Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A Calibrated Qualitative Response Estimation Approach

Erard, Brian

B. Erard Associates

24 June 2017

Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A Calibrated Qualitative Response Estimation Approach

Brian Erard
B. Erard & Associates

October 19, 2017

Abstract

Often providers of a program or service have detailed information about their clients, but only limited information about potential clients. Likewise, ecologists frequently have extensive knowledge regarding habitats where a given animal or plant species is known to be present, but they lack comparable information on habitats where they are certain not to be present. In epidemiology, comprehensive information is routinely collected about patients who have been diagnosed with a given disease; however, commensurate information may not be available for individuals who are known to be free of the disease. While it may be highly beneficial to learn about the determinants of participation (in a program or service) or presence (in a habitat or of a disease), the lack of a comparable sample of observations on subjects that are not participants (or that are non-present) precludes the application of standard qualitative response models, such as logit or probit. In this paper, we examine how one can overcome this challenge by combining a participant-only sample with a supplementary sample of covariate values from the general population. We derive some new estimators of conditional response probabilities based on this sampling scheme by exploiting the parameter restrictions implied by the relationship between the marginal and conditional response probabilities in the supplementary sample. When the prevalence rate in the population is known, we demonstrate that the choice of estimator is especially important when this rate is relatively high. Our simulation results indicate that some of our new estimators for this case rival the small sample performance of the best existing estimators. Our estimators for the case where the prevalence rate is unknown also perform comparably to the best existing estimator for this situation in our simulations. In contrast to most existing estimators, our new estimators are straightforward to apply to exogenously stratified samples (such as when the supplementary sample is drawn from a Census survey), even when the underlying stratification criteria are not available. Our new estimators also readily generalize to accommodate situations with polychotomous outcomes.

Modeling Qualitative Outcomes by Supplementing Participant Data with General Population Data: A Calibrated Qualitative Response Estimation Approach*

Brian Erard
B. Erard & Associates

October 19, 2017

1. Introduction

Often providers of a program or service have detailed information about their clients, but only very limited information about potential clients. Likewise, ecologists frequently have extensive knowledge regarding habitats where a given animal or plant species is known to be present, but they lack comparable information on habitats where they are certain not to be present. In epidemiology, comprehensive information is routinely collected about patients who have been diagnosed with a given disease; however, commensurate information may not be available for individuals who are known to be free of the disease. While it may be highly beneficial to learn about the determinants of participation (in a program or service) or presence (in a habitat or of a disease), the lack of a comparable sample of observations on subjects that are not participants (or that are non-present) precludes the application of standard qualitative response models, such as logit or probit. In fact, though, if a *supplementary* random sample can be drawn from the general population of interest, it is feasible to estimate conditional response probabilities. Importantly, this supplementary sample need not include information on whether the subjects are participants or non-participants, present or not present. Rather, it only must include measures of the relevant covariates, comparable to those collected from the *primary* sample (of subjects that are participants or that are present). This sampling scheme, involving a primary sample consisting exclusively of participants and a supplementary sample that includes both participants and non-participants, has been assigned various names in the literature, including "use control sampling", "use-availability sampling" "supplementary sampling", "case control sampling with contaminated controls", "presence pseudo-absence sampling", and "presence-background sampling".[1]

* This research was supported by Internal Revenue Service contracts TIRNO-10-D-00021-D0004, TIRNO-14-P-00157, and TIRNO-15-P-00172. I am grateful to Stephen Cosslett and Subhash Lele for their very helpful comments and suggestions. I am also grateful to John Guyton, Patrick Langetieg, Mark Payne, and Alan Plumley for helping me to refine my methodology as we worked on applying the approach to understand the determinants of taxpayer filing behavior. The views expressed in this paper are my own and do not necessarily reflect the opinions of the IRS.

In many cases, one may be able to rely on a general survey of the overall population as a supplementary sample.[2] In the U.S., a few examples include the Current Population Survey (CPS), the Survey of Income and Program Participation (SIPP), and the American Community Survey (ACS). Often, however, such data sources are created from stratified random samples, meaning that sample weights must be applied to make the surveys representative of the underlying population of interest. It is straightforward to adapt the existing statistical models for analyzing use control data by Lancaster and Imbens (1996) and Cosslett (1981) to account for a relatively simple stratified random sampling design. However, such an implementation would require knowledge not just of the sample weights, but also the underlying stratification criteria. Unfortunately, in many cases (including the three major Census surveys referred to above), the stratum definitions are not made available to the public. In any case, it would be difficult to adapt the Lancaster-Imbens and Cosslett frameworks to account for the complex sampling designs employed in many national surveys (which typically involve multi-stage sampling, clustering, post-stratification adjustment, and imputation).

In this paper, we present some new estimators that can be applied to stratified supplementary samples even when the underlying details of the sampling design are not available; all that is required are the sample weights, which are routinely available. We further show that our new estimators are easily generalized to address polychotomous response problems. Our main estimators are derived in Section 2, where we focus on the case in which the prevalence rate within the general population is known (i.e., the overall take-up rate in the case of a program, the percentage of habitats occupied by a species in the case of wildlife presence, or the share of the population that is infected in the case of a disease). In Section 3, we conduct a Monte Carlo analysis to evaluate the small sample performance of our new estimators vis à vis existing estimators. Our results indicate that the choice among estimators for this case is especially important when the prevalence rate is known to be relatively high. Under such circumstances, some of our new estimators rival the performance of the best existing estimators (Cosslett, 1981, and Lancaster and Imbens, 1996). In Section 4, we introduce some estimators for the case where the prevalence rate is unknown. In Section 5, we show that the estimator proposed by Lancaster and Imbens (1996) for this case is actually equivalent to the estimator previously proposed by Cosslett (1981). We then perform a Monte Carlo analysis to compare the small sample performance of our new estimators with that of the Cosslett-Lancaster-Imbens estimator and find that they perform comparably. In Section 6, we demonstrate how our new estimators can easily be generalized for application in stratified samples, even when the stratification criteria are unknown. In Section 7, we extend our methodology to accommodate situations with polychotomous outcomes. As an illustration of our extended methodology, we compare results from a multinomial logit study of voting behavior against those from a calibrated multinomial logit analysis in Section 8. Section 9 concludes.

---

[1] Discussions of applications of use control sampling in various fields include Breslow (1996) [epidemiology]' Keating and Cherry (2004), Royle et al. (2012), and Phillips and Elith (2013) [ecology]; Erard et al. (2016) [tax compliance]; and Rosenman, Goates, and Hill (2012) [substance abuse prevention programs].

[2] If eligibility for a program or service is limited, one may be able to restrict the supplementary sample to include only those survey respondents who are eligible, providing that eligibility can be deduced from the survey information that has been collected. For instance, the CPS has detailed income information that can be useful in assessing eligibility for means-tested programs and services.

2. Estimation Methodology

Using the notation of Lancaster and Imbens (1996), let y be a binary response variable equal to 1 (for participation/presence) or 0 (for non-participation/non-presence) and let x represent a vector of attributes with cumulative distribution function $F(x)$. We assume that the conditional probability that y is equal to 1 given x follows a known parametric form:

$$\Pr(y = 1|x;\beta) = P(\mathrm{x};\beta), \tag{1}$$

where β is an unknown parameter vector we desire to estimate. Finally, we define the prevalence rate q (the marginal probability that y equals 1 in the population) as:

$$q = \int P(x;\beta)dF(x). \tag{2}$$

Suppose we have a simple random sample of size N_1 from the subpopulation of cases with y equal to 1. The conditional probability of x given $y = 1$ is equal to:

$$\mathrm{g}(x|y = 1) = \frac{P(x;\beta)f(x)}{q}, \tag{3}$$

where $f(x)$ represents the joint marginal p.d.f. of x $\left[f(x) = \frac{dF(x)}{dx}\right]$. Therefore, if both $F(x)$ and q were known (and assuming that the marginal distribution of x does not depend directly on β), one could consistently estimate the parameter vector β by maximizing the log-likelihood function:

$$L = \sum_{i=1}^{N_1} ln[P(x_i;\beta)] \tag{4}$$

subject to the constraint on β implied by Equation (2).[3] Rather remarkably, this means that it would be possible to conduct a (constrained) maximum likelihood analysis of the propensity to participate using a data sample that consists only of participants.

In many instances, one may be able to measure (at least to some degree of confidence) the value of q. For instance, one may have a reasonably reliable estimate of the take-up rate for a particular government program or the prevalence rate for a given disease. Even allowing for the possibility that q may be known, however, the cumulative distribution of x, $F(x)$, typically will not be known. Therefore it normally will not be feasible to estimate β without some source of additional information. Below we consider how information from a supplementary sample of covariate values from the general population can be used to overcome our ignorance of the covariate distribution.

[3] See Manski and McFadden (1981, pp. 13-17) for a related discussion of choice-based estimation of qualitative response models when both the covariate distribution and prevalence rate are known. Whereas Manski and McFadden consider the case of a choice-based sample that includes participants and non-participants, the current specification involves a sample that includes only participants.

Under the supplementary sampling scheme that we consider, one draws both a primary random sample of covariate values from the subpopulation of participants and a separate random sample of covariate values from the general population. Equation (3) describes the conditional probability of x given that an observation is from the primary sample, while $f(x)$ describes the probability distribution of x among observations in the supplementary sample. As noted by Lancaster and Imbens (1996), a comparison of these two cases reveals that the function $P(x;\beta)/q$ is nonparametrically identified. If the prevalence rate q is known, then $P(x;\beta)$ is also nonparametrically identified.

When q is unknown, the relative probability $P(x;\beta)/P(y;\beta)$ continues to be nonparametrically identified. However, identification of β in this case depends on the parametric specification of the conditional response probability. For certain specifications, it is not possible to separately identify all of the elements of β. For instance, under a linear probability model, $\frac{P(x;\beta_0,\beta_1)}{q} = \frac{\beta_0+{\beta_1}'x}{q} = \left(\frac{\beta_0}{q}\right) + \left(\frac{\beta_1}{q}\right)' x$. Therefore, only the ratio of each element of β to q is identified. Ecological models of resource selection often rely on an exponential (log-linear) probability model. Under this specification, $\ln\left(\frac{P(x;\beta_0,\beta_1)}{q}\right) = \beta_0+{\beta_1}'x - lnq$ $= (\beta_0 - lnq)+{\beta_1}'x$. In this case, each of the slope coefficients of the conditional response probability is identified, but the intercept is not.[4] Fortunately, these two cases are exceptional. As discussed by Solymos and Lele (2016), all of the elements of β are identified under most qualitative choice specifications, including the logit, probit, arctan, and complementary log-log models, so long as the specification includes at least one continuous covariate. Parametric specifications that are adequate to identify β in cases where the prevalence rate is unknown will overidentify β in cases where the prevalence rate is known.

The remainder of this section focuses on qualitative response model estimation under supplementary sampling when the prevalence rate is known. The case of an unknown prevalence rate is taken up later in Section 4.

2.1 Constrained Pseudo-Maximum Likelihood Estimator

Suppose that, in addition to our primary sample of N_1 participants, we also have access to a supplementary sample of N_0 observations of covariate values from the general population of interest. Assume, for now, that this supplementary sample is a simple random sample. In Section 6 we will generalize our approach to account for exogenous stratification of both the primary and supplementary samples. With the aid of this supplementary sample, it is possible to consistently estimate β even when the cumulative distribution of x is not known.

[4] In a pure choice-based model (which is referred to as a "case-control" model by epidemiologists and ecologists), the function $\left(\frac{P(x;\beta)}{1-P(x;\beta)}\right)\left(\frac{1-q}{q}\right)$ is identified rather than $\left(\frac{P(x;\beta)}{q}\right)$. As a consequence, the intercept of the logit specification is not identified under a pure choice-based model when the prevalence rate is unknown, whereas it is the intercept of the exponential probability specification that is not identified under a supplementary sampling design.

Development of our new estimators of β follows the approach introduced by Imbens (1992) and later employed by Lancaster and Imbens (1996). Under this approach, we begin by constructing an estimator for the case where x is discrete with K known points of support. We derive this estimator by solving the above constrained maximum likelihood estimation problem based on the empirical distribution of x in the supplementary sample. We then demonstrate that our estimator can be expressed in a way that not only requires no knowledge of the points of support for x, but which remains valid even when x is continuous.

Using the supplementary sample, one can consistently estimate the probability (λ_k) that x is equal to x_k as:

$$\tilde{\lambda}_k = \frac{N_{0k}}{N_0},\ k = 1, \dots, K, \tag{5}$$

where N_{0k} represents the number of observations in the supplementary sample with covariate value $x = x_k$.[5] One can then consistently estimate β by maximizing:

$$L = \sum_{k=1}^{K} N_{1k} ln[P(x_k; \beta)] \tag{6}$$

subject to the analog of the constraint on β that is imposed by prevalence rate from Equation (2):

$$q = \sum_{k=1}^{K} \tilde{\lambda}_k\, P(x_k; \beta), \tag{7}$$

where N_{1k} represents the number of observations in the primary sample of participants with covariate value $x = x_k$. This estimator ($\tilde{\beta}_1$) can be expressed in an alternative way as the solution to:

$$\tilde{\beta}_1 = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} ln[P(x_i; \beta)] \ \ s.t.\ q = \frac{\sum_{j=1}^{N_0} P(x_j; \beta)}{N_0}. \tag{8}$$

When the parameter vector β is of dimension greater than one, there typically will be an infinite set of parameter combinations that satisfy the constraint in Equation (8). Among these alternatives, the solution to the constrained optimization problem is the parameter vector that maximizes the joint conditional probability that each of the observations in the primary sample would have an outcome of $y = 1$ given the sampled covariate values.

We refer to our estimation methodology as "calibrated qualitative response estimation", because the estimator is obtained by calibrating the response probabilities so that their average value within the supplementary sample is equal to the population prevalence rate q. Following standard terminology for

[5] Whereas our approach focuses on the unconditional probability (λ_k) of x_k and estimates it based on the supplementary sample moment (N_{0k}/N_0), the Lancaster and Imbens (1996) approach focuses on the conditional probability (π_k) that an observation with value x_k is included in the combined sample and estimates this probability using the combined sample moment $(N_{0k} + N_{1k})/(N_0 + N_1)$.

the classic qualitative response framework, we refer to our model as a "calibrated probit" when $P(x;\beta)$ is cumulative standard normal, and as a "calibrated logit" when $P(x;\beta)$ is cumulative standard logistic.

