RESEARCH REPORT SERIES
(Statistics #2020-03)

Experiments on Nonresponse using Sequential Regression Models

Andrew M. Raim[1],
Thomas Mathew[1, 2],
Kimberly F. Sellers[1, 3],
Renee Ellis[4],
Mikelyn Meyers[4]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau;
[2]Department of Mathematics and Statistics, University of Maryland Baltimore County;
[3]Department of Mathematics and Statistics, Georgetown University;
[4]Center for Behavioral Science Methods, U.S. Census Bureau

Center for Statistical Research & Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: August 17, 2020

Experiments on Nonresponse using Sequential Regression Models

Andrew M. Raim[a,*], Thomas Mathew[a,b], Kimberly F. Sellers[a,c], Renee Ellis[d], Mikelyn Meyers[d]

[a]Center for Statistical Research and Methodology, U.S. Census Bureau
[b]Department of Mathematics and Statistics, University of Maryland, Baltimore County
[c]Department of Mathematics and Statistics, Georgetown University
[d]Center for Behavioral Science Methods, U.S. Census Bureau

Abstract

Statistical agencies depend on responses to inquiries made to the public, and occasionally conduct experiments to improve contact procedures. Agencies may explicitly seek improved response rates, or may wish to assess whether or not there is significant change in response rates due to an operational improvement. The present work considers statistical experiments to assess household response rates when up to L attempts are made to contact each household. The process can be viewed as a sequence of L binary trials carried out until either the first success is observed, or failures occur in all L trials. Sequential regression models are used to associate the probabilities in such a sequence to covariates of interest. In particular, the continuation-ratio logit (CRL) model facilitates inference on the probability of success at each step of the sequence, given that failures occurred at previous steps. The CRL model is investigated as a basis for sample size determination—one of the major decisions faced by an experimenter. An adequate sample size is sought to attain a desired power for a Wald test of a general linear hypothesis. A motivating application is provided by an actual experiment being considered for nonresponse followup in the United States 2020 Decennial Census. The experiment involves assessment of a training module which provides guidance to enumerators interviewing Spanish-speaking households. Data analysis and sample size determination based on the CRL model are both addressed in detail. Taking the enumerator training experiment as an illustration, some typical features of an experiment by a statistical agency are also encountered, such as access to a portion of covariate data in advance of the experiment and constraints on the design due to the operation.

Keywords: Design of Experiments; Sample Size Determination; Continuation Ratio Logit; Generalized Linear Models; General Linear Hypothesis; Wald Test

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed are those of the authors and not those of the U.S. Census Bureau.

*For correspondence:
Andrew M. Raim (andrew.raim@census.gov)
Center for Statistical Research and Methodology
U.S. Census Bureau
Washington, DC, 20233, U.S.A.

1 Introduction

Sample surveys and censuses are heavily relied upon to measure characteristics of a population. These methods of data collection involving direct contact with members of the population provide the basis for most official statistics. A major and growing problem is nonresponse, which can occur for a variety of reasons, including inability to contact respondents or refusal to participate (e.g. Singer, 2006). Missing responses can bias inference from the data, especially when the underlying cause of nonresponse is associated with characteristics to be measured. Lohr (2010, Chapter 8) summarizes a variety of techniques developed to reduce and adjust for missing responses; these include followup operations to make further contact attempts ("callbacks"), imputing missing responses, and adjusting estimates by weights based on response probabilities. The present paper focuses on callbacks, which have been an effective strategy for improving response rates; see Hansen and Hurwitz (1946), Politz and Simmons (1949), Deming (1953), Rao (1983), and Särndal et al. (1992, Section 15.4.2). Consideration has been given to the use of administrative records and other available sources of data to augment or replace field work in official statistics (e.g. Scheuren, 1999; Morris et al., 2016; Daas et al., 2015; Brown et al., 2018). However, such use of administrative data presents its own challenges including lack of public availability and data structures that are not intended for this particular application (Davern et al., 2009; Molfino et al., 2017; Groves and Schoeffel, 2018). With field work currently the primary method of data collection, measuring and improving response rates continues to be of major interest to statistical agencies.

One of the major data collection activities of the U.S. Census Bureau is the decennial census, which seeks to contact every household and group quarters in the United States and record basic information, such as the number of residents along with ages and races. Census data are used to produce statistical summaries which are disseminated to the public. Households are initially invited to self-respond via mail or another convenient mode. Households which do not respond within a certain time period become part of the Nonresponse Followup (NRFU) operation. Here, enumerators attempt to personally contact the household and elicit a response. The specific contact strategy designed in the years leading up to the census typically includes in-person visits to the household. NRFU was the most expensive component of the 2010 decennial census, with a cost of about 1.6 billion U.S. dollars (Walker et al., 2012).

A variety of experiments are typically conducted in the years leading up to the decennial census, and also within the census itself, to test whether changes in the operation make significant changes to response rates. The National Research Council (2010) describes experiments carried out by the Census Bureau for decennial censuses between the years 1950 and 2010. For example, the 2010 Census Program of Experiments and Evaluations (CPEX) included one experiment on reducing the number of callbacks in NRFU from the 2000 decennial census. Here, decreased response rates were a concern to be weighed against the savings of decreased field work. Sample size determination is necessary in preparing such experiments, and the sequential nature of repeated callbacks does not appear to be taken into account in the planning.

This article explores use of a sequential regression model in measuring response rates where multiple callback attempts can be made to the same household. The continuation ratio logit (CRL), also referred to as the sequential logit model, is a particular parameterization of the multinomial distribution which can be interpreted as a truncated sequence of dependent Bernoulli trials. This makes it a suitable extension of logistic regression when modeling the number of attempts required for a successful contact, rather than merely the occurrence of successful contact. We consider a procedure for selecting a sample size in a study whose goal is to test a general linear hypothesis;

in particular, to detect whether two or more treatments in an experiment lead to significantly different response rates. When such effects vary over the sequence of attempts, CRL can express the situation while a model capturing only response or nonresponse can not.

An experiment under consideration for the decennial census serves as a motivating application of the CRL methodology. We emphasize that the experiment is presented to demonstrate the application of our methodology and does not reflect any official plans or position of the Census Bureau. Enumerators hired by the agency are given formal training before participating in field operations. For the 2020 Decennial Census, the Census Bureau is testing the inclusion of training for bilingual enumerators on administering the census questionnaire in their non-English ("target") language(s). The agency did not provide such training prior to the 2020 Census. Initially, it will take the form of a brief module to be added to the larger suite of training materials for bilingual, Spanish-speaking enumerators. The objective of additional training is to improve consistency in messaging and in the usage of official translations. Increased consistency may result in improved response rates and improved data quality for affected households (Pan and Lubkemann, 2013). There is thought to be little disadvantage to deploying the new training module; it does not constitute a major cost when implemented as an experimental intervention, and a negative impact to response rates is not expected. However, it is of interest whether the training significantly improves response rates for affected households. Ellis et al. (2018) describe an experiment to be carried out within the 2020 Census NRFU operation to make this assessment.[1] In the present article, we will consider the use of CRL models in two important aspects of experiment planning: to formulate a design which respects the logistics of field operations, and to select a sample size with adequate statistical power to evaluate effectiveness of the training.

Sequential models such as CRL have been widely used in a variety of applications, including survival analysis (Cox, 1972; Albert and Chib, 2001), social science (Fullerton, 2009), economics (Boes and Winkelmann, 2006), and public health (Barboza and Dominguez, 2016). CRL is also closely connected to stick-breaking processes used to fit Dirichlet process models in Bayesian analysis; e.g., see Ghosal and van der Vaart (2017, Chapter 3) and Rigon and Durante (2021). Use for nonresponse in official statistics settings, however, appears to be relatively limited. Alho (1990) formulates a model for nonresponse based on CRL for the purpose of adjusting survey estimates to avoid bias. A similar approach was taken later by Wood et al. (2006). Fienberg (2007, Chapter 6) provides an overview of CRL in the context of contingency tables, while Agresti (2013, Chapter 8) provides an overview in the context of multinomial regression. Tutz (1991) explores connections between models for sequential data (including CRL) and models for ordinal data. Tutz (1991) also establishes sequential models as multivariate generalized linear models (GLMs).

Sample size calculation is the subject of a large literature; the following brief summary features a few examples to help give context for the present work. Chow et al. (2017) provide a general reference for sample size calculation in a number of non-regression settings. Self and Mauritsen (1988) consider power calculations for a score test in the context of a GLM; there are several important features in this work which appear in later references. These authors partition the regression coefficients into a parameter of interest whose value is specified in the null hypothesis, and a nuisance parameter which is estimated. Second, covariates are treated as random variables whose distribution must be considered. In particular, Self and Mauritsen (1988) assume categorical covariates. Self et al. (1992) explore a likelihood ratio test in the setting of GLMs and make use of an asymptotic expansion to compute power. Shieh (2000) extends Self et al. (1992) and

[1] At the time of this writing, plans for NRFU and other 2020 Census operations are subject to change due to the COVID-19 pandemic. See https://2020census.gov/en/news-events/operational-adjustments-covid-19.html.

removes the restriction that covariates must be categorical. Shieh (2005) studies a Wald test in GLMs; here an adjustment is made to the significance level to account for the large sample approximation. Demidenko (2007) and Demidenko (2008) consider a Wald test, but focus on a more specific case/control setting in logistic regression with binary covariates. Lyles et al. (2007) explore Wald and likelihood ratio tests in GLMs, assuming a general linear hypothesis which subsumes the partitioning of test and nuisance parameters. These authors propose a computational approach which allows a specified distribution of the covariates to be studied without requiring derivations for each new setting. Bush (2015) summarizes many of the previously referenced works and investigates them by simulation.

The present work focuses on the CRL model. A general linear hypothesis is assumed to incorporate a range of hypotheses which may be of interest in an experimental setting. Use of the Wald test provides an explicit formula for the asymptotic power. One major departure from the referenced work is that we condition on covariates so that they are fixed throughout sample size determination. Possessing covariate information on the population of interest may be more realistic in an official statistics setting than in the clinical setting that pertains to most of the referenced literature. Another major departure is how we handle the "nuisance" part of the parameter which is not dictated by the test hypothesis; we take this to be fixed based on a priori information rather than estimated. To compute the power for a given departure from the null hypothesis, we utilize an optimization over the parameter space to ensure that the power calculations are conservative.

