines respondents' willingness to participate in other surveys (for instance face-to-face surveys) as well (e.g., Groves et al. 2006). To evaluate potential coverage error and self-selection bias, it is important to compare Internet and offline methods.

### 1.2. Comparing Internet and Offline Methods for Reaching SMIs

Very few studies have involved a research design that allows direct comparison of (near-) identical questionnaires, assessed via highly comparable administration modes, but following different sampling procedures (see e.g., Denscombe 2006). Moreover, results from existing analyses have often proven inconsistent (de Leeuw 2005).

Schillewaert and Meulemeester (2005) compared four different methods of data collection in Flanders, both offline (through a mail survey and random digit dialing) and online (through pop-ups on high traffic sites that linked to the survey and via an Internet web panel) using identical questionnaires. Differences in sociodemographic characteristics appeared widespread across both offline and online methods. Subjects in the online pop-up sample seemed to be more extroverted and outgoing, while the mail sample showed a more traditional and introverted profile. After adjusting the weight of gender and age to match national population distributions, and after randomly reselecting observations such that they no longer differed from the national population, no major differences were found between the four recruitment methods in terms of demographics and attitudes, interests and opinions. The authors concluded that for traditional research topics, online research tools are at least as externally valid as research conducted via traditional methods.

Other researchers have compared Internet-based and venue-based methods to contact and survey male SMIs. Time location sampling at venues, Internet forum-based sampling and direct marketing (placing banner advertisements on online forums frequented by male SMIs) produced samples that showed variation in terms of residence location, age, income, and self-reported HIV status, as well as prevalence of substance use, methamphetamine use, and serodiscordant partnerships. The direct marketing approaches (i.e., placing banner ads on high traffic [gay and nongay] websites) were more passive in nature and it was therefore suggested that these techniques might result in the recruitment of fewer men who engage in high-risk behaviors in comparison with more active approaches (Raymond et al. 2010). Koch and Emrey (2001) found that a sample of gay men and lesbians recruited through the Internet showed similar characteristics to a national sample of gay men and lesbians. They compared demographic data (education, income, age, race, party identification, and ideology) collected from over 10,000 gay and lesbian users of a single website with data from a sample of national voters: the 1992 Voter Research and Surveys exit poll (Edelman 1993). This national survey included information about voters' sexual orientation, thus providing a useful comparison. Although some differences were found (e.g., the Internet sample was younger), the overall distribution of responses on these demographic variables across the two samples tended to be similar.

Finally, Fernee and Keuzenkamp (2011) compared two samples with respondents who identified as SMIs. The first was a sample containing SMIs recruited through the Internet,

social networks, social media, SMI- and non-SMI specific media ( $N = 5,069$ ). Most respondents were recruited through lesbian- or gay-specific channels or by word of mouth through lesbian or gay friends/acquaintances. The second sample included randomly recruited SMIs from a large Internet panel. Neither sample was representative of the population in the Netherlands, despite the fact that the respondents from the Internet panel did not volunteer themselves. When age and gender differences were controlled for, the first sample, i.e., the group of self-selected respondents, was found to include significantly more exclusively gay/lesbian versus bisexual respondents, as well as more respondents who were open about their sexual orientation. The authors evaluated the data from the Internet panel as more reliable than the data from the sample with self-selected respondents (Ferneer and Keuzenkamp 2011). To conclude, it seems that some studies comparing different sampling methods have uncovered different sample characteristics. Other studies have found comparable samples from different sampling methods with or without controlling for sociodemographic variables.

Face-to-face CASI and web-based CASI both address the problem of interviewer effects on disclosure by SMIs. An Internet survey has the additional advantage that much larger datasets can be obtained in order to avoid unreliable parameter estimates related to small sample sizes. Internet surveys are also far less expensive than population-based probability samples, which typically entail high sampling and data collection costs. Some research shows that results from Internet surveys are comparable to results from traditional methods once differences in sociodemographic sample characteristics are controlled for. Other studies have pointed out that significant differences related to sexual health indicators or minority characteristics (e.g., type of self-identification, openness about one's sexual orientation) remain. In this article we investigate differences in sociodemographic variables, dimensions of sexual orientation, and sexual health indicators between a representative sample and a nonrepresentative (Internet) sample of SMIs.

## 2. Research Design and Methods

### 2.1. Research Procedure

The first study draws on data from the survey 'Sexual Health in Flanders' (Buysse et al. 2013). Respondents (between 14 and 80 years of age) were randomly drawn from the Belgian National Register. In order to enhance statistical power in each of the three predefined age categories we used a stratified sample, meaning that one third of the sample consisted of the youngest responders (aged 14 to 25), one third of the middle age group (aged 26 to 49) and one third of the oldest group (50 to 80 years old). Elaborate contact procedures following Dillman's Total Design Method (Dillman 1978; 2000) were used to maximize the cooperation, the (item) response rate and the quality of all the survey measures. Moreover, some refusal conversion techniques (e.g., a second contact attempt after an initial refusal, made by a different interviewer) were applied. Data were collected between February 2011 and January 2012, and the final database consisted of 1,832 respondents, 125 of whom can be identified as SMIs (response rate: 40.0% of the eligible respondents). After data collection, the data were weighted by gender, age, and schooling level in order to make them representative of the population of Flanders aged 14–80. This enabled us to

partially correct for higher nonresponse rates, which were found for older age groups and among those with a lower educational level.

All data were gathered via face-to-face interviews, using a combination of computer-assisted personal interviewing (CAPI) and computer-assisted self-interviewing (CASI). To elaborate, all sensitive information, that is, a wide range of sexual health indicators, was gathered in a CASI set-up, so that respondents never had to share private information about their sexual health with an interviewer. In this study we only used the data related to sexual health that were gathered in the CASI mode.

While we acknowledge that a response rate of 40% is somewhat lower than expected, especially when compared with other population-based probability surveys, very similar response rates have been found in other European, population-based surveys of sexual health and/or sexual behavior, such as a study in Finland (response rate of 46%), or another conducted in Estonia and Saint Petersburg (response rate of 41%) (Gronow et al.1997; Haavio-Mannila and Kontula 2001). Moreover, both the sensitivity of the survey's topic needs to be taken into consideration as well as its extent, that is, the wide range of sexual and general health indicators and the wide range of correlates covered, and consequently, the duration of the interview (80 minutes on average). Moreover, the poststratification weightings mentioned above ensured that we could, at least to some degree, adjust for higher nonresponse/refusal rates in specific sample groups.

The second study draws on data from the survey 'Click out of the bed room', a large-scale nonrepresentative survey on sexuality, sexual health and relations in SMIs in Flanders. As it was important to recruit all SMIs, including those who do not self-identify as gay, lesbian or bisexual, we set up a neutral as well as an LGB-oriented campaign. The neutral campaign refers to 10,000 posters utilizing an image that did not specifically refer to being lesbian, gay or bisexual that we distributed all over Flanders. The message on the poster presented the survey as related to sexual health in general. Banners, adverts on the Internet and press releases including this neutral image and message were also produced and circulated.

In addition, we also set up a recruitment strategy to target SMIs. To attract a relatively diverse sample we used a variety of recruitment channels and methods. We have learned from previous research that older LGBs, bisexuals, and LGBs with low levels of education are particularly difficult to reach for research purposes (Vincke and Stevens 1999). Additionally, using LGB associations to recruit potential respondents may lead to a large selection bias (see Vincke and Bleys 2003; Vincke and van Heeringen 2004). Our sample strategy was therefore oriented towards avoiding both bias and a lack of representation of specific groups.

We used the following recruitment channels to broadcast a request for respondents: specific locations such as LGB discotheques, LGB parties and LGB events; advertisements in the written press; LGB-specific and non-LGB-specific associations and organizations were invited to spread the invitation; electronic mailings were sent and the Internet was used (e.g., banners posted on LGB-specific websites). We paid for banners on two high traffic sites. The first, GayBelgium, is the largest LGB-specific website in Flanders. The second, 'Seniorennet', targets Internet users older than 50 years. A snowball method was used to recruit respondents through acquaintances, friends, family members, and so forth. Respondents who entered the survey website ([www.klikeensuitbed.be](http://www.klikeensuitbed.be)) or who finished the survey could invite a friend to participate. A promotional team (including

some of the researchers) distributed posters, flyers, and gadgets (small mint tins) that included the URL throughout all Flemish provinces, in large and small towns and villages. The team visited LGB-specific events and activities (parties, a LGBT film festival, LGBT bars) but also put in an extra effort to reach female and elderly SMIs at specific activities. Postal packages (containing posters, flyers and gadgets) were sent out to roughly 180 LGB-specific (bars, shops, associations) and 50 non-LGB-specific (including libraries, cafés, and public health centers) addresses. Key persons within LGBT associations and sexual health organizations were also approached to help us to get in contact with potential respondents through mailings and posts or banners on websites. Finally, a Facebook campaign was set up to recruit respondents. A Facebook panel was integrated into the survey website so that visitors or respondents could 'like' our Facebook page. People who 'liked' our Facebook page received posts about the progress of the research project. Friends of people who 'liked' our Facebook page were indirectly introduced to the project. Finally, we contacted several well-known LGBs who were invited to 'become friends' with the Facebook page and to post promotional messages on their Facebook walls.

Out of the total number of respondents, 35.4% found our site through a social network site (mainly Facebook), 18.5% through an electronic mailing, 15.3% through television, radio, a newspaper or a magazine, 10.6% through clicking on a banner on a website, 6.7% through their school or work, 3.7% through a gadget or flyer, 2.2% through an association or activity, 1.7% through a poster, and 5.8% through other means. Data were collected between September 2011 and March 2012. The final database consisted of 3,702 respondents, 2,468 of whom were identified as SMIs. Respondents between 13 and 86 years of age were included. At the beginning of the survey postal code of the respondents' current place of residence was registered. Respondents living in provinces outside of Flanders were removed from the dataset.

Table 1 summarizes the sampling frames and designs, the differences in (estimated) coverage of the frame, the different sampling methods, and the different contact/recruitment procedures applied in both surveys. The questionnaire used in the internet survey is similar to the first study but significantly shorter to address respondent drop out. More specifically, it only contains the questions on sexual health indicators, administered by CASI (see also Table 2 for a thorough comparison of the question wording and modes of administration).

## 2.2. Measures

We will refer to the population sample and to the Internet sample in turn. Firstly, we will elaborate on how we identified SMIs in both samples. Next, we will present an overview of sexual health measures.

In this study, we took different dimensions of sexual orientation (self-identification, sexual behavior and sexual desire, see also below) into account to construct a group of SMIs. Doing so yielded a larger proportion of female SMIs (9.8%,  $N = 90$ ) compared to male SMIs (3.9%,  $N = 35$ ). This is due to the fact that the women in our study reported more same-sex sexual desire than men. This is in line with other studies, which have shown that a higher proportion of women than men report same-sex attraction (Bajos and Bozon 2008), or report same-sex attraction without identifying as gay, lesbian, or bisexual (Laumann et al. 1994; Roberts et al. 2010). Unfortunately, in our case this meant that we

Table 1. Comparison of population frame, sampling design, sampling and recruitment procedures, response rates and coverage across surveys

	General population survey	Internet survey
<b>1. Population frame, sampling design and sampling procedure</b>		
Design and frame	Population-based probability sample of residents aged 14-80.	Targeted sampling design, via a variety of recruitment channels and methods (see below).
Sampling procedure	Sample randomly drawn from the Belgian National Register. Age-stratified sample in order to enhance statistical power in each of the three predefined age categories. After data collection, the data were weighted by gender, age, and schooling level. The selection of the SMI (sub)sample was based on a three-dimensional construct or definition of sexual orientation that included self-identification, sexual desire and sexual behavior.	The survey on sexual health was targeted towards a broader population (both heterosexual and SMI respondents). The final database included respondents between 13 and 86 years of age. Identical questions were used in both surveys.
Selection of the SMI (sub)sample		
<b>2. Contact/recruitment procedures</b>		
Contact procedure/method of recruitment	First contact: introductory letter by mail. Second contact: face-to-face, i.e., an interviewer visiting the respondent at home	Respondents were recruited via a variety of recruitment channels and methods and via a neutral as well as an LGTB-oriented campaign.
<b>3. Response rates and coverage</b>		
	High refusal rates (around 30%) compared to other surveys drawing on population-based probability samples, mostly due to the survey's topic. Post-hoc weighting to correct for higher nonresponse. No expected coverage error regarding the SMI population. High(er) refusal rates also imply that intrinsic interest in the survey's topic to some extent serves as a selection mechanism.	Unit response and nonresponse rates cannot be estimated. We cannot assume a representative sample. Access to the internet and familiarity with Internet surveys leads to a specific sample. Self-selection due to intrinsic interest in the survey's topic.

Table 1. Continued

	General population survey	Internet survey
<b>4. Costs and benefits</b>		
Financial cost	Expensive, due to: -sampling cost -costs related to interviewer recruitment, training, briefings, fees, follow-up and evaluation -operational and infrastructural costs (e.g., interviewing laptops, survey software, transport, printing and postage costs, and so on)	Estimated to be 30 times less costly than the general population survey. Here, the highest costs were related to the recruitment of respondents.
Timing	21 months: -data collection (13 months). -sampling procedures, recruitment and training of interviewers (8 months).	6 months:Recruitment of respondents and data collection
Power issues	Only a small number of male ( $N = 35$ ) and female ( $N = 90$ ) SMIs could be identified. This led to difficulties producing reliable estimates, especially for the male group	No power issues.

Table 2. Comparison of modes of administration, question wording and type between the general population survey and the Internet survey

	General population survey	Internet survey
Mode of administration	Face-to-face interviews, with CASI (i.e., electronic self-interviewing) for the more sensitive topics and CAPI for the administration of all other, non-sensitive topics.	Web survey, i.e., electronic self-interviewing in the absence of an interviewer.
Length of questionnaire/ duration	An interview took 80 minutes on average.	Respondents needed 20 minutes on average to fill in the online questionnaire.
Question wording and type	Question wording and question type (open ended versus closed) were similar in both questionnaires in terms of the sexual health indicators addressed in this article	
Dealing with sensitive topics	All sensitive questions were included in the CASI part (general population survey) or web part (Internet survey) of the questionnaire.	



could only work with the data gathered from the female group, as the smaller male group would not have produced reliable estimates.

**Sexual orientation.** In order to accurately assess the number of SMIs in each of the samples, we needed a clear definition of sexual orientation. It was deemed important to use a definition of sexual orientation that was neither too restricted nor too broad. From a sexual health perspective, for instance, this definition also had to include women who have sex with women but who do not identify as lesbian or bisexual (see e.g., [Mercer et al. 2007](#); [Kerker et al. 2006](#); [Van Kesteren et al. 2007](#)).

We conceptualized sexual orientation as a three-dimensional construct measuring self-identification, sexual behavior and sexual desire (cf. [Laumann et al. 1994](#)). Sexual self-identification was assessed with the question: “*How would you identify yourself?*” Respondents could answer on a 5-point Likert scale (i.e., straight, more straight than gay/lesbian, bisexual, more gay/lesbian than straight, gay/lesbian). An open-ended response category was added for respondents who did not identify with any of these labels (referred to as other). To measure sexual behavior, we first asked respondents: “*Throughout your life, how many people have you had sex with?*” (categorical open-ended numeric answer). Then we asked: “*Were these people men, women or both?*” Respondents could answer on a 5-point Likert scale (from 1 = only women to 5 = only men). We used two items to measure the dimension of sexual desire. We asked respondents: “*Do you sexually fantasize about men, women or both?*” and “*Do you feel sexually attracted to men, women or both?*” In both cases, respondents could answer on a 5-point Likert scale (from 1 = only about or only attracted to women to 5 = only about or only attracted to men). Respondents could also answer these questions with “*I fantasize about or I am attracted to neither*”. With the information gathered from these four items, measuring three dimensions, we created one dichotomous variable categorizing respondents as SMI (i.e., ‘0’) or heterosexual (i.e., ‘1’). They were identified as SMI when they reported identifying as gay/lesbian, bisexual or more gay/lesbian than straight, *or* when they reported having at least as many same-sex sexual fantasies as opposite-sex fantasies, *or* when they reported to feel at least as often attracted to the same sex as to the opposite sex, *or* when they reported having had at least as many same-sex sexual contacts as opposite-sex sexual contacts.

**Indicators of sexual health.** We explored seven indicators of sexual health to gain insight into the differences or similarities in sexual health outcomes between SMIs drawn from the population sample and those from the Internet sample. As some indicators of sexual health are only relevant to sexually experienced respondents (i.e., respondents who have had sex), for five out of seven variables the analyses were restricted to sexually experienced SMIs from both samples. We first discuss these five indicators (number of sexual partners, age at first sexual experience, frequency of sexual activity, perceived satisfaction, and importance of sex), followed by two indicators concerning experiences of sexual abuse.

The number of sexual partners was measured by means of a question that required a categorical open numeric answer. We asked sexually experienced respondents: “*Throughout your life, how many people have you had sex with?*” We defined sex as “*all ways of making love involving genital contact. We do not only refer to sexual intercourse*”. Because of the skewed distribution, we recoded this variable to six categories (one partner, 2 to 3 partners, 4 to 5 partners, 6 to 9 partners, 10 to 19 partners, or 20 or more partners).



Secondly, the age at first sexual experience or intercourse was probed with a categorical open numeric answer. We asked respondents: “*How old were you when you had your first sexual experience or the first time you had sexual intercourse?*” As a third indicator of sexual health, frequency of sexual activity was measured with an open numeric answer category. We asked respondents: “*In the past two weeks, how often did you have sex?*” Again, we defined sex as detailed above. Because of a skewed distribution, we recoded this variable to four categories (less than 0.5 times, 0.5 to 1.99 times, 2 to 4 times or more than 4 times in two weeks). The fourth and fifth indicators – perceived satisfaction and importance of sex – were each measured on a 5-point Likert scale (from very dissatisfied/unimportant to very satisfied/important).

Next, we presented two questions that assessed experiences of sexual abuse. We modified items from a large-scale population-based study on sexual health in the Netherlands (Bakker et al. 2009). Belgian law does not permit us to assess sexual abuse in minors without reporting any incidents to the authorities, meaning we would have had to violate the anonymity of respondents. Therefore, we choose to present these items only to respondents aged at least 18 years at the time of the survey. These final two indicators summarize information from the following items: “*Has someone forced you to masturbate against your will?*” (yes/no), “*Has someone forced you to undergo or perform oral sex against your will?*”(yes/no), “*Has someone tried to rape you?*” (yes/no) and “*Has someone raped you?*” (yes/no). For all four items, the question was split up in order to separate experiences before the age of 18 from experiences after the age of 18. Respondents who reported ‘yes’ on at least one of the aforementioned items were considered to have experienced sexual abuse. With this information, we created two dichotomous variables categorizing respondents as those who had or had not experienced sexual abuse, before or after the age of 18.

### 3. Results

Firstly, we compare the sociodemographic composition of both samples through binomial regression analyses. Secondly, as it is important to include WSW (i.e., women who have sex with women but who do not identify as lesbian or bisexual) in the sample, we focus on differences in the relationships between dimensions of sexual orientation. We will compare the proportion of respondents who reported ‘same-sex’ on one or several dimensions (i.e., sexual desire, behavior, or identity) related to sexual orientation. Thirdly, we use the ‘dataset’ (population survey versus Internet survey) as a predictor for sexual health indicators. These results can be used to deduce whether scores on sexual health indicators differ significantly between female SMIs in the population sample and the Internet sample. Finally, we explore the differences between the datasets found in the third step, this time controlling for age, level of education, occupational category, subjective evaluation of income, family situation, and existence of a current romantic relationship (see Table 3 for descriptive statistics).

#### 3.1. Differences in Sociodemographic Composition of the Samples

A broad range of respondent characteristics, available in both surveys, were included in a stepwise binomial logistic regression with sample membership as the dichotomous

Table 3. Effect of age, income, educational level, occupational status, family composition, and partner status on sample membership of SMI's (population sample versus Internet sample) (including univariate descriptives)

	Sample membership (Internet sample versus population sample)		Descriptive statistics (%)	
	B (SE)		Internet sample (N = 925)	Population sample (N = 90)
Constant	2.28 (1.01)		%	%
Age			4.2	5.6
14-17	-.83 (1.03)		55.4	26.7
18-29	1.72 (.64)**		20.1	13.3
30-39	2.39 (.70)**		12.8	20.0
40-49	1.85 (.69)**		6.6	24.4
50-64	.77 (.61)		1.0	10.0
Ref.: 65-86			2.7	18.9
Educational level			10.1	22.2
No/lower education	-3.03 (.49)***		18.8	30.0
Lower secondary	-1.66 (.38)***		68.5	28.9
Higher secondary	-1.48 (.33)***		10.0	36.8
Ref.: College/Academic			51.0	54.0
Occupational category			39.0	9.0
Inactive	-2.39 (.71)***		16.6	14.3
Paid job	-1.98 (.66)**		47.4	44.0
Ref.: Student			36.0	41.8
Subjective evaluation	1.10 (.44)*			
of income	0.38 (.30)			
(Very) difficult to live comfortably				
Not difficult/not easy to live comfortably				
Ref.: (Very) easy to live comfortably				
Family situation				
Single/living with parents	1.51 (.47)**		61.3	28.6
With partner/no children	1.17 (.42)**		22.5	28.6
With children, youngest <7 years old	.12 (.50)		8.0	17.6

Table 3. Continued

	Sample membership (Internet sample versus population sample)		Descriptive statistics (%)	
	B (SE)	Internet sample (N = 925)	Population sample (N = 90)	
Partner status			8.2	25.3
Ref.: With children, youngest > 6 years old				
In a relationship	-.41 (.40)	65.1	79.1	
Ref.: Not in a relationship		34.9	20.9	
Nagelkerke R <sup>2</sup>	36.7%			
Total N	1,015			

\*p < .05, \*\*p < .01, \*\*\*p < .001.

independent variable (population versus Internet sample). This enabled us to infer which covariates were the most likely explanations of differences in specific survey outcomes (e.g., differences between sexual health indicators). This analysis showed age, educational level, occupational category and income to be the most important correlates (Table 3). However, these four indicators only explain about one third of the variance, indicating that a lot of the explanatory factors and distinct features of self-selection and different sampling techniques remain to be explored.

### 3.2. Differences in Relationships Between Dimensions of Sexual Orientation

Because the three dimensions of sexual orientation (same-sex desire, behavior, and identity) may or may not overlap, and because this is relevant from a sexual health perspective, it is important to explore differences in the number of respondents within these intersections in both samples. Therefore, we tested whether scores on each of these combined dimensions significantly differed between the population and the Internet sample (see Figures 1 and 2).

Due to the relatively low number of female SMIs in the population sample ( $N = 90$ ), we were unable to compare every score on a specific dimension (or intersection between dimensions) between the samples. In the population sample, 60.3% of the SMI respondents reported same-sex desire and behavior without identifying as lesbian or bisexual (LB), compared to 8.2% of the respondents in the Internet sample ( $p < .001$ ). In the Internet sample 59% of the respondents reported same-sex desire, behavior, and identified as LB, compared to 23.9% of the respondents in the population sample ( $p < .001$ ). In the Internet sample we found a significant proportion (31.4%,  $N = 290$ ) of respondents who reported same-sex desire and identified as LB without reporting

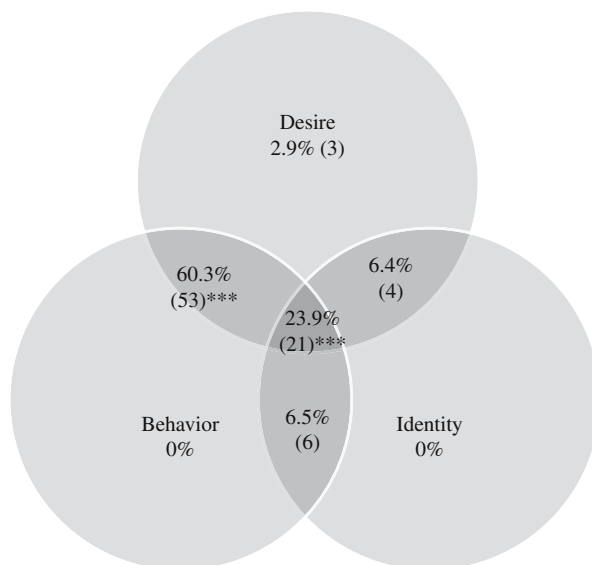


Fig. 1. Relationships between dimensions of sexual orientation in the population sample ( $N = 88$ ) (tested differences between the population sample and the Internet sample). \*\*\* $P < .001$

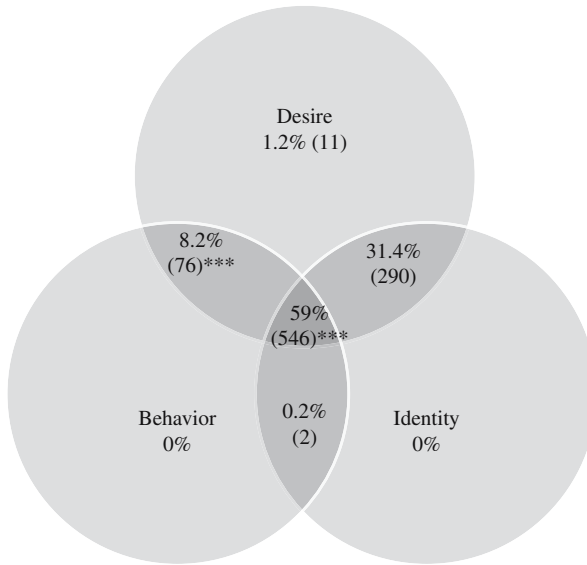


Fig. 2. Relationships between dimensions of sexual orientation in the Internet sample (N = 925) (tested differences between the population sample and the Internet sample). \*\*\*P < .001

same-sex behavior. This was not the case in the population sample (6.4%, N = 6). We conducted a binary logistic regression with age, dataset, and educational level as independent variables, and sexual orientation (1 = same-sex desire, identifying as LB but without same-sex behavior, 0 = all other categories) as a dichotomous dependent variable (see Table 4). Respondents in the age category 18 to 29 years old reported more same-sex desire while identifying as LB but without same-sex behavior than respondents

Table 4. Predictors of same-sex desire and identity without same-sex behavior

Variable	B(SE)
Constant	- 3.43 (.55)***
<b>Dataset</b>	
Population survey	- 1.71 (.47)***
Ref.: Internet survey	
<b>Educational level</b>	
No/lower education	.12 (.42)
Lower secondary	- .35 (.26)
Higher secondary	.27 (.18)
Ref.: College/Academic	
<b>Age</b>	
18-29	1.22 (.33)***
30-39	.44 (.37)
40-49	.39 (.39)
Ref: 50-86	
Nagelkerke R <sup>2</sup>	.10***
N	1,026

\*p < .05, \*\*p < .01, \*\*\*p < .001.

Table 5. Predictors of same-sex desire and behavior without identity

Variable	B(SE)
Constant	1.86 (.43)***
<b>Dataset</b>	
Population survey	3.38 (.36)***
Ref.: Internet survey	
<b>Educational level</b>	
None/primary education	− 1.41 (.62)*
Lower secondary	− .68 (.39)
Higher secondary	− .47 (.32)
Ref.: College/Academic	
<b>Age</b>	
18–29	− .03 (.35)
30–39	− 1.38 (.46)**
40–49	− .71 (.43)
Ref: 50–86	
Nagelkerke $R^2$	.31***
$N$	691

\* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ .

in the age category 50 to 86 years old. This at least partly explains the larger number of LBs with same-sex desire but without same-sex behavior in the Internet sample, since this survey included a larger group of young respondents.

The descriptive statistics mentioned above might be misleading due to different sample characteristics related to the educational level and age of respondents. To control for these different sample characteristics, we conducted a binary logistic regression (see Table 5). We created a dichotomous variable for the dependent variable. Scores of ‘1’ and ‘0’ respectively referred to women who reported same-sex desire and behavior *without* identifying as LB, and to women who report same-sex desire, behavior, *and* identified as LB. All scores related to the other categories were treated as missing values. This analysis showed that, when educational level and age were controlled for, respondents in the Internet sample were still less likely to report same-sex desire and behavior *without* identifying as LB than respondents in the population sample.

### 3.3. The Dataset As a Predictor for Sexual Health Indicators

To compare scores on sexual health indicators, we merged the two datasets. We found that the datasets only differed on one of the seven sexual health indicators (see Table 6). Furthermore, stepwise multivariate analyses (with dataset included in a first model, and dataset and sociodemographic variables included in a second model) showed that for five of the seven sexual health indicators (sex frequency, sexual satisfaction, importance of sex and experience with sexual abuse before or after the age of 18), there were no differences in scores between the datasets, independent of whether or not we added sociodemographic variables to the model (these tables have not been included in this article).

Regarding ‘Number of (lifelong) sexual partners’ (Table 7), we found that controlling for age, and especially for educational level, alters the initial difference found between

Table 6. Differences in survey measures (Internet survey versus population survey) for seven sexual health indicators

	Internet survey 870	Population survey 84
<i>N</i> (only sexually experienced respondents)		
Number of lifelong sex partners		
1	11.2%***	25.6%***
2-3	23.1%	22.0%
4-5	20.7%	17.1%
6-9	21.0%	18.3%
10-19	16.0%	13.4%
20+	8.3%	3.7%
Age at first sex/intercourse		
<i>M</i> ( <i>SE</i> )	17.78 (3.35)	17.19 (2.89)
Sex frequency (over two weeks)		
<0.5 times	29,3%	25,3%
<2 times	23,1%	24,1%
2-4	22,4%	21,7%
4+	25,2%	28,9%
Sexual satisfaction (five-point scale)		
<i>M</i> ( <i>SE</i> )	3.71 (1.08)	3.60 (1.17)
Importance of sex (five-point scale)		
<i>M</i> ( <i>SE</i> )	3.73 (.89)	3.55 (1.15)
<i>N</i> (all respondents)	925	90
Experience with sexual abuse before 18		
‘yes’	20.8%	22.2%
Experience with sexual abuse after 18		
‘yes’	8.9%	10.0%

\* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$ 

datasets. More specifically, higher numbers of sexual partners were reported in the Internet survey compared to the population survey. Since older respondents and respondents with a lower level of education are more accurately represented in the population sample and these groups report higher numbers of sexual partners, differences in sample composition might suppress differences in scores on this particular variable. Regarding ‘Age of first sexual experience/intercourse’ (Table 8) we found that when we did not control for sociodemographic variables, no differences were found between the datasets. However, when we did control for these factors, female SMIs in the Internet sample reported having had their first sexual experience/intercourse at an older age than female SMIs in the population sample. As respondents in the Internet sample were significantly younger than those in the population sample, this might reveal these respondents as a group who had their first sexual experience at a relatively older age. Concerning the latter findings, we should be aware that we cannot distinguish sampling or recruitment effects from mode effects.

#### 4. Conclusions and Discussion

In this study we focused on the differences between two survey methods to study female SMIs. This group was studied both as a subgroup of a population survey on sexual health,

Table 7. Effects of 'dataset' and sociodemographic covariates on the number of lifelong sex partners. Results of the multinomial logistic regression analysis

		Number of lifelong sexual partners <sup>b</sup>			
		2-3	4-5	6-9	10+
<b>Model 1</b>					
	<i>B (SE)</i>		<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>
Constant	.72	.61	.63	.78	
<b>Dataset</b>					
Population survey					
Ref.: Internet survey	-.88 (.35)*	-.99 (.37)**	-.97 (.36)**	-1.15 (.36)**	
Nagelkerke <i>R</i> <sup>2</sup>					1.3%
<i>N</i>					951
<b>Model 2<sup>a</sup></b>					
	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>	<i>B (SE)</i>
Constant	.38	.80	1.14	1.47	
<b>Dataset</b>					
Population survey					
Ref.: Internet survey	-1.11 (.41)*	-1.51 (.43)***	-1.63 (.44)***	-1.90 (.43)***	
<b>Educational level</b>					
None – lower secondary	.92 (.40)*	.55 (.42)	1.15 (.41)**	.70 (.42)	
Higher secondary	.13 (.32)	.51 (.32)	.15 (.33)	.28 (.32)	
Ref.: College/Academic					
<b>Age</b>					
18-29	.53 (.51)	-.10 (.48)	-.26 (.49)	-.81 (.47)	
30-39	.27 (.58)	.01 (.54)	.01 (.54)	.24 (.51)	
40-49	1.50 (.68)*	1.13 (.66)	.90 (.66)	.91 (.64)	
Ref.: 50-86					
<b>Occupational category</b>					
Inactive	-.43 (.31)	-.64 (.32)	-.98 (.32)**	-1.16 (.34)**	
Paid job	-.50 (.45)	-.11 (.44)	-.95 (.46)*	-.32 (.43)	
Ref.: Student					
Nagelkerke <i>R</i> <sup>2</sup>					13.7%
<i>N</i>					911

\**p* < .05, \*\**p* < .01, \*\*\**p* < .001.

<sup>a</sup> Model restricted to independents with significant effects on the dependent variable. <sup>b</sup>The reference category is one sexual partner.



Table 8. Predictors of age of first sexual experience/intercourse

Age at first sex/intercourse	
Model 1	<i>B (SE)</i>
Constant	17.78
<b>Dataset</b>	
Population survey (Ref.: Internet survey)	− .42 (.43)
Model 2 <sup>a</sup>	<i>B (SE)</i>
Constant	19.79
<b>Dataset</b>	
Population survey Ref.: Internet survey	− 1.24 (.44)**
<b>Educational level</b>	
None/primary education	− .74 (.59)
Lower secondary	− 1.49 (.36)***
Higher secondary Ref.: College/Academic	− .20 (.28)
<b>Age</b>	
18–29	− 2.22 (.44)***
30–39	− 1.03 (.46)*
40–49 Ref: 50–86	− .62 (.48)
<b>Occupational category</b>	
Inactive	− .72 (.28)*
Paid job Ref.: Student	.23 (.39)
<i>R</i> <sup>2</sup>	11,6%
<i>N</i>	863

\**p* < .05, \*\**p* < .01, \*\*\**p* < .001.

<sup>a</sup> Model restricted to independents with significant effects on the dependent variable

drawing from a population-based probability sample, and as a subgroup that was directly targeted and invited to participate in an Internet survey on sexual health. These surveys included different sampling and recruitment strategies while employing quasi-identical questionnaires and modes of administration.

Our analyses showed that, in terms of sociodemographic compositions, the samples differed in terms of age, educational level, occupation, and income. Consequently, these are the most plausible factors for explaining differences in survey outcomes, or at least for explaining differences that can be linked to the different sampling and recruitment techniques. Put differently, SMIs who are nonrandomly recruited through the Internet are a specific group. Our research, and that of many others (see e.g., Mathy et al. 2002; Claeys and Spee 2005), shows that these respondents often have high levels of education and are younger. However, efforts made to evaluate SMI populations are always imperfect since the population is by nature hard-to-reach, and the solutions to address this are lacking.

Of course, this does not mean that all efforts should be stopped. On the contrary, the distinction between self-identifying SMIs and WSW has become especially relevant from a sexual health perspective (Loosier and Dittus 2010; McCabe et al. 2012; Mercer et al. 2007).

We found that our Internet sample generated a smaller proportion of WSW compared to our population sample. This is not very surprising, as we can expect that a targeted recruitment strategy will at least partly appeal more to SMIs who identify as gay, lesbian, or bisexual (see e.g., Fernee and Keuzenkamp 2011). However, in absolute numbers there were more WSW in the Internet sample ( $N = 76$ ) than in the population sample ( $N = 53$ ). These WSW appeared to be younger and more likely to still be attending school than the WSW in the population sample (results not included in this article). This subgroup is therefore probably not representative of the SMI population.

Looking at specific sexual health measures, our comparison showed no differences on the scores for six of the seven variables. The datasets did differ on the number of reported lifelong sex partners or, when sociodemographic variables were controlled for, on the reported age of first sexual experience. Due to the paucity of literature on the topic of Internet use and sexual health in female SMIs, these findings are difficult to interpret. Moreover, it is hard to really distinguish sampling or recruitment (including self-selection) effects from mode effects. However, we would like to propose several solutions to address these issues, such as combining sampling methods via propensity score matching (PSM), using fully randomized studies to learn about interviewer effects, and using calibration methods.

Firstly, combining the two sampling methods through PSM, based on the most distinguishing features (e.g., age and socioeconomic status), could be one part of the solution. Including additional questions in both questionnaires or assessing other survey features could also make PSM possible. For instance, within the population survey questionnaire, questions about Internet access, the frequency and modalities of Internet use and an assessment of the respondents' willingness and the likelihood of participation in an Internet survey could generate important information about the mechanisms of self-selection in online research, sample overlap, and (non)response in Internet surveys.

Secondly, to compare CASI-on-the-web versus CASI in the presence of an interviewer, a fully randomized study could possibly help to isolate potential interviewer effects. For instance, we could construct a study where a randomly selected part of the probability sample (drawn from the National Register and using the same contact procedures) could be assigned to the Internet survey and the other part could be assigned to the CASI-with-interviewer mode. Another possibility would be that a targeted sampling design could be followed by one of these two modes of administration (CASI-on-the-web versus CASI-with-interviewer).

We also recommend further study of how different recruitment channels lead to different respondent profiles. As diversity is more important than representativeness in Internet samples, research needs to explore how different channels (Facebook, mailings, social networks, and so on) contribute to sample diversity. As our elaborate methodological description shows, we made a great effort to increase the diversity of our sample by using a variety of recruitment channels. This diversity is important as SMIs with high levels of involvement in the gay community have different psychological and risk profiles than those not involved (Ramirez-Valles 2002). Participation of SMIs within LGB venues could have characteristics that correlate with the main variables of interest in

the study (Meyer and Wilson 2009). Social network media might offer new opportunities to recruit and interact with potential research participants, especially in times where response rates are dropping (Groves 2006). Hard-to-reach populations such as SMIs might be especially accessible in these new virtual venues as they make it easier for stigmatized individuals to share delicate information. Of course, although this diversity in online samples is important and allows the generation of data on specific subpopulations (e.g., bisexuals, elderly SMIs), it does still not generate representative samples.