The Lagrangian for the optimization problem in Equation (8) is:

$$\mathcal{L}(\beta,\mu) = \sum_{i=1}^{N_1} ln[P(x_i;\beta)] \; + \mu\left[N_0 q - \sum_{j=1}^{N_0} P\left(x_j;\beta\right)\right]. \tag{9}$$

The first-order conditions are:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N_1} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \mu \sum_{j=1}^{N_0} P'_\beta\left(x_j;\beta\right) = 0. \tag{10}$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = N_0 q - \sum_{j=1}^{N_0} P\left(x_j;\beta\right) = 0. \tag{11}$$

In Equation (10), $P'_\beta(x;\beta) = \frac{\partial P(x;\beta)}{\partial \beta}$ is of order 1 x H, where H is the dimension of β. Observe that these moments do not require knowledge of the points of support and that they remain valid even when x is not discrete.

A difficulty with the above estimator is that the usual estimate of the covariance matrix of the parameter estimates that is computed from a constrained maximum likelihood algorithm will not be valid. This is because we have replaced the exact formula for $q \;\; [\int P(x;\beta) dF(x)]$ with its sample analog $\left[\frac{\sum_{j=1}^{N_0} P(x_j;\beta)}{N_0}\right]$. Intuitively, the reliance on an approximate relationship between β and q rather than the exact relationship tends to reduce the precision of our estimator to some degree. In Section 2.4, below, we rely on insights from generalized method of moments (GMM) theory to develop a covariance matrix estimator that properly accounts for this effect.

2.2 Unconstrained Pseudo-Maximum Likelihood Estimator

Equation (10) can be used to investigate the properties of the Lagrange multiplier μ. Let s be a 1/0 indicator that identifies observations from the primary sample in the combined primary and supplementary sample. Then the H first-order conditions for β can be rewritten as:

$$\frac{\partial \mathcal{L}}{\partial \beta} = \sum_{i=1}^{N} \left[s_i \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \mu(1-s_i) P'_\beta(x_i;\beta)\right] = 0, \tag{12}$$

where $N = (N_0 + N_1)$ is the size of the combined primary and supplementary sample. The conditional expectation of s given x in the combined sample is:

$$\pi_S = \frac{N_1 P(x;\beta)/q}{N_1 P(x;\beta)/q + N_0}. \tag{13}$$

The conditional expectation of the first-order conditions in Equation (12) given x may therefore be expressed as:

$$E\left(\frac{\partial \mathcal{L}}{\partial \beta}\middle| x\right) = \sum_{i=1}^{N}\left(\frac{N_1/q - \mu N_0}{N_1 P(x_i;\beta)/q + N_0}\right) P'_\beta(x_i;\beta) = 0. \tag{14}$$

This equation is satisfied when μ is set equal to $N_1/(N_0 q)$. Similar to the approach used by Manski and McFadden (1981) to develop a consistent estimator for the standard choice-based sampling problem, our second new estimator of β ($\tilde{\beta}_2$) is derived by substituting this limit value for μ in place of its actual value in Equation (9) and maximizing over β:

$$\tilde{\beta}_2 = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N}\left[s_i ln[P(x_i;\beta)] - \frac{N_1}{N_0 q}(1 - s_i)P(x_i;\beta)\right]. \tag{15}$$

The first-order conditions associated with this estimator are:

$$\sum_{i=1}^{N} s_i \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} - \frac{N_1}{N_0 q}\sum_{i=1}^{N}(1 - s_i)P'_\beta(x_i;\beta) = 0. \tag{16}$$

So while our first estimator ($\tilde{\beta}_1$) is obtained by solving a constrained optimization problem, this alternative estimator ($\tilde{\beta}_2$) requires no constraints and can be obtained using a standard maximum likelihood algorithm. However, the usual estimate of the covariance matrix of our estimator from such an algorithm (based on the estimated information matrix) will not be valid. In Section 2.4, we discuss an appropriate estimator of the asymptotic covariance matrix based on insights from GMM theory.

A potential drawback of our alternative estimator in small samples is that the average predicted probability of participation (or presence) in the supplementary sample may deviate fairly substantially from the population prevalence rate q. Our Monte Carlo simulations in Section 3 indicate that this can sometimes result in a failure to find a solution when q is relatively large (say, above 0.70) and the combined sample size is fairly small (say, 600).[6] In large samples, however, the average predicted probability of participation in the supplementary sample tends to be close to q and the estimator performs reasonably well.

[6] Typically In these failed solutions, the predicted probability of participation for all observations in the primary sample is pushed to one, while the average predicted probability of participation in the supplementary sample is kept slightly below one.

2.3 Comparison of Standard Logit and Calibrated Logit Estimators

It is instructive to compare the standard logit estimator against our new estimators. Consider a random sample of size N from a population containing N_1 observations with $\mathrm{y} = 1$ and $(N - N_1)$ observations with $\mathrm{y} = 0$. Let t serve as a 1/0 indicator for the outcome $y = 1$.

The log-likelihood function for the standard logit model is:

$$L = \sum_{i=1}^{N} t_i ln\big(P(x_i;\beta)\big) + (1 - t_i) ln\big(1 - P(x_i;\beta)\big), \tag{17}$$

where

$$P(x_i;\beta) = \frac{exp(\beta' x_i)}{1 + exp(\beta' x_i)}. \tag{18}$$

The standard logit estimator is the solution to the following first-order conditions:

$$\sum_{i=1}^{N} t_i\big(1 - P(x_i;\beta)\big)x_i' - \sum_{i=1}^{N}(1 - t_i)P(x_i;\beta)x_i' = 0, \tag{19}$$

In contrast, our unconstrained pseudo-maximum likelihood estimator for this data sample satisfies the conditions:

$$\sum_{i=1}^{N} t_i\big(1 - P(x_i;\beta)\big)x_i' - \frac{N_1}{Nq}\sum_{i=1}^{N}\big(1 - P(x_i;\beta)\big)P(x_i;\beta)\, x_i' = 0. \tag{20}$$

Observe that the expected number of observations with outcome $\mathrm{y} = 1$ in an overall sample of N randomly chosen observations is equal to Nq, so the expected value of the ratio $\frac{N_1}{Nq}$ in Equation (20) is equal to 1. Although the moment conditions in Equations (19) and (20) are both valid, the former yields a more efficient estimator, because it exploits knowledge regarding which specific observations have outcome $\mathrm{y} = 0$. In particular, this knowledge is employed when evaluating the second expression in Equation (19). In contrast, the latter moment condition replaces the term $(1 - t_i)$ in this expression with its conditional expectation $\big(1 - P(x_i;\beta)\big)$.

The Lagrangian function associated with our calibrated logit estimator for this data sample is:

$$\mathcal{L}(\beta,\mu) = \left(\sum_{i=1}^{N} t_i ln[P(x_i;\beta)]\right) + \mu\left(Nq - \sum_{j=1}^{N} P(x_j;\beta)\right). \quad (21)$$

It is well known that the sum of the predicted probabilities of a success in a sample based on the standard logit estimator is equal to the actual number of observations in the sample (N_1) with outcome $y = 1$. Since the expected value of N_1 is equal to Nq, it follows that the standard logit estimator approximately satisfies the constraint in Equation (21). However, even if N_1 were exactly equal to Nq, the standard logit estimator would not in general be equal to our constrained pseudo-maximum likelihood estimator. Among the feasible choices of β that satisfy the constraint, the choice that maximizes the objective function $\left(\sum_{i=1}^{N} t_i ln[P(x_i;\beta)]\right)$ in Equation (21) will not, in general, be the same as the choice that maximizes the objective function in Equation (17), owing to the additional expression $\left(\sum_{i=1}^{N}(1-t_i) ln[1-P(x_i;\beta)]\right)$ in the latter equation. Again, the standard logit model is more efficient, because it exploits specific information regarding which observations in the sample have outcome $y = 0$.

Intuitively, knowledge concerning which cases have outcome $y = 0$ is most valuable when there are relatively few observations that satisfy this condition (i.e., when q is large and N is small). In that case, the moment conditions for the standard logit model in Equation (19) would rely directly on a comparison of sampled participants against sampled non-participants, whereas the moment conditions in Equation (20) associated with our new calibrated qualitative response estimators would rely on a more subtle distinction between sampled participants and the overall sample that is itself composed mostly of participants. In Section 3, we perform some Monte Carlo simulations to confirm this intuition.

We also note that the primary data sample in our logit example is a proper subset of the overall sample, meaning that all of the observations in the primary sample are also present in the supplementary sample. In a more typical application involving independently drawn primary and supplementary data samples, the moment conditions for our new estimators, $\tilde{\beta}_1$ and $\tilde{\beta}_2$, will tend to be noisier owing to the random differences in the covariates across the two samples. We also explore this issue in our Monte Carlo simulations.

2.4 GMM Framework

The estimator $\tilde{\beta}_1$ is calibrated to ensure that the average predicted probability of participation in the supplementary sample is consistent with the prevalence rate, even in small samples. However, it requires solving a constrained optimization problem. While this is more complex than solving a standard maximum likelihood problem, one can obtain a solution using readily available algorithms, such as the CML application in GAUSS$^{@}$ or the nonlinear optimization routines in SAS$^{@}$/IML$^{@}$. Alternatively, one can follow Lancaster and Imbens (1996) and reframe the problem using a GMM approach. Consider the following moments:

$$g_1(x;\beta) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\frac{N_1}{N_0 q}P'_\beta(x;\beta). \quad (22)$$

$$g_2(x;\beta) = (1-s)\big(q - P(x;\beta)\big). \tag{23}$$

The moment $g_1(x;\beta)$ is the single observation score from the pseudo-log-likelihood function defined in Equation (15), while $g_2(x;\beta)$ reflects the relationship between marginal q and conditional $P(x;\beta)$. These moments have an expected value of zero when evaluated at the true value of β. Let $g(x;\beta)$ represent the vector $\begin{bmatrix} g_1(x;\beta) \\ g_2(x;\beta) \end{bmatrix}$ and $N = (N_0 + N_1)$ represent the size of the combined primary and supplementary sample. A standard GMM algorithm can then be applied to estimate β as:

$$\tilde{\beta}_{GMM} = \underset{\beta}{\operatorname{argmin}}\, g_N(x;\beta)' W_N g_N(x;\beta), \tag{24}$$

where $g_N(x;\beta) = \frac{1}{N}\sum_{n=1}^{N} g(x_n;\beta)$ is the $(H+1)$ x 1 vector of sample moment conditions, and $W_N = \frac{1}{N}\sum_{n=1}^{N} g\big(x_n;\tilde{\beta}\big)\, g\big(x_n;\tilde{\beta}\big)'$ is an estimate of the asymptotic covariance matrix of $\Big(\sqrt{N} g_N(x;\beta)\Big)$ based on $\tilde{\beta}$, a consistent estimator of β. For instance, one might rely on our estimator $\tilde{\beta}_2$ to construct W_N. The asymptotic covariance of $\tilde{\beta}_{GMM}$ can be estimated as:

$$V\big[\sqrt{N}\big(\tilde{\beta}_{GMM} - \beta\big)\big] \cong G_N\big(x;\tilde{\beta}_{GMM}\big)' \tilde{S}_N G_N\big(x;\tilde{\beta}_{GMM}\big), \tag{25}$$

where $G_N\big(x;\tilde{\beta}_{GMM}\big) = \frac{\partial g_N(x;\beta)}{\partial \beta'}\Big|_{\tilde{\beta}_{GMM}}$ and $\tilde{S}_N = \Big[\frac{1}{N}\sum_{n=1}^{N} g\big(x_n;\tilde{\beta}_{GMM}\big)\, g\big(x_n;\tilde{\beta}_{GMM}\big)'\Big]^{-1}$.

Observe that the GMM estimator $\tilde{\beta}_{GMM}$ is itself a suitable estimator of β. Alternatively, the moment conditions used to generate $\tilde{\beta}_{GMM}$ may be used to derive an appropriate large sample estimate of the covariance matrix of our constrained maximum likelihood estimator $\tilde{\beta}_1$. Specifically, one can employ Equation (25) after replacing $\tilde{\beta}_{GMM}$ with $\tilde{\beta}_1$. Alternatively, one can modify the moment conditions in Equations (22) and (23) as follows to include the Lagrange multiplier μ from the constrained optimization problem in Equation (8):

$$g_1(x;\beta,u) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\mu P'_\beta(x;\beta). \tag{26}$$

$$g_2(x;\beta,u) = (1-s)\big(q - P(x;\beta)\big). \tag{27}$$

One can then evaluate the standard GMM formula for the covariance matrix associated with these moment conditions at $\tilde{\delta} = \big(\tilde{\beta}_1, \tilde{\mu}_1\big)$, where $\tilde{\mu}_1$ is the solution for the Lagrange multiplier from our constrained optimization problem:

$$V\big[\sqrt{N}\big(\tilde{\delta} - \delta\big)\big] \cong G_N\big(x;\tilde{\delta}\big)' \tilde{S}_N G_N\big(x;\tilde{\delta}\big), \tag{28}$$

where $G_N(x;\tilde{\delta}) = \frac{\partial g_N(x;\delta)}{\partial \delta'}\Big|_{\tilde{\delta}}$ $\tilde{S}_N = \left[\frac{1}{N}\sum_{n=1}^{N} g(x_n;\tilde{\delta})\, g(x_n;\tilde{\delta})'\right]^{-1}$, and $g(x_n;\tilde{\delta}) = \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} g_1(x_n;\tilde{\delta}) \\ \frac{1}{N}\sum_{n=1}^{N} g_2(x_n;\tilde{\delta}) \end{bmatrix}$.

In the case of our unconstrained pseudo-maximum likelihood estimator $\tilde{\beta}_2$, one may construct the standard GMM formula for the covariance matrix associated only with the moment condition in Equation (22) and then evaluate it using $\tilde{\beta}_2$ as the estimator of β. It should be noted that a GMM estimator based the moment condition in Equation (22), alone, represents a suitable alternative to $\tilde{\beta}_2$ if one prefers to employ GMM estimation.

3. Monte Carlo Analysis for Case of Known q

We have undertaken some Monte Carlo simulations to compare the small sample performance of our calibrated qualitative response estimator, defined in Equation (8), and our unconstrained pseudo-maximum likelihood estimator, defined in Equation (15), with the small sample performance of a variety of alternative estimators from the existing literature on supplementary sampling. Below, we describe these alternatives, which include the Steinberg-Cardell (1992) estimator, an estimator we have derived based on Cosslett's (1981) generalized choice-based estimation framework as well as a simplified version of this estimator, and the Lancaster-Imbens (1996) estimator. Following our description of these alternative estimators, we present our Monte Carlo framework and then discuss our findings.