The article is organized as follows. Section 2 recalls the CRL model and basic inference using maximum likelihood estimation. Section 3 presents a method of sample size determination under the CRL model. Section 4 describes a detailed illustration motivated by the enumerator training experiment; here, a study design is considered and a suitable CRL model is formulated. Section 5 presents simulation results comparing empirical power of the test to the approximation described in Section 3. Section 6 presents a power study under the illustration in which a sample size can be justified. A brief discussion in Section 7 concludes the article.

2 Continuation-Ratio Logit Model

To motivate the continuation-ratio logit (CRL) model, let $\{p_\ell\}$ denote a sequence of probabilities for $\ell \in \{1, 2, \ldots\}$ with $p_\ell \in (0, 1)$. Define a discrete random variable W^* whose support is the set of positive integers $\{1, 2, \ldots\}$ with probabilities $\mathrm{P}(W^* = \ell) = p_\ell \prod_{b=1}^{\ell-1}(1 - p_b)$. The random variable W^* naturally represents a number of Bernoulli trials required to obtain the first success in a sequence of heterogeneous trials. In the special case of a common $p_\ell = p$, W^* follows a geometric distribution. In practice, it may be reasonable to assume an upper bound L for the number of trials. Here, it is natural to consider truncating W^* to $W = W^* \cdot I(W^* \leq L) + (L+1) \cdot I(W^* > L)$. With this construction, W has support $\{1, \ldots, L+1\}$ where the event $[W = L+1]$ indicates that no response was observed in the first L attempts under consideration.

By this construction, W follows a CRL distribution which we will write as $W \sim \mathrm{CRL}_L(\boldsymbol{p})$ with $\boldsymbol{p} = (p_1, \ldots, p_L)$. Define $[n]$ to be the set $\{1, \ldots, n\}$ for a given positive integer n. We may write

$$\pi_\ell \stackrel{\text{def}}{=} \mathrm{P}(W = \ell) = p_\ell \prod_{b=1}^{\ell-1}(1 - p_b), \quad \text{for } \ell \in [L+1], \tag{1}$$

with $p_{L+1} \equiv 1$. It can be shown that $\pi_1 + \cdots + \pi_{L+1} = 1$ when defined in this way. Using (1), we

can obtain a transformation from $(\pi_1, \ldots, \pi_{L+1})$ to $(p_1, \ldots, p_L, p_{L+1})$ using

$$p_\ell = \frac{\pi_\ell}{\pi_\ell + \cdots + \pi_{L+1}}, \quad \text{for } \ell \in [L+1]. \tag{2}$$

From (2), it is clear that each $p_\ell = \mathrm{P}(W = \ell \mid W \geq \ell)$ is the conditional probability of success on the ℓth trial given that trials $1, \ldots, \ell - 1$ were unsuccessful. The quantity (2) is also referred to as a discrete hazard rate in survival analysis (Ghosal and van der Vaart, 2017, Chapter 3).

Now, consider a random sample $W_i \sim \mathrm{CRL}_L(\boldsymbol{p}_i)$ for $i \in [n]$ where W_i represents the outcome for the ith subject, with a common truncation of L trials for all n subjects. We are typically interested in the relationship between response probability and a covariate $\boldsymbol{x}_{i\ell} \in \mathbb{R}^d$ which is provided for each $i \in [n]$ and may vary with trial $\ell \in [L]$. A logistic link can be used to explicitly make the connection

$$\mathrm{logit}(p_{i\ell}) = \boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta} \quad \Longleftrightarrow \quad p_{i\ell} = G(\boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta}),$$

where $G(x) = 1/(1 + e^{-x})$ denotes the inverse logit function, $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of unknown regression coefficients which are the objectives of our inference, and

$$\mathrm{logit}(p_{i\ell}) = \log\left(\frac{p_{i\ell}}{1 - p_{i\ell}}\right) \equiv \log\left(\frac{\pi_{i\ell}}{\pi_{i,\ell+1} + \cdots + \pi_{i,L+1}}\right).$$

The likelihood is

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{L+1} \left[p_{i\ell} \prod_{b=1}^{\ell-1} (1 - p_{ib}) \right]^{I(w_i = \ell)} = \prod_{i=1}^{n} \left[p_{i,w_i} \prod_{\ell=1}^{w_i - 1} (1 - p_{i\ell}) \right]. \tag{3}$$

To facilitate the upcoming discussion, let $\mathcal{I} = ((1,1), (1,2), \ldots, (n, L))$ denote pairs of indices (i, ℓ) ordered first by trial and then by observation. Write $\boldsymbol{X}$ as the $nL \times d$ design matrix with rows $\boldsymbol{x}_{i\ell}^\top$ for $(i, \ell) \in \mathcal{I}$. Denote $g(x) = e^{-x}/(1 + e^{-x})^2$ as the first derivative of $G(x)$. The following result gives the score vector and Fisher information matrix.

Result 2.1. Under likelihood (3),

a. The score vector is

$$S(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} \Big[I(w_i = \ell) \boldsymbol{x}_{i\ell} - I(w_i \geq \ell) G(\eta_{i\ell}) \boldsymbol{x}_{i\ell} \Big].$$

b. The Fisher information matrix is

$$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{X}, \quad \text{with} \quad \boldsymbol{D}_{\boldsymbol{\beta}} = \mathrm{Diag}\left\{ g(\boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta}) \prod_{b=1}^{\ell-1} [1 - G(\boldsymbol{x}_{ib}^\top \boldsymbol{\beta})] : (i, \ell) \in \mathcal{I} \right\}.$$

Using Result 2.1, maximum likelihood estimates (MLEs) can be computed using scoring iterations

$$\boldsymbol{\beta}^{(r+1)} = \boldsymbol{\beta}^{(r)} + \left[\mathcal{I}(\boldsymbol{\beta}^{(r)}) \right]^{-1} S(\boldsymbol{\beta}^{(r)}), \quad r = 1, 2, \ldots$$

until an acceptable convergence criteria has been reached. It is possible, however, to recode CRL data as a logistic regression to facilitate computations. The observed w_i can be recoded as L binary variables $(y_{i1}, \ldots, y_{iL})$, with

$$y_{i\ell} = \begin{cases} 1 & \text{if } \ell = w_i, \\ 0 & \text{if } \ell < w_i, \\ \text{NA} & \text{if } \ell > w_i, \end{cases} \tag{4}$$

so that (3) can be rewritten as

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{\ell=1}^{L} \left[p_{i\ell}^{y_{i\ell}} (1 - p_{i\ell})^{1 - y_{i\ell}} \right]^{I(y_{i\ell} \neq \text{NA})}, \tag{5}$$

where NA values are treated as missing values and excluded from the likelihood. Standard software packages, such as the `glm` function in R (R Core Team, 2020) or PROC GENMOD in SAS (SAS Institute Inc., 2018), can then be used to fit (5) via the logistic regression

$$Y_{i\ell} \sim \text{Ber}(p_{i\ell}), \quad \text{logit}(p_{i\ell}) = \boldsymbol{x}_{i\ell}^{\top} \boldsymbol{\beta}, \quad \ell \in [L] \text{ and } i \in [n],$$

and obtain the MLE $\hat{\boldsymbol{\beta}}$ for the CRL model. Such software packages also produce a Hessian $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$, from which $-\boldsymbol{H}(\hat{\boldsymbol{\beta}})$ and $-\boldsymbol{H}^{-1}(\hat{\boldsymbol{\beta}})$ can serve as an estimate of $\text{Var}(\hat{\boldsymbol{\beta}})$ and $\mathcal{I}(\hat{\boldsymbol{\beta}})$, respectively, evaluated at $\hat{\boldsymbol{\beta}}$. In a basic logistic regression setting, it can be shown that the Hessian is equivalent to the information matrix and does not depend on the sample (e.g. Agresti, 2013, Chapter 5). The logistic regression here, however, is carried out conditionally on $\{y_{i\ell} : y_{i\ell} \neq \text{NA}\}$ so that, in general, $\boldsymbol{H}(\hat{\boldsymbol{\beta}})$ is not equal to $\mathcal{I}(\hat{\boldsymbol{\beta}})$ computed by the CRL information matrix.

Remark 2.2. The CRL regression model assumes that covariates $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iL}$ are fixed during the entire process in which response W_i is generated. Covariates may vary with the attempt, as will be seen in Section 4, but cannot depend on additional data collected during the sequence of trials. This corresponds to studies which are planned in advance and not altered during the course of data collection. In contrast, work on adaptive designs seeks to adjust contact strategies during an operation for purposes such as reducing operational costs or reducing burden to respondents (e.g. Ashmead et al., 2017). This can be aided by paradata collected while attempting to contact respondents, such as the nature of previous failures (e.g., a refusal to participate or a failure to make any contact). Here, binary regression models which evolve over time and allow time-varying covariates, such as in Slud and Kedem (1994), might be considered over the CRL model. The adaptive design setting will not be considered further in this paper, but is a topic of interest for future work.