Thirdly, in our study the low numbers of male SMIs in the population sample made reliable deductions impossible for this group. Even when larger samples are available, a comparison within an SMI population (e.g., WSW versus lesbian and bisexual women) is often difficult, due to lack of power. Therefore, one could combine the strengths of both survey methods, including unique sampling and recruitment strategies and different modes of administration. Propensity matching could be used to combine the strengths of both sampling methods. Another possibility would be to weigh survey measures coming from an Internet survey using figures acquired from a reliable, highly representative population sample that includes a sufficient number of SMIs (i.e., adopting a calibration approach). Preferably, these weightings should not only incorporate a range of relevant socio-demographic variables, but also diverse indicators and dimensions of sexual orientation and other relevant features such as items on Internet access and use. One high-quality 'baseline' population study could be sufficient to supply the necessary data for many future online surveys on a variety of topics in SMI or other minority populations (e.g., ethnic minority populations). However, this would acquire large population-based samples. Estimating that 5% to 10% of a population belongs to a sexual minority, at least 4,000 respondents would be needed to obtain the minimum of 100-150 female and 100-150 male SMIs required to enable the extraction of reliable sampling sizes. This is not always possible in small communities or when limited resources are available. Also, further research has yet to prove that these adjusted and weighted samples are reliable.

To conclude, traditional methods that generate representative samples offer opportunities to gather data on the 'real' hidden SMI population (i.e., those individuals who are not inclined to participate in Web surveys). On the other hand, targeted sampling when combined with an Internet survey has some advantages, such as lower cost and the ability to generate large samples, especially in relatively small communities. The application of Internet surveys, which have often been used to reach SMIs (Dewaele et al. 2011; Aerts et al. 2012; Grov et al. 2006), could be significantly improved by combining sampling methods via propensity score matching (PSM), by further research that reveals the effect of the presence of an interviewer, and by using calibration methods. However, to gather reliable data on SMIs and to avoid self-selection bias, a population-based probability sample (of at least 4,000 respondents) remains the gold standard.

## 5. References

- Aerts, S., Van Houtte, M., Dewaele, A., Cox, N., and Vincke, J. (2012). Sense of Belonging in Secondary Schools: A Survey of LGB and Heterosexual Students in Flanders. *Journal of Homosexuality*, 59, 90–113. DOI: <http://www.dx.doi.org/10.1080/00918369.2012.638548>

- Bajos, N. and Bozon, M. (2008). *Enquête sur la sexualité en France*. Paris: Éditions La Découverte.
- Bakker, F., de Graaf, H., de Haas, S., Kedde, H., Kruijer, H., and Wijzen, C. (2009). *Seksuele gezondheid in Nederland 2009*. Utrecht: Ruthers Nisso Groep. Available at: <http://www.rutgerswvf.nl/sites/default/files/Seksuele%20Gezondheid%20in%20Nederland%202009.pdf> (accessed March 30, 2014)
- Bauermeister, J., Pingel, E., Zimmerman, M., Couper, M., Carballo-Diéguez, A., and Strecher, V.J. (2012). Data Quality in Web-Based HIV/AIDS Research: Handling Invalid and Suspicious data. *Field Methods*, 24, 272–291. DOI: <http://www.dx.doi.org/10.1177/1525822X12443097>
- Best, S.J. and Krueger, B.S. (2004). *Internet Data Collection*. Thousand Oaks, CA: Sage.
- Buysse, A., Caen, M., Dewaele, A., Enzlin, P., Lievens, J., T'Sjoen, G., Van Houtte, M., and Vermeersch, H (2013). *Seksuele Gezondheid in Vlaanderen*. Ghent: Academia Press.
- Claeys, L. and Spee, S. (2005). *Een Virtuele Illusie of Reële Kansen? – Gender in de Netwerkmatschappij*. Antwerp: Steunpunt Gelijkekansenbeleid, University of Antwerp/University of Hasselt. Available at: <http://www.steunpuntgelijkekansen.be/wp-content/uploads/18.-Een-virtuele-illusie-of-reele-kansen-L.-Claeys.pdf> (accessed March 30, 2014)
- Couper, M.P. (2000). Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly*, 64, 464–494. DOI: <http://dx.doi.org/10.1086/318641>
- Couper, M.P. (2008). *Designing Effective Web Surveys*. Cambridge: Cambridge University Press.
- de Leeuw, E.D. (2005). To Mix or not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, 233–255.
- Denscombe, M. (2006). Web-Based Questionnaires and the Mode Effect – An Evaluation Based on Completion Rates and Data Contents of Near-Identical Questionnaires Delivered in Different Modes. *Social Science Computer Review*, 24, 246–254. DOI: <http://www.dx.doi.org/10.1177/0894439305284522>
- Dewaele, A., Cox, N., van den Berghe, W., and Vincke, J. (2011). Families of Choice? Exploring the Supportive Networks of Lesbians, Gay Men and Bisexuals. *Journal of Applied Social Psychology*, 41, 312–331. DOI: <http://www.dx.doi.org/10.1111/j.1559-1816.2010.00715.x>
- Dillman, D.A. (1978). *Mail and Telephone Surveys: The Total Design Method*. New York: Wiley.
- Dillman, D.A. (2000). *Mail and Internet Surveys: The Tailored Design Method*. London: Wiley.
- Edelman, M. (1993). Understanding the Gay and Lesbian Vote in 1992. *Public Perspective*, 4, 32–33.
- Evans, J.R. and Mathur, A. (2005). The Value of Online Surveys. *Internet Research*, 15, 195–219.
- Ferneer, H. and Keuzenkamp, S. (2011). Selectiviteit van de Roze Vragenlijst. Een Vergelijking met Paneldata (paper presented at Sociology Day on May 26, 2011, Den Haque). Available at: <http://www.google.be/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0CC4QFjAA&url=http%3A%2F%2Fwww.scp.nl%2F>

- sresource%3Ftype%3Dpdf%26objectid%3Ddefault%3A29290%26versionid%3D%26subobjectname%3D&ei=9eg3U4DgMsWqhAeQg4CwAw&usg=AFQjCNG6pS0WGCZ1\_Hf4RaSfq8ILoHQYAQ&bvm=bv.63808443,d.ZG4. (accessed March 30, 2014)
- Gronow, J., Haavio-Mannila, E., Kivinen, M., Lonkila, M., and Rotkirch, A. (1997). Cultural Inertia and Social Change in Russia. Helsinki: University of Helsinki, Department of Sociology.
- Grov, C., Bimbi, D.S., Nanín, J.E., and Parsons, J.T. (2006). Ethnicity, Gender, and Generational Factors Associated with the Coming-Out Process Among Gay, Lesbian, and Bisexual Individuals. *Journal of Sex Research*, 43, 115–121.
- Groves, R.M. (2006). Non-Response Rates and Non-Response Bias in Household Surveys. *Public Opinion Quarterly*, 70, 646–675. DOI: <http://www.dx.doi.org/10.1093/poq/nfl033>
- Haavio-Mannila, E. and Kontula, O. (2001). *Seksin Trendit Meillä ja Naapureissa*. Helsinki: WSOY.
- Heerwegh, D. (2001). *Surveyonderzoek Middels het Internet: Een exploratie van het Terrein*. Leuven: Katholieke Universiteit Leuven, Departement Sociologie, Afdeling voor dataverzameling en analyse. Available at: <https://perswww.kuleuven.be/~u0034437/public/Files/Survey-onderzoek%20middels%20het%20Internet.pdf> (accessed March 30, 2014)
- Kerker, B.D., Mostashari, F., and Thorpe, L. (2006). Health Care Access and Utilization Among Women Who Have Sex With Women: Sexual Behavior and Identity. *Journal of Urban Health*, 83, 970–979. DOI: <http://www.dx.doi.org/10.1007/s11524-006-9096-8>
- Koch, S.N. and Emrey, J.A. (2001). The Internet and Opinion Measurement: Surveying Marginalized Populations. *Social Science Quarterly*, 82, 131–138. DOI: <http://www.dx.doi.org/10.1111/0038-4941.00012>
- Kreuter, F., Presser, S., and Tourangeau, R. (2008). Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly*, 72, 847–865. DOI: <http://www.dx.doi.org/10.1093/poq/nfn063>
- Laumann, E., Gagnon, J.H., Michael, R.T., and Michaels, S. (1994). *The Social Organization of Sexuality: Sexual Practices in the United States*. Chicago: University of Chicago Press.
- Lievens, J. and Waeye, H. (2009). *Participatie in Vlaanderen. Basisgegevens van de participatiesurvey*. Leuven: Acco.
- Loosier, P.S. and Dittus, P.J. (2010). Group Differences in Risk Across Three Domains Using an Expanded Measure of Sexual Orientation. *Journal of Primary Prevention*, 31, 261–272. DOI: <http://www.dx.doi.org/10.1007/s10935-010-0228-2>
- Mathy, R.M., Schillace, M., Coleman, S.M., and Berquist, B.E. (2002). Methodological Rigor With Internet Samples: New Ways to Reach Underrepresented Populations. *CyberPsychology and Behavior*, 5, 253–266.
- McCabe, S.E., Hughes, T.L., Bostwick, W., Morales, M., and Boyd, C.J. (2012). Measurement of Sexual Identity in Surveys: Implications for Substance Abuse Research. *Archives of Sexual Behavior*, 41, 649–657. DOI: <http://www.dx.doi.org/10.1007/s10508-011-9768-7>

- Mercer, C.H., Bailey, J.V., Johnson, A.M., Erens, B., Wellings, K., Fenton, K.A., and Copas, A.J. (2007). Women Who Report Having Sex With Women: British National Probability Data on Prevalence, Sexual Behaviors, and Health Outcomes. *American Journal of Public Health*, 97, 1126–1133. DOI: <http://www.dx.doi.org/10.2105/AJPH.2006.086439>
- Meyer, I.H. and Wilson, P.A. (2009). Sampling Lesbian, Gay, and Bisexual Populations. *Journal of Counseling Psychology*, 56, 23–31. DOI: <http://www.dx.doi.org/10.1037/a0014587>
- Ramirez-Valles, J. (2002). The Protective Effects of Community Involvement for HIV Risk Behavior: A Conceptual Framework. *Health Education Research*, 17, 389–403.
- Raymond, H.F., Rebchook, G., Curotto, A., Vaudrey, J., Amsden, M., Levine, D., and McFarland, W. (2010). Comparing Internet-Based and Venue-Based Methods to Sample MSM in the San Francisco Bay Area. *Aids and Behavior*, 14, 218–224. DOI: <http://www.dx.doi.org/10.1007/s10461-009-9521-6>
- Roberts, A.L., Austin, S.B., Corliss, H.L., Vandermorris, A.K., and Koenen, K.C. (2010). Pervasive Trauma Exposure Among US Sexual Orientation Minority Adults and Risk of Posttraumatic Stress Disorder. *American Journal of Public Health*, 100, 2433–2441. DOI: <http://www.dx.doi.org/10.2105/AJPH.2009.168971>
- Rothblum, E. (2007). From Science Fiction to Computer-Generated Technology: Sampling Lesbian, Gay, and Bisexual Individuals. In *The Health of Sexual Minorities: Public Health Perspectives on Lesbian, Gay, Bisexual and Transgender Populations*, I.H. Meyer and M.E. Northridge (eds). New York: Springer, 441–454.
- Schillewaert, N. and Meulemeester, P. (2005). Comparing Response Distributions of Offline and Online Data Collection Methods. *International Journal of Market Research*, 47, 163–178.
- Schonlau, M., van Soest, A., Kapteyn, A., and Couper, M. (2009). Selection Bias in Web-Surveys and the Use of Propensity Scores. *Sociological Methods and Research*, 37, 291–318.
- Schwarz, S., Spindler, H., Scheer, S., Valleroy, L., and Lansky, A. (2007). Assessing Representativeness of Sampling Methods for Reaching Men Who Have Sex With Men: A Direct Comparison of Results Obtained from Convenience and Probability Samples. *AIDS and Behavior*, 11, 596–602.
- Silenzio, V.M.B., Duberstein, P.R., Tang, W., Lu, N., Tu, X., and Homan, C.M. (2009). Connecting the Invisible Dots: Reaching Lesbian, Gay, and Bisexual Adolescents and Young Adults at Risk for Suicide Through Online Social Networks. *Social Science and Medicine*, 69, 469–474. DOI: <http://www.dx.doi.org/10.1016/j.socscimed.2009.05.029>
- Tourangeau, R., Couper, M.P., and Steiger, D.M. (2003). Humanizing Self-Administered Surveys: Experiments on Social Presence in Web and IVR Surveys: Experiments on Social Presence in Web and IVR Surveys. *Computers in Human Behavior*, 19, 1–24. DOI: [http://www.dx.doi.org/10.1016/S0747-5632\(02\)00032-8](http://www.dx.doi.org/10.1016/S0747-5632(02)00032-8)
- Van Kesteren, N.M.C., Hospers, H., and Kok, G. (2007). Sexual Risk Behavior Among HIV-Positive Men Who Have Sex With Men: A Literature Review. *Patient Education and Counseling*, 65, 5–20.
- Vincke, J. and Bleys, R. (2003). *Vitale Vragen 2001. Eindrapport*. Antwerp: Sensoa.

- Vincke, J. and Stevens, P. (1999). Een Beleidsgerichte Algemene Survey van Vlaamse Homoseksuele Mannen en Vrouwen—Basisrapport. Brussels: Ministerie van de Vlaamse Gemeenschap, Cel Gelijke Kansen, Universiteit Gent. Available at: [http://www.psw.ugent.be/cms\\_global/uploads/publicaties/personal/eindrapport.pdf](http://www.psw.ugent.be/cms_global/uploads/publicaties/personal/eindrapport.pdf). (accessed March 30, 2014).
- Vincke, J. and van Heeringen, K. (2004). Summer Holiday Camps for Gay and Lesbian Young Adults: An Evaluation of Their Impact on Social Support and Mental Well-Being. *Journal of Homosexuality*, 47, 33–46. DOI: [http://www.dx.doi.org/10.1300/J082v47n02\\_02](http://www.dx.doi.org/10.1300/J082v47n02_02)
- Wright, K.B. (2005). Researching Internet-Based Populations: Advantages and Disadvantages of Online Survey Research, Online Questionnaire Authoring Software Packages, and Web Survey Services. *Journal of Computer-Mediated Communication*, 10, article 11. Available at: [http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00259.x/full?utm\\_source](http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2005.tb00259.x/full?utm_source). (accessed March 30, 2014).

Received February 2013

Revised February 2014

Accepted March 2014



## A City-Based Design That Attempts to Improve National Representativeness of Asians

Steven Pedlow<sup>1</sup>

This article describes a case study on the potential of using smaller geographical units in an area probability design, and reports the challenges of collecting a nationally representative sample for this hard-to-reach population. The Census Integrated Communications Program Evaluation (CICPE) was designed to evaluate the promotional campaign's effect on Decennial Census participation for six race/ethnicity groups of interest. A nationally representative Core sample was designed to collect interviews for Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites. However, it was impractical to include the rarer Asian, American Indian and Alaska Native (AIAN), and Native Hawaiian and Other Pacific Islander (NHOPI) populations in the Core design. For the Asian sample, we designed a separate area probability sample.

Traditional area probability sampling designs use counties or metropolitan areas as first-stage units, but smaller geographical units can better target hard-to-reach populations. The CICPE Asian sample used cities as the first-stage units.

*Key words:* Area probability sampling; CICPE; Decennial Census; Hard-to-Reach.

### 1. Introduction

Every ten years, the U.S. Census Bureau attempts to count every American through the Decennial Census. For the 2000 Decennial Census, the Census Bureau responded to declining mail participation in the 1990 Decennial Census (which had increased the costs of in-person enumeration visits) with a greatly expanded outreach and promotion campaign called the “Partnership and Marketing Program” (PMP). NORC at the University of Chicago was contracted to conduct the 2000 Partnership and Marketing Program Evaluation (PMPE), an independent evaluation of the PMP which included a series of three face-to-face in-person surveys: before the PMP (autumn 1999); during the PMP, which was also the time period that included the mailing of Census forms to housing units and the beginning of data collection by mail (winter/spring 2000); and after the PMP during the nonresponse follow-up operation of the 2000 Decennial Census (summer 2000), which involved in-person visits to housing units that did not mail back their Census questionnaire. The Census Bureau was sensitive to differential impact of the PMP by race/ethnicity, and so the sample was equally divided among six different race/ethnicity groups, including Asians.

In 2010, the Census Bureau took the lessons learned from 2000 and designed an Integrated Communications Program (ICP) to encourage mail participation in the 2010

<sup>1</sup> NORC/University of Chicago – Statistics/Methodology, 55 E Monroe St., Suite 2000, Chicago, IL 60603, U.S.A. Email: [pedlow-steven@norc.uchicago.edu](mailto:pedlow-steven@norc.uchicago.edu)



Decennial Census. NORC at the University of Chicago again conducted an independent evaluation of the ICP called the “Census Integrated Communications Program Evaluation” (CICPE), which again utilized three waves of in-person face-to-face interviewing that matched up to before, during, and after the Census ICP. The same six race/ethnicity groups from the 2000 evaluation were again of interest. In 2000 and 2010, NORC’s sample designs included a Core sample (See Section 2 for details) that was a nationally representative area probability sample to collect interviews from Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites. However, supplemental samples were necessary for the remaining three race/ethnicity groups: Asians (the focus of this article), American Indian and Alaska Natives (AIAN), and Native Hawaiian and Other Pacific Islanders (NHOPI).

For the 2000 PMPE, NORC’s Asian Supplemental Sample collected all interviews from the five U.S. cities with the largest Asian populations. For the 2010 CICPE, we wanted a more nationally representative sample of Asians. This article describes how we used a city-based area probability sample to greatly increase the coverage of our Asian sample. The tradeoff between noncoverage error and screening cost has already been widely recognized and discussed in many books and papers, but it seems that the screening costs for Asians (4.2 percent of the national population) and other hard-to-reach groups as well as the difficulty of targeting Asians (and other hard-to-reach groups) at larger geographies has prevented such groups from being studied with nationally representative studies.

Prior to the CICPE, no independent surveys had attempted a nationally representative sample of Asians. The only two surveys that have attempted to be nationally representative for Asians previous to the CICPE were embedded within much larger studies that allow a minority of a large number of second-stage units to oversample Asians. The National Health and Nutrition Examination Survey (NHANES) started to oversample Asians in 2011 to create nationally representative estimates, but it does this by oversampling second-stage units with high Asian eligibility rates within a survey that completes 5,000 interviews per year. The second-stage units with high Asian eligibility rates are in areas where many other second-stage units with low Asian eligibility rates have been selected. The latest NHANES sample documentation (Curtin et al. 2013) describes that 1,440 second-stage units were used from 2007-2010 (sample documentation for 2011-2014 is not yet available). Only a small percentage of the second-stage units in 2011-2014 will produce Asian respondents, but the large size of the overall study still allows a sizable and representative sample of Asians. The second nationally representative survey of Asians prior to CICPE is the National Latino and Asian American Study (NLAAS), whose sample design was completely integrated with the National Comorbidity Survey Replication (NCS-R) national sample design by selecting 474 pairs of segments in the same areas: one to be nationally representative and one to oversample Asians or Latinos (Heeringa et al. 2004). The Collaborative Psychiatric Epidemiology Surveys website ([http://www.icpsr.umich.edu/icpsrweb/CPES/about\\_cpes/sample\\_design.jsp#nlaas](http://www.icpsr.umich.edu/icpsrweb/CPES/about_cpes/sample_design.jsp#nlaas)) concedes that without this pairing, survey costs would have been prohibitively high for the NLAAS. The NLAAS used only 317 second-stage units because they expected near-zero interviews in 157 of their second-stage units. Both of these surveys use counties as the smallest first-stage unit for selection. The CICPE shows that nationally representative surveys of Asians and other hard-to-reach groups can be attempted by using first-stage units smaller than counties

without need for a larger survey to support them because these groups can be better targeted with smaller first-stage units.

Looking ahead to the rest of this article, Section 2 will discuss the 2000 PMPE and 2010 CICPE sample designs, including a review of Area Probability Sampling. Section 3 will describe the details of the 2010 CICPE Asian sample design. Section 4 will present some results from fielding the 2010 CICPE Asian sample. Section 5 shows some demographic comparisons between the 2010 CICPE and the national Asian population. Section 6 discusses the limitations to our study, especially with regard to its national representativeness. Finally, Section 7 summarizes this article.

## **2. The 2000 PMPE and 2010 CICPE Sample Designs**

Both the 2000 PMPE and the 2010 CICPE had sample designs that included three waves of data collection. The first wave of data collection took place before the main campaign elements, the second wave took place while the campaign peaked, and the third wave took place after the mail participation deadline to avoid in-person follow-up. Both designs had a sample size that was an idealized 3,000 interviews per wave divided equally among six race/ethnicity groups: Hispanics, non-Hispanic African-Americans, non-Hispanic Whites, American Indian and Alaska Natives (AIAN), Native Hawaiian and Other Pacific Islanders (NHOPI), and Asians. The second wave of the 2010 CICPE design was compressed into a shorter time period, so the sample size was dropped to 2,100.

A nationally representative area probability sample called the “Core” sample was designed to collect interviews for the three largest race/ethnicity groups: Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites. National coverage as part of this Core sample was impractical for the three smaller race/ethnicity groups, so three supplemental samples were necessary. This article focuses on the Asian Supplemental Samples for the 2000 PMPE and 2010 CICPE.

At the time of the 2010 CICPE sample design, the latest source of information on local Asian populations was the 2000 Decennial Census, as the American Community Survey had not yet released small area data. According to the 2000 Census, there were 11,898,828 U.S. Non-Hispanic Asians (Barnes and Bennett 2002), alone or in combination, comprising 4.2 percent of the U.S. population at the time. This figure includes those who marked Asian, regardless of whether other race boxes were marked on the census form; the 2000 Census was the first Decennial Census where race was asked using a “mark all that apply” format.

Since there are many more Hispanics, non-Hispanic African-Americans, and non-Hispanic Whites in the U.S. population than there are Asians, collecting enough Asian interviews through the Core sample would require impractically large screening samples with heavy subsampling of the eligibles for the higher population race/ethnicity groups. In fact, national coverage itself was considered impractical during the planning of the 2000 PMPE, as shown by Wolter et al. (2002), which collected all of its supplemental Asian interviews from the five U.S. cities with the largest Asian populations: New York, Los Angeles, San Francisco, Chicago, and Seattle. At the time of the 2000 Census, 18.8 percent of the U.S. Asian population lived in these five cities. This meant that the 2000 PMPE did not attempt to achieve a nationally representative Asian sample, since the coverage of that sampling frame was only 18.8 percent of the U.S. Asian population. Within these five cities,

6.5 percent of the population was Asian. If the sample was equal probability within these cities, the eligibility rate for the Asian sample would be 6.5 percent. We can also refer to this rate as the screening “hit rate.”

For the 2010 CICPE design, our intention was to improve coverage through a national design. Most national face-to-face surveys in the United States use a multi-stage area probability (AP) sampling design that selects clusters of housing units to interview in order to reduce data-collection costs (Kish 1965). In a multi-stage AP sampling design, a set of large clusters are first selected (first-stage units). Within the selected large clusters, sets of small clusters are selected (second-stage units). Finally, within these selected small clusters, individual housing units are selected for interviewing. The basic objective for a multi-stage AP sampling design is a nationally representative equal-probability sample permitting optimal statistical efficiency. To achieve this, AP samples use probability proportional to size (PPS) sampling in which “larger” areas have a greater selection probability. The measure of size often used for the probabilities is the number of housing units, usually derived from Census data.

Most national area probability samples have first-stage units that are county based, often even using larger metropolitan statistical areas (MSAs) where present (Lohr 2009). However, the key idea in this article is that hard-to-reach populations are better targeted at small geographies. Within the large first-stage unit areas for typical area probability designs, the smaller second-stage areas are often block based, either in terms of blocks, block groups, or entire census tracts. Of course, national samples cannot use first-stage geographies as small as individual blocks; this would require too many clusters that are too spread out to be cost effective. However, NORC has a history of using smaller geographies as first-stage units to better oversample race/ethnicity groups.

The National Longitudinal Survey of Youth 1979 cohort (NLS79) obtained a nationally representative set of interviews with youths who were 14-21 years old while oversampling Hispanic and African-American youths. To do this, NORC split the task into two parts. First, a nationally representative area probability sample was used to get a nationally representative mix of Hispanic, non-Hispanic African-American, and non-Hispanic non-African-American youths. A second area probability sample was used to obtain only Hispanic and African-American youths. This “Supplemental” sample did not differ in its design from the nationally representative “Cross-Sectional” sample, but different areas were selected to better target Hispanic and African-American youths.

The National Longitudinal Survey of Youth 1997 cohort (NLSY97) took this design one step further (Moore et al. 2000). In the “Supplemental Sample” design, all first-stage units were counties rather than entire MSAs in urban areas. Remembering that our goal was to oversample Hispanic and African-American youths, MSAs often have central city counties with a high rate of minority youths surrounded by outlying, more rural areas with lower concentrations of Hispanic and African-American youths. Our strategy allowed us to separate counties with many minority youths from surrounding counties in the same MSA with fewer of them. Counties were still considered too large to target Asians, so we used places defined by the Census as the first-stage clusters in the 2010 CICPE sample design. Places defined by the Census include cities, towns, villages, as well as other “census-designated places,” but we will simplify our language in this article and refer to our first-stage clusters as “cities.”

### 3. The 2010 CICPE Asian Sample

Our first task in selecting a city-based Asian sample was to construct a sampling frame of cities. The most recent data available at the time was still the 2000 Decennial Census. Our first step was to set a threshold of 1,000 Asians for a city to be included, which led to a set of 1,261 U.S. cities that included 75.6 percent of all U.S. Asians. While our universe did not represent 100 percent coverage, it did represent a substantial increase over the 18.8 percent coverage for the 2000 PMPE Asian sample design. Within these 1,261 cities, the population is 7.8 percent non-Hispanic Asians. This population percentage of non-Hispanic Asians (7.8 percent) is our estimated eligibility rate, even though this may differ from the actual eligibility rate based on the project's protocol to interview the person most likely to handle the incoming mail. [Table 1](#) gives the coverage and estimated eligibility rates for the many different non-Hispanic Asian population thresholds that we could have used for our sampling frame of cities.

[Table 1](#) shows that as the threshold decreases, the coverage increases while the eligibility rate decreases. [Table 1](#) also shows that 12.71 percent of Asians live outside of cities (in unincorporated places, including rural areas). So it is not possible to achieve 100 percent coverage without sampling unincorporated places. As the eligibility rate decreases, more screening becomes necessary to find the same number of eligible households. Though coverage could be higher with county-level sampling, the eligibility rates would be so much lower that such a sample would be cost prohibitive.

Keeping in mind our goal of 500 Asian interviews in each wave, we needed to balance the number of cities where we would have to hire staff against the cluster size determined by the average number of interviews we would need to collect in each city. Increasing the number of cities would increase the cost, while decreasing the number of cities would increase the clustering and therefore the design effect. Balancing these two factors, we decided to select a representative sample of 25 cities (requiring an average of 20 interviews per city) with probability proportional to the city population of non-Hispanic Asians. Our design gave every Asian in our frame of 1,261 cities an equal chance of being in

*Table 1. Threshold options for the Asian frame of cities*

Minimum Number of Non-Hispanic Asians	Eligible Cities	Asian Population Coverage (%)	Eligibility Rate (%)
100,000	8	20.46	12.28
50,000	18	25.92	12.69
25,000	46	34.07	12.30
10,000	153	48.12	10.58
5,000	321	57.87	9.67
2,500	653	67.59	8.72
1,000	1,261	75.57	7.83
500	2,059	80.32	7.04
250	3,015	83.17	6.50
100	4,569	85.29	5.97
50	6,039	86.16	5.68
25	7,823	86.69	5.46
1	18,608	87.29	5.05

one of our selected cities. New York and Los Angeles, two of the cities used for the 2000 PMPE design, were selected with certainty because they each contain more than 1 in 25 of the Asians within our sampling frame (359,684). Every other city in our frame had a chance of selection equal to their Asian population divided by 359,684. Of the three other cities used in the 2000 PMPE, two (San Francisco and Chicago) were selected while Seattle was not.

We also wanted representativeness among different Asian subgroups. Census 2010 population counts are available at the place level for the following subgroups: Indian Asians, Chinese Asians (excludes Taiwanese), Filipino Asians, and Other Asians. To make sure that all four subgroups were represented, we divided the non-certainty cities into four equal groups: 1) those where at least 26.83 percent of Asians were Indian Asians, 2) those where at least 24.27 percent of Asians were Chinese Asians, 3) those where at least 24.84 percent of Asians were Filipino Asians, and 4) those where at least 54.77 percent of Asians were Other Asians. Within these four groups, the sampling frame was serpentine sorted by Census Region, State, and Asian Population Percentage. [Table 2](#) shows the distribution of respondents by Asian subgroup, which includes the three subgroups used during sampling, as well as three other major subgroups and an “Other” category. [Table 2](#) also includes the Census 2010 distribution among Asian Americans ([Hoeffel et al. 2012](#)).

Due to the Census data limitations, only the top three Asian subgroups in [Table 2](#) could be controlled in our selection process. [Table 2](#) shows that we had fewer Asian Indian respondents and more Filipino respondents than Census 2010 would lead us to expect. The surplus of Chinese respondents can be partly explained by the Census exclusion of Taiwanese Chinese (our question did not separate them). Among the “Other” subgroups, our total sum is quite close to the Census sum, though our sample did have more Japanese respondents and fewer of the rarer Other groups.

Eleven of our twenty-five selected cities were in California. Other states with more than one city selected were Hawaii, New York, and Texas. [Table 3](#) gives details on the 25 selected cities.

The Asian population sizes in our 25 cities range from the minimum of 1,000 to around 850,000, with a median of approximately 23,000. Asian population percentages range from under two percent to over 70 percent, with a median of 12.86 percent.

We then selected entire census tracts as our second-stage clusters within our selected cities. We decided to select five tracts from each noncertainty city so that we would have

*Table 2. Distribution of 2010 CICPE Asian sample by Asian subgroup*

Asian Subgroup	Wave 1 (%)	Wave 2 (%)	Wave 3 (%)	Census 2010 (%)
Asian Indian	11.9	13.1	13.7	18.4
Chinese (excl. Taiwanese)	21.3	26.5	24.2	21.8
Filipino	28.7	26.2	22.8	19.7
Japanese	20.5	20.3	20.3	7.5
Korean	7.0	4.3	7.5	9.9
Vietnamese	7.4	5.6	6.8	10.0
Other	3.3	4.0	4.6	12.7

Table 3. The 25 cities in the 2010 CICPE Asian sample

City	State	Asian Population	Asian Population Percentage (%)
New York city	New York	857,094	10.70
Los Angeles city	California	396,352	10.73
San Jose city	California	252,818	28.25
San Francisco city	California	250,364	32.23
Honolulu CDP*	Hawaii	244,698	65.84
Chicago city	Illinois	137,039	4.73
Houston city	Texas	111,511	5.71
Sacramento city	California	74,634	18.34
Philadelphia city	Pennsylvania	73,403	4.84
Stockton city	California	52,631	21.59
Anaheim city	California	42,171	12.86
Arcadia city	California	24,886	46.91
Hilo CDP*	Hawaii	23,206	56.94
Bellevue city	Washington	20,741	18.93
Riverside city	California	16,311	6.39
Bakersfield city	California	12,036	4.87
Lakewood city	California	11,870	14.96
Canton CDP*	Michigan	7,252	9.50
Pacifica city	California	6,805	17.73
Aiea CDP*	Hawaii	6,423	71.22
Syracuse city	New York	5,566	3.78
Port Arthur city	Texas	3,546	6.14
Stafford city	Texas	3,272	20.87
Marlborough city	Massachusetts	1,514	4.18
Rio Rancho city	New Mexico	1,006	1.94

\*CDP = Census Designated Place

approximately 125 tracts for 500 interviews (an average clustering of four interviews per tract). This means that every tract selected represents 359,684/5, or roughly 72,000 Asians. Since New York and Los Angeles both have more than 359,684 Asians, we selected extra tracts for them. We selected ( $5 * 857,094/359,684 = 11.9$ ) twelve tracts in New York and ( $5 * 396,352/359,684 = 5.5$ ) six tracts in Los Angeles. Only one selected city contains less than five Census tracts to select from; the Aiea CDP only contains two tracts, so both were selected (and each had  $5/2 = 2.5$  times as many housing units selected). Thus we selected a total number of 130 Census tracts, which resulted in an average of 3.8 interviews per selected Census tract. The selected tracts had an even larger range of Asian population percentages, ranging from 87 percent to less than one percent. An equal probability sample using these 130 Census tracts would have resulted in a sample eligibility rate of 7.8 percent (almost twice the national eligibility rate), but this can be increased by oversampling in tracts with a higher proportion of Asian residents. We actually designed a sample with an expected eligibility rate of 26 percent, but this required some tracts to be oversampled by a factor of 50. Differential sampling weights that result from such a skewed oversample would have created a large design effect, which would have greatly reduced our effective sample size.

Table 4. Planned and actual unweighted eligibility rates

Statistic	2000 PMPE (%)	2010 CICPE (%)
Eligibility Rate		
First-Stage	6.5	7.8
Planned	unknown	12.5
Wave 1	22.2	10.3
Wave 2	13.3	12.9
Wave 3	18.9	8.5

As in most statistical design decisions, the amount of oversampling involved a balance between lowering screening costs versus keeping variance due to differential weighting low. With help from the Census Bureau, we agreed to limit the design effect so that the loss in effective sample size due to differential sampling would be no greater than 20 percent (a design effect due to differential sampling no greater than 1.25). Our approach decreased the differential oversampling from a factor of 50 to a factor of 3. In so doing, we incurred higher screening costs, but maintained the effective sample size closer to the number of interviews. Our specific strategy was to oversample tracts with eligibility rates of at least 20 percent by a factor of 3, and to oversample tracts with eligibility rates between 10 and 20 percent by a factor of 2. With this strategy, our estimated eligibility rate was 12.5 percent, three times as large as the national eligibility rate.

#### 4. 2010 CICPE Asian Sample Field Results

We were able to meet our sample targets for the Asian sample, but our actual unweighted eligibility rates were lower than our estimate for two out of the three waves. This is not surprising when the observed eligibility rate is lower than the planned eligibility rate. Households that are eligible are the most difficult households at which to achieve cooperation (even at the screener level), so eligible households often have a lower screener response rate than ineligible households. This results in a lower eligibility rate instead of a lower response rate, so this is often referred to as “hidden interview nonresponse” within the screener nonresponse. Table 4 shows that the 2000 PMPE achieved higher eligibility rates, which were due to a higher level of oversampling. The 2000 PMPE Asian design oversampled areas with eligibility rates of at least 20 percent by a factor of 5 (Wolter et al. 2002).

Table 5 compares the weighted response rates from 2010 CICPE against the unweighted 2000 PMPE response rates.

Weighted rates are not available from the 2000 PMPE, but unweighted rates are not appropriate for the 2010 CICPE because of the mixed-mode data collection procedures

Table 5. Response rates for 2000 PMPE and 2010 CICPE

Wave	2000 PMPE (Unweighted) (%)	2010 CICPE (Weighted) (%)
Wave 1	57.2	50.7
Wave 2	71.0	64.2
Wave 3	60.8	73.8



that included subsampling of nonrespondents for in-person follow-up. Nevertheless, the average response rates for both studies are around 63 percent, and both studies have their lowest response rate in the first wave. The response rates are higher for the 2000 PMPE in the first two waves, but the response rate is much higher in the third wave for the 2010 CICPE.

Nonresponse bias is usually immeasurable, but the 2010 CICPE study is an exception. With Census Bureau cooperation, we were able to match the entire set of our selected households to 2010 Decennial Census response data. Nonresponse bias was an important issue for the 2010 CICPE since it was designed around probable/actual response to the 2010 Decennial Census, and it is logical to think that nonrespondents to our survey would be more likely to be nonrespondents to the 2010 Decennial Census. Table 6 gives the actual mail response rates to the 2010 Decennial Census by April 18 for three different types of 2010 CICPE respondents, as well as for interview nonrespondents and those households for which we could not determine eligibility (screener nonrespondents). We used the April 18 date cutoff because this marks the start of the in-person follow-up effort. Decennial forms are mailed near the end of March, Census Day is April 1, and any households returning their mail forms after April 18 may still have been visited in-person. Minimum cost is achieved for households whose mail questionnaire is received prior to April 18.

The three types of CICPE respondents are: 1) Refusers – those respondents who were (soft) refusals at one time, 2) Difficult Respondents – those respondents who had more than the median number of visits before responding, and 3) Easy Respondents – those respondents who responded after less than the median number of visits. All of the mail response rates in Table 6 are weighted. Table 6 shows that our respondents did have higher mail return rates. As expected, the easy Asian respondents had the highest mail return rate by April 18 (64.3 percent) while the nonrespondents had the lowest mail return rate by April 18 (53.0 percent). We estimated the eligible proportion among those with unknown eligibility status and counted them as nonrespondents. Combining the two nonrespondent categories together, the mail return rate was 53.6 percent. Combining the three respondent categories together, the mail return rate was 62.7 percent. Combining all five categories together, the mail return rate for our entire Asian sample was 59.4 percent.