3.1 Alternative Estimators for Supplementary Samples

Steinberg-Cardell Estimator

The Steinberg-Cardell estimator is motivated by the estimator that one might use under the hypothetical scenario where the primary sample includes all participants in the population and the supplementary sample includes all participants and non-participants in the population. Under that scenario, one could effectively estimate a standard binary choice model even if the participants and non-participants in the supplementary sample could not be distinguished:

$$\max_{\beta} \sum_{i=1}^{N_P} s_i ln\, P(x_i;\beta) \; + ln\big(1 - P(x_i;\beta)\big) - s_i\, ln\big(1 - P(x_i;\beta)\big), \tag{29}$$

where N_P represents the population size. Under the standard binary choice model, the likelihood function accumulates the values of $lnP(x_i;\beta)$ across all participants and the values of $ln\big(1 - P(x_i;\beta)\big)$ across all non-participants. The former tally is achieved by the first term in Equation (29), while the latter is achieved by the combination of the second and third terms. Rearranging terms, the optimization problem in Equation (29) can equivalently be expressed as:

$$\max_{\beta} \sum_{i=1}^{N_P} s_i ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + ln\big(1-P(x_i;\beta)\big). \tag{30}$$

Now consider the case where a simple random sample of size N_1 is drawn from the overall subpopulation of participants and a simple random sample of size N_0 is drawn from the overall population of participants and non-participants. One can approximate the objective function in Equation (30) by scaling up the sample probabilities by the inverse of the sampling rates:

$$\max_{\beta} \sum_{i=1}^{N} s_i \left(\frac{N_P q}{N_1}\right) ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + (1-s_i)\left(\frac{N_P}{N_0}\right) ln\big(1-P(x_i;\beta)\big). \tag{31}$$

After multiplying each of the terms of the objective in Equation (31) by (N_0/N_P), one arrives at the Steinberg-Cardell estimator:

$$\tilde{\beta}_{SC} = \underset{\beta}{\text{argmax}} \sum_{i=1}^{N} \left[s_i \left(\frac{N_0 q}{N_1}\right) ln\left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)}\right) + (1-s_i)\ln(1-P(x_i;\beta))\right]. \tag{32}$$

Estimators based on the Cosslett Framework

In his seminal study of discrete choice estimation under choice-based sampling, Cosslett (1981) derives a generalized framework for asymptotically efficient estimation. Although he extends his framework to consider the case of supplementary sampling when the prevalence rate is unknown, he does not derive a corresponding supplementary sampling estimator for the situation where the prevalence rate is known. We employ Cosslett's estimation framework to derive an estimator for this situation below.

The first step is to consider the optimization problem for the case where the covariate distribution follows a specified functional form $[f(x)]$:

$$\max_{\beta} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) + lnf(x_i) \;\; s.t. \;\; q = \int P(x;\beta)\, f(x)\, dx\,. \tag{33}$$

Under Cosslett's approach, one replaces the covariate density $f(x)$ in Equation (33) with an empirical density characterized by a weight factor w_i:

$$\max_{\beta, w_1, w_2, \cdots, w_N} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) + ln(w_i) \;\; s.t. \;\; q = \sum_{i=1}^{N} P(x_i;\beta)\, w_i \text{ and } \sum_{i=1}^{N} w_i = 1. \tag{34}$$

The first-order condition for w_i implies:

$$\frac{1}{w_i} = \lambda_1 P(x_i;\beta) + \lambda_0,^{7} \tag{35}$$

where λ_1 and λ_0 are the Lagrange multipliers associated with the two constraints in Equation (34). Substitution of Equation (35) into Equation (34) yields the dual optimization problem:

$$\max_{\beta} \min_{\lambda_0, \lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\, n(\lambda_1 P(x_i;\beta) + \lambda_0) - N[1 - \lambda_1 q - \lambda_0]. \tag{36}$$

Observe that, whereas the original optimization problem involved *maximization* over the weights w_i, the dual optimization problem involves *minimization* over the Lagrange multipliers. The optimization problem in Equation (36) is equivalent to the following problem:

$$\max_{\beta} \min_{\lambda_0, \lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\, n(\lambda_1\, P(x_i;\beta) + \lambda_0)\ \ s.t.\ \ \lambda_1 q + \lambda_0 = 1. \tag{37}$$

Further simplification is possible by substituting the above constraint on the multipliers into the objective function:

$$\max_{\beta} \min_{\lambda_1} \sum_{i=1}^{N} s_i ln(P(x_i;\beta)) - l\, n(\lambda_1\, P(x_i;\beta) + 1 - \lambda_1 q). \tag{38}$$

A solution to this problem can be found by maximizing the objective function over β while holding the value of λ_1 fixed at alternative values. For instance, one can perform a grid search over a range of values for λ_1 surrounding its limit value of N_1/Nq, where the range is sufficiently wide that the maximum of the objective function (over β) takes a U-shape over the range. Alternatively, one can employ a more sophisticated search over alternative choices for λ_1 using methods such as that of Brent (1973). We refer to the resulting estimator of β as the "Cosslett" estimator in our Monte Carlo simulations.

A simpler feasible estimator of β can be obtained by substituting the limit values for λ_0 and λ_1

(N_0/N and N_1/Nq, respectively) into Equation (37):

$$\max_{\beta} \sum_{i=1}^{N} \left(s_i ln[P(x_i;\beta)] - ln\left[\frac{N_1}{Nq} P(x_i;\beta) + \frac{N_0}{N}\right]\right). \tag{39}$$

[7] Note that the weights w_i must be positive, which implies that $(\lambda_1 P(x_i;\beta) + \lambda_0)$ must also be positive.

We refer to this alternative estimator as the "Simplified Cosslett" estimator in our simulations. This simplified estimator was proposed by Lancaster and Imbens (1996, p. 153) as a feasible means to obtain an initial consistent estimate for use in solving the GMM estimation problem associated with their estimator.

Lancaster-Imbens Estimator

Lancaster and Imbens develop a GMM approach to the estimation of response probabilities using a supplementary sampling scheme. In their formulation of the problem, the primary and supplementary samples are drawn using a sequence of Bernoulli trials with unknown parameter h. Specifically, with probability h a "success occurs" and an observation is randomly drawn from the subpopulation of participants. With probability (1-h) a "failure" occurs and an observation is randomly drawn from the overall population of participants and non-participants.

They begin by considering the case where the covariate values are discrete with a finite number of points of support, characterized by the p.d.f. $f(x_i;\lambda)$. The likelihood function for this problem may be expressed as:

$$L = \sum_{i=1}^{N} (s_i ln\, P(x_i;\beta) + f(x_i;\lambda)) - N_1 lnh - N_0 \ln(1-h) - N_1 lnq. \tag{40}$$

Lancaster and Imbens then reparametrize this likelihood function in terms of the sampling distribution of the covariates:

$$g(x;\lambda) = [(h/q)P(x;\beta) + (1-h)]f(x;\lambda). \tag{41}$$

The reformulated likelihood function is specified in terms of β, q, h, and π:

$$L^R = \sum_{i=1}^{N} s_i ln\, R(x_i;\beta,q,h) + (1-s_i)\ln\big(1 - R(x_i;\beta,q,h)\big) + \ln g(x_i;\pi), \tag{42}$$

where $R(x;\beta,q,h) = \frac{(h/q)P(x;\beta)}{(h/q)P(x;\beta)+(1-h)}$ and the value of π at the k^{th} point of support $\left(x^k\right)$ is equal to $\pi_k = [(h/q)P(x^k;\beta) + (1-h)]\lambda_k$.

Whereas maximization of the original likelihood function is subject to the restriction $q = \int P(x;\beta)dF(x;\lambda)$, maximization of the reformulated likelihood function is subject to the restriction $h = \int R(x;\beta)dG(x;\pi)$.[8] Rather than pursue a constrained maximum likelihood estimation

[8] Ward et al. (2009) develop an expectation-maximization (EM) algorithm that solves for the constrained maximum likelihood solution under a logistic specification for the conditional response probability distribution.

strategy, Lancaster and Imbens derive their estimator $(\tilde{\beta}, \tilde{h})_{LI}$ by applying GMM estimation based on the following three moment conditions:

$$g_1(x;\beta,h) = \frac{P'_\beta(x;\beta)}{P(x;\beta)}\big(s - R(x;\beta,q,h)\big). \tag{43}$$

$$g_2(x;\beta,h) = -\frac{1}{q}\big(s - R(x;\beta,q,h)\big). \tag{44}$$

$$g_3(x;\beta,h) = h - R(x;\beta,q,h). \tag{45}$$

The third moment condition is the sample analogue of the restriction $h = \int R(x;\beta)dG(x;\pi)$, while the first two conditions represent the single observation scores of the likelihood function in Equation (42) for β and h, respectively. Observe that these three moment conditions do not require knowledge of the points of support for x, and they remain valid even when x is continuous.

3.2 Monte Carlo Framework and Results

In our simulations, we employ a logit specification for the conditional probability of participation involving two independent standard normal regressors and an intercept. The coefficients of the two regressors are fixed at one, while the intercept is varied to achieve alternative approximate values of the prevalence rate q, including 0.125, 0.25, 0.50, 0.75, and 0.875. We perform 1,000 replications for each experiment.

Two alternative sampling designs are employed. The first is a logit sampling design consisting of N_0 overall observations, including $N_1 = N_0 q$ participants and $N_0(1-q)$ non-participants. Under this design, the "supplementary sample" is the combined sample of N_0 participants and non-participants, while the "primary sample" is the subsample of N_1 participants. This sampling design permits us to compare the relative performance of a given estimator to that of the standard logit estimator, and to explore how this relative performance varies with the value of q.

The second sampling design retains the N_1 participants from the first sampling design as the primary sample. In this case, however, a supplementary sample of size N_0 is drawn independently of the primary sample. A comparison of the results from the first and second sampling designs reveals the degree to which the relative performance of a given estimator is impacted by the reliance on an independent supplementary sample. For most of our simulations, we rely on a supplementary sample of size $N_0 = 400$. However, we also consider one case with a larger supplementary sample size of $N_0 = 1{,}600$.

The results of the Monte Carlo simulations are summarized in Table 1. For each case, we report the mean and median estimates, the mean asymptotic standard deviation of the estimates (ASD), the standard deviation of the estimates (SSD) over the 1,000 replications, and the mean absolute deviation from the

median estimates (MAD) over the 1,000 replications. In some of the simulations, certain estimators are subject to convergence problems. For such estimators, we perform our tabulations based on the subset of replications that are free from convergence problems. The number of replications where an estimator has failed to converge is reported as "#Failures".

Consider first the results for the logit sampling design. Under this design, the Steinberg-Cardell estimator is identical to the standard logit estimator.[9] For the first two cases with $q = 0.125$ and $q = 0.25$, respectively, all of the estimators perform quite similarly to the Steinberg-Cardell estimator, and hence, to the standard logit estimator. However, as q is increased beyond this point, the standard errors of these other estimators become increasingly large relative to the standard errors of the Steinberg-Cardell/standard logit estimator. Intuitively, the overall sample is made up predominantly of participants when the prevalence rate is high. When relatively few members of the sample are non-participants, specific information on identities of those non-participants becomes more valuable. As q increases, the performances of the alternative estimators also become less similar. After the Steinberg-Cardell/standard logit estimator, the calibrated logit estimates show the lowest standard errors, followed by the Lancaster-Imbens and Cosslett estimates.

All of the estimators exhibit larger standard errors under the independent primary and supplementary sampling design than under the logit sampling design. Intuitively, the lack of any overlap between the primary and supplementary samples results in greater variability in the covariates across the two samples, which leads to noisier estimates. When q is relatively low, the estimators all perform similarly under the independent primary and supplementary sampling design, with the exception of the Steinberg-Cardell estimator. Even when the prevalence rate is small, this estimator is relatively inefficient in comparison with the other estimators.

As the prevalence rate rises, the choice of estimators becomes increasingly more important under this sampling design. Unlike the other estimators under consideration, the calibrated logit, Lancaster-Imbens, and Cosslett estimators impose certain consistency requirements. The calibrated logit estimator ensures that the average predicted probability of participation in the supplementary sample is consistent with the overall prevalence rate. The Lancaster-Imbens estimator ensures that the average predicted probability that a member of the combined sample came from the primary sample of participants is consistent with the actual share of observations from the primary sample. The Cosslett estimator imposes constraints on the multipliers that correspond to a condition in the dual problem that the weighted probability of participation in the combined sample is consistent with the overall prevalence rate. When q is relatively high ($q = 0.75$ or $q = 0.875$), these three estimators substantially outperform the unconstrained pseudo-maximum likelihood estimator, the simplified Cosslett estimator, and the Steinberg-Cardell estimator,

[9] Note that, under this sampling design, the optimization problems for the estimators are somewhat different, because participants are present in both the primary and supplementary samples. So, for instance, the formula for the Steinberg-Cardell estimator in Equation (32) is modified to: $\tilde{\beta}_{SC} = \text{argmax}_{\beta} \sum_{i=1}^{N} \left[s_i \left(\frac{N_0 q}{N_1} \right) ln \left(\frac{P(x_i;\beta)}{1-P(x_i;\beta)} \right) + \ln(1 - P(x_i;\beta)) \right]$. After rearranging terms and taking into account that N_1 has been chosen to be equal to $N_0 q$ under this design, this simplifies to: $\tilde{\beta}_{SC} = \text{argmax}_{\beta} \sum_{i=1}^{N} [s_i \ln(P(x_i;\beta)) + (1 - s_i)\ln(1 - P(x_i;\beta))]$, which is recognized as the standard logit estimator.

which do not impose any consistency requirements. Not only do these latter estimators have relatively high standard errors, they also are subject to periodic convergence problems.

In Case 6 of Table 1, we explore the performance of our estimators when the prevalence rate is high ($q = 0.875$), but a larger estimation sample is employed. In particular, we quadruple the sample size (from $N_0 = 400$ and $N_1 = 350$ to $N_0 = 1{,}600$ and $N_1 = 1{,}400$). The application of a larger estimation sample largely eliminates the convergence problems associated with the unconstrained pseudo-maximum likelihood estimator, the simplified Cosslett estimator, and the Steinberg-Cardell estimator. As well, the precision of all of the estimators is substantially improved. The standard errors of our unconstrained pseudo-maximum likelihood estimator and the simplified Cosslett estimator of the slope coefficients are now reasonably similar those of the calibrated logit, Cosslett, and Lancaster-Imbens estimators. However, the standard errors of the intercept estimates remain much larger, indicating that the consistency restrictions imposed by the latter estimators remains valuable for pinning down the intercept of the conditional logit probability of participation even in larger samples.

