STUDY SERIES
(Statistics #2024-01)

A Multinomial Analysis of Bilingual Training and Nonresponse Follow-up Contact Rates in the 2020 Decennial Census

Andrew M. Raim[1],
Renee Ellis[2],
Mikelyn Meyers[2]

[1]Center for Statistical Research and Methodology, U.S. Census Bureau
[2]Center for Behavioral Science Methods, U.S. Census Bureau

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

Report Issued: June 3, 2024

A Multinomial Analysis of Bilingual Training and Nonresponse Followup Contact Rates in the 2020 Decennial Census

Andrew M. Raim[a], Renee Ellis[b], and Mikelyn Meyers[b]

[a]Center for Statistical Research and Methodology, U.S. Census Bureau
[b]Center for Behavioral Science Methods, U.S. Census Bureau

Abstract

This report analyzes data gathered from an experiment that was carried out in the 2020 Census. A new training module was designed for bilingual enumerators to aid in administering the census questionnaire to Spanish-speaking households in the nonresponse followup operation. An objective is to study the association between the training and response rate. A multinomial regression with continuation-ratio logit link is used to model the sequence of contact attempts to affected households. Several log-odds ratios are considered to provide inference about this relationship. We do not find significant evidence to conclude that the training is associated negatively with response rate; i.e., training appears not to harm response rate. We suspect that a stronger conclusion can be made about the training and improvement to response rate in applicable contact situations; however, a more controlled experiment with associated data would be needed to make this assessment. Additionally, a power analysis which was carried out in advance of the experiment is now revisited. Here, power of the selected log-odds ratios is evaluated under the experiment as carried out in practice using the realized sample size.

Keywords: Continuation-Ratio Logit; Log-Odds Ratio; Holm Procedure; Power Analysis

1 Introduction

In the years leading up to the 2020 Decennial Census, a training module was developed to aid bilingual enumerators in administering the census questionnaire to Spanish-speaking households in the nonresponse followup operation (NRFU). An experiment was carried out within the 2020 Census to study the relationship between the training and response rates of affected households (Ellis et al., 2018). An approach to capture sequential contact attempts in this setting was developed by Raim et al. (2023) using continuation-ratio logit (CRL) regression. CRL is a link function for the multinomial distribution that captures the probability of advancing to the next step in a sequence, conditional on being at a given step. This can be compared to the more commonly encountered ordinal regression, where the outcome is considered to be a discretized observation of a continuous variable. Tutz (1991) compares CRL and ordinal regression, while a more recent survey of related models is given by Tutz (2022). In the setting of sequential contact attempts, we consider CRL to be more appropriate because it can directly parameterize response rate which is of primary interest.

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

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

Raim et al. (2023) also considers an approach to evaluate the size of the experiment in advance of the census: whether an adequate number of households would be included based on power curves of a generalized linear hypothesis test. Subsequently, a set of area census offices (ACOs) was selected in which to carry out the experiment, with ACOs assigned into treatment and control pairs; all bilingual enumerators in a treatment ACO were intended to receive the training, while those in a control ACO were intended not to receive the training. To create a balance between treatment and control groups, members of the pairs were selected to be roughly similar in terms of several demographic characteristics. However, when the experiment was carried out in the 2020 Census, the training had not been assigned to most of the enumerators in treatment ACOs and was assigned to some in control ACOs.

In this report, we explore a CRL model to study the data as they were collected in practice during the 2020 Census. Section 2 discusses preparation of the raw data used in modeling. Section 3 carries out a brief model selection to obtain a regression based on several factors and their interations. Section 4 presents estimates for log-odds ratios of interest which express association between inclusion/exclusion of the training and response rates of affected households. The Holm procedure is used to carry out a joint test using several of the log-odds ratios as test statistics. Section 5 revisits the power study of Raim et al. (2023) for the experiment as carried out in practice; here, we consider power curves of tests based on the log-odds ratios and their ability to detect a negative relationship to response rate.

This report is intended to serve as a companion to Meyers et al. (2024+), which presents a holistic discussion of the experiment including: data, exploratory analysis, qualitative analysis based on feedback from focus groups, and recommendations based on findings.