The nonresponse bias is the difference between the estimate for only the respondents (62.7 percent) and the entire population of interest, represented by the entire sample (59.4 percent). Therefore, our Asian sample’s nonresponse bias is  $62.7 - 59.4 = 3.3$  percent.

*Table 6. Mail response rates for the Asian CICPE sample (weighted)*

Outcome	Return Rate (%)	Response Status (%)	ALL (%)
Unknown Eligibles	54.5	Nonrespondents: 53.6	ALL: 59.4
Nonrespondents	53.0		
Respondents – Refusers	61.9	Respondents: 62.7	
Respondents – Difficult Respondents	62.2		
Respondents – Easy	64.3		



Since this is positive, our respondents are more likely to be mail responders by April 18 than our nonrespondents, which is the expected direction of our nonresponse bias. Interestingly, of our six race/ethnicity groups, four had almost no nonresponse bias (our estimates of nonresponse bias were less than one percent), while our American Indian and Alaska Native sample had a negative nonresponse bias, meaning that respondents were less likely to be mail responders by April 18 than our entire sample of American Indian and Alaska Native housing units (Datta et al. 2012 has more details for all six race/ethnicity groups).

## 5. Demographic Comparisons to the National Asian Population

We collected demographic data from our respondents, which allows us to compare our set of respondents with national control totals to examine how representative our sample is. Table 2 (above) already showed that our sample has more Filipino and Japanese respondents and fewer Asian Indian respondents than the national totals. Table 7 compares the age and gender distribution of our respondents against Current Population Survey Annual Social and Economic Supplement data from March 2010.

We had expected age and gender to be skewed by the project's protocol that interviewed the person most likely to handle the incoming mail. However, Table 7 shows that the CICPE Asian Sample gender distribution is very close to the national average of 47.1 percent male, 52.9 percent female. Our respondents are more likely to be 65 years of age or older and less likely to be 18–29 years of age, but our distribution is consistent with most surveys that have lower response rates among this young age group and higher response rates for senior citizens.

As an additional check on the representativeness of our Asian sample, Table 8 shows the highest degree earned by our 2010 CICPE Asian Sample respondents and the

Table 7. Distribution of 2010 CICPE Asian sample by age and gender

Gender/Age	Wave 1 (%)	Wave 2 (%)	Wave 3 (%)	CPS, 2010 (%)
Male 18-29	19.7	20.1	18.0	24.6
Male 30-44	31.6	34.3	35.2	33.7
Male 45-64	32.1	29.6	29.7	30.9
Male 65 +	16.7	16.0	17.2	10.8
<b>Total Male</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
Female 18-29	16.8	11.8	13.5	22.4
Female 30-44	33.7	35.0	34.1	32.8
Female 45-64	29.4	33.6	33.0	31.4
Female 65 +	20.1	19.5	19.5	13.4
<b>Total Female</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
ALL 18-29	18.3	15.1	15.5	23.4
ALL 30-44	32.4	34.2	34.1	33.2
ALL 45-64	31.3	33.0	31.9	31.1
ALL 65 +	18.1	17.6	18.5	12.2
<b>Total</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>
<b>Percentage Male</b>	<b>46.2</b>	<b>43.4</b>	<b>49.3</b>	<b>47.1</b>

Table 8. Distribution of 2010 CICPE Asian sample by highest degree

Highest Degree	Wave 1 (%)	Wave 2 (%)	Wave 3 (%)	ACS 2009 (%)
No High School Diploma	6.6	6.2	7.7	15.0
High School Diploma	36.8	34.3	33.3	35.0
College Degree	35.8	35.8	35.3	30.0
Graduate/Professional Degree	20.8	23.7	23.7	20.0
<b>Total</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>	<b>100.0</b>

corresponding national percentages from the 2009 American Community Survey (U.S. Census Bureau 2011).

Table 8 shows that we have fewer respondents with no high school diploma than the American Community Survey, but our distributions are similar otherwise.

## 6. Limitations

Some compromises were made in order to limit the cost of the Asian sample, the largest of which is that the frame did not attain 100 percent coverage due to the minimum size requirement. Since we needed an average of 20 interviews per city per wave, we did need to set a minimum size, but we set the minimum size at 1,000 Asians in order to keep the eligibility rate to almost 8 percent. Setting a lower minimum size would have resulted in additional screening costs. In fact, Table 1 shows that using cities as the first-stage sampling units limits the coverage to 87 percent. To achieve 100 percent coverage, a sample of unincorporated areas would be necessary.

One minor limitation is that we did have larger clusters in the Aiea, Hawaii CDP since the interviews there were collected in only two Census tracts rather than five because there are only two tracts in the Aiea CDP. Clustering could also have been reduced if we had selected more than 25 cities, but adding field staff in more locations would have created significant additional costs.

Finally, our nonresponse bias does show that Asians are the only race/ethnicity group with a bias larger than one percent. One possible explanation is that we only collected interviews in English and Spanish and so the Asians were likely to have the highest language barrier among the six race/ethnicity groups. If English-speaking Asians were more likely to respond to the Census, this would result in a positive nonresponse bias as observed.

## 7. Summary

Even for a hard-to-reach population, it may be possible to attempt a nationally representative sample if the population can be targeted by local areas. Asians make up only 4.2 percent of the U.S. population, but are more clustered by city than larger first-stage sampling units such as counties or metropolitan areas. We have achieved a 75.6 percent coverage rate for a national sample of Asians for the Census Integrated

Communications Program Evaluation (CICPE) study by selecting 25 cities from a frame of 1,261 U.S. cities with a population of at least 1,000 Asians. While we could not come closer to 100 percent coverage in a cost-effective way and expensive screening was still necessary, we believe that our data is more representative of U.S. Asians than the 2000 PMPE study taking place in only five U.S. cities as well as any list-assisted telephone survey using Asian surnames or other methods with unknown biases (Davern et al. 2007). Higher coverage is possible with county-level sampling, but with eligibility rates that are lower enough to make it cost prohibitive. This is why nationally representative samples of Asians have not been previously attempted without a larger survey structure for support. While the central idea of this article may not be a theoretical breakthrough, it has not been acted on previously, greatly reducing the ability to collect nationally representative data on Asians and other hard-to-reach groups. If samplers could break away from the “rule” that first-stage units must be counties, area probability samples could more flexibly collect nationally representative data for a wider range of applications.

## 8. References

- Barnes, J. and Bennett, C. (2002). The Asian Population: 2000, a Census 2000 Brief. Washington, D.C.: Bureau of the Census. Available at: <http://www.census.gov/prod/2002pubs/c2kbr01-16.pdf> (accessed December 2013).
- Curtin, L., Mohadjer, L., Dohrmann, S., Kruszan-Moran, D., Mirel, L., Carroll, M., Hirsch, R., Burt, V., and Johnson, C. (2013). National Health and Nutrition Examination Survey: Sample Design, 2007-2010, National Center for Health Statistics. Vital and Health Statistics, Series 2, Number 160. Available at: [http://www.cdc.gov/nchs/data/series/sr\\_02/sr02\\_160.pdf](http://www.cdc.gov/nchs/data/series/sr_02/sr02_160.pdf) (accessed December 2013).
- Datta, A., Yan, T., Evans, D., Pedlow, S., Spencer, B., and Bautista, R. (2012). The 2010 Census Integrated Communications Program Evaluation (CICPE) Final Report. Washington, D.C.: Bureau of the Census. Available at: [http://www.census.gov/2010census/pdf/2010\\_Census\\_ICP\\_Evaluation.pdf](http://www.census.gov/2010census/pdf/2010_Census_ICP_Evaluation.pdf) (accessed December 2013).
- Davern, M., McAlpine, D., Ziegenfuss, J., and Beebe, T. (2007). Are Surname Telephone Oversamples an Efficient Way to Better Understand the Health and Healthcare of Minority Group Members? *Medical Care*, 45, 1098–1104.
- Heeringa, S., Wagner, J., Torres, M., Duan, N., Adams, T., and Berglund, P. (2004). Sample Designs and Sampling Methods for the Collaborative Psychiatric Epidemiology Studies (CPES). *International Journal of Methods in Psychiatric Research*, 13, 221–240, DOI: <http://www.dx.doi.org/10.1002/mpr.179>.
- Hoeffel, E., Rastogi, S., Kim, M.O., and Shahid, H. (2012). The Asian Population: 2010, a Census 2010 Brief. Washington, D.C.: Bureau of the Census. Available at: <http://www.census.gov/prod/cen2010/briefs/c2010br-11.pdf> (accessed December 2013).
- Kish, L. (1965). *Survey Sampling*. New York: Wiley.
- Lohr, S. (2009). *Sampling: Design and Analysis (Second Edition)*. Pacific Grove, CA: Duxbury Press.
- Moore, W., Pedlow, S., Krishnamurty, P., and Wolter, K. (2000). The National Longitudinal Survey of Youth 1997 (NLSY97) Technical Sampling Report. Chicago:

- NORC at the University of Chicago. Available at: <http://www.bls.gov/nls/nlsy97techsamp.pdf> (accessed December 2013).
- U.S. Census Bureau (2011). Profile America Facts for Features, Asian/Pacific American Heritage Month: May 2011. Washington, D.C.: Bureau of the Census. Available at: [http://www.census.gov/newsroom/releases/archives/facts\\_for\\_features\\_special\\_editions/cb11-ff06.html](http://www.census.gov/newsroom/releases/archives/facts_for_features_special_editions/cb11-ff06.html) (accessed December 2013).
- Wolter, K., Calder, B., Malthouse, E., Murphy, S., Pedlow, S., and Porras, J. (2002). Census 2000 Evaluation: Partnership and Marketing Program Evaluation. Washington, D.C.: Bureau of the Census. Available at: <http://www.census.gov/pred/www/rpts/D.1.PDF> (accessed December 2013).

Received February 2013

Revised November 2013

Accepted January 2014

# Recruiting an Internet Panel Using Respondent-Driven Sampling

*Matthias Schonlau<sup>1</sup>, Beverly Weidmer<sup>2</sup>, and Arie Kapteyn<sup>3</sup>*

Respondent-driven sampling (RDS) is a network sampling technique typically employed for hard-to-reach populations when traditional sampling approaches are not feasible (e.g., homeless) or do not work well (e.g., people with HIV). In RDS, seed respondents recruit additional respondents from their network of friends. The recruiting process repeats iteratively, thereby forming long referral chains.

RDS is typically implemented face to face in individual cities. In contrast, we conducted Internet-based RDS in the American Life Panel (ALP), a web survey panel, targeting the general US population. We found that when friends are selected at random, as RDS methodology requires, recruiting chains die out. When self-selecting friends, self-selected friends tend to be older than randomly selected friends but share the same demographic characteristics otherwise.

Using randomized experiments, we also found that respondents list more friends when the respondent's number of friends is preloaded from an earlier question. The results suggest that with careful selection of parameters, RDS can be used to select population-wide Internet panels and we discuss a number of elements that are critical for success.

*Key words:* Web survey; RDS.

## 1. Introduction

Respondent-driven sampling (RDS) is a chain referral sampling technique typically conducted face to face with hard-to-reach populations in individual locations such as cities. Implementing RDS on the web can be advantageous. First, web implementation is much less expensive than a face-to-face approach because the interviewer labor costs and costs associated with setting up a field operation can be avoided. Second, RDS requires the assumption that respondents will recruit from among their friends at random. This is more easily enforceable on the web, where randomization from a list of friends is easy. Third, web implementation of RDS allows recruitment across a much wider geographic area as it is not tied to one geographic location as typically happens with face-to-face implementation of RDS. This is also true for geographic spread inside a large city, where friends who live or work close to the field station may be more likely to enroll.

<sup>1</sup> University of Waterloo, Statistics and Act.Sci.200 University Ave, Waterloo, Ontario N2L3G1, Canada. Email: [schonlau@uwaterloo.ca](mailto:schonlau@uwaterloo.ca)

<sup>2</sup> RAND Corporation, Survey Research Group, Los Angeles, California, USA. Email: [weidmer@rand.org](mailto:weidmer@rand.org)

<sup>3</sup> University of Southern California and RAND Corporation. Email: [kapteyn@usc.edu](mailto:kapteyn@usc.edu)

**Acknowledgments:** Support for this research comes from grant R01AG20717 from the National Institute on Aging to RAND (Arie Kapteyn, P.I.). We thank the ALP team for all their help throughout. The project would not have been possible without their enthusiasm.

RDS studies conducted face to face have to be conducted in one location or city at a time. Fourth, in light of declining response rates to phone and mail surveys, a successful nationwide implementation of a network sample would provide a useful alternative. Fifth, populations who are routinely on the web may be more likely to participate. This may include at-risk college students, but also Internet sex workers. Finally, once the software for a web implementation is in place, it is easy to conduct additional RDS studies.

While implementing RDS on the web has several advantages, it also has different challenges. The web environment is more anonymous than face-to-face encounters, which may be useful for some hard-to-reach populations. However, recruiters may find it harder to motivate their friends to enroll and it may affect trust. Some populations do not have access to the Internet. For example, it may not be possible to reliably reach the homeless in this way. This limitation may decrease in importance over time as more and more people gain access to the Internet. Finally, operationalizing RDS through a web survey where new recruits enroll throughout the study and in turn become recruiters is not trivial.

Very few attempts have been made to conduct RDS on the Internet and many studies have either required multiple attempts or were unsuccessful. Because of the small number of studies involved, there is no conclusive evidence for any factor that may explain implementation challenges. We nonetheless find it useful to list factors we believe contributed to implementation challenges or failure. First and perhaps most importantly, requiring respondents to provide information such as email addresses of their friends appears to be a bad idea. We also note that “respondent-driven” implies that the respondents should contact their friends; not interviewers or an automated computer program. An Internet-based RDS study in Cambridge, England, about cars and the environment failed (RAND Cambridge, personal communication) because respondents were unwilling to contact friends.

Second, not providing an incentive or providing too small an incentive seems to also affect recruitment rates. The aforementioned Cambridge study did not have sufficient funds to provide incentives and failed. An attempt to recruit parents of students studying at Tilburg University in the Netherlands into a panel survey failed to generate sufficient response (personal communication, Department of Leisure Studies). Students were asked to contact and enroll their parents and were offered a small monetary incentive for doing so. Only 120 persons out of the 4,000 invited joined the panel (a 3% enrollment rate). A study of students at Wayne State University (Detroit) that invited feedback from seed respondents made it clear respondents wanted to earn more money (Bauermeister et al. 2012), among other things. In response, investigators increased the total number of referrals allowing respondents to earn money for the first five referrals (5\*\$10 for referrals plus \$20 for filling out the survey). The study succeeded in recruiting 3,426 respondents.

Incentives can be delivered in a variety of forms. A study of men who had sex with men (MSM) in Vietnam (Bengtsson et al. 2012) used credit on SIM chips or a donation to an organization the target population cared about as well a lottery draw for an iPad. The aforementioned study at Wayne State University (Bauermeister et al. 2012) provided VISA e-gift cards for filling out the questionnaire. The cards were reloaded when friends were referred successfully. A study of Muslim students, also at Wayne State University, used unspecified gift certificates (Arfken et al. 2013). In the first published Internet RDS

study (Wejnert and Heckathorn 2008) at Cornell University, respondents (or friends) had to pick up incentives in person.

Third, there is at present no evidence that the survey topic plays an important role in recruiting success. Consider two contrasting examples: The RAND Corporation studied the opinions of gays and lesbians in the military about the “Don’t Ask, Don’t Tell” policy (Berry et al. 2010). The RDS study was a resounding policy success and won the “policy impact award” from the American Association for Public Opinion Research (AAPOR) in 2011. However, the study failed as an RDS study because referral chains were not long enough. The role of gays and lesbians in the U.S. military is a topic that is presumably very important to those in the military who are gay and lesbian, and if the topic were important for recruiting it should have been more successful. In terms of speed, the most successful Internet RDS study (Wejnert and Heckathorn 2008) reached the intended sample size of 150 respondents (plus nine additional seeds) in only 72 hours. Participants were “invited to participate in a research study to empirically validate Respondent Driven Sampling (RDS) as an analytical tool for the study of social structure” (Wejnert and Heckathorn 2008, online Appendix A). This does not appear to be a ‘sexy’ topic for university students that would have contributed to the success of the study.

Fourth, adjusting the number of seeds after a study begins is an important tool to avoid recruiting chains dying out. The Vietnam study initially started with 15 seeds and then increased to 20 two weeks later (Bengtsson et al. 2012). However, increasing the number of seeds does not always work – the RAND study (Berry et al. 2010) increased the number of seeds from five to 189 and still came up short. Except for this study, the number of seeds in Internet RDS studies tends to be much smaller with numbers ranging from nine (Wejnert and Heckathorn 2008) to 22 (Bauermeister et al. 2012).

Fifth, adjusting the number of referrals allowed is a useful tool to avoid recruiting chains dying out. As mentioned above, one study (Bauermeister et al. 2012) increased the number of referrals while paying only for the first five successful referrals. This study also allowed respondents to copy the referral codes into text messages and social media (Facebook). The study eventually reached 3,448 respondents. Overall, the number of referrals in Internet RDS studies has ranged from three to five.

The purpose of our study was to explore the feasibility of recruiting respondents into a web panel using Internet-based RDS. Specifically, our goal was to recruit respondents into the American Life Panel (ALP) (<https://mmicdata.rand.org/alp/>), a probability-based Internet panel. It was hoped that the availability of a network sample would make the ALP more attractive to researchers with such needs. Already existing respondents in the ALP panel were **not** recruited over the Internet. Potential respondents without Internet access receive a laptop and broadband Internet access for free.

An overview of the sequence of experiments and recruitment efforts is given in Figure 1. In our initial pilot run we found that a large number of respondents would only list a single respondent, presumably to avoid follow-up questions. We designed Experiment 1 to find out which survey design would lead to a more successful elicitation of friends (Section 2). Next, we found respondents were listing friends but those friends would not contact us to enroll in the study. In response we designed a second experiment (Section 3), varying incentive levels and how friends were selected. Based on results from Experiment 2 we started an RDS sample in the American Life Panel (ALP) (Section 4). Section 4 also

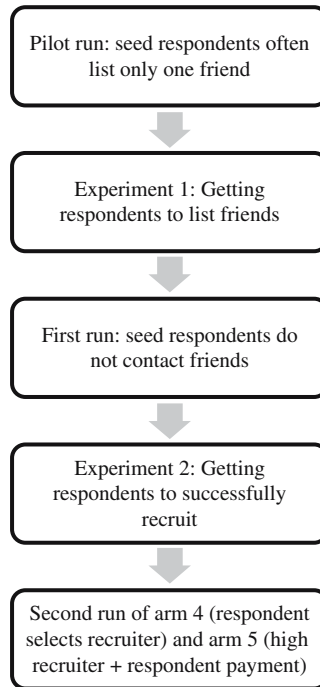


Fig. 1. Sequence of experiments and recruitment efforts.

compares the demographic composition of the recruited sample for self-selected friends and randomized friends. Section 5 concludes with a discussion. We first provide an overview of respondent-driven sampling.

### 1.1 Respondent-Driven Sampling

RDS is a chain referral sampling technique (Heckathorn 1997, 2002, 2007). A small number of seed respondents recruit additional respondents from their network of friends. The recruiting process repeats iteratively, thereby forming long referral chains. Suppose we want to estimate the percentage of males,  $p_1$ , and females,  $p_2$ , in a population. Of interest is the two-by-two gender transition matrix (male/female recruiting male/female) between the recruiter and recruit. Assuming a first-order Markov process (recruit gender depends on recruiter's gender, but not on earlier recruiters), a sample equilibrium is reached if referral chains are sufficiently long. The sample equilibrium is not the population equilibrium because well-connected people are overrepresented in the sample. For example, if women have more friends than men, the sample equilibrium would have more women than men. The assumption of reciprocity (explained below) yields an equation: The number of possible edges (links between two persons) with a male recruiter recruiting a female is the same as the number of possible edges with female recruiters recruiting a male:  $n_1 D_1 S_{12} = n_2 D_2 S_{21}$  where  $n_i$  is the number of respondents in group  $i$ ,  $D_i$  is the average group degree (e.g., average network size among females) and  $S_{ij}$  is the estimated transition probability between categories  $i$  and  $j$ . Dividing by the total sample size turns frequencies into proportions:  $p_1 D_1 S_{12} = p_2 D_2 S_{21}$ .



The average group degree of group  $i$ ,  $D_i$ , is estimated from individual degrees using the “multiplicity” formula (Salganik and Heckathorn 2004)  $D_i = n_i / [\sum_j (1/d_{ij})]$  where  $n_i$  is the number of respondents in group  $i$  and  $d_{ij}$  is an estimate of individual degree (number of friends). Unlike the arithmetic mean, this formula takes into account that respondents with a greater network are overrepresented. Because the formula relies on the inverse of self-reported degrees, this estimate is robust against large positive outliers in individual degrees.

In the example above, the equation depends on two unknown proportions,  $p_1$  and  $p_2$  (the proportion of females and males in the population), which must sum to one:  $1 = p_1 + p_2$ . Two equations with two unknowns can be solved and yield  $p_1 = S_{21}D_2 / (S_{21}D_2 + S_{12}D_1)$  and  $p_2 = 1 - p_1$ . When there are more than two categories, the estimates result from solving an over-determined system of equations. In general, RDS only allows the estimation of proportions, not absolute frequencies or totals. However, frequencies can be computed from the proportions if the population total is known from elsewhere or if the population total is estimated by capture/recapture methods (Berchenko and Frost 2011; Heckathorn et al. 2002).

RDS requires the following assumptions: *Assumption 1. Reciprocity.* If respondent A recruited respondent B, then in principle B could have recruited A also. In practice, this assumption is tested by verifying that the recruiter is part of the recruit’s social network. *Assumption 2. Networked population.* Respondents are all linked to a single component in the network (i.e., there are no isolated pockets of people without friends). *Assumption 3. Sampling is with replacement.* This assumption never holds because the same respondent is not sampled twice. In practice, this assumption is innocuous unless the sample represents a large fraction of the population. *Assumption 4. Network size.* Respondents can accurately report their degree (personal network size). Consistent under- or overestimation of network size among all respondents cancels out and is unproblematic (Wejnert 2009, sec. “Degree Estimation”). Moreover, estimates may be robust for different assessments of network size (Wejnert 2009). *Assumption 5. Random Recruitments.* Respondents recruit from their network at random. This assumption is the most controversial by far. In some arms of our experiments we have done the randomization ourselves by choosing from a list of first names or initials provided by the respondent, thus avoiding this problem.

In practice, recruitment is facilitated through a dual incentive system which includes payments for both the respondent and each referral who agrees to participate. While RDS has not yet been used to recruit a national sample, it emerges as a natural choice when social network analyses are of interest (Wejnert 2010). Respondent-driven sampling is implemented in a stand-alone package ([www.respondentdrivensampling.org](http://www.respondentdrivensampling.org)), in Stata (Schonlau and Liebau 2012), and a package is in preparation for R.

## 2. EXPERIMENT 1: Eliciting the Number of Friends

One assumption of RDS is the random recruitment of friends. Given a list of friends (first names or nicknames suffice), respondents are asked to contact specific friends selected at random by the computer software. During initial trials we found that many respondents tended to only list a single friend. The purpose of Experiment 1 was to investigate how to

ask about respondents' friends such that respondents list their friends in greater numbers. Respondents were not told why they were supposed to list the friends to avoid having respondents selectively list only friends they wanted to contact. In each case we *first ask for the number of friends* (Numerical question given in Appendix). On the following screen we then ask the same question but ask the respondent to list first names or initials. (The friends question is reproduced in Appendix) The experiment had five experimental arms:

**Experimental Arm 1 (“1 row”):** We asked the respondent to list one person at a time (Figure 2). On the same screen, we asked whether the respondent wanted to list an additional person. If this question went unanswered, we prompted for an answer on the next screen: “You did not answer the previous question(s). Your answers are important to us. Please return to the previous question and answer it to the best of your ability.”

**Experimental Arm 2 (“10 rows”):** We asked the respondent to list ten persons at a time (Figure 3). If respondents listed ten persons they were asked whether they would like to list additional persons. If respondents listed less than ten persons they proceeded to the next question (without prompt).

**Experimental Arm 3 (“prompt w preloaded #”):** As in Experimental Arm 2, respondents were asked to list ten persons at a time. However, if fewer people were listed than indicated in the preceding numerical question, we prompted for additional people: “You answered earlier you had [preloaded number] close friends and family members, but you listed a smaller number. This question is very important to us. If possible please go back and add more people.” To avoid a problem if respondents gave an unreasonably high numerical value (e.g., 100), respondents were not prompted if they listed at least ten

Please list all the close friends or family members you see, talk to or write to (via letter, email, text message, facebook, etc.) regularly. Please **do not** include people who live in your household. Please only consider people 18 years or older who live in the United States. You only need to provide their first name, nickname or initials.

**Person 1**

First Name, nickname or initials	Relationship to you	Is this person Hispanic or of Hispanic origin or descent?
<input type="text"/>	Click here ▾	Click here ▾

**Would you like to add another person?**

Yes  
 No




Fig. 2. Screenshot of Experimental Arm 1. The dropdown menu for relationship had the categories: child, parent, other relative, work friend, school friend, family friend, acquaintance, other.

Please list all the close friends or family members you see, talk to or write to (via letter, email, text message, facebook, etc.) regularly. Please **do not** include people who live in your household. Please only consider people 18 years or older who live in the United States. You only need to provide their first name, nickname or initials.

First Name, nickname or initials	Relationship to you	Is this person Hispanic or of Hispanic origin or descent?
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾




Fig. 3. Screenshot of Experimental Arm 2 (identical to Experiment Arm 3). The dropdown menu for relationship had the categories: child, parent, other relative, work friend, school friend, family friend, acquaintance, other.

friends. Respondents were allowed to list more people than indicated in the numerical question.

**Experimental Arm 4 (“ask preloaded #”):** Respondents were specifically asked for the number of friends given in the numerical question: “Please list these [preloaded number] close friends or family members. You only need to provide their first name, nickname or initials” (Figure 4). As before, if there were more than ten friends, we listed ten on each screen until the total number was exhausted. For example, if there were twelve friends, the respondents saw a screen with ten rows and a second screen with two rows. If the respondent failed to list ten friends, the second screen was not shown. There was no additional prompt if respondents listed fewer friends than indicated by the numerical question.

**Experimental Arm 5 (“single column”):** We first asked respondents to list first names only (Figure 5). The number of friends was not preloaded. In a second question, we asked respondents to list information about their friends (Figure 6). If respondents did not list anybody, they were not prompted to remind them of their earlier numerical answer. If respondents listed ten friends, they were shown an additional screen and asked to list additional friends as before.

Respondents received a \$5 incentive for responding to this survey. To avoid overly complicated programming, we did not allow listing more than 50 people.

Please list these **6** close friends or family members. You only need to provide their first name, nickname or initials.

First Name, nickname or initials	Relationship to you	Is this person Hispanic or of Hispanic origin or descent?
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾
<input type="text"/>	Click here ▾	Click here ▾

---




Fig. 4. Screenshot of Experimental Arm 4 with a listed number of rows given from numerical question. The dropdown menu for relationship had the categories: child, parent, other relative, work friend, school friend, family friend, acquaintance, other.

Please list all the close friends or family members you see, talk to or write to (via letter, email, text message, facebook, etc.) regularly. Please **do not** include people who live in your household. Please only consider people 18 years or older who live in the United States. You only need to provide their first name, nickname or initials.

First Name, nickname or initials
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>
<input type="text"/>

---



Fig. 5. Screenshot of Experimental Arm 5. First names are listed first, and then additional information is prompted.

Please list all the close friends or family members you see, talk to or write to (via letter, email, text message, facebook, etc.) regularly. Please **do not** include people who live in your household. Please only consider people 18 years or older who live in the United States. You only need to provide their first name, nickname or initials.

First Name, nickname or initials	Relationship to you	Is this person Hispanic or of Hispanic origin or descent?
john	Click here ▾	Click here ▾
sally	Click here ▾	Click here ▾
j.r.	Click here ▾	Click here ▾
karen	Click here ▾	Click here ▾
charlie	Click here ▾	Click here ▾




Fig. 6. Screenshot of Experimental Arm 5 soliciting additional information. The dropdown menu for relationship had the categories: child, parent, other relative, work friend, school friend, family friend, acquaintance, other.

*Results:* 473 respondents were invited to participate. Respondents were randomized one at a time to an experimental arm. The response rate was 86%. Respondents were randomized on the fly while they took the survey. The number of completed surveys was equal to 63 (earlier pilot experiment), 70 (Arm 1), 65 (Arm 2), 76 (Arm 3), 68 (Arm 4), 80 (Arm 5).

Box plots of the number of friends listed by experimental arm are shown in Figure 7. Listing the preloaded number of close friends and family and prompting with the preloaded number elicited the largest number of friends listed (Arms 3 and 4). Listing only one column (first names; Arm 5) instead of three columns (first names, relationship, Hispanic; Arm 2) does not affect the number of respondents listed. Asking to list one person at a time as opposed to ten persons at a time (Arm 1) does not work well at all.

### 3. EXPERIMENT 2: Getting Respondents to Contact Their Friends

#### 3.1 Motivation

Experiment 1 explored how to get respondents to list more friends. Using the most successful design from Experiment 1 (asking for the preloaded number of friends; Experimental Arm 4), we started an RDS sample with five Hispanic seed respondents looking to recruit Hispanic friends. Four of five invited Hispanic respondents completed the survey. One respondent listed only one Hispanic friend; the others had at least four Hispanic friends each. We asked respondents to invite all of the listed Hispanic friends, so between them, respondents were supposed to contact 13 friends. However, in a follow-up survey the respondents indicated they tried to contact only two out of

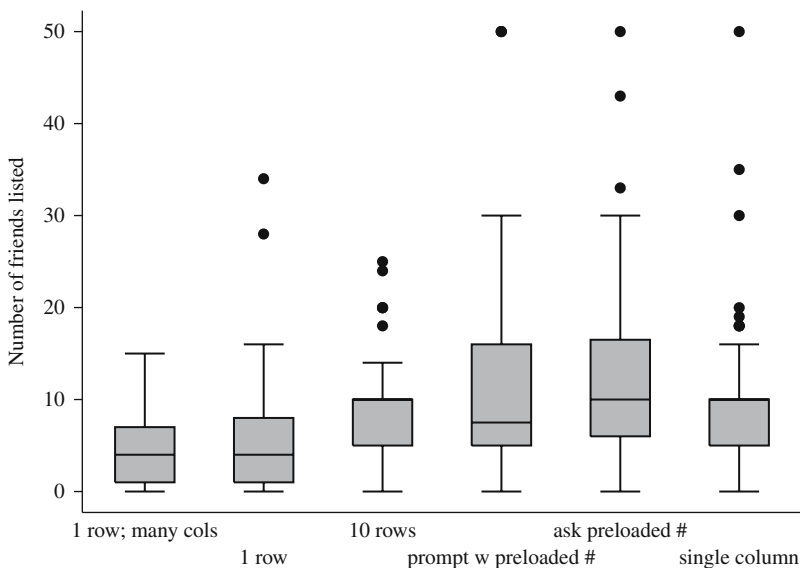


Fig. 7. Box plots of the number of close friends and family listed. (Note: “1 row; many cols” refers to an earlier pilot experiment; the five experimental arms are shown in order after that.)

13 respondents. Reasons for not contacting some respondents include “I forgot”, “I did not feel comfortable asking them”, and “He/she probably would not have participated”. The two respondents contacting one friend each thought both friends were unlikely to participate because the friend “Thinks it will be too complicated”. The referral codes were transferred by email (1) and phone (1). Overall, all four respondents thought passing on the referral code was very easy.

One of the two respondents contacted joined the panel; this person was not asked to refer additional friends because we felt we needed to conduct another experiment to test a different approach in the hope of finding a way to obtain a higher yield for the friend referrals. In conclusion, the four Hispanics who responded to the survey did not contact the friends they had listed.

### 3.2 Experiment 2

The purpose of the second experiment was to learn how to make it easier and more attractive for respondents to contact their friends. For all experimental arms, we gave respondents the option of receiving one prewritten email for each friend or one prewritten letter for each friend to be sent through the U.S. post office (Figure 8). Each email contained an explanation of the ALP, the referral code and a link to the webpage, as well as the respondent’s name in the subject line. By the end of Experiment 1, the ALP had recruited a Hispanic subsample using address-based sampling. This obviated the need to specifically recruit respondents of Hispanic ethnicity. In the second experiment we therefore no longer restricted new recruits to those of Hispanic ethnicity.

The experimental arms for Experiment 2 are displayed in Table 1. As before, we paid respondents \$5 for filling out the short referral survey (an earlier informal test revealed that

We would like to make it easy for you to forward the referral code. Please choose one or both of these options:

I would like to receive one email invitation for each referral code so that I can forward this

I would like to receive a letter for each referral code so that I can mail the invitation with the referral code already in there

Fig. 8. Option to receive one prewritten email (or postal mail letter) per referral in Experiment 2.

respondents were equally likely to respond with a \$5 payment as with a \$10 payment). We experimentally varied the amount paid for each successful respondent (\$15 vs. \$30), a sign-up bonus to the friend (\$20 vs. none), and whether the friends were selected at random by the computer or by the respondent, as is customary in RDS. Respondents were asked to nominate a maximum of four friends. With the \$30 incentive per successful referral, the respondent could earn a total of \$120 if all four friends filled out the first survey. The friends could earn the sign-up bonus as well as the regular survey payments (\$20 for each 30 minute survey). In all arms, ten friends at a time could be entered on each page.

*Results:* The response rates (86%-92%) and number of completes (44-50) are reported in Table 1. Respondents listed on average 13.5 friends (median 10, 1st quartile 6, 3rd quartile 20). The number of friends listed did not vary significantly by experimental arm (based on a Poisson regression on indicator variables for arms with robust standard errors). A histogram of the number of friends listed is shown in Figure 9. Respondents could list up to ten friends. Correspondingly, there is some heaping on the values 10, 20 and 30. The number of friends that could be listed was limited to 50 for programming reasons.

The average number of recruits differed substantially by arm (see bottom rows of Table 1). When respondents self-selected friends (Arm 4) and when both larger recruiter incentives and sign-up bonuses were provided (Arm 5), the number of recruits was significantly larger than in the control group. We conducted a logistic regression of the indicator variable “> = 1 referral” (vs. “0 referrals”) on four indicator variables for the five experimental arms. Coefficients for Arm 4 ( $p = 0.02$ ) and Arm 5 ( $p = 0.03$ ) differ significantly from the control (Arm 1).

For Arms 4 and 5, the ratio of friends recruited to the number of respondents recruiting is close to 1. A ratio greater than one would imply an increasing number of recruits from wave to wave and the referral chain would not die out. Using both larger recruiter payments and recruit sign-up bonuses (Arm 5) works much better than either one of these on its own (Arms 2 and 3). A recruit sign-up bonus by itself has little effect relative to the control group.

#### 4. Continued RDS Waves With Arms 4 and 5

We continued two separate RDS recruiting efforts corresponding to Arms 4 and 5 from Experiment 2. Monthly surveys were conducted on the second Wednesday of every month from March to October 2012. Any recruit who had responded in the previous month would be invited to recruit their friends. If the recruitment process took longer than a month for any one person, that person would simply be invited in the following month.



Table 1. Experiment 2 experimental arms, response rates, and friends recruited.

Experimental Arm	1	2	3	4	5
Name	control	high recruiter payment	respondent incentive	self-selected friends	arm 2 + arm 3
Payment to recruiter for filling out survey	5	5	5	5	5
Payment to recruiter for each successful respondent	15	30	15	15	30
Sign-up bonus payment for new respondent	0	0	20	0	20
Selection of recruits	computer	computer	computer	respondents	computer
Number invited	57	56	53	49	50
Number of respondents	49	50	46	45	44
Response rate	86%	89%	87%	92%	88%
Number of recruits	22	34	23	43	41
Ratio recruits per recruiter	0.45	0.68	0.50	0.96	0.93



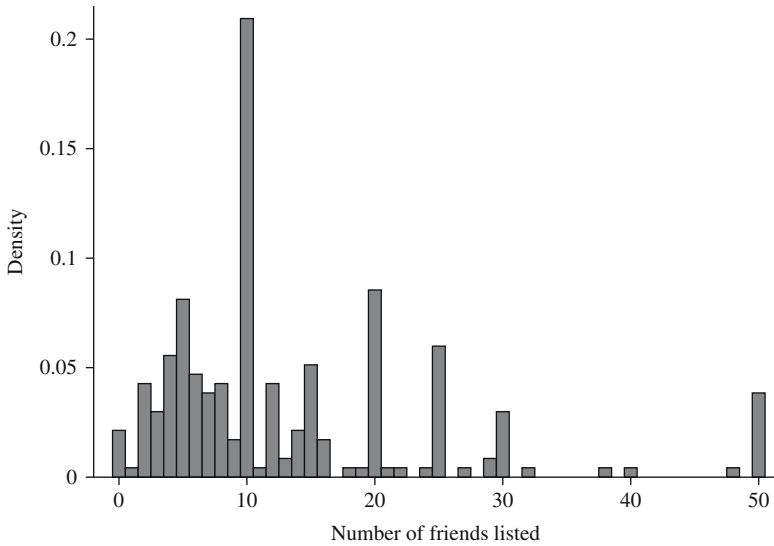


Fig. 9. Experiment 2: Histogram of the number of friends listed. Heaping effects for multiples of 10 and a ceiling effect (50) are visible.

Arm 4 (recruiter selects respondent) was only added after a two-month delay after it appeared that the enrolment in Arm 5 was slow.

Table 2 shows recruitment by wave (also referred to as depth) for both arms. Recruitment in Arm 5 decreased after Wave 1 and effectively died out in Wave 5. The recruiting strategy including incentives did not change after Wave 1. A number of respondents tried to circumvent the random recruiting assignment and gave the coupon code to self-selected friends. To avoid contamination of the ongoing experiment they were not allowed into the panel.