3 Method of Sample Size Calculation

To handle a variety of testing problems that may arise in experiments, we will assume a general linear hypothesis setting (e.g. Myers, 2000, Chapter 3). Given a known matrix $\boldsymbol{C} \in \mathbb{R}^{q \times d}$ with rank $q \leq d$ and vector $\boldsymbol{c}_0 \in \mathbb{R}^q$, consider the hypotheses

$$H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{c}_0. \tag{6}$$

A Wald test for (6) with significance level α is

$$\text{Reject } H_0 \text{ if } \mathcal{T} > \chi^2_q(1-\alpha), \quad \text{where } \mathcal{T} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}_0)^\top (\boldsymbol{C}\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}_0)$$

and $\chi^2_q(1-\alpha)$ is the $1-\alpha$ quantile of a chi-square distribution with q degrees of freedom. For large samples, we approximately have that $\hat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta}))$, so that $(\boldsymbol{C}\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}^\top)^{-1/2}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c}_0) \sim \mathrm{N}(\boldsymbol{\lambda}(\boldsymbol{\beta}), \boldsymbol{I})$ with $\boldsymbol{\lambda}(\boldsymbol{\beta}) = (\boldsymbol{C}\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}^\top)^{-1/2}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0)$. This implies $\mathcal{T}$ is distributed as a non-central chi-square with q degrees of freedom and non-centrality parameter $\psi(\boldsymbol{\beta}) = \boldsymbol{\lambda}(\boldsymbol{\beta})^\top \boldsymbol{\lambda}(\boldsymbol{\beta}) = (\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0)^\top (\boldsymbol{C}\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0)$. Let $F_{\mathcal{T}}(w; q, \psi)$ denote the cumulative distribution function (cdf) of this distribution. The power of the test, which will be denoted ϖ, is then approximately

$$\varpi = \mathrm{P}(\mathcal{T} > \chi^2_q(1-\alpha)) = 1 - F_{\mathcal{T}}(\chi^2_q(1-\alpha); q, \psi(\boldsymbol{\beta})). \tag{7}$$

Notice that $F_{\mathcal{T}}(\chi^2_q(1-\alpha); q, \psi(\boldsymbol{\beta})) = 1-\alpha$ when $\boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}_0$, which is the condition specified in H_0. The function $F_{\mathcal{T}}$ is readily computed using standard statistical software. By using (7) to express the power of the test, we can avoid more computationally demanding methods such as simulation to compute power empirically. Expression (7) was obtained using informal arguments; Cordeiro et al. (1994) provide a more rigorous justification under the closely-related setting of GLMs with $\boldsymbol{C} = (\boldsymbol{I}_q \ \boldsymbol{0}_{q\times(d-q)})$.

We make several remarks before proceeding. Although the non-centrality parameter $\psi(\boldsymbol{\beta})$ can be directly chosen to satisfy a given power ϖ, our purpose is to study ϖ through $\psi(\boldsymbol{\beta})$, as a function of the sample size. Next, H_1 may be partitioned into spheres $\mathcal{S}(\boldsymbol{c}_0, \Delta) = \{\boldsymbol{\beta} \in \mathbb{R}^d : \|\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0\| = \Delta\}$ characterized by the effect size $\Delta > 0$. Each sphere contains a set of $\boldsymbol{\beta}$ for which the power ϖ may vary. Finally, $\psi(\boldsymbol{\beta})$ is not only a function of $\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0$, but also depends on the entire vector $\boldsymbol{\beta}$ through $\mathcal{I}(\boldsymbol{\beta})$. In view of these remarks, we shall proceed as follows. Given a fixed effect size $\Delta = \|\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0\|$, we find the value $\tilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which solves the optimization problem,

$$\text{minimize} \quad \psi(\boldsymbol{\beta}) = (\boldsymbol{C}\boldsymbol{\beta})^\top (\boldsymbol{C}\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\boldsymbol{\beta}) \quad \text{subject to } \boldsymbol{\beta} \in \mathcal{S}(\boldsymbol{c}_0, \Delta), \tag{8}$$

and evaluate the power at $\psi(\tilde{\boldsymbol{\beta}})$ via (7). Other options are possible, such as drawing $\boldsymbol{\beta}$ randomly from the sphere $\mathcal{S}(\boldsymbol{c}_0, \Delta)$ and evaluating an average or quantile of attained power values, but we will make use of the optimization (8) for the remainder of the paper to ensure that the power calculation is conservative.

The constrained minimization problem (8) can be transformed to an unconstrained problem and solved using standard optimization software such as `optim` in R; to do this, we proceed as follows. Because $\boldsymbol{C}\boldsymbol{\beta} \in \mathbb{R}^q$, the number of parameters not involved in the hypothesis is $d_0 = d - q$. Let $\boldsymbol{B}$ be a $d_0 \times d$ matrix so that $\boldsymbol{A} = (\boldsymbol{B}^\top, \ \boldsymbol{C}^\top)^\top$ is a $d \times d$ nonsingular matrix. Thus $\boldsymbol{c} = \boldsymbol{C}\boldsymbol{\beta}$ is the parameter of interest, which is constrained to lie on the sphere $\mathcal{S}(\boldsymbol{c}_0, \Delta)$. Furthermore, $\boldsymbol{B}\boldsymbol{\beta}$ is the nuisance parameter whose value, say $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}_0$, is assumed to be known a priori. For example, $\boldsymbol{b}_0$ may be available from a pilot study. We can express $\boldsymbol{c}$ using spherical coordinates (e.g. Blumenson, 1960) as

$$c_1 = \Delta\cos\phi_1, \ c_2 = \Delta\cos\phi_2\sin\phi_1, \ \ldots, \quad c_{q-1} = \Delta\cos\phi_{q-1}\prod_{j=1}^{q-2}\sin\phi_j, \ c_q = \Delta\sin\phi_{q-1}\prod_{j=1}^{q-2}\sin\phi_j$$

based on $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{q-1})$, where $\phi_j \in [0, \pi]$ for $j = 1, \ldots, q-2$ and $\phi_{q-1} \in [0, 2\pi)$. Here, $\pi = 3.14159\ldots$ refers to the mathematical constant, not to be confused with (1). A second transformation $\phi_j = \pi G(\vartheta_j)$ for $j = 1, \ldots, q-2$ and $\phi_{q-1} = 2\pi G(\vartheta_{q-1})$ yields $\boldsymbol{\phi}$ from an unconstrained

$\boldsymbol{\vartheta} \in \mathbb{R}^{q-1}$, where $G(x)$ again denotes the inverse logit function. Therefore, a candidate point $\boldsymbol{\vartheta} \in \mathbb{R}^{q-1}$ from the optimizer is transformed to $\boldsymbol{\beta}$ via

$$(\boldsymbol{b}_0, \boldsymbol{\vartheta}) \quad \longrightarrow \quad (\boldsymbol{b}_0, \boldsymbol{\phi}) \quad \longrightarrow \quad \boldsymbol{\alpha} = (\boldsymbol{b}_0, \boldsymbol{c}) \quad \longrightarrow \quad \boldsymbol{\beta} = \boldsymbol{A}^{-1}\boldsymbol{\alpha}. \tag{9}$$

Such a $\boldsymbol{\beta}$ may be evaluated by the objective function in (8) with the constraint omitted.

An investigation to determine sample size can therefore be carried out as follows. Determine samples $\mathcal{J}_1, \ldots, \mathcal{J}_m \subseteq \{1, \ldots, n\}$ of increasing size which are viable for the experiment. Also determine a grid $\{\Delta_1, \ldots, \Delta_r\}$ of effect sizes to consider. For each combination of $\Delta \in \{\Delta_1, \ldots, \Delta_r\}$ and $\mathcal{J} \in \{\mathcal{J}_1, \ldots, \mathcal{J}_m\}$, solve optimization problem (8) using transformation (9). This yields $\tilde{\boldsymbol{\beta}}$, the corresponding non-centrality parameter $\psi(\tilde{\boldsymbol{\beta}})$, and the associated power via (7) for each combination. This process allows the test's power to be studied as a function of the underlying sample size. A sample may then be selected to meet testing objectives, or it can be determined that no sample under consideration meets the objectives.

4 An Illustration

We now consider an illustration based on the enumerator training experiment described in Section 1. To provide a compelling demonstration, some details anticipated for the actual experiment have been included. A number of complexities have been omitted, however: some dilute the methodology discussion and may be considered out of scope, while others present relevant complications. Section 7 discusses several of the latter.

Because the experiment is envisioned to be carried out within the decennial census, its design must be compatible with census operations. It is worthwhile to review the major components of the experiment, such as the experimental subjects, treatments, and the meaning of "sample size". A general reference for experimental design is Oehlert (2000). Experimental subjects here are Spanish-speaking households in the NRFU operation; these are not known with certainty until the actual NRFU operation is carried out, so we make use of estimates from previous operations in the planning phase. The number of households included in the study is therefore associated with the sample size, but is not something which we can directly manipulate in the design. Parameters of interest are probabilities of Spanish-speaking households to respond to the NRFU operation.

As experimenters, we can assign control ("no training") or experimental ("training") treatments to enumerators. It is impractical to assign treatments to enumerators individually, thus we instead assign treatments at the level of Area Census Office (ACO). For this discussion, an ACO is considered to be a geographic delineation used in data collection for the census. Tracts from the standard ("tabulation") geography can generally overlap with multiple ACOs; however, tracts intersecting the ACOs used in this study are contained strictly in one ACO. Enumerators associated with an experimental ACO will receive the new training, while those in a control ACO will not receive the new training. We cannot directly assign individual households to enumerators; instead, case assignments will be made dynamically based on enumerator availability and workloads (U.S. Census Bureau, 2019). Under this system, each enumerator will visit multiple households, and a household may be visited by multiple enumerators. We wish to avoid situations of "contamination" where households in the study are visited by both trained and untrained enumerators. To minimize the risk of such occurrences, we have ensured that control and experimental ACOs are geographically separated. After the data collection, any cases in which a household is visited by both trained and untrained enumerators will be discarded from the analysis.

The number of households in the sample is controlled via the ACOs we select for the experiment. This selection must be decided sufficiently in advance of field operations. To minimize impact to operations, we would prefer a small number of ACOs which will provide adequate power. We have pre-selected ACOs from several metropolitan statistical areas (MSAs) in Dallas, Houston, and Los Angeles as a starting point. Historically, these areas have had large numbers of residents who primarily speak Spanish and also a large expected workload for NRFU. Table 2 displays the fourteen pre-selected ACOs: six in the Dallas area, six in Houston, and two in Los Angeles. All ACOs in Dallas have been assigned to the control group, while Houston has been assigned to the experimental group. Of the two ACOs in Los Angeles, one has been assigned to the experimental group and the other to the control group. We have gathered some additional data from the Census Bureau Planning Database[2] for the selected ACOs, including the total number of households (HH_Total), percent of Spanish speakers (Pct_Spanish), and percent of self-responders (Pct_Selfresp). We obtain a rough estimate of the count of relevant households in each ACO using the formula

$$\text{HH\_Target} = \text{HH\_Total} \times \text{Pct\_Spanish}/100 \times \big(1 - \text{Pct\_Selfresp}/100\big), \tag{10}$$

and rounding down to the next integer. Calculation (10) is carried out at the tract level, then aggregated to the ACO level. This provides a total sample size of up to 380,018 households; although this represents a small proportion of households in the United States, it seems to be quite a large number of households to use in an experiment. A formal power analysis will reveal whether or not it is sufficient.