We have also performed some Monte Carlo simulations for our alternative GMM-based estimators. Our GMM estimator described in Equation (24) produces very similar results to our calibrated logit estimator, even in small samples. Likewise, our GMM estimator based solely on the moment condition in Equation (22) performs very comparably to our unconstrained pseudo-maximum likelihood estimator.

4 Unknown Prevalence Rate

So far, we have assumed that the prevalence rate q is known. If q was unknown but the cumulative distribution function $F(x)$ was known, one could maximize the following log-likelihood function over β and q:

$$L = \left(\sum_{l=1}^{N_1} ln\big(P(x_i;\beta)\big) \right) - N_1 ln(q) \tag{46}$$

subject to the constraint:

$$q = \int P(x;\beta) dF(x). \tag{47}$$

In practice, however, $F(x)$ is not generally known. Following our earlier approach, consider replacing the actual formula for q in Equation (47) with its analog based on a supplementary random sample of size N_0:

$$\tilde{q} = \frac{\sum_{j=1}^{N_0} P\big(x_j;\beta\big)}{N_0}. \tag{48}$$

This leads to the (concentrated) pseudo-log-likelihood function:

$$L = \left(\sum_{l=1}^{N_1} ln\big(P(x_i;\beta)\big) \right) - N_1 ln\left(\frac{\sum_{j=1}^{N_0} P\big(x_j;\beta\big)}{N_0} \right). \quad (49)$$

The parameter estimates for this specification are found as the solution to the following first-order conditions: [10]

$$\left(\sum_{i=1}^{N_1} \frac{P'_\beta(x_i;\beta)}{P(x_i;\beta)} \right) - \frac{N_1}{N_0 \tilde{q}} \left(\sum_{j=1}^{N_0} P'_\beta\big(x_j;\beta\big) \right) = 0. \quad (50)$$

The usual estimated standard errors of the coefficient estimates computed by a maximum likelihood algorithm will tend to be somewhat too small in this case, owing to the reliance on a sample analog of the population relationship between q and β. As with the known q case, asymptotically valid standard error estimates can be obtained through a GMM approach, where the moment conditions specified previously in Equations (22) and (23) now involve the unknown value of q as well as the unknown value of β:

$$g_1(x;\beta,q) = s\frac{P'_\beta(x;\beta)}{P(x;\beta)} - (1-s)\frac{N_1}{N_0 q} P'_\beta(x;\beta). \quad (51)$$

$$g_2(x;\beta,q) = (1-s)\big(q - P(x;\beta)\big). \quad (52)$$

One can either apply GMM estimation directly to Equations (51) and (52) to estimate parameters β and q as well as their standard errors, or one can estimate β by maximizing the pseudo-log-likelihood function defined in Equation (49) and then substitute the estimated values of β and $\tilde{q}$ into the GMM covariance matrix formula to estimate the standard errors.

5 Monte Carlo Analysis for Case of Unknown q

Alternative estimators for the case of an unknown prevalence rate have been proposed by Cosslett (1981) and Lancaster and Imbens (1996). We show below that these two estimators are actually equivalent. Next we compare the small-sample performance of our pseudo-maximum likelihood estimator based on Equation (49) against that of the Cosslett-Lancaster-Imbens estimator.

5.1 Cosslett-Lancaster-Imbens Estimator

[10] See Lele and Keim (2006) for a related simulation-based approach to estimation in this case .

Cosslett (1981) has derived an alternative supplementary sampling estimator for the case where q is unknown based on maximization of the following pseudo-likelihood function:

$$L = \sum_{i=1}^{N} s_i ln(\lambda P(x_i;\beta)) - ln\left(\lambda P(x_i;\beta) + \frac{N_0}{N}\right). \quad (53)$$

The above expression is maximized jointly over β and λ. If desired, an estimate of the prevalence rate can be obtained from the estimated value of λ by applying the normalization condition:

$$\left(\lambda q + \frac{N_0}{N}\right) = 1.^{11} \quad (54)$$

Alternatively, one can employ Equation (54) to re-specify the optimization problem directly in terms of β and q:

$$\max_{\beta} \max_{q\in(0,1)} \sum_{i=1}^{N} \left[s_i ln\left(\frac{N_1}{Nq} P(x_i;\beta)\right) - ln\left(\frac{N_1}{Nq} P(x_i;\beta) + \frac{N_0}{N}\right)\right]. \quad (55)$$

This is, in fact, the same as the optimization problem that Lancaster and Imbens (1996) have derived for the case where the prevalence rate is unknown.[12]

5.2 Monte-Carlo Simulations

We have undertaken some Monte Carlo simulations to compare the small sample performance of our pseudo-maximum likelihood estimator based on Equation (49) and the Cosslett-Lancaster-Imbens estimator based on Equation (55) for the case where the prevalence rate is unknown. As with our simulations for the known prevalence rate case, we employ a logit specification for the conditional probability of participation with two independent standard normal regressors and an intercept. The coefficients of the two regressors are fixed at one, while the intercept is varied to achieve alternative approximate values of the prevalence rate q, including 0.125, 0.25, 0.50, 0.75, and 0.875. We perform 1,000 replications for each experiment. For these simulations, we focus on the case where the primary and supplementary samples are independently drawn. For most of our simulations, we rely on a supplementary sample of size $N_0 = 400$. However, we also explore one case with a larger supplementary

[11] See pp. 71-73 of Cosslett (1981) for a discussion of how to estimate prevalence rates by applying scale factors based on the relevant normalization condition for a problem. In the supplementary sampling case, the normalization condition implies the relationship described by Equation (54). Although Cosslett imposed the restriction $\lambda > 0$ for the maximization of the likelihood function in Equation (53), one would also need to impose the restriction $\lambda < \frac{N_1}{Nq}$ to insure that the estimated prevalence rate is less than one.

[12] Lele (2009) has introduced a data-cloning algorithm as an alternative to standard maximum likelihood estimation routines for this problem.

sample size of $N_0 = 1{,}600$. As with our simulations for the known q case, we set the primary sample size (N_1) equal to $N_0 q$.

The results of the Monte Carlo simulations are summarized in Table 2. For each case, we report the mean and median estimates, the standard deviation of the estimates (SSD), and the mean absolute deviation from the median estimates (MAD) over the 1,000 replications. We also present the mean asymptotic standard deviation of the estimates based on the pseudo-likelihood function (LSD). In the case of the Cosslett-Lancaster-Imbens estimator, we derive the standard error estimates using the inverse of the information matrix. Lancaster and Imbens (1996) have shown that these standard error estimates are consistent for the coefficients (but not for q). For our pseudo-maximum likelihood model, we rely on the Huber-White standard errors for our LSD estimates. The LSD estimate of the standard error for our pseudo-maximum likelihood estimate of q is computed using the delta method. In large samples, these estimates will tend to be somewhat too small, because they do not account for our reliance on a sample analogue of the true relationship between marginal q and conditional $P(x;\beta)$. We compare our LSD estimates to the GMM-based standard error estimates (GSD).

In small samples, both estimators are subject to periodic convergence problems. We base our performance measures for a given estimator on the subset of replications that are free from such problems. The number of replications where an estimator has failed to converge is reported as "#Failures".

Comparing findings for the cases where the prevalence rate is and is not known, it is clear that precision suffers when q is unknown. The discrepancy in performance across these two cases is especially pronounced when q is relatively large ($q = .75$ and $q = .875$). In addition, the discrepancy is much larger for the intercept than for the slope coefficients.

Overall, our pseudo-maximum likelihood estimator performs quite comparably to the Cosslett-Lancaster-Imbens estimator in terms of mean and median performance as well as precision. Lancaster and Imbens (1996) have reported that their estimator has periodic convergence issues in small samples, particularly when the true value of q is close to zero. This problem extends to our estimator. As noted by Lancaster and Imbens, when q is close to zero, supplementary sampling is close to pure choice-based sampling, and the choice-based sampling estimator of the intercept in a logit model is not identified when q is unknown. We find that the estimated covariance matrices for both supplementary sampling estimators tend to become ill-conditioned at solutions involving estimated values of q close to zero, and the standard error of the intercept estimate becomes very large in such cases.

Our simulation results indicate that convergence problems are also prevalent when the true value of q is relatively high ($q = 0.75$ and $q = .875$). One source of such problems is that, despite the high actual prevalence rate, there are a significant number of replications where the estimated prevalence rate is actually close to zero. Another source of convergence problems when q is relatively high involves estimates of the prevalence rate that are very close to the upper bound of one. Typically in such cases, the average predicted conditional probability of participation approaches one within the primary sample, while the average predicted probability is just slightly smaller within the supplementary sample.

In Case 6 of Table 2, we explore the performance of the estimators when the prevalence rate is high ($q = 0.875$), but a larger estimation sample is employed. This not only leads to substantial improvements in precision, it also greatly reduces the incidence of convergence problems. In general, then, when the prevalence rate is not known, it is especially beneficial to employ a reasonably large overall sample in estimation.

We have also performed Monte Carlo simulations using our GMM estimator based on the moment conditions in Equation (51) and (52). The results indicate that this estimator and our pseudo-maximum likelihood estimator for the case of an unknown prevalence rate produce very similar estimates, even in small samples.

6 Stratified Samples

Each of the above estimators can be generalized to accommodate exogenously stratified primary and/or supplementary samples. Let the sample weights for the primary data source be represented by w_1 and those for the supplementary data source by w_0. We assume that these weights have been normalized so that they sum to the size of their respective samples, N_1 and N_0. If either of the samples is not stratified, the weight for each observation in that sample would be set equal to one.

The Cosslett estimator in Equation (38) for the case where the prevalence rate is known can be generalized to accommodate an exogenously stratified sampling design as follows:

$$\max_{\beta} \min_{\lambda_{11},\lambda_{12},\ldots,\lambda_{1B}} \sum_{b=1}^{B}\sum_{i=1}^{N_b} s_i ln(P(x_{ib};\beta)) - l\,n(\lambda_{1b}P(x_{ib};\beta) + 1 - \lambda_{1b}q_b), \tag{56}$$

where B represents the number of strata, N_b represents the sample size of stratum b, λ_{1b} is the stratum-specific multiplier, and q_b is the stratum-specific prevalence rate. [13]

Similarly, the simplified Cosslett estimator in Equation (39) for the case where the prevalence rate is known can be generalized to accommodate an exogenously stratified sampling design as follows:

$$\max_{\beta} \sum_{b=1}^{B}\sum_{i=1}^{N_b} \left[s_i ln(P(x_{ib};\beta)) - ln\left(\frac{N_{1b}}{N_b q_b}P(x_{ib};\beta) + \frac{N_{0b}}{N_b}\right)\right], \tag{57}$$

where N_{0b} is the number of observations in stratum b of the supplementary sample, N_{1b} is the corresponding number in the primary sample. Alternatively, the generalized estimator can be expressed in terms of the sample weights as:

[13] The stratum-specific prevalence rate q_b can be computed from the overall prevalence rate and the sample weights using the relationship: $q_b = \left(\frac{w_{1b}}{w_{0b}}\right)\left(\frac{N_{1b}/N_1}{N_{0b}/N_0}\right)q$.

$$L = \max_{\beta} \sum_{b=1}^{B} \sum_{i=1}^{N_b} \left[s_i ln(P(x_{ib};\beta)) - ln\left(\frac{1}{w_{1b}} \left(\frac{N_1}{Nq} \right) P(x_{ib};\beta) + \frac{1}{w_{0b}} \left(\frac{N_0}{N} \right) \right) \right]. \quad (58)$$

The Lancaster-Imbens estimator for the case of a known prevalence rate relies on the moment conditions specified in Equations (43) through (45). To accommodate exogenous stratification, these moment conditions can be generalized by replacing the overall sampling probability parameter h and prevalence rate q with a set of stratum-specific parameters, h_b and q_b, $b = 1, \cdots, B$.

When the prevalence rate is unknown, the generalized Cosslett-Lancaster-Imbens pseudo-likelihood function for stratified samples can be expressed as:

$$L = \sum_{b=1}^{B} \sum_{i=1}^{N_b} \left[s_{ib} ln\left(\frac{1}{w_{1b}} \left(\frac{N_1}{Nq} \right) P(x_{ib};\beta) \right) - ln\left(\frac{1}{w_{1b}} \left(\frac{N_1}{Nq} \right) P(x_{ib};\beta) + \frac{1}{w_{0b}} \left(\frac{N_0}{N} \right) \right) \right]. \quad (59)$$

Observe that implementation of these generalized estimators of Cosslett and Lancaster-Imbens requires one to be able to assign observations from both the primary and supplementary samples to the relevant sampling strata.[14] However, the requisite information is not always available. For instance, the U.S. Census Bureau does not publicly disclose its stratification criteria for national surveys such as the CPS, SIPP, and ACS.[15] An advantage of the generalized estimators discussed below is that they can be implemented even when the stratification criteria are unknown. All that is required is the sample weights, which are routinely available.

Steinberg and Cardell (1992) have proposed an extension of their estimation framework for the case of a known prevalence rate to accommodate exogenously stratified primary and/or supplementary samples. The generalized Steinberg-Cardell estimator is obtained from the following optimization problem:

[14] One needs to be able to assign the observations in each of the samples to the relevant strata in order to obtain the stratum-level counts used in Equation (57), to determine the sample weight associated with the stratum in the opposing sample in Equations (58) and (59) (i.e., one needs to know the value of w_{0b} as well as w_{1b} when an observation is in stratum b of the primary sample, and one needs to know the value of w_{1b} as well as w_{0b} when the observation is in stratum b of the supplementary sample), and to assign the correct stratum-specific parameters (h_b and q_b) in the moment conditions for a given observation under the generalized Lancaster-Imbens approach for the case of a known prevalence rate.