The response to Arm 4 was much stronger than that for Arm 5. The total number of recruits (excluding seeds) was more than twice as large. Recruiting was very slow, often taking longer than a month. The smaller number of recruits at Depth 5 does not reflect a

Table 2. Enrolment by Wave. Wave 0 refers to seed respondents; Wave 1 to friends of seed respondents; Wave 2 to friends of friends, etc.

Wave	Arm = 4	Arm = 5
0	45	44
1	46	47
2	50	19
3	63	18
4	48	10
5	16	1
Total	268	139
Total Recruits	223	95

Note: Wave 5 is not complete; only faster-recruiting chains reached this wave.

decline in enrollment but rather represents the fastest group of respondents. Some of the remaining respondents did not have the chance to reach Wave 5.

Demographics by Experimental Arm 4 (self-selected friends) vs. Arm 5 (friends selected at random) are shown in [Table 3](#). There are no statistically significant differences (based on  $\chi^2$  tests) between self-selected and randomly selected friends (recruits) with respect to gender, education, race/ethnicity, family income, or marital status. However, self-selected friends are on average eight years older than randomly selected friends (mean age = 45 vs. mean age = 37,  $p = 0.0001$  based on a t-test). There are also regional differences: Self-selected friends are more likely to live in the Midwest and less likely to live in the South or Northeast ( $p = 0.01$ ; based on a  $\chi^2$  test).

There is also a gender imbalance; roughly 70% of recruits were female. Recruits are somewhat more educated than the general public; in particular, very few recruits have less than a high school degree. Recruits are predominantly non-Hispanic whites (88%), even though there were Spanish versions of all surveys. Most of the remainder are (non-Hispanic) African Americans. Recruits have a wide range of family incomes. Thirty-nine percent of families have a household income of less than \$40,000; two thirds have a household income of less than \$75,000. Ninety percent of those under the age of 65 are working. Recruits are geographically spread across 40 U.S. states. All of the recruits had Internet access and we therefore did not have to provide Internet access in the form of a laptop and broadband for any of the recruits.

For completeness, [Table 3](#) provides the demographic characteristics of seed respondents by experimental arm. Comparisons of demographic characteristics of seeds and respondents must proceed with caution. When the equilibrium is reached, the demographic distribution of seed respondents and recruits are theoretically independent of each other. Large shifts between seed and respondent distributions only show the recruiting chain does not get stuck in any one category. Here we find that compared to the seed respondents, recruits are more often in the 18–29 age group (almost all ALP panel members are 18 years or older by design; though the occasional 17-year old is not rejected), they are more often “never married”, they are less often in the highest income category (particularly in Arm 4), and more often live in the southern part of the U.S.

## 5. Discussion

This is the first RDS study to attempt recruiting respondents throughout the U.S. rather than in individual U.S. cities. In fact, the only other RDS study we are aware of that recruited at a national level is the Vietnam study ([Bengtsson et al. 2012](#)). It was initially unclear whether respondents would spread across the United States or would remain in a confined region or state. With respondents in 40 different states, the overall geographic spread is good.

Respondents resist random selection. While incentives are important, the respondents' overriding desire was to choose whom they recruit. Respondents might have preferred self-selection to increase the probability of getting their own incentive, to channel money to specific friends and family, or because they were more comfortable contacting certain friends and family members. Self-selected respondents tend to be older than randomized

Table 3. Demographic composition of seed respondents and recruits by experimental arm.

	Arm 4 (self-selected friends)		Arm 5 (high payment + respondent incentive)	
	Seeds	Recruits	Seeds	Recruits
Gender	%	Freq.	%	Freq.
Male	40	55	29	26
Female	60	133	71	46
Age category	%		%	
<18	0	0	0	2
18-29	0	40	21	25
30-39	24	48	26	23
40-49	22	22	12	5
50-59	16	23	12	13
60-69	20	32	17	3
70 +	18	23	12	2
Education				
Less than high school	4	2	1	4
High school	11	33	18	12
Up to Bachelor	56	128	68	45
More than Bachelor	29	25	13	12
Race / Ethn.				
Non-Hispanic White	87	162	86	66
Non-Hispanic African American	7	20	11	5
Non-Hispanic Other	0	2	1	1
Hispanic	7	4	2	1
Family income				
<10,000	2	9	5	7
10,000-19,999	7	17	9	7
20,000-29,999	11	19	10	8
30,000-39,999	16	27	14	10
40,000-49,999	2	12	6	7

Table 3. Continued

	Arm 4 (self-selected friends)		Arm 5 (high payment + respondent incentive)	
	Seeds	Recruits	Seeds	Recruits
	%	Freq.	%	Freq.
				%
50,000-59,999	4	23	12	6
60,000-74,999	13	26	14	6
> 75,000	44	54	29	22
Northeast	22	10	5	10
Midwest	27	63	34	12
South	31	76	40	36
West	20	39	21	15
Married	64	119	63	40
Separated/widowed/divorced	29	28	15	11
Never married	7	41	22	22
				30

respondents, but except for increased recruiting in the Midwest, there are no discernible differences with respect to other demographic characteristics.

The slow speed of recruiting remains a challenge. A study in a single college finished recruiting in a single weekend (Wejnert and Heckathorn 2008). Web recruiting for the drug and alcohol study at a single university (Bauermeister et al. 2012) concluded after 2.5 months. Web recruiting for the study about men who have sex with men in Vietnam (Bengtsson et al. 2012) took about two months. Our recruiting effort may have been slower for several reasons: 1) We had a broader target population that was more difficult to incentivize. 2) We only mailed invitations out once a month, thereby possibly signaling a lack of urgency. The fastest Internet RDS study was automated, requiring intervention only to end recruiting and incentive payment (Wejnert and Heckathorn 2008). 3) There was no focus on a specific topic. Respondents would become ALP panel members and would be asked to participate in a variety of surveys. The ongoing recruiting effort was only one of several surveys they were asked to participate in.

It is unclear what caused the gender imbalance (Table 3). We have observed a similar gender imbalance in international face-to-face surveys ([www.itcproject.org](http://www.itcproject.org)) where women are thought to be more likely at home. Slight gender imbalances were also reported for most countries in the European Community Household panel (Behr et al. 2005) even after Wave 1 of the panel.

Once enrolled, the response rate of RDS panel members is similar to that of regular panel members recruited in the same time period. (Long-standing panel members tend to have higher response rates.) Response rates of RDS panel members ranged from 79–86% in three large surveys conducted between January and April 2013. Response rates among regular panel members recruited since July 1, 2011 for these surveys ranged from 80% to 84%.

It appears that this approach reaches low-income populations, those not working, and the elderly. It works less well for reaching racial/ethnic minorities other than African Americans, those with less than high school education or those without access to the Internet. This replicates the demographics among the seed respondents. An alternative interpretation therefore might be that the approach thus far has been unable to reach hard-to-reach populations not already represented among the seed respondents. Any Internet implementation of RDS will require careful pilot testing and experimentation. Importantly, we discovered respondents' preloaded number of friends helped to generate a longer list of friends. Other studies also had difficulties in "calibrating" Internet-based RDS (Bauermeister et al. 2012); others go unreported in the literature because they failed. While our self-selected friends chain (Arm 4) would have continued past Wave 5, the number of recruits was certainly not rising exponentially.

Our study has a number of limitations. The first may have to do with our name generator. First introduced by Laumann (Laumann 1966), name generators have become an active area of research (Marin and Hampton 2007). Our name-generating question asked for "friends", whereas many studies ask for both "friends and acquaintances". The term "friend" alone for name generating has been shown to be interpreted differently by different socioeconomic strata (Burt 1983). Therefore our name generating question may have introduced some bias, and consequently the computer-generated random recruitment was also conducted on the potentially biased list of names. Most studies work with specific

subpopulations (e.g., “Men who have sex with men”). Because we were targeting a general population we consciously decided not to include acquaintances because we felt the number of acquaintances might have been too large. Either way, the difficulties in recruiting were not due to a lack of friends listed.

Second, impersonation and duplication of respondents, while unlikely, cannot be ruled out. Duplication was probably less likely than impersonation as respondents had to provide contact information and an address in order to receive their respondent payment checks and duplicate addresses would have been discovered.

A number of issues may have affected respondents’ decision to cooperate. Recruiting respondents to an Internet panel is harder than recruiting respondents to a single survey, and the study must be seen in this context. Words such as “referral code” added to the complexity of a questionnaire and also increased respondents’ burden. Further, asking respondents to list their friends added another step to the recruiting process. While this was needed to choose friends at random, each step increases respondents’ burden and this may have contributed to the outcome.

RDS recruiting into an Internet panel also needs to be seen through the opportunity that the ALP affords. The ALP is a well-established open-access Internet panel. Open access implies that surveys cover a wide variety of topics. Survey length varies but is typically of the order of 15–30 minutes. Survey frequency also varies; 1–2 surveys a month is typical. Respondents are surely motivated by the payments structure (US \$20 for every 30 minutes of interview time). Relatively well-paid long-term panel members should contribute to successful recruitment in at least the first wave. The drop-off observed after the first wave for arm 5 (recruiter and recruit payment) might be explained by whether the recruiter is a long term panel member. Wave 0 (seed) recruiters were long-term panel members, whereas Wave 1 recruiters were not.

Finally, encouraged by an anonymous referee, we provide recommendations for the implementation of Internet RDS studies. First, experiment as much as possible. Likely fine tuning is required. Second, automate the recruiting process as much as possible; this will help to get back to potential recruits as soon as possible. Third, select seeds with great care to encourage recruiting. Fourth, allow self-selection of friends. While this is less than desirable, it is a strategy that can be used for pragmatic reasons in order to avoid the recruiting chains dying out. Fifth, for a general population, choose a high incentive/ payment. Sixth, if necessary consider letting respondents invite more friends. One study (Bauermeister et al. 2012) increased the number of friends to ten. To control costs this study paid only for the first five friends who respond. Alternatively, one can argue paying for the last five friends who respond is preferable because it motivates the recruiter to get all invited friends to participate – but whether this works in practice requires empirical study. Again, this is a pragmatic suggestion where necessary and not our first choice.

## **Appendix**

### 1) Numerical Question

“How many close friends or family members would you say you have? By close, we mean friends or family members you talk to or write to (via letter, email, text message,

Facebook, etc.) regularly. Please do not include people who live in your household. Please only consider people 18 years or older who live in the United States.”

## 2) Friends question

“Please list all the close friends or family members you see, talk to or write to (via letter, email, text message, Facebook, etc.) regularly. Please **do not** include people who live in your household. Please only consider people 18 years or older who live in the United States. You only need to provide their first name, nickname or initials.”

## 6. References

- Arfken, C.L., Ahmed, S., and Abu-Ras, W. (2013). Respondent-Driven Sampling of Muslim Undergraduate U.S. College Students and Alcohol Use: Pilot study. *Social Psychiatry and Psychiatric Epidemiology*, 48, 945–953. DOI: <http://www.dx.doi.org/10.1007/s00127-012-0588-4>.
- Bauermeister, J.A., Zimmerman, M.A., Johns, M.M., Glowacki, P., Stoddard, S., and Volz, E. (2012). Innovative Recruitment Using Online Networks: Lessons Learned from an Online Study of Alcohol and Other Drug Use Utilizing a Web-Based, Respondent-Driven Sampling (webrds) Strategy. *Journal of Studies on Alcohol and Drugs*, 73, 834–838.
- Behr, A., Bellgardt, E., and Rendtel, U. (2005). Extent and Determinants of Panel Attrition in the European Community Household Panel. *European Sociological Review*, 21, 489–512. DOI: <http://www.dx.doi.org/10.1093/esr/jci037>.
- Bengtsson, L., Lu, X., Nguyen, Q.C., Camitz, M., Hoang, N.L., Liljeros, F. and Thorron, A. (2012). Implementation of Web-Based Respondent-Driven Sampling Among Men Who Have Sex With Men in Vietnam. arXiv preprint arXiv:1206.1739. DOI: <http://www.dx.doi.org/10.1371/journal.pone.0049417>.
- Berchenko, Y. and Frost, S.D. (2011). Capture-Recapture Methods and Respondent-Driven Sampling: Their Potential and Limitations. *Sexually Transmitted Infections*, 87, 267–268. DOI: <http://www.dx.doi.org/10.1136/sti.2011.049171>.
- Berry, S., Bogdon, T., Brown, R., Corey, C., Hickey, S., Hill, L., and Schell, T. (2010). Survey of Gay Service Members. In *Sexual Orientation and U.S. Military Personnel Policy*, B.D. Rostker, S. Hosek, and J.D. Winkler (eds). Santa Monica, CA: RAND Corporation.
- Burt, R.S. (1983). Distinguishing Relational Contents. *Applied Network Analysis: A Methodological Introduction*, Ronald S. Burt, Michael J. Minor, and Richard D. Alba (eds). Thousand Oaks: Sage Publications, 35–74.
- Heckathorn, D. (1997). Respondent-Driven Sampling: A New Approach to the Study of Hidden Populations. *Social Problems*, 44, 174–199. DOI: <http://www.dx.doi.org/10.1525/sp.1997.44.2.03x0221m>.
- Heckathorn, D. (2002). Respondent-Driven Sampling ii: Deriving Valid Population Estimates from Chain-Referral Samples of Hidden Populations. *Social Problems*, 49, 11–34, DOI: <http://www.dx.doi.org/10.1525/sp.2002.49.1.11>.
- Heckathorn, D. (2007). Extensions of Respondent-Driven Sampling: Analyzing Continuous Variables and Controlling for Differential Recruitment. *Sociological*

- Methodology, 37, 151–207, DOI: <http://www.dx.doi.org/10.1111/j.1467-9531.2007.00188.x>.
- Heckathorn, D., Semaan, S., Broadhead, R., and Hughes, J. (2002). Extensions of Respondent-Driven Sampling: A New Approach to the Study of Injection Drug Users Aged 18–25. *AIDS and Behavior*, 6, 55–67, DOI: <http://www.dx.doi.org/10.1023/A:1014528612685>.
- Laumann, E.O. (1966). *Prestige and Association in an Urban Community: An Analysis of an Urban Stratification System*. New York: Bobbs-Merrill.
- Marin, A. and Hampton, K.N. (2007). Simplifying the Personal Network Name Generator Alternatives to Traditional Multiple and Single Name Generators. *Field Methods*, 19, 163–193, DOI: <http://www.dx.doi.org/10.1177/1525822X06298588>.
- Salganik, M. and Heckathorn, D. (2004). Sampling and Estimation in Hidden Populations Using Respondent-Driven Sampling. *Sociological Methodology*, 34, 193, DOI: <http://www.dx.doi.org/10.1111/j.0081-1750.2004.00152.x>.
- Schonlau, M. and Liebau, E. (2012). Respondent-Driven Sampling. *Stata Journal*, 12, 72–93.
- Wejnert, C. (2009). An Empirical Test of Respondent-Driven Sampling: Point Estimates, Variance, Degree Measures, and Out-of-Equilibrium Data. *Sociological Methodology*, 39, 73–116, DOI: <http://www.dx.doi.org/10.1111/j.1467-9531.2009.01216.x>.
- Wejnert, C. (2010). Social Network Analysis With Respondent-Driven Sampling Data: A Study of Racial Integration on Campus. *Social Networks*, 32, 112–124, DOI: <http://www.dx.doi.org/10.1016/j.socnet.2009.09.002>.
- Wejnert, C. and Heckathorn, D. (2008). Web-Based Network Sampling: Efficiency and Efficacy of Respondent-Driven Sampling for Online Research. *Sociological Methods & Research*, 37, 105–134, DOI: <http://www.dx.doi.org/10.1177/0049124108318333>.

Received February 2013

Revised October 2013

Accepted November 2013



# Locating Longitudinal Respondents After a 50-Year Hiatus

*Celeste Stone<sup>1</sup>, Leslie Scott<sup>1</sup>, Danielle Battle<sup>1</sup>, and Patricia Maher<sup>2</sup>*

Many longitudinal and follow-up studies face a common challenge: locating study participants. This study examines the extent to which a geographically dispersed subsample of participants can be relocated after 37 to 51 years of noncontact. Relying mostly on commercially available databases and administrative records, the 2011–12 Project Talent Follow-up Pilot Study (PTPS12) located nearly 85 percent of the original sample members, many of whom had not participated in the study since 1960. This study uses data collected in the base year to examine which subpopulations were the hardest to find after this extended hiatus. The results indicate that females were located at significantly lower rates than males. As expected, sample members with lower cognitive abilities were among the hardest-to-reach subpopulations. We next evaluate the extent to which biases introduced during the tracking phase can be minimized by using the multivariate chi-square automatic interaction detection (CHAID) technique to calculate tracking loss adjustments. Unlike a 1995 study that found that these adjustments reduced statistical biases among its sample of located females, our results suggest that statistical adjustments were not as effective in PTPS12, where many participants had not been contacted in nearly 50 years and the tracking rates varied so greatly across subgroups.

*Key words:* Respondent tracking; attrition bias; panel reengagement.

## 1. Introduction

While most longitudinal studies, sometimes also called panel or follow-up studies, are prospectively planned and have a definitive end date, there has been a recent resurgence in reconstituting “dormant” longitudinal or even cross-sectional studies (e.g., [Haggerty et al. 2008](#); [Hampson et al. 2001](#); [Hauser 2005](#); [Kimmel and Miller 2008](#); [Ortiz and Godinez Ballon 2007](#)). One reason for this resurgence is the higher costs associated with conducting new longitudinal studies relative to repurposing or continuing preexisting ones. Recent advances in technology provide relatively inexpensive methods for relocating sample

<sup>1</sup> American Institutes for Research’s Center for Survey Methods, 1000 Thomas Jefferson Street NW, Washington, DC 20007. Email: [cstone@air.org](mailto:cstone@air.org), [lscott@air.org](mailto:lscott@air.org), and [dbattle@air.org](mailto:dbattle@air.org)

<sup>2</sup> Institute for Social Research, University of Michigan, 426 Thompson Street, Ann Arbor, MI 48106. Email: [pmaher@umich.edu](mailto:pmaher@umich.edu)

**Acknowledgments:** This work was supported through grants provided by the National Institutes of Health’s National Institute on Aging [P30 AG012846-17S1 and U01 AG009740-21S2] and through research reinvestment funds provided by the American Institutes for Research. The authors would like to thank the following individuals for their contributions, guidance, and valuable suggestions to improve the content and quality of this article: Deanna Lyter Achorn, Martin Hahn, Sandy Eyster, and Samantha Neiman from the American Institutes for Research (AIR). We would also like to thank David Weir and Jenny Bandyk from the University of Michigan and Susan Lapham (AIR) for their contributions to the design and execution of this pilot study, as well as staff from the tracking unit of the University of Michigan’s Survey Research Operations (SRO) for their tireless efforts in locating and engaging study participants. Finally, we are indebted to Joel Devonshire from SRO, who helped compile and produce the fruitful paradata needed to conduct this research.

members after an unplanned length of noncontact. These advances – which include the compilation of information into low-cost commercial databases – have led social scientists to consider how existing data sources and samples can be reutilized. Researchers use these studies to broaden their understanding of complex causal relationships and social phenomena throughout the life span.

The feasibility of reconstituting a study lies first and foremost in the successful location (or relocation, tracing, or tracking; these terms are used interchangeably) of study sample members. It becomes particularly difficult to relocate such persons after a long period of noncontact. Successful location depends on several factors, such as the length of time since last contact, the amount of information available to locate sample members, and the budget and resources available for locating activities.

In addition, individual characteristics are known to affect tracking success. Age, lifestyle, socioeconomic status (SES), employment situation, family circumstances, geographic region, and urbanicity (i.e., urban, rural, suburban) play a role in a person's mobility and as a result can affect how difficult it is to locate someone after a long period of noncontact. Other behaviors and characteristics affect the extent to which individuals are “politically, socially, or economically engaged in their new community” (Couper and Ofstedal 2009, p.187). Lower levels of engagement decrease the likelihood that individuals will be accessible through the lower-cost commercial databases. (See Becker et al. 2012 for a discussion of contemporary low-cost and high-cost tracking databases and methods.)

These systematic differences are problematic for two reasons. First, studies with high proportions of hard-to-find individuals must be prepared to devote more time, effort, and resources to tracking activities. Second, and perhaps more problematic, is the type of bias that results when the individual behaviors and characteristics correlated with tracking propensity are also correlated with the outcomes of interest.

This article contributes to the limited but growing research on locating study sample members after a long period of noncontact (e.g., Call et al. 1982; Clarridge et al. 1978; Haggerty et al. 2008; Hampson et al. 2001; Hauser 2005; Kimmel and Miller 2008; Masson et al. 2013; Meehan et al. 2009; Ortiz and Godinez Ballon 2007; Strawn et al. 2007) by evaluating the use of widely available and used locating methods. Specifically, this study used commercial databases, which consolidate information from a variety of sources such as credit histories, voting records, property records, and voluntary registries such as the National Change of Address Register, to locate sample members. Because of the tracking methods employed, the results of this study are of practical use to a wide variety of other studies – large or small, national or regional, longitudinal or cross-sectional.

This article focuses specifically on locating the original participants of a study conducted in the United States (U.S.). Though not necessarily different from the mobility rates and frequencies for residents of other countries, the combination of Americans' geographic dispersion and mobility makes them difficult to locate in general, and particularly difficult to locate after a long period of time. If anything, this study presents a worst-case scenario for applying scalable tracking methods to other longitudinal studies worldwide. The U.S. lacks population registers common in Scandinavia and spans a large geographic area. Americans are relatively mobile. According to a recent report by the

U.S. Census Bureau (Ren 2011) approximately 50 percent of American residents age 55 and older resided in a state other than their state of birth in 2010, and data from the 2011 American Community Survey suggest that U.S. residents move as many as eleven times in a lifetime, with the majority of moves occurring between the ages of 18 and 45.

## 2. Background

Published research on locating survey sample members after a long period of noncontact is limited, but not new. The Wisconsin Longitudinal Study (WLS) was one of the first surveys that attempted to do so. In 1975, researchers used information from printed telephone directories, high schools, post offices, military locator services, employers/licensing associations, college alumni associations, parents, neighbors, siblings, and high school classmates to successfully locate 97 percent of a 10,317-person subsample of the original sample. The sample members, all around 35 years of age, had not been contacted in 10 to 17 years (Hauser 2005). Most sample members were located through their parents. WLS's next locating effort, conducted in 1992, used the then newly developed technologies of CD-ROM-based telephone and address directories, online credit agency databases, and high school reunion booklets to locate respondents, as well as their parents, siblings, classmates, and neighbors. Again, the effort achieved a 97 percent location rate (Hauser 2005).

Beginning in 1998, Hampson et al. (2001) sought to find and recontact nearly 2,000 sample members last surveyed in elementary school during the years 1959 through 1967 – with no contact in the intervening 32 to 40 years. Moreover, the study was not originally designed as a longitudinal study. Relying on the limited baseline participant information that was available, Hampson and colleagues located 75 percent of sample members by searching through printed and CD-ROM-based telephone and address directories, newspaper announcements of sample member milestone events (e.g., weddings, children's births), high school websites, Department of Education records, and Internet databases, to name a few.

Similarly, after 35 years of noncontact, the Mexican American Study Project located 79 percent of its original 1,000-plus household sample members between 18 and over 50 years old last contacted in 1965 and 1966 (Ortiz and Godinez Ballon 2007). The Mexican American Study Project also was not planned as a longitudinal study. Like Hampson and colleagues, Ortiz and Ballon employed a mix of old and new locating strategies, including manual searches of printed telephone directories and public records and computerized searches of "people-finder" databases, property and voter registration records, and the Internet database Missing Links (one of the first such people-finder databases on the Internet).

Each of these studies included relatively homogeneous samples in specific places with limited geographic dispersion, which can make tracking participants both easier and more cost efficient. The WLS was a study of twelfth graders attending high school in Wisconsin in 1957. The Hampson et al. (2001) relocating effort was limited to persons who were initially assessed as elementary students attending school on one of two Hawaiian Islands. Ortiz and Godinez Ballon (2007) focused on finding Mexican Americans in Los Angeles, California, and San Antonio, Texas.

As summarized by [Calderwood \(2012\)](#), studies such as these have two advantages over studies that attempt to recontact sample members of large-scale, national studies. First, they can make better use of social networks of friends, siblings, and neighbors, because there is a greater likelihood that those individuals will also be in the study. Second, greater geographic specificity means it is more feasible to use “localized” methods. By contacting local post offices, governmental offices, and other community groups, these studies have greater success in locating and promoting continued study participation.

Though some national studies have attempted to locate sample members after a long hiatus, only a few have published comprehensive, empirical information about the results of their efforts (see [Couper and Ofstedal 2009](#)). After over ten years of noncontact, the Longitudinal Study of American Youth (LSAY) located nearly 94 percent of a nationally representative sample of persons aged 32 to 35 who were first contacted when in seventh or 10th grade in 1987 ([Kimmel and Miller 2008](#)). The Midlife Development in the United States (MIDUS) study began in 1995 with over 7,000 individuals aged 25 to 74 in the U.S. In 2004–06, staff located approximately 90 percent of its sample members ([Ryff et al. 2006](#)). The Longitudinal Study of Adult Learning (LSAL) tracked on average 90 percent of its participants when the lag between interviews was one year (Waves 1 through 3); tracking rates dropped to 86 to 87 percent when the lag between contact was extended to two years in Waves 4 and 5 ([Strawn et al. 2007](#)). The National Longitudinal Study of Adolescent Health (Add Health) began in 1994 as a longitudinal study of adolescents aged 11 to 18. The study located 87 percent of the original sample in the unplanned third wave of data collection. Participants had not been contacted in 5 to 8 years. The study improved its locating rate to 92 percent in the fourth wave six years later. [Meehan et al. \(2009\)](#) attributed this improvement to both the reduced mobility of the sample members – who were between 24 and 32 years of age at Wave 4 – as well as improvements in tracking planning and implementation.

As noted by [Couper and Ofstedal \(2009\)](#), few studies have explored the factors that affect location propensity or have used multivariate approaches to model the tracking process. Most studies on attrition in longitudinal studies have focused on *total attrition* – the loss of sample members owing to both nonlocation and nonresponse. Only a few have disentangled nonlocation from noncooperation (e.g., [Cotter et al. 2002](#); [Hauser 2005](#); [Radler and Ryff 2010](#)). Across these studies, there are very few commonalities in the populations under study, study designs, length of time between waves, variables available for analysis, and analytical approaches. What have we learned? Certain subgroups are more difficult and expensive to find because they require more complex tracking steps: females, minorities, and individuals with lower educational attainment, higher rates of financial instability, and criminal behavior or substance abuse ([Andresen et al. 2008](#); [Cotter et al. 2005](#); [Cottler et al. 1996](#); [Haggerty et al. 2008](#); [Iannacchione 2003](#); [Jessor and Jessor 1977](#); [Passetti et al. 2000](#); [Ribisl et al. 1996](#); [Stouthamer-Loeber and van Kammen 1995](#)).

Our review found only one study ([Iannacchione 2003](#)) that researched statistical methods for reducing tracking-related biases. Using the National Health Interview Survey (NHIS), Iannacchione used sequential logistic regression models to first explore the factors related to location propensity and next compute tracking-related weighting adjustments. The NHIS wave under examination had a locating rate of 95 percent. The study found that these adjustments ultimately preserved the “location-adjusted weighted

means” among females in the study. However, we do not know if this method could be applied as effectively to studies with a lower locating rates, because they may be unable to locate a sufficient number of key groups to allow for weighting adjustments or do not have the data to allow for such adjustments.

This study contributes to the existing research on locating sample members after a long period of noncontact. It extends what is currently known by examining a longer period of noncontact – 37 to 51 years – and by investigating the effectiveness of applying a weighted adjustment to minimize tracking-related biases in such a study. Three research questions are addressed:

1. To what extent can geographically dispersed sample members of a national study be located after a very long hiatus using tracking methods that rely primarily on lower-cost commercial databases?
2. Using relatively low-cost methods to locate sample members after a lengthy hiatus, which subpopulations are the hardest to find?
3. To what extent can biases introduced through tracking failure be minimized by employing statistical tracking-loss adjustments?

### 3. Study Design and Methods

#### 3.1. Sample and Dataset

We review findings from a 2011–12 follow-up of a subsample of participants of the Project Talent longitudinal study, which began as a nationally representative sample of high school students in 1960. After three rounds of data collection, the study went on hiatus in 1974. In 2011, researchers with the American Institutes for Research (AIR), the University of Michigan’s Survey Research Center (SRC), and the University of Michigan’s Health and Retirement Study (HRS) began planning a follow-up study of the original Project Talent participants.

The large-scale pilot test was designed to gauge the feasibility of locating and persuading participants of the original 1960 study to participate in a follow-up 50 years after the initial base-year survey. A subsample of 4,879 participants was randomly selected from a ten percent random subsample of the original 1960 schools. This two-stage sample design eased the operational burden of cleaning contact information that was originally captured by some of the earliest optical scanning machines developed in the 1960s and 1970s. In addition, the sample design allowed for the possibility of using schools and classmates, who began holding their 50-year class reunions in 2010, to locate sampled individuals. Most individuals sampled for the 2011–12 Project Talent Follow-up Pilot Study (PTPS12) were between 67 and 70 years old when tracking activities began.

#### 3.2. Study Design and Tracking Methods

The follow-up began in June 2011 with a tracking phase. Information available in the Project Talent historical records was used to identify sample members’ most recent addresses. These records contain much of the information needed for locating activities, though most of it is outdated: first, middle, and last names as of 1960; updated names and

addresses through 1978; date of birth; and 1960 school name and location. The records also include Social Security Numbers (SSNs), which were collected in the year 1 and 5 follow-ups and are available for approximately 50 percent of Project Talent sample members. In addition, AIR began compiling more current contact information through outreach activities – including attendance of class reunions and participant registration through the Project Talent Website (<http://projecttalent.org>) – in 2010. Contact information collected through these outreach efforts was available for approximately five percent of sampled individuals at the start of the tracking phase.

The pilot study used a tiered approach for its retrospective tracking activities – an approach used by large-scale, national longitudinal studies such as Add Health and the Beginning Postsecondary Students Longitudinal Study (BPS) (see [Meehan et al. 2009](#) and [Wine et al. 2011](#), respectively). Two tracking methods were used to locate pilot study sample members: batch tracking and interactive tracking. *Batch tracking* refers to automated processes in which tracking vendors use client-provided inputs (e.g., first name, last name, date of birth) and hard logic algorithms to identify potential matches across multiple databases and return current contact information for those matches. Because these services process numerous cases at one time, they provide researchers with relatively quick access to address confirmations or updates at a low per-case cost and low operational burden. However, cases not matched in batch tracking will require additional effort before they can be located.

This pilot study used commercially available batch tracking services provided by LexisNexis to obtain recent telephone numbers, addresses, and vital status information on every sample member. Given the age of the target population, an important part of the tracking activities involved determining a person's vital status. Thus, the PTPS12 batch-tracking protocol ascertained sample members' vital status by also matching cases to the Social Security Administration's Death Master File (DMF) and the National Center for Health Statistics' National Death Index (NDI), which at the time of this study contained U.S. death record information for the years 1979 through 2009. Cases not located through batch tracking were sent to *interactive tracking*, where a trained staff member in SRC's centralized data collection unit reviewed each case on an individual basis and took a variety of steps to locate the sample 1 member using both free and proprietary web-based databases. SRC also carried out telephone verification for a small group of cases for which multiple recent addresses were found, conflicting information was received, or there was uncertainty that the correct person was located. Given the additional labor and costs associated with interactive tracking, only those cases not located through batch tracking were followed up using this method.

Finally, contact information obtained passively through ongoing, study-wide outreach activities was utilized as necessary. The ongoing outreach activities (which included attending or sending information to 50-year class reunions, sending press releases to local media outlets, and developing a participant registration page on the study's website) were not part of the formal pilot study tracking protocols and were completely independent of the PTPS12 batch and interactive tracking efforts. This approach relies largely on voluntary submission or active engagement on the part of the sample member. However, the pilot study made use of the information collected through the outreach activities, particularly where it allowed for contact with otherwise unlocated sample members. The tracking activities employed for the pilot study are summarized in [Table 1](#).

Table 1. 2011–12 Project Talent Follow-up Pilot Study (PTFS12) tracking activities

Tracking activity	Pilot test sample	Information collected (source)	
		Vital status	Addresses and/or telephones
Batch tracking	All pilot sample members	LexisNexis; Death Master File; National Death Index	LexisNexis; National Change of Address registry
Interactive tracking	Pilot sample members not found in batch tracking	Consolidated, proprietary tracking databases; verification of participant identity and/or address	web searches; telephone and/or address
Informal tracking	Nontargeted	Class lists, contacts	Reunions, class lists, web sign-ups, requests for 1960 test scores, contacts



### 3.3. Measures and Analysis Methods

The present article appraises the success of the tiered approach and lower-cost tracking methods used in PTPS12, which could be applied to other studies in the U.S. and elsewhere. Research question 1 uses this pilot study to assess the extent to which members of a geographically dispersed sample can be located after a very long hiatus. Based on the results from previous small-scale feasibility studies conducted for Project Talent, we expected that the tracking phase would identify twelve percent of the pilot sample as deceased and locate addresses for 74 percent of the pilot sample, for a combined tracking rate of 86 percent.

Research question 2 considers which subpopulations are the hardest to find. Because data collected in the base year (1960) were available for all cases sampled for PTPS12, base-year data were used to examine the characteristics associated with tracking success, which is defined as the identification of a sample member's current address or vital status. Previous studies of nonresponse in Project Talent follow-ups have shown that sample members in certain subgroups – particularly those in the lower quintiles for cognitive ability and family SES relative to their peers, as well as minorities and students attending high-minority schools – were less likely to have participated in at least one of the follow-ups, when key information like SSNs and name changes would have been captured (Flanagan and Cooley 1966; Orr 1963; Rossi et al. 1976). Therefore, these variables are not only indicators of subgroups who may be systematically more difficult to locate simply because the records contain less information for tracking, but are also associated with individual characteristics (e.g., educational attainment, financial instability) known from other studies to be related to lower tracking propensities. We hypothesized that the same subgroups would be difficult to locate in 2011. In addition, other studies on aging (e.g., Gale et al. 2012; Anstey et al. 2001; Ritchie and Bates 2013; Wilson et al. 2009; and Crowe et al. 2013) have shown that early life health, SES, personality, and cognitive indicators are often correlated with key aging outcomes, such as health, cognitive function, and well-being. Therefore, these variables provide good indicators of potential biases in survey estimates that may be introduced during the tracking phase of the present study.

The analyses focused on 17 measures of sample characteristics, including sex, family SES, academic performance, cognitive aptitude, three measures of early life health (health in past three years; health in first ten years; number of days sick in bed in past year), and ten measures of personality (sociability; social sensitivity; impulsiveness; vigor; calmness; tidiness; culture; leadership; self-confidence; and mature personality). The SES measure reflects students' reporting and perceptions of their family environment (e.g., number of books in home, number of rooms in home, student-reported financial well-being). Cognitive aptitude was measured using the general academic aptitude variable, a weighted composite that combines results from three informational tests (mathematics and vocabulary I and II) and six cognitive aptitude/ability tests (English, reading comprehension, creativity, abstract reasoning, and mathematics I and II). For more information on these and other measures, see Wise et al. (1979).

One shortcoming of this dataset is that individual measures of race/ethnicity were not captured in the baseline data collection in 1960. In lieu of individual measures, we included a categorical school-level measure reflecting the minority composition of each



sample member's 1960 school. Given that individual mobility and the stability of social networks over time vary by environmental factors, we also included four control variables reflecting 1960 school and regional attributes: population size of the surrounding community in 1960, Census region, school type (i.e., private versus public), and building type (junior high school versus senior high school). We also included a categorical measure of the 1960 grade cohort to control for the inherent selection bias related to school attrition, as well as to the length of time between the 1960 collection and subsequent follow-ups, which varied by grade cohort. We expected to find that even after controlling for individual characteristics such as sex, family SES, and school attributes, individuals with lower cognitive scores relative to their peers would be more difficult to find. We also anticipated that individuals with certain personality characteristics (e.g., those who were less mature, more impulsive, and less calm) when measured as adolescents would be more mobile and also less tied to local communities and therefore more difficult to locate.

Finally, research question 3 investigates the extent to which nonresponse adjustment techniques can be used and applied separately to the tracking phase to reduce attrition biases introduced solely as a result of tracking loss. A multivariate chi-square automatic interaction detection (CHAID) technique (see Kass 1980), the same method used by Iannacchione (2003), was used to calculate the tracking adjustment weighting classes. The CHAID algorithm was used to identify the variables that were the most significant predictors of being located (the dependent variable) by calculating the chi-square measure of association between the dependent and each independent variable and then using this information to successively partition the sample into subsets that are homogeneous in terms of tracking status. The predictor variable with the highest significance level for the chi-square test was used to split the sample into groups. This process was repeated for each of the predictor variables until there were no further logical splits or until there were too few observations for further splitting. The result is a tree-like structure that groups observations in the dataset into cells (or nodes) that have the greatest discrimination with respect to response status. For the purpose of this CHAID analysis, only variables having a Bonferroni-adjusted  $p$  value of less than or equal to .05 were eligible for segmentation and cells were required to have at least 50 observations. The last partitions define the tracking adjustment weighting cells to calculate the tracking adjustment factor given by each weighting cell ( $i$ ):

$$TAF_i = \frac{WLC_i + WNLC_i}{WLC_i},$$

where

The *tracking adjustment factor* ( $TAF_i$ ) is the weighted ratio of the total sampled individuals to the total located individuals for cell  $i$ , and

The *weighted located count* ( $WLC_i$ ) is the base-weighted located count for cell  $i$  and the *weighted nonlocated count* ( $WNLC_i$ ) is the base-weighted not located count for cell  $i$ .