The fourteen ACOs have been matched into $I = 7$ pairs where each pair contains one ACO for each of the $J = 2$ possible treatments. The Los Angeles ACOs form one pair, while the remaining pairs were constructed by matching an ACO from Houston with an ACO from Dallas where Pct_Spanish and Pct_Selfresp were similar. After matching, pairs were randomly assigned indices $i = 1, \ldots, I$. This defines samples using an increasing number of pairs, $\mathcal{J}_i = [i]$ for $i = 1, 2, \ldots, I$, as discussed in Section 3. Within the ith pair, the control ACO receiving no training is indexed $j = 1$, while the experimental ACO receiving training is indexed $j = 2$. Within the jth ACO of the ith pair, K_{ij} denotes the household count HH_Target from (10). Of primary concern is whether the seven available pairs will be adequate or if more are needed. A secondary interest is in plotting power curves when using one pair, two pairs, etc, up to all seven available pairs.

Let $W_{ijk} \sim \text{CRL}_L(\boldsymbol{p}_{ijk})$ indicate the number of contact attempts needed for a response for the ith pair, jth treatment, and kth household for $i \in [I]$, $j \in [J]$, $k \in [K_{ij}]$, where $\boldsymbol{p}_{ijk} = (p_{ijk1}, \ldots, p_{ijkL})$ are the associated probabilities of a response at each attempt. Recall that an observation of $w_{ijk} = L + 1$ indicates that no response was obtained in the first L attempts. We consider a basic model for response rate as

$$\begin{aligned} \text{logit}(p_{ijk\ell}) &= \zeta_{j\ell} && (11)\\ &= \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell} && (12)\\ &= \boldsymbol{s}_{j\ell}^\top \boldsymbol{\beta}. && (13) \end{aligned}$$

Model formulation (11) uses unconstrained effects $\zeta_{11}, \ldots, \zeta_{JL}$ to facilitate computations. Formulation (12) provides a more clear interpretation, with an intercept term μ, treatment effects τ_j which are of primary interest, contact attempt effects δ_ℓ, and effects $(\tau\delta)_{j\ell}$ for treatment-attempt interaction. Formulation (13) is a regression form of (11). To reparameterize from (11) to (12), we

[2] https://www.census.gov/topics/research/guidance/planning-databases.html

assume constraints

$$\sum_{j=1}^{J}\tau_j = 0, \quad \sum_{\ell=1}^{L}\delta_\ell = 0, \quad \sum_{j=1}^{J}(\tau\delta)_{j\ell} = 0, \quad \sum_{\ell=1}^{L}(\tau\delta)_{j\ell} = 0, \tag{14}$$

and let $\zeta_{j\ell} = \mu + \tau_j + \delta_\ell + (\tau\delta)_{j\ell}$ so that

$$\frac{1}{JL}\sum_{j=1}^{J}\sum_{\ell=1}^{L}\zeta_{j\ell} = \mu, \quad \frac{1}{L}\sum_{\ell=1}^{L}\zeta_{j\ell} - \mu = \tau_j, \quad \frac{1}{J}\sum_{j=1}^{J}\zeta_{j\ell} - \mu = \delta_\ell.$$

Care should be taken when interpreting μ, τ_j, and δ_ℓ, as they are averages of the raw $\zeta_{j\ell}$ parameters. There are $J-1$ distinct parameters among the τ_j's, $L-1$ among the δ_ℓ's, $(J-1)(L-1)$ among the $(\tau\delta)_{j\ell}$'s; with the addition of μ, there are a total of $(J-1)+(L-1)+(J-1)(L-1)+1 = JL$ parameters. In particular, JL is equivalent to $2L$ with $J=2$ treatments. To rewrite (12) in the form of (13), let

$$\boldsymbol{\beta} = \left(\mu, \tau_1, \delta_1, \ldots, \delta_{L-1}, (\tau\delta)_{11}, \ldots, (\tau\delta)_{1,L-1}\right)$$

with $\boldsymbol{s}_{j\ell}$ coded in the manner shown in Table 1. To emphasize the grouping of trials implied by the model, let $\mathcal{H}(j,\ell)$ represent the list of (i,j,k,ℓ) indices corresponding to the jth treatment and ℓth attempt, so that $\mathcal{H}(j,\ell)$ contains $N_{j\ell} = L\sum_{i=1}^{I}K_{ij}$ elements, and write $\boldsymbol{p}_{\mathcal{H}(j,\ell)} = (p_{ijk\ell} : (i,j,k,\ell) \in \mathcal{H}(j,\ell))$. We can then rewrite (13) as

$$\text{logit}(\boldsymbol{p}_{\mathcal{H}(j,\ell)}) = \boldsymbol{X}_{j\ell}\boldsymbol{\beta}, \quad j = 1,\ldots,J \text{ and } \ell = 1,\ldots,L,$$

where $\boldsymbol{X}_{j\ell} = \mathbf{1}_{N_{j\ell}} \otimes s_{j\ell}^{\top}$ and $\mathbf{1}_{N_{j\ell}}$ is a vector of $N_{j\ell}$ ones. Sample size determination will be based on a test of the general linear hypothesis (6) with $\boldsymbol{C} = (\mathbf{0}_{JL-1}\ \boldsymbol{I}_{JL-1})$ and $\boldsymbol{c}_0 = \mathbf{0}_{JL-1}$; i.e., a test for the presence of any treatment effects, attempt effects, or their interactions. We will assume significance level $\alpha = 0.10$ for the test, which is a standard used by the Census Bureau (U.S. Census Bureau, 2013). Section 6 will investigate the relationship between the sample size, the effect size $\Delta = \|\boldsymbol{C}\boldsymbol{\beta} - \boldsymbol{c}_0\|$, and power ϖ of the test. Some discussion will be provided to interpret the achieved Δ.

It is important to consider the number of contact attempts L to be used in the model. Too few contact attempts can fail to capture the response behavior of interest, while too many will lead to an issue of sparse observations which we will now discuss. Although a high probability of response during each contact attempt is desirable from the perspective of data collection, enumerations during later attempts will be a more rare occurrence. In turn, corresponding counts will be close to zero, large sample properties used in Section 3 will not take effect, and consequently the power expression (7) will be inaccurate unless sample sizes are taken to be very large. To make this issue concrete, suppose H_0 is true so that the probability of a successful enumeration $p_{ijk\ell} \equiv p$, given that any attempts $1,\ldots,\ell-1$ failed, depends only on μ. We may then write the overall (unconditional) probabilities of enumeration as $\pi_{ijk\ell} = p\prod_{b=1}^{\ell-1}(1-p) = p(1-p)^{\ell-1}$. Values for $\boldsymbol{\pi}_{ijk} = (\pi_{ijk1},\ldots,\pi_{ijkL})$ are shown in Table 3 for $L=5$ for several values of p under H_0. It is clear that responses occurring after two attempts are quite common under small p but become increasingly rare events when p approaches 1. In practice, many factors can influence response probability across attempts, but consideration of the model under H_0 helps to serve as a guideline.

Table 1: Coding for design matrix rows $\boldsymbol{s}_{j\ell}$ used in (13).

j	ℓ	Intercept	Treatment	Attempt				Treatment × Attempt			
1	1	1	1	1	0	$\cdots$	0	1	0	$\cdots$	0
	2	1	1	0	1	$\cdots$	0	0	1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$L-1$	1	1	0	0	$\cdots$	1	0	0	$\cdots$	1
1	L	1	1	-1	-1	$\cdots$	-1	-1	-1	$\cdots$	-1
2	1	1	-1	1	0	$\cdots$	0	-1	0	$\cdots$	0
	2	1	-1	0	1	$\cdots$	0	0	-1	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	$L-1$	1	-1	0	0	$\cdots$	1	0	0	$\cdots$	-1
2	L	1	-1	-1	-1	$\cdots$	-1	1	1	$\cdots$	1

5 Simulation

Table 3 emphasized that successful enumerations in later attempts can be quite rare in some circumstances: in particular, under H_0 with $\text{logit}^{-1}(\mu)$ approaching 1. It is anticipated that large sample approximations used in Section 3 will fail when data in later categories become too uncommon. In this section, we will compare the empirical power of the Wald test to the approximate power computed via (7). A simulation will be carried out in R (R Core Team, 2020) under the experimental design introduced in Section 4.