[15] Under a fairly simple stratified random sampling design, it may be possible to deduce the stratification criteria (at least approximately) by analyzing the characteristics of each subsample of observations with a common value for the sample weight. However, such an approach is not feasible for more complex survey designs. For instance, Census surveys often involve multi-stage sampling, clustering, post-stratification adjustment, and imputation. As a consequence, the final sample weight often varies among observations within the same initial stratum. Even when the sampling criteria for the supplementary sample can be deduced, it is only feasible to evaluate the relative sampling weights if the stratifying variables are also present in the primary sample. In cases where both the primary and supplementary data sources are stratified, one would further need to divide the existing strata for the two data samples into sub-strata that are comparable across the two samples. In such cases, the presence of sparse or empty sub-strata (i.e., cases where one of the samples contains few or no observations within a sub-stratum) would complicate estimation.

$$\max_{\beta} \sum_{i=1}^{N} \left[w_{1i} \left(\frac{N_0 q}{N_1} \right) s_i ln \left(\frac{P(x_i;\beta)}{1 - P(x_i;\beta)} \right) + w_{0i}(1 - s_i) ln\big(1 - P(x_i;\beta)\big) \right]. \tag{60}$$

However, as noted in the previous section, the Steinberg-Cardell estimator is relatively inefficient. Below, we introduce generalized versions of our more efficient calibrated qualitative response estimators.

Our generalized constrained pseudo-maximum likelihood estimator for the case of a known prevalence rate ($\tilde{\beta}_{1W}$) is constructed by incorporating the relevant sample weights into the objective function and constraint of the optimization problem described by Equation (8):

$$\tilde{\beta}_{1W} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} w_{1i} ln\big(P(x_i;\beta)\big) \;\; s.t. \;\; q = \frac{\sum_{j=1}^{N_0} w_{0j} P\big(x_j;\beta\big)}{N_0}. \tag{61}$$

Similarly, our generalized unconstrained pseudo-maximum likelihood estimator ($\tilde{\beta}_{2W}$) for this case is constructed by incorporating the sample weights into the relevant terms of Equation (15):

$$\tilde{\beta}_{2W} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N} \left[w_{1i} s_i ln\big(P(x_i;\beta)\big) - w_{0i} \left(\frac{N_1}{N_0 q} \right) (1 - s_i) P(x_i;\beta) \right]. \tag{62}$$

When the prevalence rate is unknown, the pseudo-log-likelihood function in Equation (49) is easily generalized to account for stratified sampling as follows:

$$L = \left(\sum_{l=1}^{N_1} w_{1i} ln\big(P(x_i;\beta)\big) \right) - N_1 ln \left(\frac{\sum_{j=1}^{N_0} w_{0i} P\big(x_j;\beta\big)}{N_0} \right).^{16} \tag{63}$$

Estimates of the standard errors of these generalized estimators can be obtained by appropriately weighting the moment conditions in the GMM formulae for the covariance matrices.

7 Generalization to Polychotomous Response Models

Our estimation approach readily generalizes to account for more than two outcomes. For instance, suppose there are M+1 possible outcomes indexed by the values $y = 0, 1, \cdots, M$. Define the outcome probabilities as:

[16] If the available sample weights for the primary and supplementary samples sum to their respective population totals, then the prevalence rate actually will be known since it can be computed as the ratio of the sum of the sample weights for the primary sample to the sum of the sample weights for the supplementary sample. However, if only normalized sample weights are available (which instead sum to the respective sample sizes), it will not be possible to deduce the prevalence rate from such weights.

$$\Pr(y = m|x;\beta) = P(m|x;\beta),\ \ m = 0, 1, \dots, M, \tag{64}$$

where $P(m|x;\beta)$ has a known parametric form. This framework is sufficiently general to include both ordinal and multinomial response models. Let the outcome $y = 0$ represent non-participation and let the remaining M outcomes represent alternative forms of participation. Suppose one has a random participant-only sample of size N_1 that includes observations with outcomes 1 through M. Sampling among these participants may be choice-based, meaning that the sampled number of observations (N_{1m}) for a given participation outcome m may not be representative of the incidence of this outcome within the participant population. In addition, suppose one has a supplementary random sample of size N_0 from the general population that includes observations on all types of participants as well as non-participants.

Define q_m as the prevalence rate for outcome $m,\ m = 1, \cdots, M$. Assuming these prevalence rates are known, our calibrated qualitative response estimator for the binary response case described in Equation (8) may be adapted to account for polychotomous responses as follows:

$$\tilde{\beta}_{1P} = \underset{\beta}{\operatorname{argmax}} \sum_{i=1}^{N_1} y_i ln\big(P(y_i|x_i;\beta)\big)\ \ s.t.\ q_m = \frac{\sum_{j=1}^{N_0} P\big(m|x_j;\beta\big)}{N_0},\ m = 1 \cdots, M. \tag{65}$$

Thus, the generalized form of our calibrated qualitative response estimator involves M constraints, one for each outcome in the primary sample. To estimate the covariance matrix of $\tilde{\beta}_{1P}$, one can rely on the GMM covariance matrix formula associated with the following moment conditions:[17]

$$g_0(x;\beta) = y\frac{P'_\beta(y|x;\beta)}{P(y|x;\beta)} - (1-s)\sum_{m=1}^{M}\frac{N_{1m}}{N_0 q_m}P'_\beta(m|x;\beta). \tag{66}$$

$$g_m(x;\beta) = (1-s)(q_m - P(m|x;\beta)),\ \ m = 1, \dots, M. \tag{67}$$

Alternatively, one can derive an asymptotically equivalent estimator of β by applying GMM estimation to these moment conditions.

The extension of our unconstrained pseudo-maximum likelihood estimator defined in Equation (15) to the polychotomous outcome case is similarly straightforward:

$$\tilde{\beta}_{2P} = \underset{\beta}{\operatorname{argmax}}\left[\sum_{i=1}^{N_1} y_i ln\big(P(y_i|x_i;\beta)\big) - \sum_{m=1}^{M}\left(\frac{N_{1m}}{N_0 q_m}\right)\sum_{j=1}^{N_0} P\big(m|x_j;\beta\big)\right]. \tag{68}$$

[17] As discussed in Section 2.4, an alternative asymptotically valid approach for estimating the covariance matrix is to incorporate the Lagrange multipliers into the moment conditions; Under this approach, the multipliers would replace the term $\frac{N_{1m}}{N_0 q_m}$ in Equation (66).

The moment conditions defined in Equation (66) can be used either to estimate the covariance matrix of $\tilde{\beta}_{2P}$ or to develop an asymptotically equivalent GMM estimator for β.

If the prevalence rates are unknown, the pseudo-log-likelihood function defined in Equation (49) may be generalized to:

$$L = \left(\sum_{i=1}^{N_1} y_i ln\big(P(y_i|x_i;\beta)\big)\right) - \left(\sum_{m=1}^{M} N_{1m} ln\left(\frac{\sum_{j=1}^{N_0} P\big(m|x_j;\beta\big)}{N_0}\right)\right). \quad (69)$$

The parameters q_m can then be estimated using the analogue estimator $\tilde{q}_m = \sum_{j=1}^{N_0} P\big(m|x_j;\tilde{\beta}\big)/N_0$. For estimation of the covariance matrix, one can rely on the GMM covariance matrix formula based on the moment conditions in Equations (66) and (67), where these conditions are now taken as a function of the unknown parameters q_m as well as β. Alternatively, one can derive asymptotically equivalent estimators of β and q by applying GMM estimation to these moment conditions.

To extend the above estimators to account for stratified random sampling on exogenous factors, one simply applies the appropriate primary and supplementary sample weights to the terms in equations (65) through (69).

8 An Example

Burden et al. (2014) estimate the determinants of voting behavior using data from the Current Population Survey (CPS) for 2004 and 2008 using both binary and multinomial logit specifications. In this section, we focus on their analysis for 2008. We begin by estimating similar specifications to those used in their study based on the same 2008 CPS data sample. We then compare the results against our calibrated binary and multinomial logit model estimates based on a voter-only subsample from the CPS and a supplementary sample from the overall voting-eligible population from the American Community Survey (ACS). We also apply our pseudo-MLE estimator for the case where the prevalence rate is unknown to investigate its performance.[18]

The binary logit specification employed by Burden et al. distinguished voters and non-voters. The multinomial logit specification distinguished among the following modes of voting: (1) election-day voting; (2) early voting in person; and (3) early voting by mail. Both specifications relied on the following explanatory variables:

Early: Dummy for residence in a state that permits early voting.
EDR: Dummy for residence in a state that permits one to both register and vote on Election Day.
Early*SDR: Dummy for residence in an early voting state that permits same-day registration.
Early*EDR: Interaction between Early Voting and EDR.

[18]The authors have kindly posted the Stata code they used in their analysis at https://electionadmin.wisc.edu/BCMM_AJPS_CPSanalysis.zip. This code greatly facilitated the replication of their original results.

Early*EDR*SDR: Interaction between Early Voting, SDR, and EDR.
30-Day Reg. Close: Dummy for residence in a state that requires voters to be registered 30 days in advance of an election.
ID Requirement: Dummy for residence in a state that requires voters to show some form of identification.
Education: Indicator for educational attainment (4 values ranging from less than high school diploma to Bachelor's degree or higher).
African American: Dummy for self-identified race of Black only or Black in combination with another race.
Hispanic: Dummy for self-identified race of Hispanic origin.
Naturalized Citizen: Dummy for naturalized citizenship.
Married: Dummy for married.
Female: Dummy for female.
Age: Age in years.
Age 18–24: Dummy for age between 18 and 24.
Age 75 plus: Dummy for age 75 or over.
Competitiveness: A poll-based index of campaign competitiveness (a higher value indicates a more competitive campaign).
South: Dummy for residence in a southern state.
North Dakota: Dummy for residence in North Dakota (which does not require voter registration).
Oregon: Dummy for residence in Oregon (a "vote-by-mail" state).
Washington: Dummy for residence in Washington state (a "vote-by-mail" state).
Self-Reported Vote: Dummy equal to one if voting status was self-reported and zero if reported by another family member.
Natural. 10+ Years: Dummy for naturalized citizen who entered the U.S. prior to 1998.
Residence 1 Year: Dummy for tenure of at least one year at current residence.
Income: Indicator for total family income (16 values ranging from less than $5,000 to $150,000 and over).

The estimation sample was restricted to individuals who appeared eligible to vote (age 18 or over and a U.S. citizen) and who did not reside in the District of Columbia.

In order to apply the calibrated qualitative response methodology, it is essential to have comparably defined and measured variables in the primary and supplementary data sources. This example demonstrates that this requirement can place limitations on the set of explanatory variables that one can use in an analysis. In particular, the last four variables listed above do not satisfy this requirement.

Although a comparable family income concept can be constructed from the ACS data, it turns out that the CPS family income measure is missing for approximately 20 percent of the voting-eligible sample. Based on a comparison of the ACS (which has complete income information) and the CPS, it appears that a disproportionate share of the missing responses in the CPS is attributable to lower income households. Burden et al. restrict their analysis to the portion of their CPS sample with non-missing income information. This restriction might be justified if it can be assumed that willingness to provide income

information on the CPS survey is uncorrelated with voting behavior. However, the validity of this assumption is uncertain. Note that even if this assumption were valid, it would not be feasible to include the income measure as a regressor in the calibrated qualitative response model. Since the income measure in the CPS is not missing at random, but rather systematically with the level of income, the (weighted) subsample with non-missing information is not representative of the overall population and therefore cannot be validly compared against the (weighted) ACS sample.

Similarly, information regarding tenure at the current address is missing for approximately 13 percent of the CPS voter-eligible sample. The authors set the Residence 1 Year dummy equal to one when this information was missing, which resulted in an unknown number of instances of misclassified residential tenure status. Such an approach introduces bias into the binary and multivariate logit findings. Moreover, it invalidates the comparison against ACS data employed under our calibrated qualitative response approach.

The Naturalized 10+ Years dummy is based on information concerning the date of entry to the U.S. The ACS inquires about the date of naturalization but not the date of entry (which typically occurs many years earlier). So, a comparable dummy variable cannot be constructed using the ACS. Similarly, it is not feasible to construct a Self-Reported Vote indicator using the ACS sample.[19]

For purposes of illustration and comparison of methodologies, we have therefore estimated specifications that exclude these four variables from the analysis. Tables 3 and 4, respectively, compare the standard binary and multinomial logit estimates based on the CPS to the corresponding estimates of our alternative models based on a supplementary sampling scheme that includes the subsample of voters in the CPS as our primary sample and a 10 percent random subsample of voting-eligible individuals in the ACS as our supplementary sample. Both the CPS and the ACS rely on stratified sampling designs, so we incorporate the publicly available sample weights from both surveys in our analysis as discussed in Section 7.[20] For our calibrated binary logit model, we have relied on the weighted mean value of the voting indicator in the CPS sample, inclusive of those observations with missing income information, as our measure of the prevalence rate. Similarly, for the calibrated multinomial logit model, we have relied on the weighted mean values of reported shares of individuals voting on election day in person, voting early in person, and voting early by mail (inclusive of observations with missing income information) as our measures of the prevalence rates for these three different voting methods.

Overall, our calibrated qualitative response estimation methodology produces qualitatively quite similar results to the standard binary and multinomial logit approaches. Differences in the relative magnitudes of certain coefficients across methods are largely attributable to moderate differences in the weighted mean

[19] The number of individuals who report voting on surveys tends to be higher than the actual number of voters based on election statistics. The Self-Reported Vote variable was included by the authors to control for the possibility that this tendency to misreport voting status is more pronounced when an individual is asked directly about his or her voting behavior than when voting status is reported by proxy.

[20] In the case of the standard binary and multinomial logit specifications based on the CPS data sample, we have followed the authors in performing an unweighted analysis, followed by the computation of cluster-robust standard errors by state.

values of the underlying regressors (such as the dummies for marital status, age range, and residence in certain states with different voting requirements) across the two data sources.

Our pseudo-maximum likelihood estimates based on unknown prevalence rates are very similar to our calibrated binary and multinomial logit results based on specified values for the prevalence rates. In addition, the pseudo-maximum likelihood estimates of the prevalence rates are reasonably close to measures computed using the weighted CPS statistics. Overall, our combined estimation sample is quite large (273,933). Although we do lose some precision when we do not specify a prevalence rate in estimation, the large overall size of the combined sample (273,933) ensures that we are still able to obtain reasonably precise estimates of the conditional response probability parameters.

9 Summary and Conclusion

Frequently, researchers have access to detailed information on the relevant characteristics of participants in a program, patients suffering from a disease, or habitats where a species is known to be present. However, their lack of comparable information about households that do not participate in the program, individuals who are free of the disease, or habitats where the species is not present precludes the application of standard qualitative response models to analyze the determinants of the outcome under investigation.