The final *tracking adjusted weight* is the product of the TAF and the PTPS12 base weight. The base weight is the product of the student-level 1960 Project Talent base weight and the probability of selection for the PTPS12 sample.

The CHAID algorithm used the same candidate variables used as predictor and control variables in the logistic regression analysis. In addition, the CHAID algorithm included four additional categorical variables that were either of substantive interest to the study or were believed to be related to location propensity: 1960 school size, proximity to largest cities as of the 1950 census, 1960 self-reported number of doctor visits in the past year, and a flag indicating if an SSN was provided in one of the previous Project Talent follow-ups.

## 4. Results

### 4.1. Tracking Rates

At the end of the tracking phase, the project team had identified 14.7 percent of the pilot sample members as deceased and found updated address information for 71.5 percent of the pilot sample members – yielding an initial locating rate of 86.2 percent. These rates are in line with the expected locating rates for this study. The initial locating rate is nearly ten percentage points higher than the locating rates reported by [Hampson et al. \(2001\)](#) and [Ortiz and Godinez Ballon \(2007\)](#), who faced a similar challenge of locating participants after an extended hiatus but with little geographic dispersion. However, this locating rate is between three and nine percentage points lower than those reported by LSAY, MIDUS, LSAL, and Add Health, which used similar tracking methods, but with only two to ten years of elapsed time since the last contact.

Approximately 71 percent of sample members were located at the batch-tracking stage: about 87 percent of all decedents and 82 percent of all located, presumed living sample members were found at this tracking stage. Interactive tracking increased the overall tracking rate by nearly 14 percentage points. Fewer than two percent of all cases were located through other means (e.g., outreach activities), or had multiple tracking sources and could not be categorized.

Survey materials were mailed to all presumed surviving cases for whom an address was found ( $n = 3,462$ ). To evaluate the accuracy of the address information obtained during the tracking phase, we measured the extent to which we were able to verify that the correct person was located for the 3,462 mailing cases. A case was considered to be verified if the name, school, and date of birth collected for a located individual as part of the tracking and data collection activities matched the corresponding information available in the study's historical files. About 71 percent of the contacted individuals either returned a survey or were verified to be the correct person through some type of direct contact logged in the Project Talent participant database. Only about two percent of the mailing cases were classified as erroneous matches because there was some indication that tracking had obtained either the wrong address (e.g., an outdated address) or had located the wrong person, and interviewers were unable to find the correct address during the data collection period. About one percent of the mailed cases were later identified as deceased. The analysis was unable to verify that the correct person was located for the remaining 26 percent of the mailing cases. At the end of data collection (and after these changes had been accounted for), 84.8 percent of the sample had been located: 15.5 percent as deceased, 50.3 percent as located with an address and verified, and 19.0 percent as located with an address and not verified. About 15.2 percent of the sample was not located.

#### 4.2. Identification of Hard-to-Find Subpopulations

We used a multivariate logistic regression model to compare the individuals who were not located (15.2 percent) to those who were located (84.8 percent) on a subset of variables available from the 1960 baseline data collection. As shown in [Table 2](#) below, 22 predictor and control variables – covering student characteristics, cognitive aptitude and personality, as well as school and regional attributes – were used in the analysis. All continuous measures were categorized into quintiles. Those in the middle three quintiles were combined and used as the referent group in the logistic regression.

Analyses of tracking propensity consistently showed that sex was the single biggest predictor of tracking success, such that male respondents were 3.6 times more likely to be located than females. This is partially due to the substantial number of sampled females with no SSN and no new (e.g., married) last name in the historical records (39.0 percent of all sampled females). Because SSN is a unique identifier that rarely changes over the course of a person's life, it is particularly helpful for locating people whose names have changed. For example, locating rates for females with an SSN or married name available were comparable to those for males (males: 92.0 percent; females with SSN or married name: 94.3 percent), whereas the average locating rate for females for whom no SSN or married name were available was 52.7 percent.

Differences like this could suggest that the mechanisms underlying locating propensity operate differently for males and females. This is due in part or in whole to name changes. For example, certain types of females may have been more likely to get married right after high school. Name changes, as well as the increased likelihood of housing and banking records being in their partners' names rather than their own, make tracking these females successively more difficult relative to males or females who did not get married or change their names shortly after high school. Males are generally easier to find regardless of their marital status or age at first marriage, because by comparison men rarely change their names.

To investigate this further, we analyzed males and females separately. Such an approach can be used to answer the following question: Within a given subpopulation (i.e., males or females), what characteristics are associated with tracking failure? To control for the unobserved heterogeneity across models, we report the predicted probabilities obtained from the separate logistic regression models (see [Mood 2010](#)). [Table 2](#) presents the predicted probabilities and  $p$  values for the significant results from the multivariate analyses of tracking failure for males and females separately (complete results are reported in the Appendix).

For males, the analysis indicated that the parameters associated with six of the 22 variables examined were significantly associated with tracking propensity ( $p < .05$  for the Wald chi-square test for parameters): 1960 family SES, percent minority for the school attended in 1960, general academic aptitude, and three personality measures (impulsivity, tidiness, and leadership). We confirmed our suspicion that males scoring in the top quintile for the impulsivity measure were harder to locate than those scoring in the middle quintiles. The regression showed that we were also less likely to locate males who scored in the bottom quintiles for tidiness and leadership (relative to those in the middle or top quintiles). The predicted probabilities were highest for males who attended a school in which minorities

Table 2. Predicted probability estimates of tracking failure for significant predictors from multivariate logistic regression model, by selected 1960 variables

Selected characteristics (1960)	Males		Females	
	Predicted probability	p value	Predicted probability	p value
Grade in 1960: 9th/10th vs. 11th/12th grade	0.80	.0877	0.81	.0005
Family SES: Bottom quintile vs. higher quintiles	0.80	.0256	0.81	.0052
General academic aptitude:	0.82	.0435	0.86	<.0001
Bottom quintile vs. middle quintiles				
General academic aptitude:	0.66	.1562	0.62	.0002
Top quintile vs. middle quintiles				
Sociability: Bottom quintile vs. middle quintiles	0.73	.9219	0.79	.0477
Impulsivity: Bottom quintile vs. middle quintiles	0.82	.0179	0.72	.4671
Tidiness: Bottom quintile vs. middle quintiles	0.84	.0096	0.75	.5564
Leadership: Bottom quintile vs. middle quintiles	0.75	.0383	0.75	.5058
Self-confidence: Bottom quintile vs. all higher quintiles <sup>1</sup>	0.67	.1576	0.68	.0266
Percent minority in school: (50–79%) vs. 0–49%	0.95	.0465	0.63	.0585
Census region: Mid-Atlantic, South Atlantic, or Mountain regions vs. all other regions	0.78	.1794	0.79	.0153
Population size of surrounding community: Under 5,000 vs. 1.5 million or more	0.69	.6683	0.64	.0384
Population size of surrounding community: 5,000–249,999 vs. 1.5 million or more	0.83	.2874	0.64	.0170
Population size of surrounding community: 250,000 vs. 1.49 million or more	0.89	.0951	0.63	.0192

<sup>1</sup> The parameter estimates indicated that those in the bottom and top quintiles for the psychological measure of self-confidence were tracked at higher rates than were those in the middle three quintiles, although the difference was only significant for those in the bottom quintile.

Note: Only significant results are reported from the multivariate logistic regression. Regression models were run separately for males and females. In addition, the models included the following categorical measures: health in past three years; health in first ten years; number days sick in bed in past year; class rank; personality (creativity, social sensitivity, vigor, calmness, culture, mature personality); school type and building type (junior high school or senior high school).

made up 50–79 percent of the student body. It is interesting to note that the predicted probabilities of tracking failure for males attending high-minority schools (80–100 percent of the school body) were lower than the predicted probabilities for those attending schools where minorities constituted 50–79 percent of the student body, and that there was no significant difference in tracking propensity for males attending high-minority schools relative to males who attended a school with 0–49 percent minorities ( $p = .11$ ). We also confirmed that even after controlling for other individual characteristics, males with lower cognitive scores relative to their peers would be more difficult to find. As [Table 2](#) shows, we were significantly less likely to locate males scoring in the bottom quintile for the composite measures of general academic aptitude (relative to those in the middle or top quintiles).

As expected, the patterns associated with tracking failure were quite different for females, with one notable similarity. Consistent with the findings for males and our expectations, we were significantly less likely to locate females scoring in the bottom quintile for the composite measure of general academic aptitude (relative to those in the middle or top quintiles). In addition, there were significant differences in tracking propensity for parameters associated with grade cohort in 1960, family SES quintile, and self-confidence as well as Census region and population size of the surrounding community for the school attended in 1960. The tracking efforts employed for this pilot study were significantly less likely to locate females from the 9th- or 10th-grade cohorts or those in the bottom quintile for the family SES measure. Finally, the analysis suggested that there was a U-shaped relationship between self-confidence and tracking propensity, such that those in the bottom and top quintiles of the self-confidence measure were tracked at higher rates relative to those in the middle three quintiles. However, the difference was only significant for those in the bottom quintile.

There were also significant regional differences. We were significantly less likely to locate females who attended schools in very large urban communities. Tracking propensities for those in the Mid-Atlantic, South Atlantic, or Mountain regions in 1960 were also lower relative to females from other regions.

#### 4.3. Implementation of Tracking-loss Adjustments

To assess the extent to which biases introduced through tracking failure can be minimized by employing tracking-loss adjustments, we first created a tracking-loss adjustment using the CHAID algorithm. We then computed the weighted estimates and percent relative bias before and after the tracking loss adjustment and compared the estimates for a selected subset of the base-year variables available for all of the cases sampled for the pilot study. The percent relative bias is the estimated bias divided by the estimates produced after the analysis is restricted to include only those who were located. We also computed the estimated bias, which is the difference between all those who were sampled for PTPS12 and the subset that were located. A t-test was used to determine whether these differences were statistically significant. Estimates from PTPS12 sample that are significantly different ( $p < .05$ ) from those located in PTPS12 estimate suggest that a potential for tracking bias may be present.

[Table 3](#) shows the results for five of the 41 variables selected for this analysis. The nontracking adjustment was able to reduce the significant tracking loss bias for the

estimated proportion of individuals whose family SES in 1960 was in the bottom quintile, but not for the estimated proportion of individuals who had missing data for this variable. The adjustment reduced the tracking loss bias for the variables measuring health until age ten, general academic aptitude, and impulsivity; however, the t-tests indicate that even after this adjustment, these estimates remain significantly biased.

## 5. Discussion

The Project Talent pilot study was used to evaluate the use of relatively low-cost methods to locate and obtain cooperation from a large, national follow-up of longitudinal study sample members after an extended hiatus. Though the age of the target population and the extended hiatus of 37 to 51 years may limit the generalizability of these results to other studies, the study offered a unique opportunity to explore correlates related to tracking success. In addition, we were able to use the base-year measures to examine the use of statistical adjustments of tracking-related biases.

This study confirmed our expectations that even without expensive in-person tracking activities we would be able to locate a relatively high proportion of sample members (85 percent). As might be expected, the tracking rate for this study was slightly lower than those obtained in similar studies of less geographically dispersed sample members, which are able to apply more localized tracking methods (e.g., [Hampson et al. 2001](#); [Ortiz and Godinez Ballon 2007](#)), as well as those obtained in large, national studies using similar tracking methods, but with shorter periods of noncontact (e.g., LSAY, MIDUS, LSAL, and Add Health). AIR recently applied the same tiered tracking protocol used in this study to a geographically dispersed group of minorities who participated in an aquatic science program as undergraduate and graduate students between 1990 and 2011. Though many of the participants had not been involved with the program in as many as 22 years, 93.4 percent were located ([Sandoval and Stone 2013](#)).

However, the results presented here suggest that even with relatively high tracking rates, certain subpopulations may be systematically harder to find (research question 2). We confirmed that even after controlling for individual characteristics such as sex, family SES, and school attributes, individuals with lower cognitive scores relative to their peers would be more difficult to find and that certain personality characteristics are associated with higher rates of tracking failure. Our findings suggest that males who scored low on the tidiness measure in 1960 tend to be harder to find – perhaps because these males are less likely to update their addresses with creditors, voluntary registers, or other databases after moves or to engage in other behaviors that may leave traces to their new address and make them easier to locate.

Contrary to our expectations, this study also found that males who scored lower on the impulsivity measure (i.e., those who were less impulsive) were actually harder to locate than those who scored in the middle quintiles. We are not sure why this may be, but one explanation could be that less impulsive males exhibit more deliberate purchasing behavior (e.g., saving and then purchasing), particularly when purchasing expensive items such as cars or electronics. As a result, these less impulsive males may be less likely to open lines of credit, and hence will have less information in consolidated tracking databases.

Table 3. Comparison of base weighted estimates for the pilot study population and population located during the tracking phase for the 2011–12 Project Talent Follow-up Pilot Study, before and after nontracking weighting adjustments, for key sample characteristics measured in 1960

Selected sample characteristics (1960)	Before nontracking adjustment (base weights only)			After nontracking adjustment and base weights		
	Estimated percent for pilot sample <sup>1</sup>	Estimated percent for located <sup>2</sup>	Percent relative bias	Estimated bias	Estimated percent for located <sup>2</sup>	Percent relative bias
Sex (1960)						
Male	48.8	49.2	0.01	0.5	46.2	-2.5
Female	51.2	50.8	-0.01	-0.5	53.8	2.5
Family SES						
Missing	2.4	1.4	-0.72*	-1.0	1.4	-0.9
Bottom 20%	19.6	15.0	-0.31*	-4.6	15.7	-3.9
Middle 60%	58.3	60.1	0.03	1.8	59.6	1.3
Top 20%	19.7	23.6	0.16	3.8	23.3	3.6
Usual health before age ten						
Missing	9.2	6.0	-0.53*	-3.2	6.2	-3.0
Very poor or poor	4.8	4.8	-0.01	0.0	5.0	0.1
Good or average	29.6	28.1	-0.05	-1.5	28.1	-1.5
Very good or excellent	56.4	61.1	0.08*	4.7	60.8	4.3
General academic aptitude						
Missing	12.7	10.3	-0.23	-2.4	10.2	-2.5
Q1	19.1	11.2	-0.71*	-8.0	11.5	-7.6
Q2	18.2	16.2	-0.12	-2.0	16.7	-1.5
Q3	17.8	19.2	0.07	1.4	19.5	1.7
Q4	16.8	21.2	0.21*	4.4	21.3	4.5
Q5	15.4	22.0	0.30*	6.6	20.8	5.4
Impulsivity						
Missing	1.2	0.6	-0.94*	-0.6	0.6	-0.6
Q1	29.9	29.7	0.00	-0.1	29.5	-0.4
Q2	15.7	16.0	0.02	0.3	16.2	0.5
Q3	22.7	22.7	0.00	0.1	23.0	0.4



Table 3. Continued

Selected sample characteristics (1960)	Before nontracking adjustment (base weights only)			After nontracking adjustment and base weights			
	Estimated percent for pilot sample <sup>1</sup>	Estimated percent for located <sup>2</sup>	Estimated bias	Percent relative bias	Estimated percent for located <sup>2</sup>	Estimated bias	Percent relative bias
Q4	14.5	14.1	-0.5	-0.03	13.8	-0.7	-0.05
Q5	16.0	16.8	0.8	0.05	16.8	0.8	0.05

<sup>1</sup> All cases sampled for the 2011–12 Project Talent Follow-up Pilot Study.

<sup>2</sup> Located includes only those who were located during the tracking phase.

Note: An asterisk (\*) indicates where there is a statistically significant difference at the  $p = 0.05$  level between the estimates for the full pilot study sample and the estimates produced after the analysis is restricted to include only those who were located. The percent relative bias is the estimated bias divided by the estimates produced after the analysis is restricted to include only those who were located.



Another unexpected finding was that among females, those scoring in the middle quintiles of the self-confidence measure were located at significantly lower rates relative to those scoring in the bottom quintile (those who were less self-confident). One possible explanation is that females who were less confident in high school were also less likely to relocate after completing high school. This reduced mobility would make them easier to find.

Regardless, these findings clearly support existing concerns that certain groups are located at disproportionately low rates that, if left uncorrected, could bias the survey results. For example, the pilot study was significantly less successful at locating females as well as males and females scoring in the bottom quintile of the cognitive measure of general academic aptitude. This is of particular concern because lower cognitive ability is also associated with different decision-making processes, different risk factors, and higher mortality rates. As a result, these individuals are often of particular interest for longitudinal studies that focus on health, financial, psychosocial, and general well-being outcomes throughout the life course.

Like [Iannacchione \(2003\)](#), we applied a tracking-loss adjustment to evaluate the extent to which such a statistical adjustment could be used to reduce biases introduced as a result of systematic differences in tracking propensity across subgroups (research question 3). The tracking weighting adjustment reduced, but did not remove, all significant biases. Our findings suggest that where differential tracking rates are expected to vary greatly across subgroups (e.g., where the amount and quality of information needed for tracking may be uneven across subgroups), studies attempting to locate individuals after a long hiatus should not rely solely on statistical adjustments to reduce and remove biases. Rather, researchers should identify the hardest-to-locate subgroups and use this information to develop a sample design and tracking protocol that can improve the representation and statistical efficiency of the resulting study. This could include the stratification of samples using variables that are expected to be correlated with tracking propensity and the outcomes of interest to ensure that a sufficient number of individuals from key subgroups will be located and can be used for tracking-loss adjustments.

Studies should also reflect on whether shifting study resources into the development and implementation of more extensive, tailored tracking protocols – methods that would simply be too costly to implement for the full sample – can be applied on a smaller scale for historically underlocated subgroups. For example, it seems likely that individuals with certain personality or cognitive characteristics may be more isolated or lead elusive lives that make them more difficult to locate using methods that rely on consolidated tracking databases. Tracking plans may want to include protocols for contacting local government offices (e.g., marriage bureaus or city halls), or even classmates and siblings, in order to find hard-to-find sample members. Such work would be most effective if implemented early in the tracking phase, as these individuals will likely take more time and resources (i.e., tracking steps) to locate. Longitudinal studies, including those that have been revived after an extended hiatus, should make good use of their existing data and consider using key measures and indicators to prioritize cases with lower tracking propensities to receive longer and more intensive tracking methods.

**Appendix**  
*Appendix. Percentage of persons not found during tracking and predicted probability estimates of tracking failure, from multivariate logistic regression model, by selected 1960 variables*

Selected characteristics (1960)	Males				Females			
	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP
Grade cohort								
9-10	1,328	9.3	0.80	(0.72, 0.87)	1,329	25.9	0.81***	(0.76, 0.86)
11-12	1,049	6.4	-	-	1,173	17.6	-	-
Socioeconomic status								
Low SES	452	10.8	0.80	(0.72, 0.88)	492	31.1	0.81**	(0.75, 0.86)
Not low SES	1,850	7.0	-	-	1,963	19.7	-	-
Missing	75	17.3	0.99*	(0.77, 1.00)	47	25.5	0.78	(0.61, 0.98)
Health in last three years								
Very poor or poor	49	10.2	0.78	(0.61, 0.97)	56	28.6	0.76	(0.65, 0.9)
Good or average	525	7.0	0.70	(0.63, 0.79)	729	23.9	0.74	(0.69, 0.79)
Very good or excellent	1,517	7.4	-	-	1,557	20.4	-	-
Missing	286	12.6	0.93	(0.63, 1.00)	160	27.5	0.81	(0.61, 0.99)
Health in first 10 years								
Very poor or poor	102	7.8	0.74	(0.61, 0.91)	131	19.8	0.70	(0.63, 0.8)
Good or average	682	6.6	0.69	(0.63, 0.77)	774	23.3	0.73	(0.69, 0.78)
Very good or excellent	1,300	7.9	-	-	1,433	20.9	-	-
Missing	293	11.9	0.61	(0.52, 0.9)	164	27.4	0.68	(0.57, 0.91)
Number of days sick in bed in past year								
None	655	8.1	-	-	523	20.8	-	-
1-2	887	7.2	0.70	(0.64, 0.78)	906	20.1	0.75	(0.69, 0.81)
3 or more	565	7.3	0.70	(0.63, 0.79)	922	23.8	0.77	(0.72, 0.84)
Missing	270	12.2	0.64	(0.53, 0.9)	151	27.2	0.71	(0.57, 0.96)
Class rank								
Bottom 20%	407	9.6	0.77	(0.69, 0.87)	284	24.3	0.72	(0.66, 0.78)
Middle 60%	1,498	7.5	-	-	1,588	23.2	-	-
Top 20%	374	7.5	0.75	(0.66, 0.85)	559	17.2	0.71	(0.66, 0.76)

Appendix. Continued

Selected characteristics (1960)	Males				Females			
	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP
Missing	98	12.2	0.62	(0.54, 0.83)	71	23.9	0.70	(0.59, 0.88)
General academic aptitude								
Bottom 20%	456	12.7	0.82*	(0.73, 0.91)	407	35.4	0.86	(0.79, 0.91)
Middle 60%	1,206	7.2	—	—	1,358	21.2	—	—
Top 20%	407	4.2	0.66	(0.59, 0.76)	444	11.5	0.62***	(0.59, 0.67)
Missing	308	9.4	0.76	(0.66, 0.87)	293	23.2	0.73	(0.67, 0.80)
Sociability								
Bottom 20%	778	9.0	0.73	(0.67, 0.82)	503	25.8	0.79*	(0.73, 0.85)
Middle 60%	1,266	7.5	—	—	1,381	21.4	—	—
Top 20%	297	6.7	0.76	(0.66, 0.88)	590	19.5	0.72	(0.67, 0.77)
Missing <sup>1</sup>	36	16.7	0.84	(0.64, 0.99)	28	35.7	0.84	(0.66, 0.98)
Social sensitivity <sup>1</sup>								
Bottom 20%	745	8.7	0.72	(0.65, 0.81)	387	23.5	0.71	(0.66, 0.77)
Middle 60%	1,427	7.6	—	—	1,681	22.5	—	—
Top 20%	169	7.1	0.78	(0.65, 0.92)	406	17.5	0.72	(0.66, 0.78)
Impulsivity <sup>1</sup>								
Bottom 20%	1,116	9.4	0.82*	(0.75, 0.90)	1,117	21.3	0.72	(0.68, 0.76)
Middle 60%	874	6.3	—	—	947	21.8	—	—
Top 20%	351	7.1	0.76	(0.67, 0.87)	410	23.7	0.76	(0.7, 0.82)
Vigor <sup>1</sup>								
Bottom 20%	468	9.0	0.76	(0.68, 0.85)	553	22.6	0.72	(0.67, 0.78)
Middle 60%	1,627	8.0	—	—	1,676	22.3	—	—
Top 20%	246	5.3	0.69	(0.6, 0.81)	245	17.6	0.71	(0.65, 0.78)
Calmness <sup>1</sup>								
Bottom 20%	810	9.9	0.78	(0.71, 0.87)	721	24.3	0.75	(0.7, 0.81)
Middle 60%	1,300	7.2	—	—	1,421	22.0	—	—
Top 20%	231	4.8	0.69	(0.6, 0.83)	332	16.3	0.71	(0.65, 0.78)

Appendix. Continued

Selected characteristics (1960)	Males				Females			
	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP
Tidiness <sup>1</sup>								
Bottom 20%	844	10.0	0.84**	(0.76, 0.92)	495	23.6	0.75	(0.69, 0.81)
Middle 60%	1,222	7.0	—	—	1,440	22.4	—	—
Top 20%	275	5.5	0.69	(0.63, 0.76)	539	18.9	0.73	(0.68, 0.79)
Culture <sup>1</sup>								
Bottom 20%	937	8.2	0.81	(0.69, 0.93)	444	23.6	0.72	(0.67, 0.78)
Middle 60%	1,159	7.6	—	—	1,459	22.4	—	—
Top 20%	245	8.2	0.66	(0.61, 0.73)	571	19.1	0.74	(0.69, 0.80)
Leadership <sup>1</sup>								
Bottom 20%	913	6.8	0.75*	(0.66, 0.85)	842	22.8	0.75	(0.70, 0.80)
Middle 60%	1,036	8.9	—	—	1,134	21.8	—	—
Top 20%	392	7.9	0.68	(0.63, 0.75)	498	20.5	0.73	(0.68, 0.79)
Self-confidence <sup>1</sup>								
Bottom 20%	702	7.8	0.67	(0.6, 0.77)	714	20.3	0.68*	(0.65, 0.72)
Middle 60% and top 20%	1,639	7.9	—	—	1,323	22.5	—	—
Mature personality <sup>1</sup>								
Bottom 20%	612	8.8	0.70	(0.63, 0.78)	483	21.7	0.69	(0.64, 0.74)
Middle 60%	1,416	8.0	—	—	1,547	23.4	—	—
Top 20%	313	5.8	0.69	(0.61, 0.82)	444	16.7	0.71	(0.66, 0.78)
School type								
Public	2,163	8.3	0.90	(0.68, 1.00)	2,209	22.1	0.76	(0.60, 0.96)
Private	214	5.1	—	—	293	21.2	—	—
Building type								
Not junior high school	2,111	7.5	—	—	2,245	21.6	—	—
Junior high school	266	12.0	0.73	(0.65, 0.84)	257	25.7	0.71	(0.65, 0.78)

Appendix. Continued

Selected characteristics (1960)	Males			Females				
	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP	Number	Percent not found	Predicted probability (PP)	95% confidence interval for PP
Percent minority in school								
0–49 percent	2,055	7.3	—	—	2,071	21.4	—	—
50–79 percent	27	22.2	0.95*	(0.73, 1.00)	78	23.1	0.63	(0.57, 0.74)
80–100 percent	279	12.5	0.78	(0.68, 0.88)	336	26.2	0.70	(0.65, 0.76)
Missing	16	6.3	0.70	(0.53, 1.00)	17	11.8	0.60	(0.52, 0.87)
Census region (school)								
Mid-Atlantic, South Atlantic, Mountain	890	10.2	0.78	(0.71, 0.86)	1,004	25.4	0.79*	(0.74, 0.84)
Other region	1,487	6.7	—	—	1,498	19.8	—	—
Population size of surrounding community								
Rural	262	7.3	0.77	(0.62, 0.95)	239	26.8	0.66	(0.60, 0.76)
Under 5,000	259	4.6	0.69	(0.57, 0.9)	280	19.6	0.64*	(0.58, 0.73)
5,000–249,999	1,155	8.7	0.83	(0.67, 0.97)	1,188	21.3	0.64*	(0.59, 0.71)
250,000–1,499,999	323	10.5	0.89	(0.71, 0.99)	337	19.6	0.63*	(0.58, 0.71)
1,500,000 or more	109	7.3	—	—	140	30.7	—	—
Not classified	269	6.7	0.92	(0.68, 1.00)	318	22.0	0.71	(0.58, 0.93)

Odds ratio notes: Adjusted odds ratio that person will not be located based on multivariate logistic regression adjusting for the other factors shown in the table. <sup>1</sup>N = 36 observations were missing data for all 10 of the personality measures under examination. As a result, only one of the parameters associated with these variables could be estimated, since the corresponding parameters associated with the remaining variables would be perfectly correlated. The odds ratio estimate for the “Missing” parameter is reported only for the “sociability” variable. The redundant parameters for the remaining personality variables (social sensitivity, impulsivity, vigor, calmness, tidiness, culture, leadership, self-confidence, and mature personality) have been removed from the table. The number and percent not found are not reported for these variables; therefore, the estimates will not sum to totals. — indicates reference group.

\*\*\*p < .001, \*\*p < .01, \*p < .05.

## 6. References

- Andresen, E.M., Renea Machuga, C., van Booven, M.E., Egel, J., Chibnall, J.T., and Tait, R.C. (2008). Effects and Costs of Tracing Strategies on Nonresponse Bias in a Survey of Workers With Low-Back Injury. *Public Opinion Quarterly*, 72, 40–54. DOI: <http://www.dx.doi.org/10.1093/poq/nfm055>
- Anstey, K.J., Luszcz, M.A., Giles, L.C., and Andrews, G.R. (2001). Demographic, Health, Cognitive, and Sensory Variables as Predictors of Mortality in Very Old Adults. *Psychology and Aging*, 16, 3–11. DOI: <http://www.dx.doi.org/10.1037/0882-7974.16.1.3>
- Becker, K., Berry, S., Orr, N., and Perlman, J. (2012). Finding the Hard to Reach and Keeping Them Engaged in Research. Presented at the 2012 International Conference on Methods for Surveying and Enumerating Hard-to-Reach Populations, New Orleans, LA., October 31–November 3.
- Calderwood, L. (2012). Tracking Sample Members in Longitudinal Studies. *Survey Practice*, 5(4), 1. Available at: <http://www.surveypractice.org/index.php/SurveyPractice/article/view/34> (accessed August 2, 2013).
- Call, V.R.A., Otto, L.B., and Spenner, K.I. (1982). *Tracking Respondents: A Multi-Method Approach*. Lexington, MA: Lexington Books.
- Clarridge, B.R., Sheehy, L.L., and Hauser, T.S. (1978). Tracing Members of a Panel: A 17-Year Follow-up. *Sociological Methodology*, 9, 185–203.
- Cotter, R.B., Burke, J.D., Loeber, R., and Navratil, J.L. (2002). Innovative Retention Methods in Longitudinal Research: A Case Study of the Developmental Trends Study. *Journal of Child and Family Studies*, 11, 485–498.
- Cotter, R.B., Burke, J.D., Stouthamer-Loeber, M., and Loeber, R. (2005). Contacting Participants for Follow-up: How Much Effort Is Required to Retain Participants in Longitudinal Studies? *Evaluation and Program Planning*, 28, 15–21. DOI: <http://www.dx.doi.org/10.1016/j.evalprogplan.2004.10.002>
- Cottler, L.B., Compton, W.M., Ben-Abdallah, A., Horne, M., and Claverie, D. (1996). Achieving a 96.6 Percent Follow-up Rate in a Longitudinal Study of Drug Users. *Drug and Alcohol Dependence*, 41, 209–217. DOI: [http://www.dx.doi.org/10.1016/0376-8716\(96\)01254-9](http://www.dx.doi.org/10.1016/0376-8716(96)01254-9)
- Couper, M.P. and Ofstedal, M.B. (2009). Keeping in Contact With Mobile Sample Members. In *Methodology of Longitudinal Surveys*, P. Lynn (ed.). New York: Wiley, 183–203.
- Crowe, M., Clay, O.J., Martin, R.C., Howard, V.J., Wadley, V.G., Sawyer, P., and Allman, R.M. (2013). Indicators of Childhood Quality of Education in Relation to Cognitive Function in Older Adulthood. *Journals of Gerontology: Biological and Medical Sciences*, 68, 198–204. DOI: <http://www.dx.doi.org/10.1093/gerona/gls122>
- Flanagan, J.C. and Cooley, W.W. (1966). *Project TALENT: One-Year Follow-up Studies*. Pittsburgh, PA: University of Pittsburgh, School of Education.
- Gale, C.R., Cooper, R., Craig, L., Elliott, J., Kuh, D., Richards, M., Starr, J.M., Whalley, L.J., and Deary, I.J. (2012). Cognitive Function in Childhood and Lifetime Cognitive Change in Relation to Mental Well-Being in Four Cohorts of Older People. *PLOS ONE*, 7, e44860.
- Haggerty, K.P., Fleming, C.B., Catalano, R.F., Petrie, R.S., Rubin, R.J., and Grassley, M.H. (2008). Ten Years Later: Locating and Interviewing Children of Drug Abusers.

- Evaluation and Program Planning, 31, 1–9. DOI: <http://www.dx.doi.org/10.1016/j.evalprogplan.2007.10.003>
- Hampson, S.E., Dubanoski, J.P., Hamada, W., Marsella, A.J., Matsukawa, J., Suarez, E., and Goldberg, L.R. (2001). Where Are They Now? Locating Former Elementary-School Students After Nearly 40 Years for a Longitudinal Study of Personality and Health. *Journal of Research in Personality*, 35, 375–387. DOI: <http://www.dx.doi.org/10.1006/jrpe.2001.2317>
- Hauser, R.M. (2005). Survey Response in the Long Run: The Wisconsin Longitudinal Study. *Field Methods*, 17, 3–29. DOI: <http://www.dx.doi.org/10.1177/1525822X04272452>
- Iannacchione, V.G. (2003). Sequential Weight Adjustments for Location and Cooperation Propensity for the 1995 National Survey of Family Growth. *Journal of Official Statistics*, 19, 31–45.
- Jessor, R. and Jessor, S.L. (1977). *Problem Behavior and Psychological Development: A Longitudinal Study of Youth*. New York: Academic Press.
- Kass, G.V. (1980). An Exploratory Technique for Investigating Large Quantities of Categorical Data. *Applied Statistics*, 29, 119–127.
- Kimmel, L.G. and Miller, J.D. (2008). The Longitudinal Study of American Youth: Notes on the First 20 Years of Tracking and Data Collection. *Survey Practice*, 19. Available at: <http://www.surveypractice.org/index.php/SurveyPractice> (accessed August 2013).
- Masson, H., Balfe, M., Hackett, S., and Phillips, J. (2013). Lost Without a Trace? Social Networking and Social Research With a Hard-to-Reach Population. *British Journal of Social Work*, 43, 24–40. DOI: <http://www.dx.doi.org/10.1093/bjsw/bcr168>
- Meehan, A., Saleska, E.L., Kinsey, N.L., Hinsdale-Shouse, M.A., and Tischner, C. (2009). The Challenges of Locating Young Adults for a Longitudinal Study: Improved Tracing Strategies Implemented for the National Longitudinal Study of Adolescent Health, Wave IV. *Proceedings of the Annual Conference of the American Association for Public Opinion Research*, Hollywood, FL. Available at: <http://www.amstat.org/sections/srms/proceedings/y2009/Files/400056.pdf> (accessed August 2013).
- Mood, C. (2010). Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do About It. *European Sociological Review*, 26, 67–82. DOI: <http://www.dx.doi.org/10.1093/esr/jcp006>
- Orr, D.B. (1963). *A Study of Nonrespondents to the First Project TALENT One-Year Follow-Up Mail Questionnaire*. Philadelphia, PA: Presented at the Meetings of the American Psychological Association.
- Ortiz, V. and Godinez Ballon, E. (2007). Longitudinal Research at the Turn of the Century: Searching for the Mexican American People. *Social Methods & Research*, 36, 112–137.
- Passetti, L.L., Godley, S.H., Scott, C.K., and Siekmann, M. (2000). A Low-Cost Follow-up Resource: Using the World Wide Web to Maximize Client Location Efforts. *American Journal of Evaluation*, 21, 195–203. DOI: [http://www.dx.doi.org/10.1016/S1098-2140\(00\)00072-2](http://www.dx.doi.org/10.1016/S1098-2140(00)00072-2)
- Radler, B.T. and Ryff, C.D. (2010). Who Participates? Accounting for Longitudinal Retention in the MIDUS National Study of Health and Well-Being. *Journal of Aging and Health*, 22, 307–331. DOI: <http://www.dx.doi.org/10.1177/0898264309358617>

- Ren, P. (2011). Lifetime Mobility in the United States: 2010 (ACSB/10-07). U.S. Census Bureau, American Community Survey Briefs. Available at: <http://www.census.gov/prod/2011pubs/acsbr10-07.pdf> (accessed August 2013).
- Ribisl, K.M., Walton, M.A., Mowbray, C.T., Luke, D.A., Davidson, W.S., and BootsMiller, B.J. (1996). Minimizing Participant Attrition in Panel Studies Through the Use of Effective Retention and Tracking Strategies: Review and Recommendations. *Evaluation and Program Planning*, 19, 1–25. DOI: [http://www.dx.doi.org/10.1016/0149-7189\(95\)00037-2](http://www.dx.doi.org/10.1016/0149-7189(95)00037-2)
- Ritchie, S.J. and Bates, T.C. (2013). Enduring Links From Childhood Mathematics and Reading Achievement to Adult Socioeconomic Status. *Psychological Sciences*, 24, 1301–1308. DOI: <http://www.dx.doi.org/10.1177/0956797612466268>
- Rossi, R.J., Wise, L.L., Williams, K.L., and Carrel, K.S. (1976). *Methodology of the Project TALENT 11-Year Follow-Up Study*. Washington, DC: American Institutes for Research.
- Ryff, C., Almeida, D.M., Ayanian, J.S., Carr, D.S., Cleary, P.D., Coe, C., Davidson, R., Krueger, R.F., Lachman, M.E., Marks, N.F., Mroczek, D.K., Seeman, T., Seltzer, M.M., Singer, B.H., Sloan, R.P., Tun, P.A., Weinstein, M., and Williams, D. (2006). National Survey of Midlife Development in the United States (MIDUS II), 2004–2006 (ICPSR 4652). Field Report for MIDUS 2 Longitudinal Sample. Available at: <http://www.icpsr.umich.edu/icpsrweb/NACDA/studies/04652/version/6> (accessed August 3, 2013).
- Sandoval, A. and Stone, C. (2013). Tracking and Re-engaging Respondents for Follow-up Research: A Methodological Examination of Two Research Studies. Presented at the 68th Annual Conference of the American Association for Public Opinion Research, Boston, MA.
- Stouthamer-Loeber, M. and van Kammen, W.B. (1995). Participant Acquisition and Retention. In *Data Collection and Management: A Practical Guide (Applied Social Research Methods, Vol. 39)*, M. Stouthamer-Loeber and W.B. van Kammen (eds). Thousand Oaks, CA: Sage, 62–80.
- Strawn, C., Lopez, C., and Setzler, K. (2007, revised). *It Can Be Done: Sample Retention Methods Used by the Longitudinal Study of Adult Learning*. Portland, OR: Portland State University.
- Wilson, R.S., Hebert, L.E., Scherr, P.A., Barnes, L.L., Mendes de Leon, C.F., and Evans, D.A. (2009). Educational Attainment and Cognitive Decline in Old Age. *Neurology*, 72, 460–465. DOI: <http://www.dx.doi.org/10.1212/01.wnl.0000341782.71418.6c>
- Wine, J., Janson, N., and Wheelless, S. (2011). 2004/09 Beginning Postsecondary Students Longitudinal Study (BPS:04/09) Full-Scale Methodology Report (NCES 2012-246). Washington, DC: U.S. Department of Education, National Center for Education Statistics. Available at [http://nces.ed.gov/pubs2012/2012246\\_1.pdf](http://nces.ed.gov/pubs2012/2012246_1.pdf) (accessed August 2013).
- Wise, L.L., McLaughlin, D.H., and Steel, L. (1979). *The Project TALENT Data Bank Handbook (Revised)*. Washington, DC: American Institutes for Research.