Suppose there is $I = 1$ pair with K households in the experimental ACO and K households in the control ACO; therefore, $J = 2$ treatments are assumed. We take $K \in \{10, 50, 200\}$. We consider CRL models of the form (12) which include $L \in \{1, 2, 3, 4\}$ attempts. For the baseline effect, we take $\text{logit}^{-1}(\mu) \in \{0.60, 0.75, 0.90\}$. For the departure from H_0, we consider $\Delta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1\}$. Here, we explicitly choose the parameters to be

$$\boldsymbol{\beta} = \Big(\mu, \tau_1 = \Delta, \delta_1 = 0, \ldots, \delta_{L-1} = 0, (\tau\delta)_{11} = 0, \ldots, (\tau\delta)_{1,L-1} = 0\Big).$$

so that Δ is entirely allocated to τ_1. The simulation proceeds by drawing a sample $W_{ijk} \sim \text{CRL}_L(\boldsymbol{p}_{ijk})$ for $i \in [1]$, $j \in [2]$, and $k \in [K]$, recoding W_{ijk}'s to $Y_{ijk\ell}$'s via (4), then fitting the (correctly specified) data-generating model (12) by a logistic regression with the `glm` function. This is repeated $R = 1{,}000$ times for each simulation setting, yielding coefficient estimates $\hat{\boldsymbol{\beta}}^{(r)}$ and corresponding covariance estimates $\hat{\boldsymbol{V}}^{(r)} = \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}^{(r)})$ for $r = 1, \ldots, R$. We then compute Wald statistics

$$W^{(r)} = (\boldsymbol{C}\hat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{c}_0)^\top (\boldsymbol{C}\hat{\boldsymbol{V}}^{(r)}\boldsymbol{C}^\top)^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}}^{(r)} - \boldsymbol{c}_0),$$

to obtain an empirical probability of rejection $\frac{1}{R}\sum_{r=1}^{R} I(W^{(r)} \geq \chi^2_q(1-\alpha))$. Here, $\chi^2_q(1-\alpha)$ denotes the $1 - \alpha = 0.90$ quantile of the χ^2 distribution with $q = JL - 1$ degrees of freedom which is the critical value of the test. For some repetitions, the coefficients or the associated covariance estimates could not be fully computed. For example, this occurred when no outcomes were observed for an attempt ℓ in one or both of the treatments. These were recorded as $W^{(r)} = \text{NA}$ and excluded from

the empirical power calculation. The approximate rejection probability (7) is also computed for each simulation setting; note that this does not make use of the simulation draws.

Tables 4 and 5 display the empirical power and approximated power, respectively, after carrying out the simulation. Respective entries across the two tables can be compared to check their agreement. Table 6 displays frequencies of $W^{(r)} = \text{NA}$ from the empirical power calculation; e.g., a count of zero indicates that all samples in the given setting could be estimated.

When $L = 1$, the empirical and approximate power closely agree when $\mu = \text{logit}(0.6)$, for all sample sizes K and all Δ. When μ is increased to $\text{logit}(0.75)$, $K = 10$ becomes too small, and the empirical power is systematically smaller than the approximation. For this value of μ, $K = 50$ appears to be a sufficient number of households. When we further increase μ to $\text{logit}(0.9)$, $K = 50$ is no longer sufficient, but increasing to $K = 200$ is enough for the two power calculations to agree.

If we increase L to 2, $K = 10$ is no longer a sufficient number of households for any displayed setting of μ. $K = 50$ gives a sufficient power approximation when $\mu = \text{logit}(0.6)$, but not the two larger values of μ. $K = 200$ is enough when $\mu = \text{logit}(0.6)$ or $\mu = \text{logit}(0.75)$. When $\mu = \text{logit}(0.90)$, however, we need a larger sample to use the approximation reliably.

The pattern becomes more severe as L increases, with larger K needed for a reasonably good approximation of the power for larger μ. Referring to Table 6, we notice that NA counts increase accordingly when L and μ are both larger. For example, in the case of $L = 3$ and $\mu = \text{logit}(0.90)$, it is rare to obtain valid estimates under $K = 10$, but slowly becomes more frequent as the number of households increases to $K = 50$ and to $K = 200$. Referring back to Table 3, we see that Attempt 3 for $\mu = \text{logit}(0.90)$ has probability of about 0.009 under H_0. Therefore, we expect that a sample size of approximately 100 will be needed to observe third attempts in both treatments, which is a minimum requirement to be able to use a model with $L = 3$.

6 Sample Size for Illustration

With some insight into the quality of the approximation (7), we now present a power study using the fourteen ACOs from Table 2. For each $\mathcal{J} \in \{[1], \ldots, [7]\}$ and each $\Delta \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1.0\}$, the optimization problem (8) is solved to yield the minimizer $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}(\Delta, \mathcal{J})$ and associated power $\varpi(\Delta, \mathcal{J})$. We repeat this using $L \in \{2, \ldots, 5\}$ contact attempts and baseline response effect $\text{logit}^{-1}(\mu) \in \{0.75, 0.90\}$. Figure 1 displays the results as a grid of power curves. For this discussion, we will consider $\varpi = 0.80$ as a rough target for the power.

First, we give an upper bound on μ to decide on the largest L that can be supported by the model. Internal discussions with Census Bureau personnel have suggested that the baseline response probability μ might be larger than $\text{logit}(0.75)$ but should be no greater than $\text{logit}(0.90)$; therefore, Table 3 suggests modeling at most $L = 3$ attempts. With $L = 3$, using all seven pairs, we achieve nearly $\varpi = 1$ when $\mu = \text{logit}(0.75)$. Under $\mu = \text{logit}(0.90)$, we also achieve $\varpi \approx 1$ except under the smallest effect size in the study, $\Delta = 0.1$, where $\varpi \approx 0.77$ is achieved.

Therefore, $\Delta = 0.1$ represents the smallest effect size we can detect using all seven pairs, modeling $L = 3$ contact attempts, achieving power $\varpi \approx 0.77$, and assuming $\mu = \text{logit}(0.90)$. Stakeholders of the experiment will likely need an intuitive interpretation of $\Delta = 0.1$ to decide if this provides a level of detection precise enough to be practically useful. To assist with interpretation, we can consider the extreme cases of the alternative hypothesis with effect size Δ, namely

$$\boldsymbol{\beta} \in \Big\{(\mu, \Delta, 0, \ldots, 0),\ \ldots,\ (\mu, 0, \ldots, 0, \Delta)\Big\}, \tag{15}$$

so that Δ is completely allocated to one of the coordinates of $\boldsymbol{\beta}$ aside from the intercept. Table 7 shows the $p_{ijk\ell}$ and $\pi_{ijk\ell}$ corresponding to each of the values in (15), along with the value $\boldsymbol{\beta} = (\mu, 0, \ldots, 0)$ under H_0. A comparison of each case (b)–(f) in Table 7 to case (a) suggests that $\Delta = 0.1$ corresponds to rather small changes in probabilities. Presented with this information, stakeholders may determine whether this level of detection is sufficiently precise for the experiment.

7 Discussion and Conclusions

Experiments assessing changes to response rates may involve multiple attempts to establish contact with households, persons, businesses, or other entities. Sequential models such as the continuation-ratio logit (CRL) provide a statistical framework for such experiments. Through an illustration based on an actual experiment for a new enumerator training module, we have explored use of the CRL model in an experimental design to measure changes in response rates. The presented methodology was used to justify a sample size and provide intuition on effect sizes which could be detected in the experiment with a desired level of power.

A number of extensions can be considered in future work, which may be relevant to practical applications. A likelihood ratio test can be considered in place of the Wald test using an approximate power expression (e.g. Self et al., 1992). Test procedures relying less on asymptotic approximation could also be considered, but may be onerous to scale to larger datasets if they rely heavily on computation. In the illustration, all covariates have been treated as known ahead of the experiment, but it would be desirable to account for uncertainty in the counts of housing units. This work has focused solely on unit-level nonresponse; item-level nonresponse may also be of interest in sample size calculation.

The illustration featured several notable simplifications which may need to be addressed in a real-life experiment. The illustration assumed a common maximum number of attempts L across all households. Section 4 mentioned plans to dynamically assign enumerators to households during the 2020 Census NRFU operation until attempts are exhausted; however, L itself is also subject to dynamic adjustment (U.S. Census Bureau, 2019). To account for uncertainty during planning, it may be conceivable to formulate a model for L and extend the sample size methodology accordingly. Experimenters may also wish to define "success" more broadly than in-person contact by an enumerator, and may include contact by another mode such as phone call, contact with a proxy, or an implicit response via administrative records in lieu of contact. For example, Ashmead et al. (2017) consider a more holistic contact process in the context of the American Community Survey. Therefore, it may be necessary to generalize the outcome model beyond simple sequences of trials to provide a more comprehensive notion of response.

The ability to support mixed effects would be a desirable extension to this work. For example, our illustration grouped the ACOs into pairs, with one element in the pair receiving the experimental treatment and the other receiving the control treatment. Such a design would be especially desirable if ACOs within a pair exhibit more similar response behavior than ACOs across pairs. Here, a random intercept for each pair may be appropriate to reduce overall uncertainty in the fixed effects of interest. Other random effects such as enumerator and enumerator-attempt interaction could be considered as well; however, their use in sample size determination would be complicated in a setting with dynamic workload allocation.

Acknowledgements

The authors thank Luke Larson, Kathleen Kephart, and Marcus Berger (Center for Behavioral Science Methods, U.S. Census Bureau) for useful discussions on the enumerator training experiment. We are also grateful to Jennifer Hutnick (Decennial Statistical Studies Division, U.S. Census Bureau) and Eric Slud (Center for Statistical Research and Methodology, U.S. Census Bureau) for insightful feedback regarding the manuscript.

References

Alan Agresti. *Categorical Data Analysis*. Wiley, 3rd edition, 2013.

James H. Albert and Siddhartha Chib. Sequential ordinal modeling with applications to survival data. *Biometrics*, 57(3):829–836, 2001.

Juha M. Alho. Adjusting for nonresponse bias using logistic regression. *Biometrika*, 77(3):617–624, 1990.

Robert Ashmead, Eric Slud, and Todd Hughes. Adaptive intervention methodology for reduction of respondent contact burden in the American Community Survey. *Journal of Official Statistics*, 33(4):901–919, 2017.

Gia Elise Barboza and Silvia Dominguez. A sequential logit model of caretakers' decision to vaccinate children for the human papillomavirus virus in the general population. *Preventive Medicine*, 85:84–89, 2016.

L. E. Blumenson. A derivation of n-dimensional spherical coordinates. *The American Mathematical Monthly*, 67(1):63–66, 1960.

Stefan Boes and Rainer Winkelmann. Ordered response models. *Allgemeines Statistisches Archiv*, 90: 167–181, 2006.

J. David Brown, Misty L. Heggeness, Suzanne M. Dorinski, Lawrence Warren, and Moises Yi. Understanding the quality of alternative citizenship data sources for the 2020 Census. CES Working Paper Series: CES 18–38, Center for Economic Studie, U.S. Census Bureau, 2018. URL https://www2.census.gov/ces/wp/2018/CES-WP-18-38.pdf.