2 Data

This section describes how the analysis data are derived from raw data on contact attempts which were obtained in the 2020 Census. Raw data consist of records of each NRFU contact attempt to each household. Several filtering steps are taken to focus on relevant records. Table 1 displays the area census offices (ACOs) which were originally chosen for the experiment; households with records outside of these ACOs are filtered out of the analysis. Counts of households which received one, two, and three or more contact attempts are displayed for each ACO; these counts are the result of applying several other filters described in the present section. The raw data contain records on contacts which occurred either after a successful enumeration or a stop-work order had been issued for a household; these records are dropped from the analysis, while previous records for the household are retained. The analysis is intended to focus on Spanish-speaking households but an exact indicator of a household's language is not available in the data. We consider three indicators as a surrogate for household language: (1) whether a household was selected to receive mailed materials in Spanish, which is determined by region; (2) an annotation from the enumerator that most of interview was conducted in Spanish, though this was not used consistently; and (3) an indicator of whether the respondent is Hispanic, but this does not necessarily mean the interview was carried out in Spanish. Households having no records that meet at least one these crtieria are excluded from the analysis.

The steps described up to this point are largely the same as Section 3.2 of Meyers et al. (2024+); we take some additional steps for the present analysis. To establish the order of contact attempts, the remaining records were then grouped by household and ordered by timestamp. A number of households received one or more NRFU contacts and responded via a mode other than in-person NRFU; these have also been filtered out of the analysis, as the model in Section 3 does not adjust for effects which might potentially differ between response modes. The model requires a cutoff L with a successful contact may be distinguished within the the first L attempts; any further attempts are regarded as the outcome "no successful contact withoin the first L attempts." We take $L = 3$ for this study and exclude records corresponding to attempts four and onwards for each household; there are two primary reasons. First, many households became eligible for contact via proxy respondent on the fourth attempt in the 2020 Census (Meyers et al., 2024+). Proxies in such cases may have a Spanish language profile or response behavior differing from the household for which they were reporting; therefore, such cases would complicate the analysis. Second, contact attempts later in the sequence are relatively rare so that small counts in multinomial categories may lead to normal

Table 1: ACO locations originally selected for the experiment. The group column indicates the original designation as a treatment (T) or control (C) group, which was not necessarily adhered to in practice. The total number of households within each ACO are displayed, along with counts of those which received one, two, and three or more contact attempts.† Note to reviewer: "Label" column was added to this table and corresponding labels are used in Table 3.

			Contact Attempt			
Label	Location	Group	1	2	3+	Households
0	Phoenix	T	5,000	550	70	5,600
1	Phoenix	T	6,100	800	100	7,000
2	Phoenix	T	13,000	1,600	250	14,000
3	Phoenix	T	22,000	1,800	300	24,500
4	El Paso	T	13,000	1,700	400	15,000
5	El Paso	T	48,000	5,800	1,200	55,000
6	Houston	T	21,000	2,800	400	24,000
7	Houston	T	14,000	2,000	350	16,500
8	Houston	T	4,900	950	200	6,100
9	Houston	T	12,500	1,900	350	14,500
10	Houston	T	6,500	1,100	200	7,800
11	Houston	T	11,500	1,600	250	13,500
12	Dallas	C	3,600	400	60	4,100
13	Dallas	C	2,100	150	20	2,300
14	Dallas	C	10,000	1,500	250	12,000
15	Dallas	C	9,100	1,300	200	10,500
16	Dallas	C	6,600	700	100	7,400
17	Dallas	C	6,000	500	70	6,600
18	Los Angeles	T	23,000	4,300	950	28,500
19	Los Angeles	C	22,000	4,400	900	27,500
	Total		261,000	36,000	6,600	303,000

†These data are rounded—varying by magnitude—in accordance with approved disclosure avoidance practices of the Census Bureau. Totals are based on summing the unrounded counts then applying rounding rules.

approximations used in this work to be inappropriate (Raim et al., 2023). To construct the outcome for the model, each record is coded with an indicator of whether a successful contact was made on this attempt. Additional factors used as independent variables in the regression are: an indicator $\text{Bil}_{i\ell} \in \{0, 1\}$ of whether the enumerator for the ℓth attempt at household i is bilingual, an indicator $\text{Train}_{i\ell} \in \{0, 1\}$ of whether the enumerator has taken the training, a factor ACO_i encoding the ACO for the ith household, and the contact attempt number $\ell \in \{1, 2, 3\}$ of the record.