Received February 2013

Revised November 2013

Accepted November 2013



## Evaluating the Efficiency of Methods to Recruit Asian Research Participants

*Hyunjoo Park<sup>1</sup> and M. Mandy Sha<sup>2</sup>*

Few empirical studies have evaluated the efficiency of recruitment methods to recruit non-English-speaking research participants. We attempt to fill this research gap by conducting a systematic evaluation using recruitment data from a large cognitive testing study that pretested the translations of the American Community Survey. In our study we contacted 1,084 Chinese and Korean speakers to identify those who spoke little or no English. We measured the efficiency of the recruitment methods (newspaper advertisements, flyers, online communication, and word of mouth) using four criteria: time spent, outreach capacity, screener completion, and eligibility rate. We also examined differences in recruitment efficiencies by recruiters and sublanguage groups. Among the recruitment methods examined, newspaper advertisements were most efficient in reaching a larger number of Asians while using the least amount of recruiters' time. For recruiting non-English speakers, word of mouth by recruiters with strong ties to the ethnic community worked best.

*Key words:* Community-based recruitment; recruitment efficiency; cognitive interview.

### 1. Introduction

Language barriers can prevent non-English speakers from responding to surveys that are available only in English. As the population in the United States becomes more diverse, the inclusion of non-English speakers may have potential implications for the estimates of major national surveys. Therefore, translating English language survey questionnaires can reduce language barriers and encourage survey participation of non-English speakers (native speakers of non-English languages, regardless of their English language proficiency). To ensure that the translations convey the intended meanings, researchers are conducting cognitive testing of the translations to pretest and evaluate the quality of the translations. These cognitive interviews require the successful recruitment of participants who speak a language other than English and also speak little or no English, because recruiting participants who speak English well does not represent the intended population, namely non-English speakers who are likely to use the translated survey questionnaires.

Asians form a significant segment of the population who may benefit from translated survey questionnaires and materials. Among non-English languages spoken in the United States, Chinese and Korean were among the top five languages for which the U.S. Census Bureau provided language assistance in the 2010 Census ([Kim and Zapata 2012](#)).

<sup>1</sup>Center for Survey Methodology, Survey Research Division at RTI International, 351 California Street, Suite 500, San Francisco, CA 94104, U.S.A. Email: [mpark@rti.org](mailto:mpark@rti.org)

<sup>2</sup>RTI International, 230 W. Monroe, #2100, Chicago, IL., U.S.A.

When Asian non-English speakers were recruited to participate in cognitive testing of the translations, past studies cited several challenges to recruitment, including a lack of trust and unfamiliarity with research studies in general (Pan et al. 2007; Yuan et al. 2009), stemming from the participants' lack of English fluency.

To date, very few studies in the literature address the efficiency of methods for recruiting Asian research participants or non-English speakers overall. Although a wealth of public health literature has investigated the topic of participant recruitment, most studies have concentrated on English-speaking and specific clinical populations (Lai et al. 2006; Yancey et al. 2006). This line of literature has recommended a community-based approach, such as using personal referrals, to address the recruits' possible concerns and to encourage the participation of minority populations. These findings have tended to come from the reporting of recruiting practices and recruiters' debriefings. In contrast, empirical studies that test hypotheses regarding the efficiency of the recruitment methods have been less common. This article attempts to fill this gap by investigating the efficiency of four common recruitment methods: non-English language newspaper advertisements, flyers, online communication, and word of mouth. Using recruitment data collected for a large cognitive testing study with Asian research participants (Chinese and Korean) in the United States, we improved upon the exploratory efficiency measures used in Liu et al. (2013) by refining the existing measures of time efficiency, outreach capacity, and eligibility rate and by adding screener completion rate, and we conducted a systematic evaluation of the efficiency of these recruitment methods. The measures we used were time efficiency, outreach capacity, screener completion rate, and eligibility rate determined from the screening process. We also examined the effect of the community-based approach when recruiters with strong ties to the Asian communities were present.

## 2. Related Research

Most of the published research on participant recruitment comes from the public health literature. A key concern in this literature is to identify factors that influence participant recruitment (e.g., Areán and Gallagher-Thompson 1996; Yancey et al. 2006), and the results tend to be similar. Representative of this line of literature, Fujimoto (1998) categorized three barriers that hindered research participation by minority populations: (1) simple logistical barriers that can easily be addressed, such as transportation, child- or eldercare, and scheduling; (2) complex logistical barriers, such as a fear of institutional settings, research staff who lack cultural diversity or sensitivity, or inappropriately written recruitment materials; and (3) complex barriers that pertain to recruits' knowledge, beliefs, and attitudes (e.g., general distrust of research studies).

Much of the literature has focused on the African American population group because African Americans tend to have a traditionally lower research participation rate than the general population. Several studies found that a common reason for the lower participation rate among African Americans was their distrust of research conducted by white-dominated research communities. To address the distrust issue, researchers recommended network-based recruitment approaches, such as the use of referrals in a church-based community (Coleman et al. 1997; Holcombe et al. 1999; Gorelick et al. 1996; Reed et al. 2003; Stoy et al. 1995).

However, recruitment of other minority populations, especially non-English-speaking populations, has not been studied as extensively. Similarly to the recruitment of African American research participants, researchers also observed distrust as the main hurdle to participation in recent cognitive testing studies involving non-English speakers (Pan et al. 2007; Yuan et al. 2009). The reasons for this distrust seemed to have stemmed from non-English speakers' limited exposure to research studies and social experience in the United States.

A handful of research articles focusing on recruitment of non-English speakers are available in both public health and exploratory methodological research. In addition to common recruitment methods, such as print and online advertising or flyer posting, recruiters often used community organizations serving the target minority populations to recruit Latino and Asian immigrants (Berrigan et al. 2010; Lau and Gallagher-Thompson 2002; Forsyth et al. 2007; Wellens 1994). Sha and her colleagues (2010) reported community-based word of mouth as the most successful method for recruiting Spanish speakers to pretest a large housing survey. Similarly to research on African American recruitment, they noted that in-person recruiting in the community helped to establish trust. After examining basic efficiency measures of time, outreach, and eligibility, Liu et al. (2013) also found word of mouth to be efficient in identifying eligible non-English speaking Asians. However, these findings lacked statistical testing. Like many other studies that used a community-based approach, Maxwell et al. (2005) also reported successfully recruiting Filipinos for a cancer screening study through a female project liaison. Most of these findings are based on descriptive reporting of recruitment experience, rather than a systematic analysis of recruitment data.

Regardless of the various recruitment targets, recruiters in the literature we reviewed used similar recruitment methods such as printed or online advertisements, flyers, and word of mouth to implement snowball recruitment. However, depending on the population, the method used most frequently and successfully to reach a certain population varied (Appel et al. 1999; Bistricky et al. 2010; Gilliss et al. 2001; Harris et al. 2003; Hughes et al. 2004; McLean and Campbell 2003; The DPP Research Group 2002; Wisdom et al. 2002). From the published recruitment literature, Yancey et al. (2006) concluded that when study eligibility criteria were general, reactive methods (e.g., a newspaper ad that asked interested people to call a researcher) were likely to reach target population groups, whereas proactive methods (e.g., in-person appeals by study staff) tended to be more productive when eligibility criteria were very specific. Several studies found that broadcasting or printed media reached a larger group of potential participants and worked particularly well for the general population, especially white participants. However, they tended to render a high ineligibility rate (Appel et al. 1999; Gilliss et al. 2001; McLean and Campbell 2003). Referrals or word of mouth seemed to work better for identifying minority participants who qualify for the study. However, they reached a far smaller number of potential participants (Bistricky et al. 2010; Gilliss et al. 2001; Wisdom et al. 2002).

Based on past literature, we observed that a community-based approach, mostly implemented via word of mouth, was frequently used to recruit minority populations. However, most of these studies involved public health research with English-speaking populations, or lacked systematic analysis. To attempt to fill this research gap, we used

recruitment data from a large non-English language cognitive testing study to examine the following research questions:

- (1) Which is the most efficient recruitment method (a) for reaching the target population quickly (“time efficiency”), (b) for reaching a larger number of people (“outreach capacity”), (c) for completing a larger number of screening questionnaires without breakoffs (“screener completion rate”), and (d) for identifying qualified non-English speakers for the purpose of the research (“eligibility rate”)?
- (2) Are there differences in recruiting efficiency by sublanguage groups (Chinese speakers vs. Korean speakers) and by recruiters (strong ties to ethnic community vs. weaker ties)?
- (3) What is the effect of having recruiters with strong ethnic ties?

Because we do not know how many people actually saw the recruitment message (we only know for certain for the ones that contacted us), our use of the term “reach” indicates how many successful contacts were made after contacting the target population. These measures affected the success of recruiting eligible research participants and complemented one another in the recruitment process. For example, a recruitment method may render a high eligibility rate, but also be less time efficient. More details about how the four measures were derived can be found in the Methods section.

### 3. Methods

#### 3.1. Recruiters and Recruits

To study the efficiency of methods to reach potential Asian participants, we used the recruitment data from a large cognitive testing study conducted by the U.S. Census Bureau. This study pretested the Chinese and Korean translations of the American Community Survey (ACS) questionnaire with Chinese and Korean speakers who spoke little or no English. The recruiters contacted 1,084 potential participants and administered 845 screening questionnaires that collected information on English language proficiency and additional information to ensure that a significant cross-section of the population was included (e.g., demographics, years in the United States, etc.). Among these 845 screened participants, 497 potential participants met the eligibility criterion of being non-English speakers of the target languages. Ultimately, 258 cognitive interview participants were selected to represent the diverse demographics desired for the testing, a figure that mirrored the targeted Korean or Chinese populations who were likely to need the translated ACS language guide based on the previous ACS statistics. A team of cognitive interviewers conducted the cognitive interviews in the greater Washington, DC area, Illinois, and North Carolina.

This ACS study used eight recruiters who spoke both English and a target language (Chinese or Korean). Three of the eight recruiters (two Chinese and one Korean) had strong ties to their ethnic communities; they interacted regularly with Chinese or Korean speakers at their workplace (e.g., a social service agency serving immigrants and a language school). We categorized the rest of the recruiters as having weaker ethnic ties. They had access to the Chinese and Korean communities through their personal social networks, but the interactions were not as regular or extensive as those with strong ties.

### 3.2. Recruitment Methods

The recruitment materials were written in Chinese and Korean and all contained the same recruitment message. Directed at Chinese and Korean speakers, the message stated the purpose of the study, the length of the interview, and the monetary incentive. It also specified the study sponsor. Each recruiter was responsible for three of the four recruitment methods: flyers, online communication, and word of mouth. One recruiter per language managed the fourth method, advertising in Chinese and Korean language newspapers.

Recruiters carried out similar specific activities for each recruitment method. For example, they posted flyers at locations that potential participants frequented, such as ethnic business areas or churches with large congregations of Chinese and Korean origins. They also sent electronic messages using ethnic group e-mail lists that key informants suggested and posted online advertisements on Chinese and Korean language websites. These websites were intended for immigrants to exchange information. They usually contained an electronic bulletin board for questions and answers and an internet-based classified section for jobs, housing, and announcements relevant to Asian immigrants. Both the group e-mailing list and the websites could be considered “cold calls”, because the recruiters did not usually know who the recipients were. For the word-of-mouth method, recruiters contacted local community centers, engaged community leaders or key informants to spread the recruitment message, recruited participants in group gatherings, and also asked people to spread the message. Therefore, the specific activities conducted for each recruitment method were the same across all recruiters, regardless of their level of ethnic ties. To control cost, the recruiters visited community centers and business areas closer to them; thus recruitment sites differed somewhat. The activities of posting flyers, online communication, and placing newspaper advertisements were most active at the beginning of the recruitment period. After they delivered the recruitment message using those three methods, recruiters answered incoming calls from potential recruits (those who responded to the recruitment message) and then administered the screening questionnaire. As part of the screening process, recruiters asked where the potential participant first saw the recruitment message, which helped to identify the specific recruitment method.

One method – word-of-mouth activities – continued throughout the recruitment period because it required relationship building and continued interactions with key informants and others. Sometimes the recruiters had to travel to meet with the key informants or screen participants in person when they recruited participants in group gatherings. Usually, potential participants recruited through word of mouth were screened on the telephone, as were those potential participants recruited from flyers, online communication, and newspaper advertisements.

### 3.3. Recruitment Data

To support the systematic analysis of recruitment methods, study researchers documented detailed information about all recruitment activities. First, the recruiters recorded the answers to the screening questionnaire (“screener”) to determine a recruit’s eligibility for the study. They also recorded the demographic characteristics of the potential participants, the recruitment source (where the potential participants first heard about the study), and a record of calls (“contact history”). The contact history recorded the time each recruiter

spent in an attempt to establish the contact, the mode of contact (phone vs. in person), and the contact result (whether a screener was completed). In addition, the recruiters kept a detailed list of activities that led to the contact with a recruit (“recruitment activity record”), including the type of activity (e.g., travel to ethnic business areas, calling community centers), date and time spent for the recruitment activity, and the financial costs, such as newspaper advertisement fees.

### 3.4. Efficiency Measures

To improve on the exploratory efficiency measures used by Liu et al. (2013), we developed a complete set of measures as shown in Table 1: time efficiency, outreach capacity, screener completion rate, and eligibility rate. Using these measures, we conducted statistical tests to evaluate the efficiency of the common recruitment methods used to recruit this large group of Asian research participants.

**Time efficiency**, the first measure, indicated the recruitment time spent for a potential participant. Recruitment time refers to the time recruiters spent on conducting recruitment activities. For example, for a newspaper advertisement a recruiter developed the ad, identified the appropriate newspapers, communicated with the newspapers, and coordinated the details to ensure that the ad was published as requested. Because the time efficiency measure was meant for recruitment activities only, we did not include time spent administering screeners.

**Outreach capacity**, the second efficiency measure, assessed the power of a recruitment method by how many potential participants and associated calls (i.e., number of people) were generated following one particular recruitment attempt. A recruitment attempt was defined as each time a recruiter tried to establish contact with a potential participant.

**Screener completion rate**, the third efficiency measure, showed how many interested potential participants actually completed the screener without breaking off. If potential participants resulting from a particular recruitment method broke off at a higher rate (low screener completion rate), the value of the recruitment method decreased, even though that method may have reached a large number of people (outreach capacity) while spending less time (time efficiency). We reason that the ultimate purpose of recruitment is to secure eligible research participants. Without completed screeners, the time spent in

Table 1. Efficiency measures and calculation formula

Efficiency measures	Formula
Time efficiency	$\frac{\text{Total recruitment time}}{\text{Total number of potential participants}}$
Outreach capacity	$\frac{\text{Total number of potential participants}}{\text{Total number of recruitment attempts}}$
Screener completion rate	$\frac{\text{Total number of potential participants who completed screeners}}{\text{Total number of potential participants}}$
Eligibility rate	$\frac{\text{Total number of non-English speakers}}{\text{Total number of screened participants}}$

recruitment was essentially wasted because recruiters could not determine the potential participants' eligibility.

**Eligibility rate**, the final efficiency measure, showed the proportion of non-English speakers among screened potential participants. If potential participants from a particular recruitment method were ineligible at a higher rate, the value of that recruitment method decreased, even though the methods tested well for the other three measures. Because the purpose of the study was to pretest the translation with potential users of the translation, those who spoke English very well did not qualify and were screened out. Therefore, the eligibility rate was critical in assessing the efficiency of a recruiting method.

The formula in [Table 1](#) provides more details: For time efficiency, a lower number indicates greater efficiency, meaning the recruiters spent less time (either for incoming calls or outgoing calls) to produce one call from potential participants. For outreach capacity, screener completion rate, and eligibility rate, a higher number indicates greater efficiency. In other words, more people were reached per each recruitment attempt (outreach capacity), more people recruited via a certain recruitment method completed screeners (screener completion rate), and more people who fit our eligibility criteria (i.e., being a non-English speaker) were recruited (eligibility rate). A summary of potential participant demographics, number of potential participants recruited by each recruitment method, and the recruiter characteristics is presented in the appendix.

## 4. Analysis and Findings

### 4.1. Efficiency of Recruitment Methods

To analyze the efficiency of the recruitment methods in locating Asian research participants, we began by comparing the four methods (newspaper ads, flyers, online communication, and word of mouth) with the four efficiency measures: time efficiency, outreach capacity, screener completion rate, and eligibility rate. As shown in [Table 2](#), newspaper advertisements achieved the highest time efficiency and outreach capacity. Specifically, potential participants responding to newspaper ads took 0.8 minutes of recruitment time compared with 5.4 minutes for online communication, 11.3 minutes for word of mouth, and 13.9 minutes for flyers. In addition, the outreach of newspaper advertisements brought 39.5 calls per one recruitment attempt, followed by online (3 calls), flyer (2 calls), and word of mouth (1.8 calls).

Table 2. Recruitment efficiency across four recruitment methods

Recruitment method	Newspaper advertisements ( <i>n</i> = 237)	Flyers ( <i>n</i> = 192)	Online communication ( <i>n</i> = 140)	Word of mouth ( <i>n</i> = 276)
Time efficiency (minutes)	0.8	13.9	5.4	11.3
Outreach capacity (frequency of calls)	39.5	2.0	3.0	1.8
Screener completion rate (% calls)	95	91	74	93
Eligibility rate (% screened callers)	54	59	49	86



For the third measure of efficiency (screener completion rate), online communication produced the lowest screener completion rate (74%). We found that potential participants failed to complete the screener more often when they were recruited online than when recruited by any other method. Almost three out of ten calls originating from online communication broke off during screening, thus preventing the recruiters from gathering enough information to determine whether the recruits were eligible to participate. As a comparison, the screener completion rates of other recruitment methods were much higher, above 90%.

Finally, for the fourth efficiency measure (eligibility rate), word of mouth had the highest eligibility rate. About 86% of screened individuals recruited via word of mouth were non-English speakers, whereas eligibility rates of recruits via other methods remained between 49% and 59%.

#### 4.2. *Efficiency of Recruitment Methods by Language Groups and by Recruiters*

The next step in analyzing the efficiency of the recruitment methods was to further examine whether the same patterns held true for language groups (Chinese vs. Korean). Our analysis showed an overall similarity between the efficiency pattern as a whole and at the subgroup levels. As shown in [Table 3](#), regardless of language, newspaper advertisements ranked as the most efficient method for time efficiency and outreach capacity, and word of mouth showed the highest eligibility rate. Screener completion rates, on the other hand, varied somewhat across languages. Screener completion rate of potential participants recruited by physical flyers ranked the lowest for Chinese speakers (which had a small number of recruits  $n = 17$ ), while the screener completion rate of potential participants recruited by via online communication was lowest for Korean speakers. Word of mouth and newspaper advertisements were efficient in generating high screener completion rates for both languages.

To further analyze the effectiveness of our recruitment methods, we examined the efficiency measures to see whether a general efficiency pattern of the methods emerged among recruiters who had varying levels of connection to the Chinese or Korean communities. Our analysis revealed that regardless of whether the recruiters had strong ties to ethnic communities, ethnic newspaper advertising reached a larger number of people in less time than other methods, and word of mouth reached more non-English speakers than other methods. Again, online communication had the lowest screener completion rates, while the rates were about the same (above 90%) for potential participants recruited via other methods, regardless of the level of recruiters' ethnic ties. These results corroborate the efficiency pattern observed in the language subgroup analyses. Specific efficiency scores are shown in [Table 3](#).

#### 4.3. *Statistical Testing of the Efficiency of Recruitment Methods*

The next crucial step in our analysis was to conduct statistical testing of these findings. We conducted analysis of variance (ANOVA) and post-hoc analyses (Tukey test) to investigate the statistical significance of our findings and to understand more about the differences among the recruiting methods. This testing was only available for two (screener completion rate and eligibility rate) of the four efficiency measures because the



Table 3. Efficiency measures by language and strength of recruiters' ethnic ties

Measure	Total (N = 845)	Language		Recruiters ties	
		Chinese (n = 261)	Korean (n = 584)	Strong (n = 253)	Weaker (n = 592)
<i>Time Efficiency (minutes)</i>					
Newspaper advertisements	0.8	3.0	0.3	NA	1.0
Physical flyers	13.9	27.1	12.7	21.3	12.3
Online communication	5.4	13.6	1.5	18.6	2.2
Word of mouth	11.3	14.3	7.8	15.7	6.4
<i>Outreach Capacity (frequency of calls)</i>					
Newspaper advertisements	39.5	12.5	93.5	NA	31.8
Physical flyers	2.0	2.4	2.0	1.3	2.4
Online communication	3.0	1.6	5.3	0.8	8.1
Word of mouth	1.8	1.6	2.3	1.2	4.1
<i>Screening Completion Rate (% calls)</i>					
Newspaper advertisements	95	98	94	100	93
Physical flyers	91	71	93	91	91
Online communication	74	80	72	85	72
Word of mouth	93	90	98	91	96
<i>Eligibility Rate (% screened callers)</i>					
Newspaper advertisements	54	80	47	70	50
Physical flyers	59	83	57	72	56
Online communication	49	72	37	56	47
Word of mouth	86	87	85	89	83

other two measures (time efficiency and outreach capacity) were based on the sum of all recruitment time, recruitment attempts, and number of calls by potential participants, and we therefore constructed measures consisting of different entities to facilitate comparisons of recruitment methods in different aspects, which prevented further statistical testing. For this reason we used screener completion rate and eligibility rate only as separate dependent variables and recruitment methods as the independent variables.

As shown in [Table 4](#), the differences of screener completion rates across recruitment methods were statistically significant as a whole ( $F = 16.9, p < .0001$ ) ( $F = 28.4, p < .0001$ ) as well as at the subgroup levels ( $F = 4.6, p = .0037$  for Chinese,  $F = 18.8, p < .0001$  for Korean,  $F = 16.6, p < .0001$  for recruiters with weaker ties). The only exception was for recruiters with strong ties. The post-hoc analyses showed that the difference between the online and other recruitment methods drove the overall statistical significance of the ANOVA test. The differences among the other three methods (flyers, newspaper ads, and word of mouth) appeared equally efficient in terms of screener completion rates, and none of them reached statistical significance except for one difference between flyers and newspaper ads in the Chinese data.

Our analyses also revealed that the pattern shown via the ANOVA and post-hoc analysis of eligibility rate was somewhat similar to the outcome of the screener completion rate analysis in one aspect – one particular recruitment method was quite different from other methods and it drove the statistical significance. The differences in the eligibility rates across recruitment methods were statistically significant as a whole ( $F = 28.4, p < .0001$  for eligibility rate) as well as all the subgroups, with the exception of the Chinese data ( $F = 21.4, p < .0001$  for Korean,  $F = 7.0, p = .0002$  for recruiters with strong ties,  $F = 14.4, p < .0001$  for recruiters with weaker ties). The post-hoc analyses indicated that the difference between word of mouth and other recruitment methods drove the overall statistical significance of the ANOVA test. The other three methods (flyers, newspaper ads, and online) appeared equally efficient in reaching Chinese or Korean speakers who also spoke little or no English, and none of these differences reached statistical significance. The group differences that had statistical significance ( $p = .05$ ) are summarized with an asterisk (\*) in [Table 4](#).

#### 4.4. Efficiency Measures by Recruiters

As an additional examination of the methods used, we compared the efficiency measures according to recruiters' ethnic ties. As show in [Table 5](#), recruiters with strong ties to the ethnic community had completed screeners at a higher rate (92% vs. 89%) and recruited a higher number of non-English speakers (80% vs. 59% in eligibility rate). However, recruiters with strong ties exhibited less efficiency in two respects. Compared with their counterparts who had weaker ties to the ethnic community, these recruiters tended to be less time-efficient in making contacts (13.9 minutes vs. 5.4 minutes) and had a much lower outreach capacity compared with recruiters with weaker ties (1.4 potential participants vs. 5.0 potential participants).

Because recruiters with strong ethnic ties had more frequent and regular access to non-English speakers than those with weaker ties, we suspected that recruiters with strong ethnic ties relied on the word-of-mouth method more often. If that is the case, the effect of

Table 4. ANOVA and Tukey test of screener completion and eligibility rate

Source comparison	Difference between means by					
	Total (n = 845)	Language			Recruiters ties	
		Chinese (n = 261)	Korean (n = 584)	Strong (n = 253)	Weaker (n = 592)	
<b>Screener Completion Rate (% calls)</b>						
Word of mouth vs. flyers	0.02	0.19	0.04	0.00	0.05	
Word of mouth vs. newspapers	-0.01	-0.08	0.04	-0.09	0.03	
Word of mouth vs. online	0.19*	0.10	0.26*	0.06	0.25*	
Flyers vs. newspapers	-0.03	-0.27*	0.00	-0.09	-0.02	
Flyers vs. online	0.17*	-0.10	0.22*	0.06	0.19*	
Newspapers vs. online	0.20*	0.18*	0.22*	0.15	0.22*	
<i>F</i>	16.9	4.6	18.8	2.0	16.6	
<i>p</i> -value	<.0001	0.0037	<.0001	0.114	<.0001	
<b>Eligibility Rate (% screened callers)</b>						
Word of mouth vs. flyers	0.28*	0.04	0.28*	0.18	0.27*	
Word of mouth vs. newspapers	0.33*	0.08	0.38*	0.20*	0.33*	
Word of mouth vs. online	0.38*	0.16	0.48*	0.33*	0.36*	
Flyers vs. newspapers	0.05	0.04	0.11	0.03	0.06	
Flyers vs. online	0.1	0.12	0.21*	0.16	0.1	
Newspapers vs. online	0.05	0.08	0.11	0.14	0.04	
<i>F</i>	28.4	1.7	21.4	7.0	14.4	
<i>p</i> -value	<.0001	0.1618	<.0001	0.0002	<.0001	

Table 5. Recruitment efficiency by strength of recruiters' ethnic ties

Measures	Recruiters ties	
	Strong (n = 253)	Weaker (n = 592)
Time Efficiency (minutes)	13.9	5.4
Outreach Capacity (frequency of calls)	1.4	5.0
Screener Completion Rate (% of calls)	92	89
Eligibility Rate (% screened callers)	80	59

a recruiter's ethnic ties was commingled with the effect of word of mouth, and we should not simply interpret that they recruited at a higher eligibility rate, but were far less efficient in other measures. By running a cross-tabulation of the recruitment methods and recruiters' levels of community ties, we confirmed that recruiters with strong ethnic ties depended heavily on word of mouth: They used word of mouth most often (57.3%) among the four methods, and their dependence on word of mouth was conspicuous compared with other recruiters (22.1%), reaching statistical significance (Chi-square = 99.84,  $p < .0001$ ). This tendency was true for both Chinese (78.5% of recruiters with strong ties versus 40% of recruiters with weak ties, Chi-square = 69.5 [ $p < .0001$ ]) and Korean data (39.4% of recruiters with strong ties vs. 16.3 % of recruiters with weak ties, Chi-square = 40.5 [ $p < .0001$ ]).

As explained in the methods section, the word-of-mouth method sometimes involved traveling to a specific location and having face-to-face interaction. It is therefore not surprising that this method was less efficient in terms of time efficiency and outreach capacity. What is less clear, however, is why word of mouth worked better for recruiting non-English speakers compared with the other recruitment methods. To tease out the effect of word of mouth and recruiters with strong ethnic ties, we set up a logistic regression to predict whether screened individuals were non-English speakers, and we used recruitment methods and recruiters' ethnic ties as independent variables. To explain the different effects of recruiter's ethnic ties on the word-of-mouth recruitment method, an interaction term between word of mouth and recruiters with strong ties was included. Potential participants' language (Chinese or Korean) was also included in the model to explain language differences. All of the independent variables were included as dummy variables: online communication, recruiters with weaker ties, and Korean speakers were the reference groups to be used as the baseline of the parameter interpretation. The Logistics regression model is as follows:

$$f(z) = \frac{e^z}{e^z + 1} = \frac{1}{1 + e^{-z}},$$

$$Z = \beta_0 + \beta_1 * \text{newspaper} + \beta_2 * \text{Flyer} + \beta_3 * \text{Word-of-mouth} + \beta_4 * \text{Recruiter with strong ties} + \beta_5 * \text{Interaction of recruiters with strong ties and Word-of-mouth} + \beta_6 * \text{Chinese speakers}$$

First, we ran the model with the entire data set. The overall model was adequate (Wald Chi-Square = 90.96 [ $p < .0001$ ]) to predict the dependent variable – whether the

screened potential participants were non-English speakers – and we can reject the global null hypothesis that none of the independent variables in the model were related to changes in the probability of reaching non-English speakers. As seen in [Table 6](#), through the Wald Chi-Square test of the parameter estimates, we found statistically significant positive estimates of flyers (0.9), word of mouth (2.13), recruiters with strong ties (1.04), and Chinese speakers (1.04). These findings show the positive contribution of these variables to the dependent variable. That is, potential participants recruited via flyers, word of mouth, or potential participants recruited by recruiters with strong ethnic ties had an increased probability of being non-English speakers compared with the reference groups. These results confirm that using word of mouth and having recruiters with strong ethnic ties have independent effects on recruiting qualified non-English speaking research participants.

## 5. Discussion

In this study, we analyzed the efficiency of four recruitment methods (newspaper advertisements, flyers, online communication, and word of mouth) using four criteria: time spent, outreach capacity, screener completion, and eligibility rate. We also examined differences in recruitment efficiencies by recruiters and sublanguage groups. Our findings show that newspaper advertisements are the most efficient method for reaching a larger number of Asian research participants in less time, whereas word of mouth works best for reaching non-English-speaking participants. This finding holds true for both Chinese and Korean speakers, regardless of whether the recruiters had strong ties to the ethnic community. These findings echoed prior recruitment research that shows media broadcasting or printed materials reach a larger portion of the general population ([Appel et al. 1999](#); [Gilliss et al. 2001](#); [McLean and Campbell 2003](#)), and that word of mouth was effective for recruiting Hispanics who spoke little or no English ([Sha et al. 2010](#)).

We also found that the differences of screener completion rates across the recruitment methods were statistically significant. As a whole and at the subgroup levels, online communication was less efficient in persuading interested potential participants to complete the screening questionnaire compared to other recruitment methods (newspaper advertisements, physical flyers, word of mouth). Almost three out of ten calls originating from online communication broke off during screening, thus preventing the recruiters

*Table 6. Analysis of maximum likelihood estimates from logistic model predicting non-English speaking research participants*

Parameter	Estimates	Wald Chi-Square	Pr > ChiSq
Intercept	-0.74	7.97	0.0048
Newspapers	0.34	1.20	0.561
Flyers	0.9	<b>8.75</b>	0.0031
Word of mouth	2.13	<b>28.62</b>	< .0001
Recruiters with strong ties	1.04	<b>14.88</b>	0.0001
Interaction of word of mouth and strong ties	-0.14	0.06	0.8113
Chinese speakers	1.04	<b>23.33</b>	< .0001

from gathering enough information to determine whether the recruits were eligible non-English speakers. As a comparison, the screener completion rates of other recruitment methods were much higher, above 90%. A possible explanation is that messages from the Internet are commonly considered less credible than traditional media platforms, such as newspapers (Flanagin and Metzger 2000; Koo and Skinner 2005). As such, people who realized that they had to answer a screening questionnaire might have broken off the call rather than spend time to complete the screener. Completion of screeners is important because we cannot determine potential participants' eligibility or obtain their contact information without information gathered during the screening process. We may also gauge potential participants' willingness to participate based on whether they broke off or completed the screeners. To our knowledge, however, no literature supports or dismisses the notion that people recruited via online methods are less likely to complete the screener.

It should be noted that the online recruitment methods used in this study were "cold calls" because they were limited to the use of ethnic group emailing lists and postings on Chinese and Korean language websites. However, "online" can be just a medium for communication and recruiters could have also sent the recruitment message to their acquaintances. If such a personal contact was involved in online methods, our finding of low screener completion rate via online recruits may not hold true.

The differences among recruitment efficiencies, eligibility rate in particular, across the recruitment methods were statistically significant, and this statistical significance was driven by the unique nature of the word-of-mouth method. These findings were reported previously by Yancey et al. (2006): Word of mouth is distinct from the other reactive recruitment methods (newspaper advertisements, flyers, and online communication) that require potential participants to take the initiative to respond. By contrast, word of mouth is a proactive method for which recruiters actively seek qualified research participants among the target groups. However, we also found that the word-of-mouth method demands significantly more time to deliver the necessary recruitment message because of its in-person communication format.

Because our study analyzed the recruitment of Chinese and Korean speakers, the finding that recruiters spend more time using the word-of-mouth method during a recruitment attempt may not be surprising. Asians may have a culturally based expectation that the recruiters will create a friendly context before stating their needs (i.e., delivering the actual recruitment message). The emphasis on harmonious interpersonal relationships and politeness in the Asian culture (Markus and Kitayama 1991) most likely explains this expectation and finding. As described in the methods section, the word-of-mouth method in this study may necessitate in-person visits, and the level of personal interactions was relatively high compared with other recruitment methods. The cultural expectations that come with personal interactions could have contributed to widening the gap between word of mouth and the rest of the recruitment methods in this study.

In addition, our analysis showed that recruiters with strong ties to the Asian communities used the word-of-mouth method quite extensively. As corroborated in prior literature on participant recruitment, recruiters with strong ties were successful in reaching eligible participants (i.e., non-English speakers in this study). Compared with recruiters with weaker ties, recruiters with strong ties spent relatively more time recruiting (and thus were less time efficient) and reached a smaller number of people at each recruitment

attempt (lower outreach capacity). The logistic regression analysis teased out the commingled effect of recruiters with strong ties and word of mouth by controlling for recruitment methods. These results demonstrate that recruiters with strong ethnic ties were more successful in reaching non-English speakers regardless of the recruitment methods they used. Not surprisingly, recruiters with strong ethnic ties had more success in recruiting non-English speakers because they interacted with the target populations at a higher frequency and intensity. Compared to their counterparts, recruiters with strong ties are more likely to know the target population better simply because of more frequent interactions, and they could focus their recruitment efforts on these highly eligible participants group with this prior knowledge. We can interpret the high eligibility rate for recruiters with strong ties as the combined results of easy access to the target populations, prior knowledge about the target population, and the trust they could more easily establish by the virtue of their access. The relatively low time efficiency and low outreach capacity results can be explained easily by these recruiters' heavy dependence on word of mouth among the four recruitment methods.

The measures developed in this study may be applicable to future studies even when they are not cross-cultural. We evaluated the four recruitment methods commonly used in many research studies, and our findings strengthened prior literature on participant recruitment reported for the general population as well as minorities. The unique issues facing recruitment of non-English speakers may be identifying the most popular outlets for print and online advertisements as well as identifying and retaining recruiters with strong ties to ethnic communities.

In addition to its contributions to the research, this study has several limitations. First, we only recruited non-English speaking Chinese and Koreans in Illinois, North Carolina, and the greater Washington, DC area. These sites were selected because of their proximity to highly skilled Chinese- and Korean-speaking cognitive interviewers and because many people of Chinese and Korean origin live in these areas. For example, we were able to choose from several Korean language newspapers in the greater Washington, DC area, which resulted in the successful recruitment of eligible Korean speakers. The same result might not be achieved in other areas of the country or among other population groups. Second, in reality people may see the recruitment message disseminated by multiple recruitment methods (e.g., in a newspaper advertisement and a flyer) and then decide to participate. In our analysis, we had to assume that a potential participant was contacted by only one method because there was no systematic way to determine which method encouraged participation.

## 6. Conclusion

This article has discussed common methods to recruit research participants, which include placing print or online advertisements and getting help from the potential participants' community. Few research studies have examined the efficiency of recruiting non-English speaking participants, in particular, because empirical studies were not conducted or they were lacking in systematic analysis. We attempted to fill this research gap by examining the efficiency of common recruitment methods using the measures of time efficiency, outreach capacity, screener completion rate, and eligibility rate. We used large-scale

recruitment data of Asian research participants – that is, Chinese and Korean speakers – in a cognitive testing study.

Our findings show that the word-of-mouth method identified more non-English speakers, and newspaper advertisements reached a larger number of people in less time. In general, callers recruited via online communication completed screener questionnaires at a lower rate. We also looked into the effect of using recruiters with strong ethnic ties and found that when controlling for recruitment methods, they had a positive effect on recruiting for non-English speakers. In addition, the four recruitment methods we evaluated are commonly used in many research studies; therefore, we believe that the efficiency measures we developed in this study may inform systematic analysis of the recruitment methods to recruit other minority populations or the general population. Clearly, this area could benefit from future research.

Based on the findings of this study, we recommend choosing recruitment methods appropriate for the particular needs of the study. If a study needs to recruit eligible Asian non-English speakers quickly, newspaper advertisements are most efficient. However, researchers may not wish to use both newspapers and flyers because the demographics (i.e., age, education, home ownership status, immigration year) of the individuals recruited via these two methods are similar. For example, our earlier analysis in [Park et al. \(2011\)](#) showed that the average age and the gender distribution of the recruits who responded to newspapers and flyers were very similar. In addition, when a study needs to recruit under very specific eligibility criteria, the word-of-mouth method focusing on the targeted group of people and desired characteristics for the research would likely render more success. Online methods do not seem advantageous for recruiting non-English speaking Asian participants, particularly since online methods have low screening completion rates as well as low eligibility rates. Furthermore, involving recruiters with strong ties to the target community will likely recruit more eligible non-English speakers. Through their strong ties, these recruiters are usually trusted members of the community and can increase the study's credibility.