Stephen Bush. Sample size determination for logistic regression: A simulation study. *Communications in Statistics - Simulation and Computation*, 44(2):360–373, 2015.

Shein-Chung Chow, Jun Shao, Hansheng Wang, and Yuliya Lokhnygina. *Sample Size Calculations in Clinical Research*. Chapman and Hall/CRC, 3rd edition, 2017.

Gauss M. Cordeiro, Denise A. Botter, and Silvia L. De Paula Ferrari. Nonnull asymptotic distributions of three classic criteria in generalised linear models. *Biometrika*, 81(4):709–720, 1994.

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

Piet J. H. Daas, Marco J. Puts, Bart Buelens, and Paul A. M. van den Hurk. Big data as a source for official statistics. *Journal of Official Statistics*, 31(2):249–262, 2015.

Michael Davern, Marc Roemer, and Wendy Thomas. Investing in a data quality research program for administrative data linked to survey data for policy research purposes is essential. In *Federal Committee on Statistical Methodology Research Conference, Washington, DC.*, 2009. URL https://nces.ed.gov/FCSM/pdf/2009FCSM_Davern_IX-A.pdf.

Eugene Demidenko. Sample size determination for logistic regression revisited. *Statistics in Medicine*, 26 (18):3385–3397, 2007.

Eugene Demidenko. Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine*, 27(1):36–46, 2008.

W. Edwards Deming. On a probability mechanism to attain an economic balance between the resultant error of non-response and the bias of non-response. *Journal of the American Statistical Association*, 48: 743–772, 1953.

Renee Ellis, Patricia Goerman, Kathleen Kephart, Aleia Clark Fobia, Anna Sandoval Giron, Mikelyn Meyers, Rodney Terry, Leticia Fernandez, Fane Lineback, Marcus Berger, Antonio Bruce, and Eric Jensen. Research on coverage of underrepresented populations in anticipation of a records-based census, 2018. 2020 Census: Evaluation, Experiment, and Research and Testing Study.

Stephen E. Fienberg. *The analysis of cross-classified categorical data*. Springer Science & Business Media, 2nd edition, 2007.

Andrew S. Fullerton. A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, 38(2):306–347, 2009.

Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference*. Cambridge University Press, 2017.

Robert M. Groves and George J. Schoeffel. Use of administrative records in evidence-based policymaking. *The ANNALS of the American Academy of Political and Social Science*, 678(1):71–80, 2018.

Morris H. Hansen and William N. Hurwitz. The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41(236):517–529, 1946.

Sharon L. Lohr. *Sampling: Design and Analysis*. Brooks/Cole, Boston, MA, 2nd edition, 2010.

Robert H. Lyles, Hung-Mo Lin, and John M. Williamson. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Statistics in Medicine*, 26(7):1632–1648, 2007.

Emily Molfino, Gizem Korkmaz, Sallie A. Keller, Aaron Schroeder, Stephanie Shipp, and Daniel H. Weinberg. Can administrative housing data replace survey data? *Cityscape*, 19(1):265–292, 2017.

Darcy Steeg Morris, Andrew Keller, and Brian Clark. An approach for using administrative records to reduce contacts in the 2020 Decennial Census. *Statistical Journal of the IAOS*, 32(2):177–188, 2016.

Raymond H. Myers. *Classical and Modern Regression with Applications*. Duxbury Press, 2nd edition, 2000.

National Research Council. *Envisioning the 2020 Census*. The National Academies Press, Washington, DC, 2010. doi: https://dx.doi.org/10.17226/12865. Lawrence D. Brown and Michael L. Cohen and Daniel L. Cork and Constance F. Citro, editors.

Gary W. Oehlert. *A first course in design and analysis of experiments*. W. H. Freeman, 2000. URL http://users.stat.umn.edu/~gary/Book.html.

Yuling Pan and Stephen Lubkemann. Observing census enumeration of non-English speaking households in the 2010 Census: Evaluation report. Research Report Series: Survey Methodology #2013-02, Center for Survey Measurement, U.S. Census Bureau, 2013. URL https://www.census.gov/library/working-papers/2013/adrm/ssm2013-02.html.

Alfred Politz and Willard Simmons. An attempt to get the "not at homes" into the sample without callbacks. *Journal of the American Statistical Association*, 44(245):9–16, 1949.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

P. S. R. S. Rao. Callbacks, follow-ups, and repeated telephone calls. In W. G. Madow, I. Olkin, and D. B. Rubin, editors, *Incomplete Data in Sample Surveys*, volume 2, pages 33–44. Academic Press, New York, 1983.

Tommaso Rigon and Daniele Durante. Tractable Bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, 211:131–142, 2021.

Carl-Erik Särndal, Bengt Swensson, and Jan Wretman. *Model Assisted Survey Sampling*. Springer-Verlag New York, Inc., New York, 1992.

SAS Institute Inc. *The GENMOD Procedure*, chapter 48, pages 3407–3607. SAS Publishing, 2018. URL

http://support.sas.com/documentation/onlinedoc/stat/151/genmod.pdf.

Fritz Scheuren. Administrative records and census taking. *Survey Methodology*, 25(2):151–160, 1999.

Steven G. Self and Robert H. Mauritsen. Power/sample size calculations for generalized linear models. *Biometrics*, 44(1):79–86, 1988.

Steven G. Self, Robert H. Mauritsen, and Jill Ohara. Power calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 48(1):31–39, 1992.

Gwowen Shieh. On power and sample size calculations for likelihood ratio tests in generalized linear models. *Biometrics*, 56(4):1192–1196, 2000.

Gwowen Shieh. On power and sample size calculations for Wald tests in generalized linear models. *Journal of Statistical Planning and Inference*, 128(1):43–59, 2005.

Eleanor Singer. Introduction: Nonresponse Bias in Household Surveys. *Public Opinion Quarterly*, 70(5): 637–645, 2006.

Eric Slud and Benjamin Kedem. Partial likelihood analysis of logistic regression and autoregression. *Statistica Sinica*, 4(1):89–106, 1994.

Gerhard Tutz. Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11 (3):275–295, 1991.

U.S. Census Bureau. U.S. Census Bureau statistical quality standards, July 2013. URL https://www.census.gov/content/dam/Census/about/about-the-bureau/policies_and_notices/quality/statistical-quality-standards/Quality_Standards.pdf.

U.S. Census Bureau. 2020 Census detailed operational plan for: 18. nonresponse followup operation (NRFU), July 2019. URL https://www.census.gov/programs-surveys/decennial-census/2020-census/planning-management/planning-docs/NRFU-detailed-op-plan.html. Version 2.0 Final.

Shelley Walker, Susanna Winder, Geoff Jackson, and Sarah Heimel. 2010 census nonresponse followup operations assessment. Technical Report 190, 2010 Census Planning Memoranda Series, 2012. URL https://www.census.gov/2010census/pdf/2010_Census_NRFU_Operations_Assessment.pdf.

Angela M. Wood, Ian R. White, and Matthew Hotopf. Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169(3):525–542, 2006.

A Appendix

Proof of Result 2.1. Write $\eta_{i\ell} = \boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta}$. To derive (a), first note that

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log p_{i\ell} = \frac{1}{p_{i\ell}} g(\eta_{i\ell}) \boldsymbol{x}_{i\ell} = (1 + e^{-\eta_{i\ell}}) \frac{e^{-\eta_{i\ell}}}{(1 + e^{-\eta_{i\ell}})^2} \boldsymbol{x}_{i\ell} = [1 - G(\eta_{i\ell})] \boldsymbol{x}_{i\ell}$$

and

$$\frac{\partial}{\partial \boldsymbol{\beta}} \log(1 - p_{ib}) = -\frac{1}{1 - p_{ib}} g(\eta_{ib}) \boldsymbol{x}_{ib} = -\frac{1 + e^{-\eta_{ib}}}{e^{-\eta_{ib}}} \frac{e^{-\eta_{ib}}}{(1 + e^{-\eta_{ib}})^2} \boldsymbol{x}_{ib} = -G(\eta_{ib}) \boldsymbol{x}_{ib}.$$

We then have

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}} \log \mathcal{L}(\boldsymbol{\beta}) &= \frac{\partial}{\partial \boldsymbol{\beta}} \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} I(w_i = \ell) \left[\log p_{i\ell} + \sum_{b=1}^{\ell-1} \log(1 - p_{ib}) \right] \\ &= \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} I(w_i = \ell) \left[[1 - G(\eta_{i\ell})] \boldsymbol{x}_{i\ell} - \sum_{b=1}^{\ell-1} G(\eta_{ib}) \boldsymbol{x}_{ib} \right] \\ &= \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} I(w_i = \ell) \boldsymbol{x}_{i\ell} - \sum_{i=1}^{n} \sum_{\ell=1}^{L+1} I(w_i \geq \ell) G(\eta_{i\ell}) \boldsymbol{x}_{i\ell}.\end{aligned}$$

For (b), let us first write

$$\begin{aligned}\boldsymbol{D}_{\boldsymbol{w}} &= \text{Diag} \left\{ I(w_i \geq \ell) g(\boldsymbol{x}_{i\ell}^{\top} \boldsymbol{\beta}) : (i, \ell) \in \mathcal{I}, \right\}, \\ \boldsymbol{D}_{\boldsymbol{\beta}} &= \text{Diag} \left\{ \text{P}(W_i \geq \ell) g(\boldsymbol{x}_{i\ell}^{\top} \boldsymbol{\beta}) : (i, \ell) \in \mathcal{I} \right\} = \text{Diag} \left\{ g(\boldsymbol{x}_{i\ell}^{\top} \boldsymbol{\beta}) \prod_{b=1}^{\ell-1} [1 - G(\boldsymbol{x}_{ib}^{\top} \boldsymbol{\beta})] : (i, \ell) \in \mathcal{I} \right\}. \quad (16)\end{aligned}$$

so that $\boldsymbol{D}_{\boldsymbol{\beta}} = \text{E}[\boldsymbol{D}_{\boldsymbol{w}}]$. The last equality in (16) can be justified by

$$p_{i\ell} = \frac{\pi_{i\ell}}{\pi_{i\ell} + \cdots + \pi_{i,L+1}} = \frac{\pi_{i\ell}}{\text{P}(W_i \geq \ell)} = \frac{p_{i\ell} \prod_{b=1}^{\ell-1} (1 - p_{ib})}{\text{P}(W_i \geq \ell)} \quad \Longleftrightarrow \quad \text{P}(W_i \geq \ell) = \prod_{b=1}^{\ell-1} (1 - p_{ib}).$$

Now the second derivative of the log-likelihood is

$$\frac{\partial^2}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\top}} \log \mathcal{L}(\boldsymbol{\beta}) = - \sum_{i=1}^{n} \sum_{\ell=1}^{L} I(w_i \geq \ell) g(\boldsymbol{x}_{i\ell}^{\top} \boldsymbol{\beta}) \boldsymbol{x}_{i\ell} \boldsymbol{x}_{i\ell}^{\top} = - \boldsymbol{X}^{\top} \boldsymbol{D}_{\boldsymbol{w}} \boldsymbol{X}. \quad (17)$$

Taking the negative expectation of (17) yields the desired information matrix. □

Table 2: ACOs under consideration for the experiment.