3 Analysis Model

Let $w_i \in \{1, 2, \ldots, L, L+1\}$ be the number of NRFU contact attempts required for a successful enumeration, where up to L attempts are observed. The outcome $w_i = L + 1$ denotes that no attempt within the first L was successful. We assume that $w_1, \ldots, w_n$ are observations from a CRL model

$$W_i \sim \text{CRL}_L(\boldsymbol{p}_i), \quad \text{logit}\, p_{i\ell} = \boldsymbol{x}_{i\ell}^\top \boldsymbol{\beta},$$

where $p_{i\ell}$ is the probability of enumeration in the ℓth attempt at the ith household, given that the previous $\ell-1$ attempts have been unsuccessful, for $\ell = 1,\ldots,L$ and $i = 1,\ldots,n$. Here, logit is the function $\text{logit}(p) = \log\{p/(1-p)\}$. This model assumes that W_i has distribution

$$\mathrm{P}(W_i = \ell) = \begin{cases} p_{i\ell}\displaystyle\prod_{b=1}^{\ell-1}(1-p_{ib}), & \ell = 1,\ldots,L, \\ \displaystyle\prod_{b=1}^{L}(1-p_{ib}), & \ell = L+1. \end{cases}$$

This CRL model was used by Raim et al. (2023) to justify that enough households were included in the experiment, but the form of $\boldsymbol{x}_{i\ell}$ now needs to be modified to accommodate how the experiment was carried out in practice. We compare four variations of increasing complexity. Model I "Intercept Only" has only an intercept,

$$\boldsymbol{x}_{i\ell}^\top\boldsymbol{\beta} = \lambda_0,$$

to be used as a basis for comparison with other models. The notation $\boldsymbol{\beta}$ is used to denote the vector of all model coefficients, which is solely λ_0 in Model I. Model II "Main Effects" has main effects for training, bilingualness of the enumerator, and the attempt number:

$$\boldsymbol{x}_{i\ell}^\top\boldsymbol{\beta} = \lambda_0 + \lambda^{\text{Bil}}\cdot\text{Bil}_{i\ell} + \lambda^{\text{Train}}\cdot\text{Train}_{i\ell} + \sum_{b=2}^{L}\lambda_b^{\text{Att}}\cdot\mathrm{I}(\ell = b).$$

Model III "All Interactions" includes main effects from Model II and their interactions:

$$\begin{aligned}\boldsymbol{x}_{i\ell}^\top\boldsymbol{\beta} = {}& \lambda_0 + \lambda^{\text{Bil}}\cdot\text{Bil}_{i\ell} + \lambda^{\text{Train}}\cdot\text{Train}_{i\ell} + \lambda^{\text{Train}\times\text{Bil}}\cdot\text{Bil}_{i\ell}\cdot\text{Train}_{i\ell} + \sum_{b=2}^{L}\lambda_b^{\text{Att}}\cdot\mathrm{I}(\ell = b) \\ &+ \sum_{b=2}^{L}\lambda_b^{\text{Bil}\times\text{Att}}\cdot\text{Bil}_{i\ell}\cdot\mathrm{I}(\ell = b) + \sum_{b=2}^{L}\lambda_b^{\text{Train}\times\text{Att}}\cdot\text{Train}_{i\ell}\cdot\mathrm{I}(b = \ell) \\ &+ \sum_{b=2}^{L}\lambda_b^{\text{Train}\times\text{Att}}\cdot\text{Bil}_{i\ell}\cdot\text{Train}_{i\ell}\cdot\mathrm{I}(\ell = b).\end{aligned}$$

Model IV "Interactions & ACOs" includes the effects from Model III and an additional main effect for ACO:

$$\begin{aligned}\boldsymbol{x}_{i\ell}^\top\boldsymbol{\beta} = {}& \lambda_0 + \lambda^{\text{Bil}}\cdot\text{Bil}_{i\ell} + \lambda^{\text{Train}}\cdot\text{Train}_{i\ell} + \lambda^{\text{Train}\times\text{Bil}}\cdot\text{Bil}_{i\ell}\cdot\text{Train}_{i\ell} + \sum_{b=2}^{L}\lambda_b^{\text{Att}}\cdot\mathrm{I}(\ell = b) \\ &+ \sum_{b=2}^{L}\lambda_b^{\text{Bil}\times\text{Att}}\cdot\text{Bil}_{i\ell}\cdot\mathrm{I}(\ell = b) + \sum_{b=2}^{L}\lambda_b^{\text{Train}\times\text{Att}}\cdot\text{Train}_{i\ell}\cdot\mathrm{I}(b = \ell) \\ &+ \sum_{b=2}^{L}\lambda_b^{\text{Train}\times\text{Att}}\cdot\text{Bil}_{i\ell}\cdot\text{Train}_{i\ell}\cdot\mathrm{I}(\ell = b) + \sum_{b=2}^{I}\lambda_b^{\text{ACO}}\cdot\mathrm{I}(\text{ACO}_i = b). \qquad (1)\end{aligned}$$