**Appendix. Characteristics of Recruits ( $n = 845$ )**

	Categories	Frequency (%)		Categories	Frequency (%)
Language	Chinese	261 (30.9%)	Gender	Female	476 (62.6%)
	Korean	584 (69.1%)		Male	285 (37.5%)
Recruited from	Illinois	394 (46.6%)	Age	18–24	59 (7.8%)
	VA/DC/MD	306 (36.2%)		25–34	119 (15.6%)
	North Carolina	145 (17.2%)		35–44	139 (18.3%)
				45–54	194 (25.5%)
		55–64		127 (16.7%)	
Recruited by Recruiters Whose tie to Ethnic Community was:	Strong	253 (29.9%)			
	Weaker	592 (70.1%)			
Recruitment Source	Newspaper	237 (28.1%)	Education	65 +	123 (16.2%)
	Flyer	192 (22.7%)		Less than High School	94 (12.4%)
	Online	140 (16.6%)		High School	243 (31.9%)
	Word of Mouth	276 (32.7%)		College or above	424 (55.7%)
Eligibility	English Speaking	264 (34.7%)	Entry to the U.S.	Before 1980	56 (7.4%)
	Non-English Speaking (eligible)	497 (65.3%)		1980–1989	125 (16.4%)
		1990–1999		177 (23.3%)	
		2000–2009		310 (40.7%)	
		After 2010		93 (12.2%)	
Number of Contacts per potential participant	1 time	665 (78.7%)			
	2 times	146 (17.3%)			
	3 + times	34 (4.0%)			

**7. References**

- Appel, L.J., Vollmer, W.M., Obarznek, E., Aicher, K.M., Conlin, P.R., Kennedy, B.M., Chaleston, J.B., Reams, P.M., and the DASH Collaborative Research Group (1999). Collaborative Research Group Recruitment and Baseline Characteristics of Participants in the Dietary Approaches to Stop Hypertension Trial. *Journal of the American Dietetic Association*, 99, S69–S75, DOI: [http://www.dx.doi.org/10.1016/S0002-8223\(99\)00419-8](http://www.dx.doi.org/10.1016/S0002-8223(99)00419-8).
- Aréán, P.A., and Gallagher-Thompson, D. (1996). Issues and Recommendations for the Recruitment and Retention of Older Ethnic Minority Adults into Clinical Research. *Journal of Consulting and Clinical Psychology*, 64, 875–880.
- Berrigan, D., Forsyth, B.H., Helba, C., Levin, K., Norberg, A., and Willis, G. (2010). Cognitive Testing of Physical Activity and Acculturation Questions in Recent and Long-Term Latino Immigrants. *BioMed Central Public Health*, 10(481). Available at: <http://www.biomedcentral.com/content/pdf/1471-2458-10-481.pdf> (accessed March 2012).

- Bistricky, S.L., Mackin, R.S., Chu, J.P., and Arean, P.A. (2010). Recruitment of African Americans and Asian Americans with Late Life Depression and Mild Cognitive Impairment. *American Journal of Geriatric Psychiatry*, 18, 734–742.
- Coleman, E.A., Tyll, L., LaCroix, A.Z., Allen, C., Leveille, S.G., Wallace, J.I., Buchner, D.M., Grothaus, L.C., and Wagner, E.H. (1997). Recruiting African-American Older Adults for a Community Based Health Promotion Intervention: Which Methods Are Effective? *American Journal of Preventive Medicine*, 13, 51–56.
- Flanagin, A.J., and Metzger, M.J. (2000). Perceptions of Internet Information Credibility. *Journalism and Mass Communication Quarterly*, 73, 515–540.
- Forsyth, B.H., Kudela, M.S., Levin, K., Lawrence, D., and Willis, G. (2007). Methods for Translating an English-Language Survey Questionnaire on Tobacco Use into Mandarin, Cantonese, Korean and Vietnamese. *Field Methods*, 19, 264–283.
- Fujimoto, W.Y. (1998). Community Involvement and Minority Participation in Clinical Research. *Diabetes Spectrum*, 11, 161–166.
- Gilliss, C.L., Lee, K.A., Gutierrez, Y., Taylor, D., Beyene, Y., Neuhaus, J., and Murrell, N. (2001). Recruitment and Retention of Healthy Minority Women into Community-based Longitudinal Research. *Journal of Women's Health Gender Based Medicine*, 10, 77–85, DOI: <http://www.dx.doi.org/10.1089/152460901750067142>.
- Gorelick, P.B., Richardson, D., Hudson, E., Perry, C., Robinson, D., Brown, N., and Harris, Y. (1996). Establishing a Community Network for Recruitment of African Americans into a Clinical Trial. The African American Antiplatelet Stroke Prevention Study (AAASPS) Experience. *Journal of the National Medical Association*, 88, 701–704.
- Harris, K.J., Ahluwalia, J.S., Catley, D., Okuyemi, K.S., Mayo, M.S., and Resnicow, K. (2003). Successful Recruitment of Minorities into Clinical Trials: The Kick It at Swope Project. *Nicotine Tobacco Research*, 5, 575–584.
- Holcombe, R.F., Jacobson, J., Li, A., and Moinpour, C.M. (1999). Inclusion of African Americans in Oncology Clinical Trials. *American Journal of Clinical Oncology*, 22, 18–21.
- Hughes, C., Peterson, S., Ramirez, A., Gallion, K., McDonald, P.G., Skinner, C.S., and Bowen, D. (2004). Minority Recruitment in Hereditary Breast Cancer Research. *Cancer Epidemiology Biomarkers Prevention*, 13, 1146–1155.
- Kim, J., and Zapata, J. (2012). 2010 Census Language Program Assessment Report. 2010 Census Planning Memoranda Series, No. 204. Washington, DC: U.S. Census Bureau. Available at: [http://www.census.gov/2010census/pdf/2010\\_Census\\_Language\\_Program\\_Assessment.pdf](http://www.census.gov/2010census/pdf/2010_Census_Language_Program_Assessment.pdf) (accessed March 2013).
- Koo, M., and Skinner, H. (2005). Challenges of Internet Recruitment: A Case Study with Disappointing Results. *Journal of Medical Internet Research*, 7, E6, DOI: <http://www.dx.doi.org/10.2196/jmir.7.1.e6>.
- Lau, A., and Gallagher-Thompson, D. (2002). Ethnic Minority Older Adults in Clinical and Research Programs: Issues and Recommendations. *The Behavior Therapist*, 25, 10–11.
- Lai, G.Y., Gary, T.L., Tilburt, J., Bolen, S., Baffid, C., Wilson, R.F., Howerton, M.W., Gibbson, M.C., Tanpitukpongsee, T.P., Powe, N.R., Bass, E.B., and Ford, J.G. (2006). Effectiveness of Strategies to Recruit Underrepresented Populations into Cancer

- Clinical Trials. *Clinical Trials*, 3, 133–141. DOI: <http://dx.doi.org/10.1191/1740774506cn143oa>.
- Liu, L., Sha, M., and Park, H. (2013). Exploring the efficiency and utility of methods to recruit non-English speaking qualitative research participants. *Survey Practice*, 6(3), 1–8. Available at: <http://www.surveypractice.org/index.php/SurveyPractice> (accessed December 2013).
- Markus, H.R., and Kitayama, S. (1991). Culture and the Self: Implications for Cognition, Emotion, and Motivation. *Psychological Review*, 20, 568–579.
- Maxwell, A.E., Bastani, R., Vida, P., and Warda, S. (2005). Strategies to Recruit and Retain Older Filipino-American Immigrants for a Cancer Screening Study. *Journal of Community Health*, 30, 167–179. DOI: <http://www.dx.doi.org/10.1007/s10900-004-1956-0>.
- McLean, C., and Campbell, C. (2003). Locating Research Informants in a Multi-ethnic Community: Ethnic Identities, Social Networks and Recruitment Methods. *Ethnicity and Health*, 8, 41–61.
- Pan, Y., Landreth, A., Hinsdale, M., Park, H., and Schoua-Glusberg, A. (2007). Methodology for Cognitive Testing of Translations in Multiple Languages. In *Proceedings of the Section on Survey Research Methods, Annual Conference of the American Association*, 3801–3808 (Alexandria, VA, July 29–August 2, 2007).
- Park, H., Liu, L., and Sha, M. (2011). Do Different Recruitment Methods Reach Different People? *Annual Conference of the American Association* (Phoenix, AZ, May 12–15, 2011).
- Reed, P.S., Foley, K.L., Hatch, J., and Mutran, E.J. (2003). Recruitment of Older African Americans for Survey Research: A Process Evaluation of Community and Church-based Strategy in the Durham Elders Project. *The Gerontologist*, 43, 52–61. DOI: <http://www.dx.doi.org/10.1093/geront/43.1.52>.
- Sha, M., McAvinchey, G., Reed, L., Rodriguez, S., and Carter, G. (2010). Respondent Recruitment, Interviewing, and Training: Lessons Learned From a Spanish Language Cognitive Interviewing Project. In *Proceedings of the Section on Survey Research Methods, Annual Conference of the American Association*, 6372–6381 (Alexandria VA, May 13–16, 2010).
- Stoy, D.B., Curtis, R.C., Dameworth, K.S., Dowdy, A.A., Hegland, J., Levin, J.A., and Sousoulas, B.G. (1995). The Successful Recruitment of Elderly Black Subjects in a Clinical Trial: The CRISP Experience. *Cholesterol Reduction in Seniors Program. Journal of the National Medical Association*, 87, 280–287.
- The DPP Research Group (2002). The Diabetes Prevention Program: Recruitment Methods and Results. *Controlled Clinical Trials*, 23, 157–171. DOI: [http://www.dx.doi.org/10.1016/S0197-2456\(01\)00184-2](http://www.dx.doi.org/10.1016/S0197-2456(01)00184-2).
- Wellens, T. (1994). The Cognitive Evaluation of the Nativity Questions for the Current Population Survey. In *Proceedings of the Section on Survey Research Methods, Annual Conference of the American Association*, 1204–1209 (Alexandria, VA, May 11–15, 1994).
- Wisdom, K., Neighbors, K., Williams, V.H., Havstad, S.L., and Tilley, B.C. (2002). Recruitment of African Americans with Type 2 Diabetes to a Randomized Controlled Trial Using Three Sources. *Ethnic Health*, 7, 267–278.

- Yancey, A.K., Ortega, A.N., and Kumanyika, S.K. (2006). Effective Recruitment and Retention of Minority Research Participants. *Annual Review of Public Health*, 27, 1–28, DOI: <http://www.dx.doi.org/10.1146/annurev.publhealth.27.021405.102113>.
- Yuan, Y.M., Wake, V., Park, H., and Nguyen, L. (2009). Conducting Cognitive Interviews with Linguistically Isolated Asian Populations. Presented in the 43rd International Field Directors and Technologies conference (Delray Beach, FL, May 17–20, 2009).

Received February 2013

Revised November 2013

Accepted November 2013

## Reaching Hard-to-Survey Populations: Mode Choice and Mode Preference

*Marieke Haan<sup>1</sup>, Yfke P. Ongena<sup>1</sup>, and Kees Aarts<sup>2</sup>*

This study assesses the effect of response-mode choices on response rates, and response-mode preferences of hard-to-survey populations: young adults, full-time workers, big city inhabitants, and non-Western immigrants. Using address-based sampling, a stratified sample of 3,496 households was selected. The first group of sample members was contacted face to face and could choose between a CAPI and web response mode. The second group, contacted by telephone, could choose between CATI and web. The third group, contacted by telephone, was randomly allocated to a response mode. Our address-based sampling technique was successful in reaching most of the hard-to-survey groups. Insufficient numbers of non-Western immigrants were reached; therefore this group was excluded from our analyses. In our mixed-effect models, no significant effects on the willingness to participate were found for mode choice. We found that full-time workers and young adults were significantly more likely to choose web over CAPI when contacted face to face.

*Key words:* Hard-to-survey groups; response-mode choice; mixed mode experiment.

### 1. Introduction

Collecting data from hard-to-survey populations is challenging; they are hard to reach and known for low cooperation rates after having been contacted (Stoop 2005). To address these data collection difficulties, survey designs have been adjusted to increase response rates for hard-to-survey groups (e.g., increasing the number of contact attempts, Feskens 2009). Obviously such designs need to be carefully selected. To achieve contact with and cooperation of hard-to-survey populations, designs need to be tailored to characteristics of the sample (Groves et al. 1992; Haan and Ongena 2014). To identify such targeted approaches is therefore an important task for survey researchers. However, there is no guarantee whatsoever that designs can actually be found that perform better for specific groups. The existing literature is ambiguous about the effects of variations in contact modes and response modes. With this article, we aim to contribute to the existing literature by simultaneously investigating the effects of contact modes and response-mode choices on response rates, and analyzing the response-mode choices of several hard-to-survey groups.

<sup>1</sup> University of Groningen, Faculty of Arts, PO BOX 716 Groningen 9700 AS, Groningen, the Netherlands. Email: [marieke.haan@rug.nl](mailto:marieke.haan@rug.nl) and [y.p.ongena@rug.nl](mailto:y.p.ongena@rug.nl)

<sup>2</sup> University of Twente – Political Science and Research Methods, PO Box 217, Enschede 7522, AE Overijssel, the Netherlands. Email: [c.w.a.m.aarts@utwente.nl](mailto:c.w.a.m.aarts@utwente.nl)

**Acknowledgments:** This research is part of a project that was funded by the Netherlands Organization for Scientific Research (NWO), grant #471-09-002.

Stoop (2005; 2007) identified groups in society that are hard to survey. In the experiment described in this article, four of Stoop's hard-to-survey groups were selected. First, young adults can be difficult to contact due to, for example, unlisted cell phone numbers (Holbrook et al. 2003) and outdoor obligations (Stoop 2005), but they are generally willing to cooperate (De Leeuw and Hox 1998). Second, households with more than one full-time worker can be difficult to contact because of their at-home pattern. However, when contacted, they are generally willing to cooperate (Goyder 1987). Third, inhabitants of highly urbanized cities may be reluctant to let strangers enter their homes and are therefore hard to reach and their attitude towards survey research can be more negative (Campanelli et al. 1997; Goyder et al. 1992; Groves and Couper 1998). Finally, non-Western first and second-generation immigrants (henceforth referred to as 'non-Western immigrants') are often thought to be difficult to survey. However, in some studies, their contact rates are low – perhaps because of periods spent abroad (Blohm and Diehl 2001) – but they can have higher cooperation rates than natives (Feskens et al. 2007; Feskens 2009). To specify the group of non-Western immigrants, the following definitions were used (Statistics Netherlands 2013a):

- Someone with a first-generation foreign background*: “Someone born abroad with at least one parent who was born abroad.”
- Someone with a second-generation foreign background*: “Someone born in the Netherlands who has at least one parent born abroad.”
- Someone with a non-Western background*: “Someone originating from a country in Africa, South America or Asia (excluding Indonesia and Japan) or Turkey.”

Previous research has shown that different sample members may have divergent preferences for modes of contact (De Leeuw and van der Zouwen 1992), or favor different modes of responding (Dillman et al. 1994; Groves and Kahn 1979). Researchers have been offering multiple response-mode options to enable contacted sample members to select the response mode of their choice (Dillman et al. 2009; Shih and Fan 2007). However, mixed-mode experiments have shown deviating results with respect to the effects that response-mode choices have on response rates: increasing response rates (e.g., Schneider et al. 2005), decreasing response rates (e.g., Millar and Dillman 2011), or no influence on response rates (e.g., Friese et al. 2010). In addition, not many studies have focused on the effects of response-mode choices on response rates of hard-to-survey populations.

The deviating results and the survey design possibilities of presenting mode choices led us to believe that offering response-mode choices still can have positive effects on response rates. Combinations other than mail/web response-mode choices and varying contact modes could increase response rates. In addition, only a few studies have explored the effects of contact modes and response modes on response rates of hard-to-survey populations. To fill in some of these blanks in survey research literature, we designed and conducted an experiment to address the following two key questions:

1. What are the effects of offering response-mode choices on the willingness to participate of hard-to-survey populations and sample members in general?
2. To what extent do hard-to-survey populations differ in response-mode choice?

Before presenting our experiment and results, we start by providing the necessary background on the advantages and disadvantages of using response-mode choices in a survey design. Then we describe different ways of how response-mode choices are implemented in concurrent survey designs. This is followed by a section on contact mode preferences of hard-to-survey groups. Finally, we discuss literature on response-mode preferences of hard-to-survey populations.

## 2. Theoretical Background

### 2.1. *The Advantages and Disadvantages of Response-Mode Choices*

Preferences for response modes have been expressed for face-to-face interviews (Groves and Kahn 1979), telephone interviews (Smyth et al. 2009), mail surveys (Millar et al. 2009), and web surveys (Miller et al. 2002; Ryan et al. 2002; Tarnai and Paxson 2004). Therefore it can be worth the effort to create survey designs in which hard-to-survey populations are offered specific response modes. Offering sample members a response-mode choice not only makes it possible for them to cooperate in their favorite response mode, but they are also more involved in the decision to participate in the survey. This involvement can create goodwill, resulting in greater willingness to participate in the survey (De Leeuw 2005).

However, a choice in response mode could also prove overly cognitively challenging (Medway and Fulton 2012; Schwartz 2004). Too many choices can lead to ‘choice overload’ or ‘overchoice situations’, which lead to difficulties in the decision-making process (Dhar 1997; Iyengar and Lepper 2000; Toffler 1971). It is not clear whether choice overload problems also play a role in survey participation, since most studies focusing on choice overload are conducted in the field of marketing research (i.e., consumers choosing products). However, researchers have speculated on the effect the number of response-mode choices could have on response rates (Gillian et al. 2010; Martin 2011).

Furthermore, some researchers argue that if response-mode preferences really exist, then sample members should choose their preferred mode when choice is offered (Diment and Garrett-Jones 2007; Millar and Dillman 2011). Of course, being offered a choice does not imply that people also consciously make a choice. Sample members may select the response mode they have been approached in, simply because they do not want to weigh the pros and cons of alternatives or they might not have a mode preference. Moreover, it can be too much of a burden for them to switch modes and therefore they choose the response mode in which they were contacted (Lynn 2013).

### 2.2. *Offering Response-Mode Choices in Concurrent Survey Designs*

As the Internet is a widely-used medium and the low costs of offering web response modes are attractive (Dillman 2007), many organizations want to offer a web response mode to sample members in addition to the survey mode already existing (e.g., mail). When two (or more) response modes are offered simultaneously as an actual choice during the first contact moment, a so-called concurrent survey design is in use. However, according to a meta-analysis of Medway and Fulton (2012), offering concurrent web/mail response mode choices does not have positive effects on response rates. They found significantly lower



response rates for designs in which a concurrent web option was included in a mail survey than for designs in which the web option was not added. Nineteen studies were included in their meta-analysis. Two of these studies reported increased response rates when a concurrent choice was offered between mail and web (Brady et al. 2003; Schneider et al. 2005). Some experiments found almost no effects on response rates when web options were offered alongside a mail questionnaire (Friese et al. 2010; Lesser et al. 2010), but many studies conclude that the concurrent choice between web and mail reduces response rates (Brøgger et al. 2007; Gentry and Good 2008; Griffin et al. 2001; Hardigan et al. 2012; Israel 2010; Millar and Dillman 2011; Radon et al. 2002; Schmuhl et al. 2010; Smyth et al. 2010; Turner et al. 2010; Werner and Forsman 2005; Ziegenfuss et al. 2010). In an attempt to explain this outcome, Medway and Fulton (2012) argue that response-mode choices might make the survey participation process too complex and that they are a distracting factor in the response process.

However, by presenting a mode choice in a certain way, survey designers can try to ensure that the survey participation process is not overly cognitively challenging. Tancreto et al. (2012) experimented with response-mode choices to determine the best method to present the new web mode of the American Community Survey. They tested both concurrent designs and sequential designs. Their so-called prominent choice strategy, with a concurrent choice between a mail questionnaire and highlighted web mode, achieved the highest response rates. Within the concurrent choice designs, more people responded by web in the prominent choice condition than in the nonprominent choice condition (no highlighting of the web mode possibility). However, in the designs with the sequential choices, more people responded by web than in the concurrent choice designs. After asking sample members about their choice behavior, the researchers found no strong motivational indicators to explain why sample members chose a specific mode. Some did indicate that they like mail questionnaires better than web surveys, but mostly choices were made for practical reasons such as a lack of Internet access or computer issues.

Furthermore, different varieties of response-mode choices in concurrent designs could affect response rates in another way. Offering other combinations of response modes to choose from than mail and web may be more successful (e.g., a computer-assisted telephone interview (CATI) and web). Furthermore, sample members can be questioned about their response-mode preferences in advance and their preferred mode can be offered when they are approached again for another survey or a follow-up questionnaire (Hoffer et al. 2007; Olson et al. 2012). Overall, in these studies, higher response rates were found for sample members who were immediately assigned to their preferred response mode. However, when studying mode preferences in a sequential design, Olson et al. (2012) did not find differences in response rates between the sample members who were given their preferred response mode immediately and those who received their preferred mode when recontacted.

The survey design possibilities of presenting response-mode choices and the goodwill that mode choice can create led us to believe that offering response-mode choices still can have positive effects on response rates. Although concurrent web options in mail surveys seem to have negative effects on response rates (Medway and Fulton 2012), other combinations of response-mode choices in a concurrent design could increase the



willingness of sample members to participate in the survey. Therefore we expect that sample members will be more willing to participate in a survey when they can choose their response mode than when they cannot choose a response mode (*Hypothesis 1*).

### 2.3. Contact Mode Preferences of Hard-to-Survey Populations

Before sample members can be offered a response-mode choice, they first have to be reached by a contact mode. Not much is known about sample members' contact mode preferences for the request to participate. For a variety of reasons, it is harder to obtain high response rates in telephone surveys than in face-to-face surveys (e.g., Groves 1977; Holbrook et al. 2003; Weeks et al. 1983). This argument is supported by results of a meta-analysis of 45 studies in which the highest completion rates were found for face-to-face surveys compared to telephone and mail surveys (Hox and de Leeuw 1994). Concentrating only on the contact moment, De Leeuw and van der Zouwen (1992) found lower response rates for telephone contacts than for face-to-face contacts. However, other researchers report no differences in response rates between telephone and face-to-face contacts (Scherpenzeel and Toepoel 2012) and between telephone and mail contacts (Wilkins et al. 1997).

Focusing on the four hard-to-survey groups studied in this article, according to Stoop (2005) young adults can be difficult to reach in general because of their outdoor obligations. Furthermore, this group is also known for having an unlisted cell phone number instead of a listed landline number (Blumberg and Luke 2007; Holbrook et al. 2003). Therefore it is likely that young adults are easier reached face to face than by telephone. Households with more than one full-time worker are difficult to reach in both of the contact modes (face-to-face and telephone) because of their at-home pattern. For this group the timing is more important, for example calling in the evenings (Weeks et al. 1983). According to the literature, big city inhabitants may be reluctant to let strangers enter their homes (Campanelli et al. 1997; Goyder et al. 1992; Groves and Couper 1998), so it is likely that this group is easier reached by telephone than by a house visit. In general, non-Western immigrants can be difficult to reach because of periods spent abroad (Blohm and Diehl 2001). In addition, landline telephone coverage among this ethnic-minority group is relatively low (Feskens 2009). Therefore it is likely that this group is easier reached by a house visit than by telephone. Nevertheless, reaching this group might be difficult either way.

### 2.4. Response-Mode Preferences of Hard-to-Survey Populations

Only a few studies have reported information about response-mode choices made by hard-to-survey sample members. As a consequence, not much is known about the mode preferences of most difficult-to-survey groups and how to target them. The one hard-to-survey group that does get attention in literature on response-mode preferences is the young adult population. Schneider et al. (2005) found that when a choice between a mail questionnaire and a web response mode was offered, young individuals in the sample preferred web. Furthermore, many other studies found that young adults prefer the web response mode (Diment and Garrett-Jones 2007; Kaplowitz et al. 2004; Millar and Dillman 2011; Stoop's 2005; Vehovar et al. 2002). An explanation for this preference can

be that young adults use web on a daily basis (De Leeuw and Hox 1998) and therefore it is a very convenient response mode for them. However, this group is also known for not owning a landline number (Blumberg and Luke 2007; Holbrook et al. 2003), so when offered response-mode choices (e.g., CATI and web) their choice can also be based on the response mode that is available to them. Other studies have not found evidence for a web preference of young adults but have found that older people prefer non-web response modes (Millar et al. 2009; Smyth et al. 2009). In a US study, senior citizens of 65 and over are assumed to be less likely to have Internet access in their homes and to use the Internet less frequently because they think it is not relevant to them or their web proficiency level is too low (Zickuhr and Smith 2012). Therefore their preferences for non-web response modes can derive from mode availability as well as mode proficiency. In another mixed mode study by Tancreto et al. (2012) in which a choice between mail and web was offered, choices for web were predominantly made by young adults but also by highly educated sample members and non-English-speaking households. These results can be explained when looking at other studies on Internet use in which age, education and income positively correlated with use of the web (Couper et al. 2007; Loges and Jung 2001). Based on these studies, we thus expect that young adults will choose the web response mode more often than older adults (*Hypothesis 2*), as this mode suits this group in terms of mode proficiency and mode availability.

It is harder to predict the response-mode preference for the other three hard-to-survey groups (full-time workers, big-city inhabitants, and non-Western immigrants) as the literature on them is less extensive. It is assumed that full-time workers are less likely to spend their time on surveys as they are busy working and prefer to spend their time on other activities after work (Groves et al. 2002). However, Stoop (2007) has argued that this group is used to multitasking, and is therefore willing to find time for survey participation when reached. In addition, Vercruyssen et al. (2013) found that survey participation among 'busy people' is mainly affected by 'feeling busy' regardless of the time spent on working. Based on this, we expect that households with more than one full-time worker will choose a self-administered mode more often than households with one full-time worker or less (*Hypothesis 3*). Choosing a self-administered mode, such as web, enables this hard-to-survey group to fill in the questionnaire at their own convenience (i.e., they can decide on their own how to manage their time). However, there is a risk that this choice may not translate to higher response rates due to the procrastination factor (i.e., putting off survey participation until the last possible moment).

With regard to inhabitants of urbanized cities, studies have shown that they are reluctant to let strangers enter their homes (Campanelli et al. 1997; Groves and Couper 1998), therefore it can be easier to reach this group by telephone than through a house visit. Furthermore, their attitude towards survey research can be more negative compared to inhabitants of other areas. For this reason, we also expect that this group will choose the web response mode more often than households from other areas (*Hypothesis 4*), as web is the response mode with the least interviewer interference and therefore suits this group the best. As this article describes an experiment that is conducted in The Netherlands, we did not include hypotheses about Internet penetration, as opposed to studies from the United States (Sylvester and McGlynn 2010). The level of Internet penetration in The Netherlands is not only very high (95% of all households), but also dominated by broadband access

(88% of all households with an Internet connection). There might be slight differences between the penetration in urbanized cities and rural areas, but overall the Internet has become a very common mode of communication in the Netherlands (Deutskens et al. 2004; Statistics Netherlands 2013b). Therefore we do not expect that differences in response-mode preferences between big city households and households from other areas are attributable to broadband access or speed, but are attributable to people's willingness to let strangers enter their homes or their attitudes towards survey research.

According to Feskens et al. (2006), non-Western immigrants are known for their low education level. As lower education levels often correlate negatively with Internet use (Couper et al. 2007; Loges and Jung 2001; Tancreto et al. 2012), it is likely that this hard-to-survey group will not choose the web response mode. Furthermore, language barriers can also constitute a problem for this group. Therefore we expect non-Western immigrants to choose interviewer-administered response modes over self-administered modes more often than natives (*Hypothesis 5*). If language problems arise, the interviewer can clarify the questionnaire if necessary (Blohm and Diehl 2001), and interviewer-administered response modes are more convenient to non-Western immigrants because of a possible low web proficiency.

Another factor that can affect the mode choice of sample members in general is the contact mode. We did not find studies in which this was empirically tested, but there are studies in which the effect of mode switching is analyzed. For example, Lynn (2013) found lower response rates for telephone-contacted sample members who were asked to participate in a computer-assisted personal interview (CAPI) than for face-to-face-contacted sample members who were asked to participate in the same mode (CAPI). So it would seem that sample members prefer to continue survey cooperation in the mode they were contacted in, although it is of course also possible that it was the face-to-face contact mode that increased response rates as compared to the telephone contacts. It is possible that in our study the contact mode will influence the response-mode choice of sample members.

### 3. Method

#### 3.1. Sampling of Households

Based on the literature of Stoop (2005; 2007), we focused on four hard-to-survey groups: young adults (ages 15 to 34), households with more than one full-time worker, inhabitants of big cities, and non-Western immigrants. This experiment was based on a multistage sample. The fieldwork of this study was carried out from March to June 2012 by GfK Panel Services Benelux. In the first sampling stage, all 441 municipalities of the Netherlands were compared on location (12 provinces) and urbanization levels. In the second sampling stage, data from the European Social Survey (ESS) 2010 round was used to study in which municipalities the respondents lived who fulfilled at least one of the criteria of the four selected hard-to-survey groups. Based on this analysis, 283 municipalities were selected. In the third stage, 169 municipalities were selected based on the location of GfK's employed interviewers. Finally, 40 municipalities from these

169 were selected, again taking into account the location (equal selection within 12 provinces) and the urbanization levels of the municipalities.

Enriched address-based sampling was applied using databases with information on population characteristics using ZIP code areas within the 40 municipalities. The addresses with ZIP codes based on the three selection variables were obtained from Cendris (Cendris is a commercial organization that provides addresses for marketing or research purposes with specific information about households). Three selection variables were used to oversample the four hard-to-survey populations. Many full-time working couples and young adults live in newly-built neighborhoods (Raets 2008), and non-Western immigrants predominantly reside in low-income neighborhoods and urbanized areas (Feskens et al. 2007; Nicolaas et al. 2010; Statistics Netherlands 2010). Therefore the selection variables were: newly-built neighborhoods (which is likely to oversample households with more than one full-time worker and people aged 15 to 34), low-income neighborhoods (which is likely to oversample non-Western immigrants), and random selection of remaining ZIP codes to vary in location and urbanization levels so as to reach big city inhabitants and to make an effort to include more members of the hard-to-survey groups. After the data collection was completed, we defined the four hard-to-survey groups using self-reports of participating respondents.

In this way, a total of 3,496 households were randomly selected for this study within an additional round of the ESS. An adapted form of the last-birthday method was used when there was more than one individual living in the household. This standard method of the ESS entails that the interviewer asks which person in the household had his or her birthday closest to a randomly chosen date. The identified individual should then be selected for the survey (no substitute can be taken). This method entailed a difference in sample member selection in single-person households and multi-person households.

### 3.2. *Experimental Design*

To study the effects of the response-mode choices on response rates, a concurrent design was used. First, the sample members received a letter sent to their home, in which the goal of the survey was introduced and their selection for this study was explained. The sampled households were randomly allocated to three experimental groups. One group was contacted face to face and was given the choice between CAPI and a web survey. Another group was contacted by telephone and was given the choice between CATI and a web survey. Sample units in the third group were randomly allocated to CAPI, CATI, or web after being contacted by telephone. Interviewers who paid house visits were compensated per interview and telephone interviewers were compensated per hour. To take into account possible interviewer effects due to this compensation difference, sample members who were contacted face to face and wanted to participate through CAPI could not be interviewed on the spot, but had to make an appointment with the interviewer. The telephone-contacted sample members could participate on the spot or make an appointment.

In several other studies, mail/web choices are offered by mail (see meta-analysis by Medway and Fulton 2012). In our opinion, this does not present a very attractive

combination of contact and response modes to obtain high response rates. First, in a mail contact mode there is no interviewer present who can convince the sample member to participate. In our design we want to compare two contact modes with interviewers, as interviewer presence can have an impact on sample members (Bethlehem et al. 2011; Groves et al. 1992). Second, both response modes (mail/web) are self-administered, while some sample members might prefer interviewer-administered response modes. Therefore we included both interviewer-administered and self-administered response modes in the design. Due to budget constraints, there was no group randomly allocated to a response mode after being contacted face to face. Accordingly, random assignment and contact mode were partially confounded in the design.

### 3.3. Participating Households

Of the 3,496 households selected for this study, 824 participated in the survey. Furthermore, 327 households indicated that they were willing to cooperate. Willing households confirmed they would like to take part in the survey, but were not contacted again for interview or were not sent a link to the web survey because of budget reasons or expiration of the data-collection period. The number of willing respondents was larger for the web conditions because it was harder to find respondents that wanted to participate in CAPI or CATI. Interviewers continued visiting and calling respondents until a sufficient number was reached for all response modes. The number of households that refused to cooperate was 1,579. This group includes hard refusals as well as soft refusal households that indicated they were not interested in participating in the survey after a call back. Additionally, 428 households were known to be ineligible (e.g., due to language barriers, or selected address is business office), and for 208 households their eligibility was unknown (e.g., unable to locate address, or technical problems). Furthermore, 130 households were approached without making contact with the sample member, the so-called noncontacts.

## 4. Results

### 4.1. Outcome Rates

Table 1 shows the response rates and cooperation rates of the experimental groups.

Table 1. Response rates and cooperation rates

Contact modes	1. Face-to-face: Choice between CAPI or web	2. Telephone: Choice between CATI or web	3. Telephone: No choice (random: CAPI, CATI, or web)	All contact modes combined
<b>Outcome Rates</b>				
AAPOR RR1	54.9	34.8	28.8	37.5
AAPOR COOP1	60.6	40.6	32.0	42.1

The following definitions of the American Association for Public Opinion Research (AAPOR 2011) were used to calculate the response rates and cooperation rates:

$$\text{Response Rate 1 (RR1)} = \frac{\text{Complete interviews}}{(\text{Complete interviews} + \text{Partial interviews}) + (\text{Refusal and Break off} + \text{Noncontact} + \text{Other}) + (\text{Unknown if household occupied} + \text{Unknown Other})}$$

$$\text{Cooperation Rate 1 (COOP1)} = \frac{\text{Complete interviews}}{(\text{Complete interviews} + \text{Partial interviews}) + (\text{Refusal and Break off} + \text{Other})}$$

In the calculation, the willing sample members were included in both the numerator and the denominator, as these sample members did agree to cooperate with the interview which is the variable of interest in our analyses. The highest outcome rates were found for the households in Group 1, who were contacted face to face and could choose a response mode. We found significant differences for the response rates obtained in Group 1 compared to the response rates in Group 2, who were contacted by telephone and could choose a response mode ( $\chi^2(1, N = 1,696) = 69.57, p = .00$ ), as well as for the cooperation rates ( $\chi^2(1, N = 1,493) = 59.95, p = .00$ ). However, it is difficult to determine if these differences are the result of the contact mode or the offered response-mode choices. Significant differences were also found comparing the response rates of Groups 2 and 3 ( $\chi^2(1, N = 2,251) = 8.83, p = .00$ ) and the cooperation rates ( $\chi^2(1, N = 1,990) = 15.24, p = .00$ ). Therefore it would seem that offering response-mode choices has a positive effect on the willingness to participate, corroborating Hypothesis 1; however, this will be analyzed further in Subsection 4.2.

Table 2 shows the numbers and proportions of sample members in each experimental group. When comparing Groups 1 and 2, face-to-face-contacted households seem to be more prepared to cooperate than households approached by telephone, as the proportion of refusers is lower in Group 1 (33.1%) than in Group 2 (43.5%). We should take into account that these proportions can be affected both by the contact mode and by the combination of response-mode choices. The same can be found when comparing the refusal proportions of Group 1 (33.1%) and Group 3 (allocated to CAPI 58.8%, allocated to CATI 54.9%, and allocated to web 44.5%), but again this result can also be influenced by the fact that Group 3 was not offered a response-mode choice and by the different contact mode. Looking at Groups 2 and 3, higher refusal proportions were found in the telephone-contact groups in which sample members were allocated to CAPI (58.8%), CATI (54.9%), or web (44.5%) in comparison to Group 2 (43.5%). As households were contacted with telephone in both groups, offering a choice in response mode seems to positively affect the refusal proportions (however, see Subsection 4.2).

Table 3 shows the proportions of the hard-to-survey sample members that were reached in the neighborhoods that were used in the selection process, and it shows the proportions of sample members in the three experimental groups.