If the joint probability distribution of the underlying covariates were known, we have demonstrated how a constrained maximum likelihood procedure could be used to estimate the parameters of the conditional response probability distribution based solely on an available sample of participants. This approach exploits the parameter restrictions implied by the relationship between the marginal and conditional probabilities of participation: $q = \int P(x;\beta)\, dF(x)$, where q is the marginal probability of participation (i.e., the prevalence rate), $P(x;\beta)$ is the conditional probability of participation, and $F(x)$ is the joint distribution function of the covariates. In practice, however, this approach is not generally feasible to implement, because $F(x)$ is unknown.

To overcome this problem , we have shown that one can replace the unknown relationship between the marginal and conditional response probability distributions with its analogue based on a supplementary sample of size N_0 from the general population: $\tilde{q} = \frac{1}{N_0}\sum_{i=1}^{N_0} P(x_i;\beta)$. Using this analogue relationship, we have derived some feasible new constrained and unconstrained pseudo-maximum likelihood estimators of the parameters of the conditional response probability distribution. Following Lancaster and Imbens (1996), we show how our optimization problem can be recast under a GMM framework. This leads to some additional new estimators as well as a straightforward way to obtain appropriate standard errors for our pseudo-maximum likelihood estimators. We also demonstrate that our framework is readily generalized to accommodate polychotomous responses.

We have conducted some Monte Carlo simulations to compare the small sample performance of our new estimators against that of existing estimators, including those proposed by Cosslett (1981) [including some estimators for the known prevalence rate case that we have derived based on his generalized choice-based estimation framework], Lancaster and Imbens (1996), and Steinberg and Cardell (1992). Our

Monte Carlo simulations reveal several insights. When the prevalence rate is known, our calibrated qualitative response estimator rivals the performance of the best existing estimators (Lancaster-Imbens and Cosslett) in small samples. A common feature among these top-performing estimators is that they impose certain consistency requirements. The estimators without this feature exhibit less precision in our Monte Carlo simulations, and they are also subject to convergence issues, particularly when the sample size is small and q is relatively large.

When the prevalence rate is unknown, our pseudo-maximum likelihood estimator performs comparably to the Cosslett-Lancaster-Imbens estimator. Our Monte Carlo simulations reveal that both estimators are relatively imprecise in small samples and are subject to convergence problems, particularly when q is fairly close to either of its boundaries (0 or 1). Both of these problems are alleviated by using a larger estimation sample.

An important advantage of our new estimators over those proposed by Cosslett and Lancaster-Imbens is that the latter estimators require detailed knowledge of the sampling criteria when the primary and/or supplementary sample is exogenously stratified. This precludes their use when the relevant sampling criteria have not been made available, such as when the supplementary sample has been drawn from a Census survey. In contrast, our estimators require knowledge only of the sample weights, which are routinely available.

References

Brent, R.P. (1973) *Algorithms for Minimization Without Derivatives,* Englewood Cliffs, NJ: Prentice-Hall.

Breslow, N.E. (1996) "Statistics in Epidemiology: The Case-Control Study," *Journal of the American Statistical Association* (91:433) 14-28.

Burden, B.C., D.T. Canon, K.R. Mayer, and D.P. Moynihan (2014) "Election Laws, Mobilization, and Turnout: The Unanticipated Consequences of Election Reform," *American Journal of Political Science* (58:1) 95-109.

Cosslett, S.R. (1981) "Efficient Estimation of Discrete Choice Models," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden, Cambridge: MIT Press, 51-111.

Erard, B., J. Guyton, P. Langetieg, M. Payne, and A. Plumley (2016) "What Drives Income Tax Filing Compliance? *IRS Research Bulletin, Publication 1500,* Washington, DC: Internal Revenue Service, 32-37.

Imbens, G.W. (1992) "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice –Based Sampling", *Econometrica* (60:5) 1187-1214.

Keating, K.A. and S. Cherry (2004) "Use and Interpretation of Logistic Regression in Habitat Selection Studies," *Journal of Wildlife Management* (68:4) 774-789.

Lancaster, T. and G. Imbens (1996) "Case Controlled Studies with Contaminated Controls," *Journal of Econometrics* (71) 145-160.

Lele, S.R. (2009) "A New Method for Estimation of Resource Selection Probability Function," *Journal of Wildlife Management* (73:1) 122-127.

Lele, S.R. and J.L. Keim (2006) "Weighted Distributions and Estimation of Resource Selection Probability Functions," *Ecology* (87:12) 3021-3028.

Manski, C.F. and D. McFadden (1981) "Alternative Estimators and Sample Designs for Discrete Choice Analysis," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C. Manski and D. McFadden, Cambridge: MIT Press, 2-49.

Phillips, S.J. and J. Elith (2013) "On Estimating Probability of Presence from Use-Availability or Presence-Background Data," *Ecology* (94:6) 1409-1419.

Rosenman, R., S. Goates, and L. Hill (2012) "Participation in Universal Prevention Programs," *Applied Economics* (44:2) 219-28.

Royle, J.A., R.B. Chandler, C. Yackulic, and J.D. Nichols (2012) "Likelihood Analysis of Species Occurrence Probability from Presence-Only Data for Modelling Species Distributions," *Methods in Ecology and Evolution* (3) 545-554.

Solymos, P. and S.R. Lele (2016) "Revisiting Resource Section Probability Functions and Single-Visit Methods: Clarifications and Extensions," *Methods in Ecology and Evolution* (7:2), 196-205.

Steinberg, D. and N.S. Cardell (1992) "Estimating Logistic Regression Models When the Dependent Variable Has No Variance," *Communication in Statistics –Theory and Methods* (21:2) 423-450.

Ward, G., T. Hastie, S. Barry, J. Elith, and J.R. Leathwick (2009) ``Presence-Only Data and the EM Algorithm," *Biometrics* (65) 554-563.

Table 1: Monte Carlo Simulation Results, Prevalence Rate Known

Case 1: q **= 0.125,** N_0 **= 400,** N_1 **= 50**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Actual	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00
Logit Sample																		
Mean	-2.60	1.02	1.01	-2.62	1.05	1.03	-2.61	1.03	1.02	-2.61	1.03	1.02	-2.61	1.03	1.02	-2.61	1.03	1.02
Median	-2.59	1.01	1.00	-2.61	1.04	1.03	-2.59	1.01	1.01	-2.59	1.01	1.01	-2.59	1.02	1.02	-2.59	1.02	1.02
ASD	0.24	0.19	0.19	0.18	0.20	0.20	0.19	0.20	0.20	0.19	0.21	0.21	0.24	0.21	0.21	0.19	0.20	0.20
SSD	0.18	0.20	0.19	0.20	0.22	0.22	0.19	0.21	0.20	0.20	0.22	0.21	0.19	0.22	0.21	0.19	0.21	0.21
Mad	0.14	0.16	0.15	0.15	0.17	0.17	0.15	0.16	0.16	0.15	0.17	0.17	0.15	0.17	0.17	0.15	0.17	0.16
#Failures	0			0			0			0			0			0		
Independent Primary and Supplementary Samples																		
Mean	-2.64	1.05	1.04	-2.58	1.00	1.00	-2.61	1.03	1.02	-2.61	1.03	1.03	-2.61	1.04	1.03	-2.61	1.03	1.02
Median	-2.59	1.01	1.01	-2.56	0.99	0.98	-2.58	1.01	1.01	-2.59	1.01	1.01	-2.59	1.02	1.01	-2.59	1.03	1.02
ASD	0.30	0.32	0.32	0.18	0.23	0.24	0.20	0.25	0.25	0.20	0.25	0.25	0.26	0.26	0.26	0.20	0.25	0.25
SSD	0.26	0.31	0.30	0.21	0.26	0.26	0.20	0.26	0.25	0.21	0.26	0.25	0.21	0.26	0.25	0.21	0.26	0.25
Mad	0.19	0.24	0.23	0.16	0.21	0.20	0.16	0.20	0.19	0.16	0.20	0.19	0.16	0.20	0.19	0.16	0.20	0.19
#Failures	0			0			0			0			0			0		

Case 2: q **= 0.25,** N_0 **= 400,** N_1 **= 100**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Actual	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00
Logit Sample																		
Mean	-1.51	1.02	1.02	-1.51	1.04	1.04	-1.51	1.02	1.03	-1.51	1.03	1.03	-1.51	1.03	1.03	-1.51	1.03	1.03
Median	-1.50	1.01	1.01	-1.50	1.03	1.03	-1.50	1.01	1.02	-1.50	1.01	1.02	-1.50	1.02	1.02	-1.50	1.01	1.02
ASD	0.16	0.16	0.16	0.11	0.17	0.17	0.11	0.17	0.17	0.11	0.18	0.18	0.16	0.18	0.18	0.11	0.17	0.17
SSD	0.10	0.15	0.16	0.11	0.17	0.18	0.11	0.16	0.17	0.11	0.17	0.18	0.11	0.17	0.18	0.11	0.16	0.17
Mad	0.08	0.12	0.12	0.09	0.13	0.14	0.09	0.13	0.13	0.09	0.13	0.14	0.08	0.13	0.14	0.08	0.13	0.13
#Failures	0			0			0			0			0			0		
Independent Primary and Supplementary Samples																		
Mean	-1.53	1.04	1.05	-1.50	1.00	1.01	-1.51	1.02	1.03	-1.51	1.03	1.04	-1.51	1.03	1.04	-1.51	1.02	1.03
Median	-1.50	0.99	1.00	-1.49	0.99	0.99	-1.50	1.00	1.02	-1.50	1.01	1.01	-1.50	1.01	1.01	-1.50	1.01	1.02
ASD	0.21	0.32	0.32	0.10	0.22	0.22	0.11	0.23	0.23	0.11	0.23	0.23	0.19	0.23	0.23	0.11	0.23	0.23
SSD	0.15	0.31	0.30	0.11	0.22	0.23	0.11	0.22	0.23	0.11	0.22	0.23	0.11	0.22	0.23	0.11	0.22	0.23
Mad	0.11	0.23	0.23	0.08	0.17	0.18	0.08	0.17	0.18	0.08	0.17	0.18	0.09	0.17	0.18	0.09	0.17	0.18
#Failures	0			0			0			0			0			0		

Case 3: q **= 0.50,** N_0 **= 400,** N_1 **= 200**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Actual	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00
Logit Sample																		
Mean	0.00	1.01	1.01	0.01	1.03	1.03	0.00	1.02	1.02	0.00	1.02	1.03	0.01	1.03	1.03	0.01	1.02	1.03
Median	0.00	1.01	1.00	0.01	1.02	1.02	0.00	1.01	1.01	0.00	1.02	1.01	0.01	1.02	1.01	0.01	1.01	1.01
ASD	0.12	0.14	0.14	0.07	0.17	0.17	0.06	0.16	0.16	0.07	0.18	0.18	0.13	0.17	0.17	0.07	0.16	0.16
SSD	0.06	0.14	0.14	0.07	0.17	0.16	0.06	0.16	0.16	0.07	0.18	0.17	0.08	0.18	0.17	0.07	0.17	0.16
Mad	0.05	0.11	0.11	0.05	0.14	0.13	0.05	0.13	0.12	0.05	0.14	0.13	0.06	0.14	0.13	0.06	0.14	0.13
#Failures	0			0			0			0			0			0		
Independent Primary and Supplementary Samples																		
Mean	0.02	1.10	1.08	0.01	1.02	1.01	0.01	1.03	1.02	0.01	1.03	1.02	0.02	1.05	1.04	0.01	1.04	1.03
Median	0.01	1.05	1.02	0.00	1.01	1.00	0.01	1.03	1.01	0.01	1.03	1.01	0.01	1.04	1.02	0.00	1.04	1.02
ASD	0.28	0.48	0.47	0.07	0.23	0.23	0.08	0.24	0.24	0.07	0.24	0.24	0.24	0.26	0.26	0.08	0.25	0.25
SSD	0.09	0.42	0.41	0.07	0.25	0.23	0.07	0.25	0.23	0.07	0.25	0.23	0.10	0.25	0.24	0.09	0.25	0.23
Mad	0.06	0.30	0.29	0.05	0.20	0.19	0.06	0.20	0.18	0.05	0.19	0.18	0.08	0.20	0.19	0.07	0.20	0.18
#Failures	2			0			0			0			0			0		

Case 4: q **= 0.75,** N_0 **= 400,** N_1 **= 300**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2
Actual	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00
Logit Sample																		
Mean	1.50	1.02	1.02	1.53	1.04	1.04	1.51	1.03	1.03	1.53	1.04	1.04	1.57	1.06	1.06	1.55	1.05	1.05
Median	1.50	1.01	1.01	1.51	1.03	1.03	1.50	1.02	1.02	1.51	1.02	1.03	1.52	1.02	1.04	1.52	1.02	1.03
ASD	0.16	0.16	0.16	0.15	0.20	0.20	0.14	0.20	0.20	0.16	0.21	0.22	0.25	0.22	0.23	0.22	0.21	0.21
SSD	0.10	0.17	0.15	0.14	0.21	0.20	0.12	0.19	0.19	0.15	0.21	0.21	0.24	0.24	0.24	0.20	0.22	0.21
Mad	0.08	0.13	0.12	0.11	0.17	0.16	0.10	0.15	0.15	0.11	0.17	0.16	0.17	0.18	0.18	0.15	0.17	0.16
#Failures	0			0			0			0			0			0		
Independent Primary and Supplementary Samples																		
Mean	1.71	1.16	1.19	1.55	1.01	1.02	1.56	1.04	1.05	1.57	1.05	1.06	1.65	1.09	1.10	1.62	1.08	1.09
Median	1.53	1.01	1.01	1.52	1.01	1.02	1.54	1.03	1.03	1.54	1.03	1.04	1.53	1.05	1.07	1.52	1.05	1.05
ASD	1.33	1.20	1.21	0.23	0.35	0.36	0.24	0.34	0.35	0.24	0.35	0.35	0.66	0.44	0.44	0.27	0.36	0.36
SSD	0.58	0.76	0.75	0.26	0.38	0.38	0.24	0.34	0.36	0.25	0.35	0.36	0.47	0.41	0.42	0.45	0.40	0.40
Mad	0.38	0.54	0.53	0.19	0.30	0.30	0.18	0.27	0.28	0.18	0.27	0.28	0.31	0.30	0.31	0.28	0.30	0.30
#Failures	30			0			0			0			5			1		