				Percent		HH Counts	
Pair	Area	Group	Tracts	Spanish	Selfresp	Total	Target
1	Dallas	Ctrl	176	6.8	62.8	352,347	11,900
1	Houston	Expt	136	21.0	44.1	253,932	33,305
2	Dallas	Ctrl	163	14.2	48.5	293,170	24,847
2	Houston	Expt	148	10.1	47.9	278,782	18,412
3	Dallas	Ctrl	180	10.4	57.4	337,574	19,828
3	Houston	Expt	140	15.8	44.0	282,424	31,434
4	Dallas	Ctrl	170	24.9	41.2	277,452	43,271
4	Houston	Expt	122	21.6	41.3	240,950	36,575
5	Dallas	Ctrl	194	11.6	55.6	335,557	23,521
5	Houston	Expt	146	20.0	40.7	238,144	32,587
6	Dallas	Ctrl	235	4.0	66.3	482,153	8,084
6	Houston	Expt	91	8.0	61.3	268,572	9,525
7	LA	Ctrl	304	13.9	49.5	441,726	35,989
7	LA	Expt	355	16.1	48.5	496,564	50,740
Total			2,560			4,579,347	380,018

[1] Total HH Counts, Percent Spanish, and Percent Self-Response are based on Planning Database variables `Tot_Occp_Units_ACS_13_17`, `pct_Age5p_Spanish_ACS_13_17`, and `Self_Response_Rate_ACS_13_17`, respectively, which are sourced from American Community Survey 5-year estimates for the year 2017.

[2] Percentages are based on ACOs counts which have been aggregated from tract data; Target HH Count cannot be reproduced via (10) from here.

Table 3: Probabilities $\pi_{ijk\ell}$ under H_0 of a successful enumeration for attempts $\ell = 1, \ldots, 5$. Category 6+ contains the leftover probability that enumeration occurs after attempt 5.

	Attempt					
p	1	2	3	4	5	6+
0.05	0.05	0.0475	0.0451	0.0429	4.073E-2	7.738E-1
0.10	0.10	0.0900	0.0810	0.0729	6.561E-2	5.905E-1
0.15	0.15	0.1275	0.1084	0.0921	7.830E-2	4.437E-1
0.20	0.20	0.1600	0.1280	0.1024	8.192E-2	3.277E-1
0.25	0.25	0.1875	0.1406	0.1055	7.910E-2	2.373E-1
0.30	0.30	0.2100	0.1470	0.1029	7.203E-2	1.681E-1
0.35	0.35	0.2275	0.1479	0.0961	6.248E-2	1.160E-1
0.40	0.40	0.2400	0.1440	0.0864	5.184E-2	7.876E-2
0.45	0.45	0.2475	0.1361	0.0749	4.118E-2	5.033E-2
0.50	0.50	0.2500	0.1250	0.0625	3.125E-2	3.125E-2
0.55	0.55	0.2475	0.1114	0.0501	2.255E-2	1.845E-2
0.60	0.60	0.2400	0.0960	0.0384	1.536E-2	1.024E-2
0.65	0.65	0.2275	0.0796	0.0279	9.754E-3	5.253E-3
0.70	0.70	0.2100	0.0630	0.0189	5.670E-3	2.430E-3
0.75	0.75	0.1875	0.0469	0.0117	2.930E-3	9.766E-4
0.80	0.80	0.1600	0.0320	0.0064	1.280E-3	3.200E-4
0.85	0.85	0.1275	0.0191	0.0029	4.303E-4	7.594E-5
0.90	0.90	0.0900	0.0090	0.0009	9.000E-5	1.000E-5
0.95	0.95	0.0475	0.0024	0.0001	5.938E-6	3.125E-7

Table 4: Empirical power computed by simulation. A dash (—) means that no samples in this setting yielded valid estimates of the coefficient and variance.

L	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
1	10	0.60	0.1090	0.1020	0.1100	0.1840	0.1950	0.2970	0.4480	0.5810
		0.75	0.0450	0.0440	0.0670	0.0580	0.0860	0.1110	0.2200	0.2820
		0.90	0.0010	0.0020	0.0000	0.0020	0.0020	0.0050	0.0070	0.0240
1	50	0.60	0.0900	0.1280	0.2470	0.4290	0.6230	0.7710	0.9800	0.9990
		0.75	0.0940	0.1270	0.2300	0.3410	0.5420	0.6830	0.9450	0.9950
		0.90	0.0590	0.0800	0.1220	0.1550	0.2670	0.3600	0.6380	0.7730
1	200	0.60	0.0900	0.2520	0.6450	0.9030	0.9820	1.0000	1.0000	1.0000
		0.75	0.1160	0.2290	0.5380	0.8090	0.9680	0.9980	1.0000	1.0000
		0.90	0.0840	0.1520	0.3320	0.5700	0.7720	0.9160	0.9960	1.0000
2	10	0.60	0.0071	0.0213	0.0163	0.0123	0.0392	0.0484	0.1065	0.2326
		0.75	0.0034	0.0000	0.0022	0.0046	0.0082	0.0120	0.0294	0.0656
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0034
2	50	0.60	0.0790	0.1290	0.2050	0.3670	0.5640	0.7400	0.9770	1.0000
		0.75	0.0580	0.0820	0.1270	0.2170	0.3650	0.5200	0.8880	0.9900
		0.90	0.0283	0.0263	0.0276	0.0443	0.0730	0.1313	0.2913	0.6145
2	200	0.60	0.0950	0.2190	0.5760	0.9120	0.9910	1.0000	1.0000	1.0000
		0.75	0.0970	0.1890	0.4680	0.7810	0.9520	0.9940	1.0000	1.0000
		0.90	0.0540	0.0870	0.1780	0.3290	0.5830	0.7630	0.9870	1.0000
3	10	0.60	0.0000	0.0000	0.0000	0.0000	0.0017	0.0017	0.0068	0.0408
		0.75	0.0000	0.0000	0.0000	0.0000	0.0045	0.0000	0.0000	0.0088
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
3	50	0.60	0.0430	0.0631	0.1474	0.2590	0.4789	0.6839	0.9621	1.0000
		0.75	0.0087	0.0294	0.0496	0.0938	0.1553	0.3211	0.7169	0.9594
		0.90	0.0351	0.0272	0.0210	0.0210	0.0000	0.0227	0.0962	0.2537
3	200	0.60	0.0900	0.2080	0.5580	0.8750	0.9930	1.0000	1.0000	1.0000
		0.75	0.0480	0.1260	0.3490	0.6770	0.9179	0.9890	1.0000	1.0000
		0.90	0.0384	0.0369	0.0865	0.1832	0.3450	0.5611	0.9657	1.0000
4	10	0.60	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		0.75	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		0.90	—	—	—	—	—	0.0000	—	—
4	50	0.60	0.0150	0.0352	0.0506	0.1226	0.2513	0.4449	0.8755	0.9909
		0.75	0.0033	0.0071	0.0206	0.0326	0.1188	0.1741	0.5946	0.8889
		0.90	0.0000	0.0000	0.0000	0.0000	0.0000	—	0.0000	1.0000
4	200	0.60	0.0810	0.1650	0.5055	0.8660	0.9850	0.9990	1.0000	1.0000
		0.75	0.0242	0.0631	0.2335	0.5271	0.8302	0.9548	1.0000	1.0000
		0.90	0.1800	0.1282	0.0645	0.2250	0.3333	0.6129	1.0000	1.0000

Table 5: Approximate power in each simulation setting.