Focusing on the neighborhoods, for young adults and households with more than one full-time worker we found the highest proportions in the newly-built neighborhoods (54% and 46% respectively), and for the non-Western immigrants we found the highest

Table 2. Proportion of sample members per group

	Group 1		Group 2		Group 3		
	Letter by mail Face to face CAPI or Web		Letter by mail Telephone CATI or Web		Letter by mail Telephone No: randomly allocated to CAPI	Letter by mail Telephone No: randomly allocated to CATI	Letter by mail Telephone No: randomly allocated to Web
<b>Respondents:</b>							
CAPI	117	13.3	-	-	100	15.8	-
CATI			100	9.7	-	-	106
Web	171	19.4	125	12.2	-	-	105
<b>Other:</b>							
CAPI willing	17	1.9	-	-	15	2.4	-
CATI willing	-	-	10	1.0	-	-	6
Web willing	144	16.3	71	6.9	-	-	64
Refusal	291	33.1	447	43.5	373	58.8	229
No contact	65	7.4	34	3.3	16	2.5	4
Known ineligibility	63	7.2	149	14.5	80	12.6	59
Unknown eligibility	12	1.4	92	8.9	50	7.9	13
Subtotal	N = 880	100	N = 1,028	100	N = 634	100	N = 417
Total	3,496						N = 537

Table 3. *Hard-to-survey sample members in neighborhoods and experimental groups*

	Proportions of sample members				
	Young adults (<35)	Full-time workers	Big-city inhabitants	Non-Western immigrants	Other respondents
<b>Neighborhoods</b>					
Low income	23.0	18.0	37.7	45.8	29.4
Newly built	54.0	46.0	26.2	29.2	31.8
Other	23.0	36.0	36.1	25.0	38.8
	100	100	100	100	100
<b>Experimental groups</b>					
Group 1 (12f + choice)	43.2	40.0	34.4	50.0	41.1
Group 2 (phone + choice)	23.6	36.0	29.5	20.8	25.4
Group 3 (phone + no choice)	33.5	24.0	36.1	29.2	33.5
	100	100	100	100	100



proportions in the low-income neighborhoods (45.8%). This is what we aimed for when using the neighborhood selection variables in our design. The highest proportions of big-city inhabitants were found in low-income neighborhoods (37.7%) and neighborhoods in the 'other' category (36.1%).

When looking at the proportions in the experimental groups, we found the highest proportions for young adults (43.2%) in Group 1. This is in line with the literature, which states that this group is easier to reach face to face than by telephone. The difference between Group 2 (23.6%) and Group 3 (33.5%) is more difficult to interpret. It is possible that telephone calls are harder when an interviewer needs to explain the mode choice to the sample member compared to calls to randomly-allocated sample members where only the request to participate needs to be communicated. Furthermore, young adults might be too impatient to listen carefully to what is being said over the telephone; mode choice may be not attractive enough to them. However, it is also possible that these differences between Groups 2 and 3 are caused by coincidence because of differences in the sample. For households with more than one full-time worker, we did not find major differences between Group 1 (40.0%) and Group 2 (36.0%). According to the literature, this group is hard to contact but the ones reached are willing to cooperate. The proportions for full-time workers were lower in Group 3 (24.0%), so it is possible that mode choice had a positive effect on this group. For big-city inhabitants, we found no major differences comparing the three groups. The highest proportion of non-Western immigrants was found in Group 1 (50.0%); this is in accordance with the literature discussed. However, just as with the young adult group, we found higher proportions in Group 3 (29.2%) than in Group 2 (20.8%). It is possible that non-Western immigrants have no mode-choice preferences when contacted over the telephone because they may have language difficulties.

#### 4.2. *The Effect of Mode Choice and Neighborhood on the Willingness to Participate*

In the present study, we deal with the dichotomous dependent variable 'willingness to participate' (non-normally distributed), and a partially-crossed data structure. Sample members ( $n = 1,502$ , including: actual respondents, willing sample members, and refusers) are nested within interviewers ( $n = 21$ ) and interviewers are crossed with municipalities ( $n = 40$ ). Based on reviews of statistical software for mixed-effect (multilevel) modeling (e.g., Li et al. 2011; Quené and Van den Bergh 2008), generalized linear mixed models (GLMM) were fit using the *lmer* function in package *lme4* (Bates 2005), which is an extension package to R.

Our models in Table 4 only include the sample members that were contacted by telephone. We excluded the group that was asked to participate in the CAPI response mode contacted by telephone (the no-choice group) to make the choice group and no-choice group more comparable (i.e., including only telephone and web as response modes). Our dichotomous variable 'willingness to participate', coded 'one' when willing to participate and 'zero' when not, is included in our model as the dependent variable. Fixed-effect and random-effect factors are distinguished in the models. We have three fixed-effect factors of interest. The first fixed factor is mode choice. By including this fixed factor in our model, we can report some results on the effect of mode choice in our experiment, and

Table 4. Prediction of the likelihood of being willing to participate in the survey

	Model 1	Model 2	Model 3
<b>Fixed-effect factors</b>			
Intercept	- 0.4781*** (0.0994)	- 0.3380* (0.1614)	- 0.3310* (0.1670)
Choice	0.1584 (0.1067)	0.0551 (0.1210)	0.04621 (0.1216)
Low income	- 0.3614** (0.1228)	- 0.3763** (0.1357)	- 0.3933** (0.1364)
Newly built	0.1793 (0.1255)	0.1469 (0.1289)	0.1487 (0.1295)
<b>Random-effect factors</b>			
Intercept (interviewer)		0.4743	0.4821
Intercept (municipality)			0.2237
<b>Model fit</b>			
AIC	2000	1967	1967
<b>Empty models (intercept only)</b>			
Intercept (fixed effect)	- 0.4439*** (0.05288)	- 0.3634** (0.1360)	- 0.3653** (0.1410)
Intercept (interviewer)		0.4855	0.4910
Intercept (municipality)			0.2019

The significance of the fixed-effect factors was evaluated by means of the Wald test for the coefficients in the models, \* $p < .05$ , \*\* $p < .01$ , \*\*\* $p < .001$

contribute some additional insights to the mode choice discussion. The second and third fixed factor describe the neighborhood situation of the sample member, living in a low-income neighborhood, or living in a newly-built neighborhood. The living environment of sample members is associated with willingness to participate (Groves and Couper 1998), which is why we were interested in neighborhood effects in our sample.

All our fixed factors were binary, coded 'one' when having a mode choice, living in a low-income neighborhood, or living in a newly-built neighborhood, and coded 'zero' when not having a mode choice or not living in a low-income or newly-built neighborhood.

The random factors in our model are interviewers and municipalities. By including these two random factors, the structural variability associated with these factors can be taken into account. Sample members are not included as a random factor, since there is no variance linked to sample members as there is only one value for each sample member. Including the interviewer as a random factor in our model was necessary, because an individual interviewer can cause systematic variation on the willingness of a respondent to participate in the survey (Couper and Groves 1996). Although sample members could have been in contact with multiple interviewers, each sample member was only linked to the interviewer who undertook the last telephone contact attempt for this sample member. Furthermore, we also included the municipality as a random factor to take into account possible geographical systematic variance (e.g., sample members from a certain municipality could be more willing than others). Since interviewers conducted interviews in multiple municipalities, interviewers are associated with multiple municipalities, whereas sample members are nested under the interviewer that last reached them.

Furthermore, there can be variability in the effect the three fixed factors (slopes) have. For example, interviewers (random-effect factor) could have been more successful in obtaining cooperation in one neighborhood than in another neighborhood. It is important to take into account these so-called random slopes, as they make the formula of the

mixed-effect model as precise as possible. However, the random slopes in our analyses did not affect our model. Therefore we excluded the random slopes from the models.

In [Table 4](#) three models are presented: a single-level model with fixed factors (Model 1), a two-level model (Model 2) with fixed factors and one random factor, and a three-level model (Model 3) with fixed factors and two random factors. For the fixed-effect factors, each coefficient is shown with its accompanying standard error in parentheses. For the random-effect factors, the standard deviation is presented. We found that offering a choice in response mode did not have a significant effect on the willingness to participate in the survey. This result was compared with our results from Subsection 4.1 where we found a significant difference for cooperation rates ( $\chi^2(1, N = 1,990) = 15.24, p = .00$ ) between experimental Groups 2 and 3, and lower refusal rates in Group 2 than in Group 3. When removing the telephone-contacted sample members from the analysis that were randomly allocated to CAPI, the effect of mode choice on the willingness to participate is no longer significant. A possible explanation for this finding is that the sample becomes too small in the analysis. Nevertheless, based on our mixed-effect models we did not find support for Hypothesis 1. The likelihood of being willing to participate in the survey was significantly lower for households in low-income neighborhoods than for households in other neighborhoods. No significant effects on willingness to participate were found for households in newly-built neighborhoods compared to households in other neighborhoods. Models including interactions did not yield additional significant effects. We saw a decrease in the Akaike Information Criterion (AIC) when the random factors were included in Models 2 and 3, which means the goodness of fit improved. The random factors did not yield additional significant effects. According to the AIC, Models 2 and 3 perform equally well. However, the municipality random effect factor does not cause much systematic variance. Therefore Model 2 is preferred.

### 4.3. Response-Mode Choices

Analyses were performed to assess the response-mode preferences of hard-to-survey respondents in comparison with other respondents. As shown in [Table 5](#), supported by a marginally significant difference ( $\chi^2(1, N = 288) = 3.60, p = .06$ ), the hard-to-survey households (young adults, full-time workers, and big-city inhabitants combined) were more likely to choose web when contacted face to face than the other respondents. Furthermore, young adults were more likely to choose web over CAPI when contacted face to face compared to the older respondents – a significant difference was found ( $\chi^2(1, N = 288) = 5.33, p = .02$ ). This provides evidence for Hypothesis 2 for the face-to-face-contacted young adults. Households with more than one full-time worker were also more likely to choose the web response mode when contacted face to face than households with other work hours, this is also supported by a marginally significant difference ( $\chi^2(1, N = 288) = 3.79, p = .05$ ). This result is in accordance with Hypothesis 3 for the face-to-face-contacted households with more than one full-time worker. No significant difference was found for big-city inhabitants contacted face to face in comparison with non-big-city inhabitants ( $\chi^2(1, N = 287) = 2.51, p = .11$ ). Therefore, Hypothesis 4 does not hold for the face-to-face-contacted big-city inhabitants. When sample members were contacted by telephone, no significant differences were

Table 5. Response-mode choices of hard-to-survey populations

	Response-mode choices					
	Face-to-face contact			Telephone contact		
	CAPI	Web		CATI	Web	
Hard-to-survey (all groups combined)	32.6 %	67.4 %	100 %	36.8 %	63.2 %	100 %
Other respondents	44.4 %	55.6 %	100 %	47.0 %	53.0 %	100 %
Chi square	$\chi^2 (1, N = 288) = 3.60, p = .06$			$\chi^2 (1, N = 225) = 1.79, p = .18$		
Young adults (< 35)	28.1 %	71.9 %	100 %	34.3 %	65.7 %	100 %
35 and older	44.2 %	55.8 %	100 %	46.3 %	53.7 %	100 %
Chi square	$\chi^2 (1, N = 288) = 5.33, p = .02$			$\chi^2 (1, N = 225) = 1.73, p = .18$		
Full-time workers	20.0 %	80.0 %	100 %	33.3 %	66.7 %	100 %
Non-full-time workers	42.2 %	57.8 %	100 %	45.4 %	54.6 %	100 %
Chi square	$\chi^2 (1, N = 288) = 3.79, p = .05$			$\chi^2 (1, N = 225) = 0.97, p = .32$		
Big-city inhabitants	57.1 %	42.9 %	100 %	38.9 %	61.1 %	100 %
Non-big-city inhabitants	39.5 %	60.5 %	100 %	44.9 %	55.1 %	100 %
Chi square	$\chi^2 (1, N = 287) = 2.51, p = .11$			$\chi^2 (1, N = 225) = 0.24, p = .62$		

found (respectively for all four hard-to-survey groups ( $\chi^2(1, N = 225) = 1.79, p = .18$ ), for young adults ( $\chi^2(1, N = 225) = 1.73, p = .18$ ), for full-time workers ( $\chi^2(1, N = 225) = 0.97, p = 0.32$ ), and for big-city inhabitants ( $\chi^2(1, N = 225) = 0.24, p = .62$ )). Thus we only found support for Hypotheses 2-4 for the face-to-face-contacted groups. In comparison with other ESS rounds conducted in the Netherlands ([www.europeansocialsurvey.org](http://www.europeansocialsurvey.org)), only a very low number of non-Western immigrants were reached for this experiment. Consequently, this group was excluded from our analyses. Thus we did not find support for Hypothesis 5 for the non-Western immigrants, neither in the face-to-face-contacted group nor in the telephone-contacted group.

## 5. Discussion

In this article, an experiment was described in which the effects of offering response-mode choices on the willingness to participate were examined and response-mode choices of hard-to-survey households were studied: respectively young adults, households with more than one full-time worker, big-city inhabitants, and non-Western immigrants.

### 5.1. *The Effect of Response-Mode Choice and Neighborhoods on the Willingness to Participate*

Our first research question focused on the effects of offering response-mode choices on the willingness to participate of hard-to-survey populations and sample members in general (Hypothesis 1). We expected that sample members would be more willing to participate in a survey when they could choose a response mode. Although at first response-mode choice seemed to have a positive effect on the willingness to participate in our analysis of Subsection 4.1, this effect disappeared when we excluded the CAPI group that was contacted by telephone from the analysis in Subsection 4.2. It is possible that this result was caused by the sample size. For future studies, besides including more sample members, we would recommend including similar response modes in both experimental groups to further study the effect of response-mode choice on the willingness to participate.

Our mixed-effect models did show that sample members from low-income neighborhoods were less likely to be willing to participate than sample members from other neighborhoods. We did not find such an effect for sample members from newly-built neighborhoods. As we have no more specific data on the nonrespondents we can only speculate about these effects. [Groves and Couper \(1998\)](#) argued that the living environment of sample members has been associated with the willingness to participate. We think it is likely that the personal characteristics of people in the neighborhoods differ, and that these characteristics correlate with the willingness to participate in surveys. For example, it is likely that the number of less-educated persons in low-income neighborhoods is higher than in other neighborhoods, since education is positively associated with income ([De Gregorio and Lee 2002](#)). Moreover, many low-income neighborhoods in the Netherlands are close to big cities ([Ament 2008](#)). As less-educated people and big-city inhabitants are known for low response rates, such personal characteristics could explain why the willingness to participate is lower in low-income neighborhoods than in other neighborhoods.

### 5.2. *Response-mode Choices of Hard-to-Survey Populations*

Regarding our second research question on the extent to which hard-to-survey populations differ in response-mode choices (Hypotheses 2-5), we only found significant results for the face-to-face-contacted groups. For young adults and households with more than one full-time worker, we found response-mode preferences for web. It seems it would be useful to offer these hard-to-survey groups their preferred single response mode in future surveys. However, we think it is likely that our positive outcomes are not only influenced by response-mode preferences but also by the opportunity of getting involved in the survey by making personal choices. Future studies should investigate if it is the response-mode preference or the choice offering that makes it attractive for sample members to cooperate. This could be tested in experiments using public records (such as Municipal Basic Administrations) for stratified sampling, offering subgroups the dominantly-preferred mode within a stratum, compared with subgroups that are offered response-mode choices.

The possible burden of switching from a contact mode to a different response mode (e.g., sample members are approached face to face and are asked to participate in a web survey) might suggest a larger proportion of sample members staying in the same mode when provided a choice. However, we found higher proportions of sample members choosing the web response mode in experimental groups 1 and 2 than choosing the response mode that was similar to the contact mode (face-to-face or telephone). It seems that the preference for the web response mode is stronger than the effect of the possible burden to switch modes. However, we found higher refusal rates for the telephone-contacted sample members that were allocated to CAPI than those that were allocated to the CATI response mode. Still, the lowest refusal rates were found for sample members that were allocated to the web response mode. Thus, it seems that switching to another interviewer-administered mode leads to lower response rates as was found by [Lynn \(2013\)](#), but switching to web (or maybe other self-administered modes) is not a burden. However, to draw more firm conclusions on the effects of mode switching, we recommend that new experiments be conducted.

### 5.3. *Directions for Future Studies*

To reach hard-to-survey populations we used neighborhood selection variables in the sampling design; young adults, households with more than one full-time worker, and non-Western immigrants were reached in the expected neighborhood. We want to propose studying this sampling possibility further. However, it is likely that these neighborhood-selection variables are country specific. Therefore, researchers should obtain information from governmental institutions or statistical agencies on neighborhood characteristics in their country of interest. Furthermore, interviewer observations of household characteristics (e.g., presence of children or non-natives) can be used to adapt strategies for contacting specific groups ([Durrant and Steele 2009](#)). However, interviewer observations in nonresponse adjustments and the targeting of survey features based on such observations should be used with care, since observations may be prone to error ([West 2012](#)). In addition, we suggest studying which contact mode is the most effective in reaching hard-to-survey populations, and whether there are contact-mode preferences for specific difficult-to-survey groups.

Another important arena for additional research would be investigating the costs of concurrent designs versus the obtained response rates and possible errors. Whereas many studies focus on response rates' expenses, the cost of the contact mode is an important outcome-rates indicator that is worth investigating (Porter and Whitcomb 2007; Sinclair et al. 2012; Tse 1998). Furthermore, mixed-mode designs are used because of the lower risk of selection error in comparison to single-mode designs. However, according to Vannieuwenhuyze (forthcoming 2014) the higher fixed costs and risks of higher measurement errors in mixed-mode designs could counteract the advantage of lower selection errors. For future research, it might be interesting to consider a study that concentrates on the trade-off between selection error, measurement error, and costs (Vannieuwenhuyze forthcoming 2014). When focusing on hard-to-survey populations in the sample, such a study can be particularly interesting. Researchers interested in hard-to-survey groups might be willing to accept certain risks in their survey design to reach these populations and to obtain their cooperation. Furthermore, a cost-benefit assessment could be conducted including experimental groups with two or more response-mode choices and no choice in order to assess the outcome rates.

As mode choice can create goodwill (De Leeuw 2005), it could have positive effects on respondents' answering behavior. Conrad et al. (2013) found less satisficing (rounding numerical answers and nondifferentiation) when sample members could choose a response mode than when they were allocated to a mode, so the data quality of the survey improved. Furthermore, the choice group enjoyed participating in the survey more than the no-choice group. In addition, it is also possible that mode choice could decrease social-desirability effects. Sample members who are offered a mode choice might be more willing to give honest answers to sensitive questions, since they were able to choose the mode that feels most comfortable for responding. Therefore we recommend exploring the effects of mode choice on data quality further.

## 6. References

- Ament, P. (2008). Most People on Long-Term Low Incomes Live in Major Cities. Statistics Netherlands. Available at: <http://www.cbs.nl> (accessed August 2013).
- American Association for Public Opinion Research (2011). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. Ann Arbor, MI: AAPOR.
- Bates, D. (2005). Fitting Linear Mixed Models in R. *R News*, 5, 27–30.
- Bethlehem, J., Cobben, F., and Schouten, B. (2011). *Handbook of Nonresponse in Household Surveys*. Hoboken, NJ: Wiley.
- Blohm, M. and Diehl, C. (2001). Wenn Migranten Migranten befragen Zum Teilnahmeverhalten von Einwanderern bei Bevölkerungsbefragungen. *Zeitschrift für Soziologie*, 30, 223–242.
- Blumberg, S.J. and Luke, J.V. (2007). Coverage Bias in Traditional Telephone Surveys of Low-Income and Young Adults. *Public Opinion Quarterly*, 71, 734–749. DOI: <http://www.dx.doi.org/10.1093/poq/nfm047>
- Brady, S.E., Stapleton, C.N., Bouffard, J.A., and Imel, J.D. (2003). Effect of Alternative Data Collection Modes on Cooperation Rates and Data Quality. Proceedings of the American Statistical Association, Joint Statistical Meetings, Section on Survey



- Research Methods, 693-700, San Francisco, August 3-7, 2003, <http://www.amstat.org/sections/srms/Proceedings/> (accessed November 2013).
- Brøgger, J., Nystad, W., Cappelen, I., and Bakke, P. (2007). No Increase in Response Rate by Adding a Web Response Option to a Postal Population Survey: A Randomized Trial. *Journal of Medical Internet Research*, 9(5), e40. DOI: <http://www.dx.doi.org/10.2196/jmir.9.5.e40>
- Campanelli, P., Sturgis, P., and Purdon, S. (1997). *Can you Hear Me Knocking: an Investigation into the Impact of Interviewers on Survey Response Rates*. London: The Survey Methods Centre SCPR.
- Conrad, F.G., Schober, M.F., Zhang, C., Yan, H.G., Vickers, L., Johnston, M., Hupp, A., Hemingway, L., Fail, S., Ehlen, P., and Antoun, C. (2013). Mode Choice on an iPhone Increases Survey Data Quality. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Boston., May 16-19, 2013.
- Couper, M.P. and Groves, R.M. (1996). Social Environmental Impacts on Survey Cooperation. *Quality & Quantity*, 30, 173–188. DOI: <http://www.dx.doi.org/10.1007/BF00153986>
- Couper, M.P., Kapteyn, A., Schonlau, M., and Winter, J. (2007). Noncoverage and Nonresponse in an Internet Survey. *Social Science Research*, 36, 131–148. DOI: <http://www.dx.doi.org/10.1016/j.ssresearch.2005.10.002>
- De Gregorio, J. and Lee, J.W. (2002). Education and Income Inequality: New Evidence from Cross-Country Data. *Review of Income and Wealth*, 48, 395–416. DOI: <http://www.dx.doi.org/10.1111/1475-4991.00060>
- De Leeuw, E.D. (2005). To Mix or Not to Mix Data Collection Modes in Surveys. *Journal of Official Statistics*, 21, 233–255.
- De Leeuw, E.D. and Hox, J.J. (1998). Nonresponses in Surveys: Een Overzicht. *Kwantitatieve Methoden*, 19, 31–53.
- De Leeuw, E.D. and van der Zouwen, J. (1992). Data Quality and Mode of Data Collection: Methodology and Explanatory Model. *La qualité de l'information dans les enquetes*, L. Lebart (ed.). Paris: Dunod, 11–31.
- Deutskens, E., Ruyter, K., Wetzels, M., and Oosterveld, P. (2004). Response Rate and Response Quality of Internet-Based Surveys: An Experimental Study. *Marketing Letters*, 15, 21–36. DOI: <http://www.dx.doi.org/10.1023/B:MARK.0000021968.86465.00>
- Dhar, R. (1997). Consumer Preference for a No-Choice Option. *Journal of Consumer Research*, 24, 215–231.
- Dillman, D.A. (2007). *Mail and Internet Surveys: The Tailored Design Method*. Hoboken, NJ: Wiley.
- Dillman, D.A., Phelps, G., Tortora, R., Swift, K., Kohrell, K., Berck, J., and Messer, B.L. (2009). Response Rate and Measurement Differences in Mixed-Mode Surveys Using Mail, Telephone, Interactive Voice Response (IVR) and the Internet. *Social Science Research*, 38, 1–18.
- Dillman, D.A., West, K.K., and Clark, J.R. (1994). Influence of an Invitation to Answer by Telephone on Response to Census Questionnaires. *Public Opinion Quarterly*, 58, 557–568. DOI: <http://www.dx.doi.org/10.1086/269447>



- Diment, K. and Garrett-Jones, S. (2007). How Demographic Characteristics Affect Mode Preference in a Postal/Web Mixed Mode Survey of Australian Researchers. *Social Science Computer Review*, 25, 410–417. DOI: <http://www.dx.doi.org/10.1177/0894439306295393>
- Durrant, G.B. and Steele, F. (2009). Multilevel Modeling of Refusal and Non-Contact in Household Surveys: Evidence from Six UK Government Surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172, 361–381. DOI: <http://www.dx.doi.org/10.1111/j.1467-985X.2008.00565.x>
- Feskens, R.C.W. (2009). Difficult Groups in Survey Research and the Development of Tailor-Made Approach Strategies. Utrecht: University of Utrecht.
- Feskens, R.C.W., Hox, J.J., Lensvelt-Mulders, G.J.L.M., and Schmeets, J.J.G. (2007). Non-Response Among Ethnic Minorities: a Multivariate Analysis. *Journal of Official Statistics*, 23, 387–408.
- Feskens, R.C.W., Hox, J.J., Lensvelt-Mulders, G.J.L.M., and Schmeets, J.J.G. (2006). Collecting Data Among Ethnic Minorities in an International Perspective. *Field Methods*, 18, 284–304. DOI: <http://www.dx.doi.org/10.1177/1525822X06288756>
- Friese, C.R., Lee, C.S., O'Brien, S., and Crawford, S.D. (2010). Multi-Mode and Method Experiment in a Study of Nurses. *Survey Practice*, 3. Available at: <http://surveypractice.org/index.php/SurveyPractice/issue/view/32> (accessed August 2013).
- Gentry, R. and Good, C. (2008). Offering Respondents a Choice of Survey Mode: Use Patterns of an Internet Response Option in a Mail Survey. Paper presented at the Annual Conference of the American Association for Public Opinion Research, New Orleans., May 15-18, 2008.
- Gillian, E., Loosveldt, G., Lynn, P., Martin, P., Revilla, M., Saris, W., and Vannieuwenhuyze, J. (2010). ESS Prep6 – Mixed-Mode Experiment. Deliverable 21 Final Mode Report. Available at: [www.europensocialsurvey.org](http://www.europensocialsurvey.org)
- Griffin, D.H., Fischer, D.P., and Morgan, M.T. (2001). Testing an Internet Response Option for the American Community Survey. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Montreal., May 17–20, 2001.
- Groves, R.M. (1977). An Experimental Comparison of National Telephone and Personal Interview Surveys. *Proceedings of the Section on Social Statistics: American Statistical Association*, 232–241.
- Groves, R.M., Cialdini, R.B., and Couper, M.P. (1992). Understanding the Decision to Participate in a Survey. *Public Opinion Quarterly*, 56, 475–495. DOI: <http://www.dx.doi.org/10.1086/269338>
- Groves, R.M. and Couper, M.P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.
- Groves, R.M., Dillman, D.A., Eltinge, J.L., and Little, R.J.A. (2002). *Survey Nonresponse*. New York: Wiley.
- Groves, R.M. and Kahn, R.L. (1979). *Surveys by Telephone: A National Comparison with Personal Interviews*. New York: Academic Press.
- Goyder, J. (1987). *The Silent Minority. Nonsample Members on Sample Surveys*. Cambridge: Polity Press.
- Goyder, J., Lock, J., and McNair, T. (1992). Urbanization Effects on Survey Non-Response: A Test Within and Across Cities. *Quality and Quantity*, 26, 39–48.

- Haan, M. and Ongena, Y.P. (2014). Tailored and Targeted Designs for Hard-to-Survey Populations. In *Hard to Survey Populations*, R. Tourangeau et al. (eds). Cambridge: Cambridge University Press. (in press).
- Hardigan, P.C., Succar, C.T., and Fleischer, J.M. (2012). An Analysis of Response Rate and Economic Costs Between Mail and Web-Based Surveys Among Practicing Dentists: A Randomized Trial. *Journal of Community Health*, 37, 383–394. DOI: <http://www.dx.doi.org/10.1007/s10900-011-9455-6>
- Hoffer, T., Grigorian, K., and Fesco, R. (2007). Effectiveness of Using Respondent Mode Preference Data. Paper presented at the Joint Statistical Meetings of the American Statistical Association, Salt Lake City., July 29–August 2, 2007.
- Holbrook, A.L., Green, M.C., and Krosnick, J.A. (2003). Telephone vs. Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias. *Public Opinion Quarterly*, 67, 79–125. DOI: <http://www.dx.doi.org/10.1086/346010>
- Hox, J.J. and de Leeuw, E.D. (1994). A Comparison of Nonresponse in Mail, Telephone, and Face-to-Face Surveys. *Quality and Quantity*, 28, 329–344. DOI: <http://www.dx.doi.org/10.1007/BF01097014>
- Israel, G.D. (2010). Using Web-Hosted Surveys to Obtain Responses from Extension Clients: A Cautionary Tale. *Journal of Extension*, 48, <http://www.joe.org/joe/2010august/a8.php> (accessed November 2013).
- Iyengar, S.S. and Lepper, M.R. (2000). When Choice Is Demotivating: Can One Desire Too Much of a Good Thing? *Journal of Personality and Social Psychology*, 79, 995–1006. DOI: <http://www.dx.doi.org/10.1037/0022-3514.79.6.995>
- Kaplowitz, M.D., Hadlock, T.D., and Levine, R. (2004). A Comparison of Web and Mail Survey Response Rates. *Public Opinion Quarterly*, 68, 94–101. DOI: <http://www.dx.doi.org/10.1093/poq/nfh006>
- Lesser, V.M., Newton, L., and Yang, D. (2010). Does Providing a Choice of Survey Modes Influence Response? Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Chicago., May 13-16, 2010.
- Li, B., Lingsma, H.F., Steverberg, E.W., and Lesaffre, E. (2011). Logistic Random Effects Regression Models: A Comparison of Statistical Packages for Binary and Ordinal Outcomes. *BMC Medical Research*, 11, Article 77. DOI: <http://www.dx.doi.org/10.1186/1471-2288>
- Loges, W.E. and Jung, J. (2001). Exploring the Digital Divide: Internet Connectedness and Age. *Communication Research*, 28, 536–562. DOI: <http://www.dx.doi.org/10.1177/009365001028004007>
- Lynn, P. (2013). Alternative Sequential Mixed-Mode Designs: Effects on Attrition Rates, Attrition Bias, and Costs. *Journal of Survey Statistics and Methodology*, 1, 183–205. DOI: <http://www.dx.doi.org/10.1093/jssam/smt015>
- Martin, P. (2011). What Makes a Good Mix? Chances and Challenges of Mixed Mode Data Collection in the ESS. Working Paper No. 02. Centre for Comparative Social Surveys, City University, London.
- Medway, R.L. and Fulton, J. (2012). When More Gets You Less: a Meta-Analysis of the Effect of Concurrent Web Options on Mail Survey Response Rates. *Public Opinion Quarterly*, 76, 733–746. DOI: <http://www.dx.doi.org/10.1093/poq/nfs047>

- Millar, M.M. and Dillman, D.A. (2011). Improving Response to Web and Mixed-Mode Surveys. *Public Opinion Quarterly*, 75, 249–269. DOI: <http://www.dx.doi.org/10.1093/poq/nfr003>
- Millar, M.M., O'Neill, A.C., and Dillman, D.A. (2009). Are Mode Preferences Real? Technical Report 09-003, Social and Economic Sciences Research Center. Pullman: Washington State University.
- Miller, T.I., Kobayashi, M.M., Caldwell, E., Thurston, S., and Collett, B. (2002). Citizen Surveys on the Web: General Population Surveys of Community Opinion. *Social Science Computer Review*, 20, 124–136. DOI: <http://www.dx.doi.org/10.1177/089443930202000203>
- Nicolaas, H., Wobma, E., and Ooijevaar, J. (2010). Demografie van (Niet-Westerse) Allochtonen in Nederland. Statistics Netherlands. Available at: <http://www.cbs.nl> (accessed August 2013).
- Olson, K., Smyth, J.D., and Wood, H. (2012). Does Providing Sample Members with Their Preferred Survey Mode Really Increase Participation Rates? *Public Opinion Quarterly*, 76, 611–635.
- Porter, S.R. and Whitcomb, M.E. (2007). Mixed-Mode Contacts in Web Surveys: Paper Is Not Necessarily Better. *Public Opinion Quarterly*, 71, 635–648. DOI: <http://www.dx.doi.org/10.1093/poq/nfm038>
- Quené, H. and van den Bergh, H. (2008). Examples of Mixed-Effects Modeling With Crossed Random Effects and With Binomial Data. *Journal of Memory and Language*, 59, 413–425. DOI: <http://www.dx.doi.org/10.1016/j.jml.2008.02.002>
- Radon, K., Goldberg, M., Becklake, M., Pindur, U., Hege, I., and Nowak, D. (2002). Low Acceptance of an Internet-Based Online Questionnaire by Young Adults. *Epidemiology*, 13, 748–749.
- Raets, B. (2008). Vinex-Bewoners zijn Geen Doorsnee Stedelingen. Statistics Netherlands. Available at: <http://www.cbs.nl> (accessed August 2013).
- Ryan, J.M., Corry, J.R., Attewell, R., and Smithson, M.J. (2002). A Comparison of an Electronic Version of the SF-36 General Health Questionnaire to the Standard Paper Version. *Quality of Life Research*, 11, 19–26. DOI: <http://www.dx.doi.org/10.1023/A:1014415709997>
- Scherpenzeel, A. and Toepoel, V. (2012). Recruiting a Probability Sample for an Online Panel. Effects of Contact Mode, Incentives and Information. *Public Opinion Quarterly*, 76, 470–490. DOI: <http://www.dx.doi.org/10.1093/poq/nfs037>
- Schmuhl, P., van Duker, H., Gurley, K.L., Webster, A., and Olson, L. (2010). Reaching Emergency Medical Services Providers: Is One Survey Mode Better Than Another? *Prehospital Emergency Care*, 14, 361–369.
- Schneider, S.J., Cantor, D., Malakhoff, L., Arieira, C., Segel, P., Nguyen, K., and Tancreto, J.G. (2005). Telephone, Internet and Paper Data Collection Modes for the Census 2000 Short Form. *Journal of Official Statistics*, 21, 89–101.
- Schwartz, B. (2004). *The Paradox of Choice: Why More Is Less*. New York: Harper Perennial.
- Shih, T. and Fan, X. (2007). Response Rates and Mode Preferences in Web-Mail Mixed-Mode Surveys: A Meta-Analysis. *International Journal of Internet Science*, 2, 59–82.

- Sinclair, M., O'Toole, J., and Malawaraarachchi, M. (2012). Comparison of Response Rates and Cost-Effectiveness for a Community-Based Survey: Postal, Internet and Telephone Modes with Generic or Personalized Recruitment Approaches. *BMC Medical Research Methodology*, 12, Article 132. DOI: <http://www.dx.doi.org/10.1186/1471-2288-12-132>
- 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. DOI: <http://www.dx.doi.org/10.1177/0002764210361695>
- Smyth, J.D., Olson, K., and Richards, A. (2009). Are Mode Preferences Real? Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Hollywood, Florida., May 14-17, 2009.
- Statistics Netherlands. (2013a). Definitions. Available at: <http://www.cbs.nl> (accessed November 2013).
- Statistics Netherlands. (2013b). ICT Gebruik van Huishoudens naar Huishoudkenmerken. Available at: <http://statline.cbs.nl> (accessed November 2013).
- Statistics Netherlands. (2010). Laag en Langdurig Laag Inkomen; Particuliere Huishoudens naar Kenmerken. Available at: <http://statline.cbs.nl> (accessed August 2013).
- Stoop, I. (2007). No time, Too Busy: Time Strain and Survey Cooperation. In *Measuring Meaningful Data in Social Research*, G. Loosveldt, M. Swyngedouw, and B. Cambré (eds). Leuven: Acco, 301–314.
- Stoop, I. (2005). *The Hunt for the Last Respondent. Non-Response in Sample Surveys*. The Hague: Social and Cultural Planning Agency.
- Sylvester, D.E. and McGlynn, A.J. (2010). The Digital Divide, Political Participation, and Place. *Social Science Computer Review*, 28, 64–74. DOI: <http://www.dx.doi.org/10.1177/0894439309335148>
- Tancreto, J.G., Zelenak, M.F., Davis, M., Ruiters, M., and Matthews, B. (2012). 2011 American Community Survey Internet Tests: Results from First Test in April 2011. Final Report. Washington, DC: US Census Bureau.
- Tarnai, J. and Paxson, M.C. (2004). Survey Mode Preferences of Business Respondents. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, Phoenix., May 13-16, 2004.
- Toffler, A. (1971). *Future Shock*. United States: Bantam Books.
- Tse, A. (1998). Comparing the Response Rate, Response Speed, and Response Quality of Two Methods of Sending Questionnaires: Email vs. Mail. *Journal of the Market Research Society*, 40, 353–361.
- Turner, S., Viera, L., and Marsh, S. (2010). Offering a Web Option in a Mail Survey of Young Adults: Impact on Survey Quality. Poster presented at the Annual Meeting of the American Association for Public Opinion Research, Chicago., May 13-16, 2010.
- Vannieuwenhuyze, J. (forthcoming 2014). On the Relative Advantage of Mixed-Mode Surveys. *Survey Research Methods*.
- Vehovar, V., Batagelj, Z., Lozar Manfreda, K., and Zalatel, M. (2002). Nonresponse in Web Surveys. In *Survey Nonresponse*, R.M. Groves, D.A. Dillman, J.L. Eltinge, and R.J.A. Little (eds). New York: John Wiley and Sons, 229–242.

- Vercruyssen, A., Roose, H., Carton, A., and van de Putte, B. (2013). The Effect of Busyness on Survey Participation: Being Too Busy or Feeling Too Busy to Cooperate? *International Journal of Social Research Methodology*. DOI: <http://www.dx.doi.org/10.1080/13645579.2013.799255>
- Weeks, M.F., Kulka, R.A., Lessler, J.T., and Whitmore, R.W. (1983). Personal versus Telephone Surveys for Collecting Household Health Data at the Local Level. *American Journal of Public Health*, 73, 1389–1394. DOI: <http://www.dx.doi.org/10.2105/AJPH.73.12.1389>
- Werner, P. and Forsman, G. (2005). Mixed Mode Data Collection Using Paper and Web Questionnaires. *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 4015–4017.
- West, B.T. (2012). An Examination of the Quality and Utility of Interviewer Observations in the National Survey of Family Growth. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176, 211–225. DOI: <http://www.dx.doi.org/10.1111/j.1467-985X.2012.01038.x>
- Wilkins, J.R., Hueston, W.D., MacCrawford, J., Steele, L.L., and Gerken, D.F. (1997). Mixed-Mode Survey of Female Veterinarians Yields High Response Rate. *Occupational Medicine*, 47, 458–462. DOI: <http://www.dx.doi.org/10.1093/occmed/47.8.458>
- Ziegenfuss, J.Y., Beebe, T.J., Rey, E., Schleck, C., Locke, III, G.R., and Talley, N.J. (2010). Internet Option in a Mail Survey: More Harm Than Good? *Epidemiology*, 21, 585–586. DOI: <http://www.dx.doi.org/10.1097/EDE.0b013e3181e09657>
- Zickuhr, K. and Smith, A. (2012). *Digital Differences*. Pew Internet Project Report. Washington, DC: Pew Research Center. Available at: <http://www.pewInternet.org/Reports/2012/Digital-differences.aspx> (accessed August 2013).

Received February 2013

Revised November 2013

Accepted January 2014