Case 5: q **= 0.875,** N_0 **= 400,** N_1 **= 350**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2				β_0	β_1	β_2
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00
Logit Sample																		
Mean	2.61	1.03	1.02	2.65	1.03	1.04	2.64	1.04	1.05	2.67	1.06	1.06	2.85	1.12	1.13	2.77	1.09	1.10
Median	2.60	1.02	1.01	2.60	1.02	1.02	2.61	1.03	1.03	2.63	1.03	1.04	2.67	1.04	1.06	2.66	1.04	1.05
ASD	0.24	0.19	0.19	0.28	0.27	0.26	0.27	0.26	0.26	0.30	0.28	0.28	0.64	0.36	0.38	0.30	0.28	0.28
SSD	0.18	0.19	0.20	0.30	0.29	0.29	0.24	0.25	0.26	0.30	0.28	0.29	0.70	0.39	0.43	0.51	0.34	0.34
Mad	0.14	0.15	0.15	0.21	0.22	0.22	0.19	0.20	0.20	0.22	0.22	0.23	0.43	0.28	0.29	0.34	0.24	0.25
#Failures	0			0			0			0			10			2		
Independent Primary and Supplementary Samples																		
Mean	2.96	1.02	1.10	2.75	0.97	1.01	2.81	1.02	1.06	2.82	1.04	1.08	3.02	1.10	1.13	2.97	1.10	1.13
Median	2.63	0.88	0.95	2.65	0.94	1.00	2.72	1.03	1.07	2.72	1.02	1.08	2.66	1.04	1.09	2.66	1.03	1.10
ASD	3.90	2.25	2.44	0.50	0.60	0.62	0.54	0.55	0.55	0.61	0.63	0.63	1.94	0.99	0.93	0.64	0.56	0.57
SSD	1.02	0.94	1.02	0.63	0.63	0.63	0.55	0.61	0.61	0.64	0.60	0.63	1.07	0.74	0.73	1.00	0.71	0.71
Mad	0.70	0.70	0.74	0.46	0.49	0.50	0.41	0.45	0.47	0.44	0.44	0.47	0.73	0.53	0.53	0.68	0.51	0.52
#Failures	181			0			0			0			86			53		

Case 6: q **= 0.875,** N_0 **=1,600,** N_1 **= 1,400**

	Steinberg-Cardell			Lancaster-Imbens			Calibrated Logit			Cosslett			Simplified Cosslett			Unconstrained Pseudo-MLE		
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2				β_0	β_1	β_2
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00
Logit Sample																		
Mean	2.61	1.03	1.02	2.65	1.03	1.04	2.64	1.04	1.05	2.67	1.06	1.06	2.85	1.12	1.13	2.77	1.09	1.10
Median	2.60	1.02	1.01	2.60	1.02	1.02	2.61	1.03	1.03	2.63	1.03	1.04	2.67	1.04	1.06	2.66	1.04	1.05
ASD	0.24	0.19	0.19	0.28	0.27	0.26	0.27	0.26	0.26	0.30	0.28	0.28	0.64	0.36	0.38	0.30	0.28	0.28
SSD	0.18	0.19	0.20	0.30	0.29	0.29	0.24	0.25	0.26	0.29	0.28	0.29	0.70	0.39	0.43	0.51	0.34	0.34
Mad	0.14	0.15	0.15	0.21	0.22	0.22	0.19	0.20	0.20	0.22	0.23	0.23	0.43	0.28	0.29	0.34	0.24	0.25
#Failures	0			0			0			0			10			2		
Independent Primary and Supplementary Samples																		
Mean	2.96	1.02	1.10	2.75	0.97	1.01	2.81	1.02	1.06	2.80	1.04	1.08	3.02	1.10	1.13	2.97	1.10	1.13
Median	2.63	0.88	0.95	2.65	0.94	1.00	2.72	1.03	1.07	2.72	1.03	1.07	2.66	1.04	1.09	2.66	1.03	1.10
ASD	3.90	2.25	2.44	0.50	0.60	0.62	0.54	0.55	0.55	0.59	0.61	0.62	1.94	0.99	0.93	0.64	0.56	0.57
SSD	1.02	0.94	1.02	0.63	0.63	0.63	0.55	0.61	0.61	0.63	0.58	0.61	1.07	0.74	0.73	1.00	0.71	0.71
Mad	0.70	0.70	0.74	0.46	0.49	0.50	0.41	0.45	0.47	0.44	0.43	0.46	0.73	0.53	0.53	0.68	0.51	0.52
#Failures	181			0			0			16			86			53		

Table 2: Monte Carlo Simulation Results, Prevalence Rate Unknown

Case 1: q **= 0.125,** N_0 **= 400,** N_1 **= 50**

	q **Known**						q **Unknown**							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	-2.574	1.00	1.00	-2.574	1.00	1.00	-2.574	1.00	1.00	0.125	-2.574	1.00	1.00	0.125
Logit Sample														
Mean	-2.62	1.05	1.03	-2.61	1.03	1.02	-2.59	1.16	1.15	0.15	-2.63	1.13	1.14	0.14
Median	-2.61	1.04	1.03	-2.59	1.01	1.01	-2.58	1.07	1.08	0.14	-2.61	1.06	1.06	0.13
GSD	0.18	0.20	0.20	0.19	0.20	0.20	0.81	0.39	0.38	0.08	3.66	1.18	1.16	0.44
LSD							3.00	0.54	0.54	0.23	1.35	0.47	0.49	0.11
SSD	0.20	0.22	0.22	0.19	0.21	0.20	0.72	0.49	0.43	0.08	0.70	0.57	0.75	0.07
Mad	0.15	0.17	0.17	0.15	0.16	0.16	0.51	0.28	0.27	0.06	0.48	0.26	0.27	0.05
#Failures	0			0			51				37			
Independent Primary and Supplementary Samples														
Mean	-2.58	1.00	1.00	-2.61	1.03	1.02	-2.42	1.26	1.25	0.18	-2.47	1.24	1.24	0.17
Median	-2.56	0.99	0.98	-2.58	1.01	1.01	-2.40	1.15	1.17	0.16	-2.43	1.16	1.17	0.16
GSD	0.18	0.23	0.24	0.20	0.25	0.25	1.30	0.49	0.49	0.12	2.93	1.16	1.13	0.38
LSD							1.33	0.47	0.46	0.12	1.53	0.47	0.45	0.11
SSD	0.21	0.26	0.26	0.20	0.26	0.25	0.93	0.60	0.57	0.10	0.94	0.52	0.51	0.10
Mad	0.16	0.21	0.20	0.16	0.20	0.19	0.68	0.34	0.34	0.08	0.68	0.32	0.32	0.08
#Failures	0			0			288				297			

Case 2: **q = 0.25, N_0 = 400, N_1 = 100**

	q Known						**q Unknown**							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	-1.492	1.00	1.00	-1.492	1.00	1.00	-1.492	1.00	1.00	0.125	-1.492	1.00	1.00	0.125
Logit Sample														
Mean	-1.51	1.04	1.04	-1.51	1.02	1.03	-1.51	1.07	1.08	0.26	-1.51	1.06	1.07	0.25
Median	-1.50	1.03	1.03	-1.50	1.01	1.02	-1.49	1.03	1.03	0.26	-1.51	1.02	1.03	0.25
GSD	0.11	0.17	0.17	0.11	0.17	0.17	0.49	0.28	0.28	0.07	2.52	0.82	0.82	0.41
LSD							1.98	0.51	0.51	0.31	0.70	0.31	0.31	0.11
SSD	0.11	0.17	0.18	0.11	0.16	0.17	0.48	0.29	0.31	0.08	0.43	0.26	0.28	0.07
Mad	0.09	0.13	0.14	0.09	0.13	0.13	0.35	0.22	0.23	0.06	0.31	0.20	0.21	0.05
#Failures	0			0			2				1			
Independent Primary and Supplementary Samples														
Mean	-1.50	1.00	1.01	-1.51	1.02	1.03	-1.45	1.14	1.16	0.27	-1.49	1.12	1.15	0.27
Median	-1.49	0.99	0.99	-1.50	1.00	1.02	-1.41	1.09	1.09	0.27	-1.44	1.08	1.09	0.27
GSD	0.10	0.22	0.22	0.11	0.23	0.23	0.95	0.39	0.40	0.13	2.69	0.98	1.00	0.42
LSD							0.91	0.37	0.37	0.13	0.91	0.35	0.35	0.11
SSD	0.11	0.22	0.23	0.11	0.22	0.23	0.76	0.38	0.40	0.11	0.79	0.37	0.38	0.11
Mad	0.08	0.17	0.18	0.08	0.17	0.18	0.58	0.28	0.29	0.09	0.59	0.27	0.28	0.09
#Failures	0			0			138				136			

Case 3: **q = 0.5, N_0 = 400, N_1 = 200**

	q Known						**q Unknown**							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	0.00	1.00	1.00	0.00	1.00	1.00	0.00	1.00	1.00	0.50	0.00	1.00	1.00	0.50
Logit Sample														
Mean	0.01	1.03	1.03	0.00	1.02	1.02	0.08	1.08	1.08	0.51	0.05	1.06	1.06	0.50
Median	0.01	1.02	1.02	0.00	1.01	1.01	0.06	1.05	1.04	0.51	0.04	1.03	1.04	0.51
GSD	0.07	0.17	0.17	0.06	0.16	0.16	0.41	0.27	0.27	0.06	1.39	0.55	0.55	0.26
LSD							1.89	0.65	0.65	0.42	0.60	0.33	0.33	0.10
SSD	0.07	0.17	0.16	0.06	0.16	0.16	0.45	0.33	0.30	0.07	0.35	0.27	0.26	0.06
Mad	0.05	0.14	0.13	0.05	0.13	0.12	0.32	0.23	0.23	0.05	0.27	0.21	0.20	0.05
#Failures	0			0			0				0			
Independent Primary and Supplementary Samples														
Mean	0.01	1.02	1.01	0.01	1.03	1.02	0.09	1.13	1.12	0.50	0.02	1.11	1.10	0.49
Median	0.00	1.01	1.00	0.01	1.03	1.01	0.04	1.04	1.05	0.50	0.01	1.03	1.04	0.50
GSD	0.07	0.23	0.23	0.08	0.24	0.24	0.96	0.46	0.44	0.15	2.63	1.03	1.00	0.45
LSD							0.83	0.40	0.40	0.14	0.85	0.39	0.38	0.12
SSD	0.07	0.25	0.23	0.07	0.25	0.23	0.89	0.47	0.45	0.13	0.89	0.45	0.43	0.14
Mad	0.05	0.20	0.19	0.06	0.20	0.18	0.65	0.32	0.32	0.11	0.65	0.31	0.31	0.11
#Failures	0			0			56				57			

Case 4: **q = 0.75, N_0 = 400, N_1 = 300**

	q Known						**q Unknown**							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	1.492	1.00	1.00	1.492	1.00	1.00	1.492	1.00	1.00	0.75	1.492	1.00	1.00	0.75
Logit Sample														
Mean	1.53	1.04	1.04	1.51	1.03	1.03	1.77	1.14	1.16	0.76	1.67	1.09	1.11	0.75
Median	1.51	1.03	1.03	1.50	1.02	1.02	1.59	1.05	1.05	0.76	1.55	1.04	1.04	0.75
GSD	0.15	0.20	0.20	0.14	0.20	0.20	0.70	0.37	0.37	0.05	1.23	0.48	0.49	0.13
LSD							3.61	1.21	1.23	0.59	1.01	0.47	0.48	0.09
SSD	0.14	0.21	0.20	0.12	0.19	0.19	1.07	0.51	0.60	0.05	0.80	0.42	0.48	0.04
Mad	0.11	0.17	0.16	0.10	0.15	0.15	0.56	0.31	0.32	0.04	0.44	0.26	0.26	0.03
#Failures	0			0			0				0			
Independent Primary and Supplementary Samples														
Mean	1.55	1.01	1.02	1.56	1.04	1.05	2.10	1.30	1.33	0.72	1.91	1.25	1.25	0.72
Median	1.52	1.01	1.02	1.54	1.03	1.03	1.62	1.06	1.10	0.75	1.58	1.05	1.07	0.75
GSD	0.23	0.35	0.36	0.24	0.34	0.35	2.34	0.92	0.93	0.17	3.26	1.22	1.22	0.36
LSD							1.80	0.70	0.75	0.15	1.51	0.61	0.61	0.12
SSD	0.26	0.38	0.38	0.24	0.34	0.36	2.77	1.07	1.34	0.15	2.30	0.95	1.09	0.15
Mad	0.19	0.30	0.30	0.18	0.27	0.28	1.39	0.59	0.59	0.11	1.26	0.55	0.52	0.11
#Failures	0			0			67				61			

Case 5: **q = 0.875, N_0 = 400, N_1 = 350**

	q Known						q Unknown							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	0.875	2.574	1.00	1.00	0.875
Logit Sample														
Mean	2.65	1.03	1.04	2.64	1.04	1.05	3.24	1.25	1.27	0.88	3.08	1.19	1.21	0.88
Median	2.60	1.02	1.02	2.61	1.03	1.03	2.77	1.06	1.08	0.88	2.71	1.04	1.05	0.88
GSD	0.28	0.27	0.26	0.27	0.26	0.26	1.57	0.64	0.65	0.04	1.71	0.59	0.63	0.09
LSD							8.51	2.60	2.62	0.82	2.36	0.88	0.92	0.09
SSD	0.30	0.29	0.29	0.24	0.25	0.26	2.55	1.03	1.00	0.04	1.81	0.74	0.77	0.04
Mad	0.21	0.22	0.22	0.19	0.20	0.20	1.07	0.48	0.48	0.03	0.89	0.41	0.41	0.03
#Failures	0			0			38				12			
Independent Primary and Supplementary Samples														
Mean	2.75	0.97	1.01	2.81	1.02	1.06	4.40	1.51	1.73	0.83	4.31	1.48	1.68	0.83
Median	2.65	0.94	1.00	2.72	1.03	1.07	2.86	1.11	1.08	0.88	2.89	1.10	1.08	0.88
GSD	0.50	0.60	0.62	0.54	0.55	0.55	4.61	1.40	1.63	0.19	4.26	1.44	1.48	0.26
LSD							5.38	1.59	2.02	0.17	2.93	0.99	1.04	0.13
SSD	0.60	0.62	0.63	0.55	0.61	0.61	7.44	2.15	3.25	0.14	7.30	2.23	3.11	0.15
Mad	0.45	0.48	0.50	0.41	0.45	0.47	2.86	1.02	1.16	0.09	2.83	1.01	1.14	0.10
#Failures	1			0			220				181			