L	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
1	10	0.60	0.1000	0.1081	0.1321	0.1707	0.2220	0.2830	0.4551	0.6142
		0.75	0.1000	0.1063	0.1250	0.1552	0.1951	0.2429	0.3796	0.5118
		0.90	0.1000	0.1030	0.1120	0.1264	0.1455	0.1683	0.2345	0.3013
1	50	0.60	0.1000	0.1403	0.2558	0.4248	0.6084	0.7665	0.9622	0.9963
		0.75	0.1000	0.1315	0.2226	0.3597	0.5183	0.6697	0.9098	0.9820
		0.90	0.1000	0.1152	0.1595	0.2290	0.3171	0.4149	0.6461	0.8025
1	200	0.60	0.1000	0.2570	0.6198	0.8963	0.9859	0.9990	1.0000	1.0000
		0.75	0.1000	0.2237	0.5308	0.8205	0.9586	0.9942	1.0000	1.0000
		0.90	0.1000	0.1602	0.3270	0.5491	0.7515	0.8868	0.9917	0.9996
2	10	0.60	0.1000	0.1061	0.1246	0.1554	0.1980	0.2510	0.4117	0.5713
		0.75	0.1000	0.1043	0.1170	0.1381	0.1669	0.2025	0.3111	0.4245
		0.90	0.1000	0.1018	0.1071	0.1158	0.1274	0.1415	0.1838	0.2284
2	50	0.60	0.1000	0.1312	0.2281	0.3871	0.5776	0.7514	0.9656	0.9975
		0.75	0.1000	0.1216	0.1881	0.2991	0.4428	0.5953	0.8735	0.9713
		0.90	0.1000	0.1090	0.1363	0.1814	0.2429	0.3169	0.5192	0.6857
2	200	0.60	0.1000	0.2293	0.5922	0.8993	0.9899	0.9996	1.0000	1.0000
		0.75	0.1000	0.1891	0.4572	0.7703	0.9442	0.9922	1.0000	1.0000
		0.90	0.1000	0.1368	0.2508	0.4325	0.6358	0.8036	0.9776	0.9983
3	10	0.60	0.1000	0.1051	0.1206	0.1467	0.1833	0.2298	0.3759	0.5286
		0.75	0.1000	0.1034	0.1134	0.1301	0.1531	0.1820	0.2725	0.3711
		0.90	0.1000	0.1014	0.1054	0.1119	0.1207	0.1315	0.1642	0.1991
3	50	0.60	0.1000	0.1261	0.2100	0.3550	0.5399	0.7193	0.9581	0.9967
		0.75	0.1000	0.1170	0.1704	0.2630	0.3899	0.5338	0.8304	0.9550
		0.90	0.1000	0.1068	0.1275	0.1623	0.2109	0.2711	0.4475	0.6083
3	200	0.60	0.1000	0.2111	0.5556	0.8831	0.9881	0.9995	1.0000	1.0000
		0.75	0.1000	0.1712	0.4035	0.7153	0.9197	0.9869	1.0000	1.0000
		0.90	0.1000	0.1279	0.2172	0.3698	0.5587	0.7351	0.9590	0.9959
4	10	0.60	0.1000	0.1044	0.1177	0.1403	0.1723	0.2133	0.3457	0.4900
		0.75	0.1000	0.1028	0.1112	0.1252	0.1446	0.1690	0.2471	0.3345
		0.90	0.1000	0.1011	0.1044	0.1099	0.1172	0.1261	0.1533	0.1827
4	50	0.60	0.1000	0.1225	0.1958	0.3266	0.5015	0.6814	0.9458	0.9952
		0.75	0.1000	0.1142	0.1592	0.2387	0.3513	0.4850	0.7883	0.9360
		0.90	0.1000	0.1056	0.1228	0.1518	0.1926	0.2440	0.4005	0.5525
4	200	0.60	0.1000	0.1968	0.5169	0.8584	0.9836	0.9993	1.0000	1.0000
		0.75	0.1000	0.1599	0.3636	0.6647	0.8916	0.9793	1.0000	1.0000
		0.90	0.1000	0.1231	0.1980	0.3303	0.5044	0.6804	0.9385	0.9924

Table 6: Count of NAs in each simulation setting when calculating empirical power.

L	K	$\text{logit}^{-1}(\mu)$	$\Delta = 0$	0.1	0.2	0.3	0.4	0.5	0.75	1.0
1	10	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
1	50	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
1	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
2	10	0.60	11	15	16	26	30	29	61	110
		0.75	109	111	106	132	145	167	252	314
		0.90	576	589	572	587	583	613	649	708
2	50	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	4
		0.90	10	10	23	29	28	48	80	144
2	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	0	0	0	0
		0.90	0	0	0	0	0	0	0	0
3	10	0.60	303	324	339	359	400	414	557	681
		0.75	804	800	794	773	776	795	853	887
		0.90	995	989	993	985	988	991	986	996
3	50	0.60	0	1	3	4	4	13	49	150
		0.75	77	81	91	158	182	265	410	532
		0.90	827	852	855	856	866	868	896	933
3	200	0.60	0	0	0	0	0	0	0	0
		0.75	0	0	0	0	1	3	24	102
		0.90	244	269	306	356	371	435	592	747
4	10	0.60	772	754	794	789	803	832	892	930
		0.75	979	977	976	985	968	993	983	983
		0.90	1000	1000	1000	1000	1000	999	1000	1000
4	50	0.60	69	90	110	168	236	292	510	669
		0.75	697	720	709	785	798	799	889	946
		0.90	993	998	998	997	999	1000	998	999
4	200	0.60	0	0	1	0	3	5	59	203
		0.75	91	97	135	207	317	380	625	765
		0.90	950	961	969	960	973	969	980	991

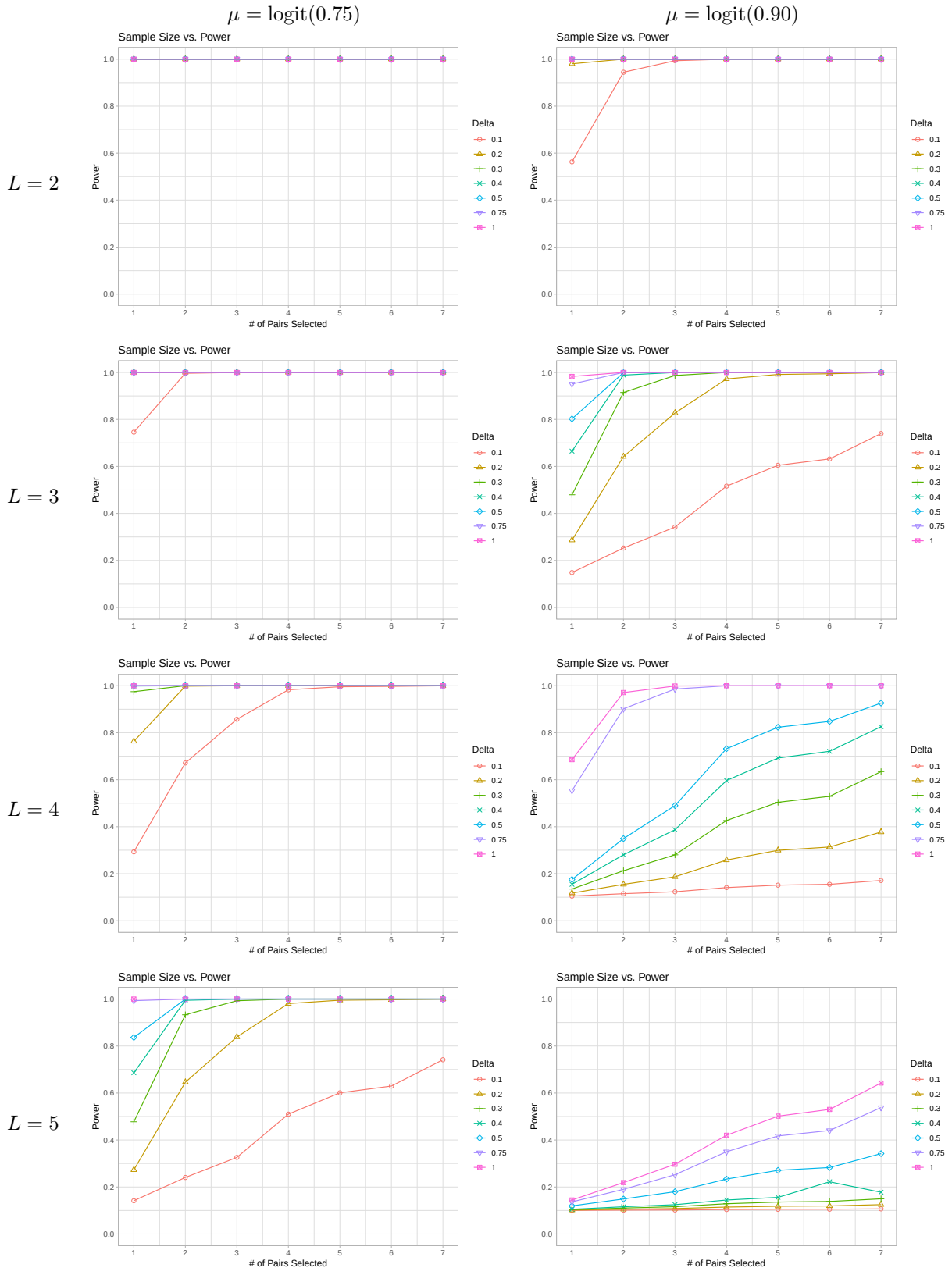

Figure 1: Power study using the fourteen pre-selected ACOs in Dallas, Houston, and Los Angeles.

Table 7: An aid to interpret effect size $\Delta = 0.1$ with $\mu = \text{logit}(0.9)$. Case (a) represents H_0, while cases (b)–(f) place all of effect size Δ on one particular coordinate of $\boldsymbol{\beta}$. Trial probabilities $\boldsymbol{p}_{ijk}$ from case (a) can be compared to each case (b)–(f) to visualize the differences that can be detected by the experiment. Similarly, overall enumeration probabilities $\boldsymbol{\pi}_{ijk}$ from case (a) can be compared to each case (b)–(f).

(a) H_0

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9000	0.9000	0.9000	0.0900	0.0090	0.0010
2	0.9000	0.9000	0.9000	0.9000	0.0900	0.0090	0.0010

(b) $\tau_1 = \Delta$

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9086	0.9086	0.9086	0.0830	0.00758	0.000762
2	0.8906	0.8906	0.8906	0.8906	0.0974	0.01065	0.001308

(c) $\delta_1 = \Delta$

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999
2	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999

(d) $\delta_2 = \Delta$

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999
2	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999

(e) $(\tau\delta)_{11} = \Delta$

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9086	0.9000	0.8906	0.9086	0.0822	0.00814	0.000999
2	0.8906	0.9000	0.9086	0.8906	0.0984	0.00994	0.000999

(f) $(\tau\delta)_{12} = \Delta$

j	p_{ijk1}	p_{ijk2}	p_{ijk3}	π_{ijk1}	π_{ijk2}	π_{ijk3}	π_{ijk4}
1	0.9000	0.9086	0.8906	0.9000	0.0909	0.00814	0.000999
2	0.9000	0.8906	0.9086	0.9000	0.0891	0.00994	0.000999