Case 6: **q = 0.875, N_0 = 1,600, N_1 = 1,400**

	q Known						**q Unknown**							
	Lancaster-Imbens			Calibrated Logit			Cosslett-Lancaster-Imbens				Pseudo-MLE			
	β_0	β_1	β_2	β_0	β_1	β_2	β_0	β_1	β_2	q	β_0	β_1	β_2	q
Actual	2.574	1.00	1.00	2.574	1.00	1.00	2.574	1.00	1.00	0.875	2.574	1.00	1.00	0.875
Logit Sample														
Mean	2.59	1.01	1.01	2.59	1.01	1.01	2.69	1.05	1.04	0.88	2.66	1.04	1.03	0.88
Median	2.58	1.00	1.01	2.58	1.00	1.00	2.63	1.02	1.01	0.88	2.62	1.02	1.01	0.88
GSD	0.13	0.13	0.13	0.12	0.13	0.13	0.44	0.21	0.21	0.02	0.43	0.23	0.23	0.04
LSD							3.21	0.95	0.95	0.37	0.74	0.31	0.30	0.04
SSD	0.09	0.10	0.10	0.11	0.12	0.12	0.48	0.23	0.23	0.02	0.40	0.20	0.20	0.02
Mad	0.07	0.08	0.08	0.09	0.10	0.10	0.36	0.18	0.17	0.02	0.30	0.16	0.15	0.01
#Failures	0			0			0				0			
Independent Primary and Supplementary Samples														
Mean	2.59	0.98	0.98	2.61	1.01	1.00	2.81	1.09	1.08	0.86	2.75	1.07	1.07	0.86
Median	2.57	0.99	0.99	2.59	1.00	1.00	2.62	1.00	1.00	0.87	2.58	1.01	1.00	0.87
GSD	0.22	0.25	0.25	0.23	0.25	0.25	1.26	0.46	0.46	0.08	2.05	0.71	0.70	0.13
LSD							1.08	0.41	0.41	0.07	0.96	0.36	0.36	0.70
SSD	0.24	0.28	0.27	0.23	0.24	0.25	1.79	0.60	0.64	0.07	1.76	0.62	0.60	0.08
Mad	0.19	0.21	0.21	0.17	0.19	0.19	0.92	0.35	0.35	0.05	0.91	0.35	0.34	0.05
#Failures	0			0			15				10			

Table 3: Standard Logit vs. Supplementary Sampling Estimators of the Decision to Vote

Variable	Original Specification		Restricted Specification					
	Standard Logit		Standard Logit		Calibrated Logit		Pseudo-MLE *q* unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	-0.1845	-3.32	-0.1283	-2.18	-0.1083	-2.75	-0.1108	-2.63
EDR	0.1870	2.07	0.2392	3.31	0.2745	3.65	0.2825	3.31
Early*SDR	0.0037	0.08	0.0004	0.01	0.0336	0.71	0.0328	0.67
Early*EDR	-0.0723	-0.57	0.0283	0.25	0.0218	0.17	0.0198	0.15
Early*EDR*SDR	0.1292	1.58	0.2033	2.68	0.1778	2.31	0.1807	2.22
30-Day Reg. Close	-0.1220	-2.51	-0.1048	-2.46	-0.0581	-1.54	-0.0596	-1.50
ID Requirement	0.0036	0.06	-0.0090	-0.16	-0.0042	-0.10	-0.6029	-0.13
Education	0.6002	28.64	0.6277	31.93	0.7074	41.17	0.7322	5.91
African American	0.7181	11.83	0.4030	7.09	0.6192	11.34	0.6429	4.84
Hispanic	-0.0489	-0.48	-0.1068	-1.00	0.0600	1.11	0.0650	1.06
Naturalized Citizen	-1.0275	-5.88	-0.5793	-8.31	-0.5242	-8.34	-0.5319	-7.30
Married	0.4258	18.04	0.4619	19.06	0.8235	24.01	0.8515	6.03
Female	0.1489	8.26	0.1693	12.08	0.2353	7.57	0.2424	5.21
Age	0.0254	21.29	0.0237	21.89	0.0248	17.98	0.0256	5.92
Age 18–24	0.4257	11.37	0.2141	6.23	0.3308	6.14	0.3455	3.82
Age 75 plus	-0.1085	-2.03	-0.2443	-6.12	-0.3448	-4.95	-0.3564	-3.96
Competitiveness	0.0119	4.33	0.0095	3.86	0.0121	5.22	0.0126	4.17
South	-0.0760	-1.25	-0.0457	-0.87	-0.1154	-2.68	-0.1205	-2.34
North Dakota	-0.3501	-4.28	-0.2542	-3.23	-0.2570	-1.16	-0.2579	-1.11
Oregon	0.1872	4.01	0.0912	1.62	0.2453	1.89	0.2467	1.84
Washington	-0.0204	-0.34	0.0305	0.51	0.0814	0.69	0.0818	0.67
Self-Reported Vote	0.8231	28.51						
Natural. 10+ Years	0.4565	2.76						
Residence 1 Year	0.2681	7.58						
Income	0.0828	25.57						
Constant	-4.9878	-19.83	-3.4479	-14.49	-4.2386	-19.72	-4.3398	-8.34
q								
# Overall Sample	73,333		91,161		274,172		274,172	
# Partic. Sample	50,362		59,090		59,090		59,090	
# Suppl. Sample					215,082		215,082	

Table 4: Standard Multinomial Logit vs. Supplementary Sampling Estimators of the Decision to Vote

Vote on Election Day in Person

Variable	Original Specification		Restricted Specification					
	Standard MNL		Standard MNL		Calibrated MNL		Pseudo-MLE q Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	-0.5173	-5.07	-0.4576	-4.14	-0.4294	-10.92	-0.3653	-2.18
EDR	0.1368	1.52	0.1906	2.21	0.2148	2.89	0.2194	1.93
Early*SDR	-0.3858	-3.94	-0.3932	-3.64	-0.3541	-7.30	-0.2634	-1.41
Early*EDR	-0.3898	-2.38	-0.2845	-1.82	-0.3011	-2.36	-0.2240	-1.08
Early*EDR*SDR	-0.1721	-1.69	-0.0928	-0.91	-0.1036	-1.34	-0.0374	-0.30
30-Day Reg. Close	-0.1394	-1.68	-0.1231	-1.55	-0.0959	-2.51	-0.0987	-2.10
ID Requirement	-0.0749	-0.81	-0.0895	-1.03	-0.0888	-2.03	-0.0657	-1.09
Education	0.5522	24.37	0.5724	26.95	0.6470	37.01	0.6985	3.47
African American	0.6633	9.57	0.3625	5.38	0.5597	10.18	0.6056	3.01
Hispanic	-0.0501	-0.39	-0.1046	-0.77	0.0676	1.21	0.0695	0.84
Naturalized Citizen	-1.0241	-5.53	-0.5721	-7.12	-0.5060	-7.90	-0.5230	-6.07
Married	0.4624	16.67	0.4903	16.08	0.8455	24.19	0.8792	3.28
Female	0.1166	6.65	0.1440	10.10	0.2058	6.52	0.2202	3.34
Age	0.0189	12.39	0.0178	12.50	0.0186	13.28	0.0216	4.27
Age 18–24	0.2708	6.36	0.0605	1.48	0.1877	3.40	0.2483	2.80
Age 75 plus	-0.1958	-3.22	-0.3312	-7.68	-0.4067	-5.67	-0.3987	-2.60
Competitiveness	0.0068	1.40	0.0051	1.09	0.0097	4.12	0.1108	3.46
South	-0.2331	-1.95	-0.2056	-1.84	-0.2776	-6.40	-0.2638	-2.07
North Dakota	-0.2383	-1.92	-0.1588	-1.38	-0.1888	-0.85	-0.2158	-0.90
Oregon	-1.9307	-23.43	-1.9537	-19.93	-1.6540	-10.21	-1.3481	-2.46
Washington	-1.5068	-17.79	-1.4311	-16.34	-1.2843	-9.33	-1.0219	-2.20
Self-Reported Vote	0.8387	27.56						
Natural. 10+ Years	0.4540	2.66						
Residence 1 Year	0.3311	8.77						
Income	0.0770	19.10						
Constant	-4.0707	-9.40	-2.9532	-6.20	-3.5269	-16.06	-3.9669	-8.85
Estimated value of q							0.4384	3.96
CPS-based value of q	0.4455							
# Overall Sample	73,333		91,161		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082	

Vote Early in Person

Variable	Original Specification		Restricted Specification					
	Standard MNL		Standard MNL		Calibrated MNL		Pseudo-MLE *q* Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	1.6829	4.13	1.7419	4.15	1.8617	29.16	1.9331	11.52
EDR	-0.1572	-0.32	-0.1359	-0.27	0.4320	2.83	0.4392	2.25
Early*SDR	1.5449	3.05	1.5848	3.06	1.8319	25.89	1.9296	10.45
Early*EDR	1.6352	3.47	1.7582	3.65	2.0800	13.13	2.1590	9.73
Early*EDR*SDR	1.8385	3.92	1.9231	4.07	2.2332	22.00	2.3007	17.26
30-Day Reg. Close	0.2923	1.29	0.2945	1.26	0.4199	8.81	0.4188	7.31
ID Requirement	-0.4379	-1.12	-0.3991	-1.02	-0.3684	-6.23	-0.3439	-4.49
Education	0.7469	20.50	0.8114	22.07	0.8968	38.90	0.9529	4.87
African American	1.1944	9.54	0.8339	7.80	1.1334	17.16	1.1909	6.18
Hispanic	-0.0047	-0.03	-0.0413	-0.22	0.1637	2.19	0.1629	1.63
Naturalized Citizen	-1.0959	-3.61	-0.7775	-6.12	-0.7515	-8.23	-0.7637	-6.32
Married	0.4207	8.34	0.4805	12.46	0.8676	19.19	0.9034	3.44
Female	0.2119	5.72	0.2072	7.56	0.2903	7.05	0.3061	4.30
Age	0.0380	14.07	0.0345	13.30	0.0363	19.94	0.0394	7.73
Age 18–24	0.5669	5.65	0.3272	3.44	0.4198	5.01	0.4871	4.27
Age 75 plus	-0.3148	-3.96	-0.4794	-6.95	-0.6119	-6.61	-0.6009	-3.64
Competitiveness	0.0422	2.01	0.0363	1.72	0.0371	11.43	0.3848	9.30
South	1.0992	4.09	1.1374	4.07	1.2982	23.37	1.3129	10.43
North Dakota	-0.0975	-0.31	0.0152	0.05	0.0638	0.26	0.0438	0.16
Oregon	-0.9134	-2.04	-1.0924	-2.39	-0.2739	-0.57	-0.0091	-0.01
Washington	-0.7455	-1.76	-0.7733	-1.80	-0.2162	-0.49	0.0520	0.08
Self-Reported Vote	0.8745	20.42						
Natural. 10+ Years	0.3071	1.04						
Residence 1 Year	0.0659	1.09						
Income	0.1066	12.15						
Constant	-12.7293	-8.01	-10.8073	-6.64	-11.9484	-38.76	-12.2765	-17.16
Estimated value of *q*							0.1029	8.89
CPS-based value of *q*	0.0911							
# Overall Sample	73,333		91,161		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082	

Vote Early by Mail

Variable	Original Specification		Restricted Specification					
	Standard MNL		Standard MNL		Calibrated MNL		Pseudo-MLE *q* Unknown	
	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.	Coeff.	t-Stat.
Early	0.6610	1.72	0.6854	1.76	0.6341	11.67	0.7022	4.29
EDR	-0.0728	-0.14	-0.0241	-0.05	-0.1503	-1.42	-0.1497	-1.03
Early*SDR	1.5043	3.27	1.4634	3.07	1.5338	23.21	1.6298	8.87
Early*EDR	1.1422	2.79	1.1938	2.87	1.3157	8.46	1.3952	6.47
Early*EDR*SDR	1.2403	3.52	1.2763	3.59	1.1877	12.38	1.2502	9.97
30-Day Reg. Close	-0.5417	-1.66	-0.4977	-1.51	-0.6267	-11.66	-0.6329	-10.27
ID Requirement	0.8577	2.71	0.8069	2.63	1.0427	16.95	1.0710	13.75
Education	0.7245	22.38	0.7658	26.52	0.8615	38.15	0.9183	4.70
African American	0.3408	2.52	-0.0272	-0.23	0.1493	1.91	0.2067	1.06
Hispanic	-0.0909	-0.74	-0.2001	-1.74	-0.1276	-1.66	-0.1236	-1.29
Naturalized Citizen	-0.9739	-4.18	-0.5166	-4.92	-0.4301	-5.29	-0.4327	-3.89
Married	0.2802	6.05	0.3513	8.78	0.7246	16.50	0.7571	2.86
Female	0.2638	10.09	0.2803	11.91	0.3640	9.03	0.3818	5.52
Age	0.0511	16.95	0.0477	15.66	0.0510	27.06	5.4712	10.88
Age 18–24	1.2234	8.06	1.0245	6.59	1.1358	13.77	1.2215	11.10
Age 75 plus	0.1292	1.69	-0.0007	-0.01	-0.0654	-0.76	-0.0554	-0.34
Competitiveness	0.0131	1.11	0.0107	0.89	0.0055	1.87	0.0712	1.92
South	-0.8552	-2.36	-0.8136	-2.31	-0.9299	-16.07	-0.9226	-7.32
North Dakota	-1.1583	-3.50	-0.9977	-2.99	-1.0984	-4.59	-1.1334	-4.35
Oregon	3.2773	10.37	3.0915	9.65	3.3895	22.45	3.7038	7.20
Washington	2.0571	5.46	2.1073	5.51	2.0251	15.00	2.2957	5.15
Self-Reported Vote	0.7550	19.95						
Natural. 10+ Years	0.5019	1.96						
Residence 1 Year	0.1443	2.44						
Income	0.0980	13.46						
Constant	-9.9828	-9.61	-8.3723	-7.68	-8.7091	-31.10	-9.0592	-13.06
Estimated value of *q*							0.1146	8.37
CPS-based value of *q*	0.0986							
# Overall Sample	73,333		91,161		273,933		273,933	
# Election Day	36,027		42,468		42,468		42,468	
# Early in Person	6,518		7,473		7,473		7,473	
# Early by Mail	7,667		8,910		8,910		8,910	
# Supp. Sample					215,082		215,082	