Note that the baseline level of the variable ACO_i is taken to be the ACO with label zero in Table 1. The design matrix $\boldsymbol{X}$ for each model consists of rows $\boldsymbol{x}_{i\ell}$ ordered by attempt $\ell = 1,\ldots,L$ and then by household $i = 1,\ldots,n$. As discussed in Raim et al. (2023), the CRL model may be fitted using statistical software for logistic regression by recoding $\boldsymbol{w} = (w_1,\ldots,w_n)$ to $\boldsymbol{y} = (y_{11},\ldots,y_{1L},\ldots,y_{n1},\ldots,y_{nL})$ with

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

Table 2: Comparison between models I–IV using AIC.

	Model	AIC
I	Intercept Only	344,226
II	Main Effects	326,741
III	All Interactions	315,741
IV	Interactions & ACOs	312,485

Outcomes coded to NA are dropped when fitting the model. The four models are therefore fitted in R using the following calls to the `glm` function (R Core Team, 2023).

```
glm1_out = glm(y ~ 1, data = tbl, family = binomial)                    # Model I
glm2_out = glm(y ~ bil + train + att, data = tbl, family = binomial)    # Model II
glm3_out = glm(y ~ bil*train*att, data = tbl, family = binomial)        # Model III
glm4_out = glm(y ~ bil*train*att + aco, data = tbl, family = binomial)  # Model IV
```

Fitting Models I–IV to the data yields Akaike information criteria (AIC) shown in Table 2. The AIC reduces substantially at each stage which suggests that added terms help to improve the fit. Therefore, we proceed with Model IV for the analysis. Estimated coefficients and their associated p-values for Model IV are given in Table 3.

4 Results

It is anticipated that the training is associated with improved response rates for Spanish-speaking households, and at least should not have a negative impact on response rates. To investigate this association, we consider log-odds ratios and compare the odds of a response in cases where the enumerator received the training versus when they did not receive the training. Log-odds ratios are iterated over cases where enumerators are bilingual and non-bilingual and over the three contact attempts which are considered in the model.

First, let $\boldsymbol{x}_{0\ell}$ be the covariate of a trained bilingual enumerator on the ℓth attempt under (1) for Model IV, and let $\tilde{\boldsymbol{x}}_{0\ell}$ be the corresponding covariate for a bilingual enumerator who did not receive the training. A common ACO must be selected for the coding of $\boldsymbol{x}_{0\ell}$ and $\tilde{\boldsymbol{x}}_{0\ell}$; however, the choice is arbitrary as its effect will cancel. Define log odds ratios

$$\begin{aligned}\vartheta_\ell^{\text{Bil}} &= \log \frac{\mathrm{P}(W=\ell \mid W \geq \ell, \boldsymbol{x}_{0\ell})/\{1-\mathrm{P}(W=\ell \mid W \geq \ell, \boldsymbol{x}_{0\ell})\}}{\mathrm{P}(W=\ell \mid W \geq \ell, \tilde{\boldsymbol{x}}_{0\ell})/\{1-\mathrm{P}(W=\ell \mid W \geq \ell, \tilde{\boldsymbol{x}}_{0\ell})\}} \\ &= \operatorname{logit} p_{i\ell}(\boldsymbol{x}_{0\ell}) - \operatorname{logit} p_{i\ell}(\tilde{\boldsymbol{x}}_{0\ell}) \\ &= (\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell})^\top \boldsymbol{\beta}\end{aligned}$$

for $\ell = 1, 2, 3$. Similarly, for the setting of non-bilingual enumerators, define

$$\vartheta_\ell^{\neg\text{Bil}} = (\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell})^\top \boldsymbol{\beta}$$

for $\ell = 1, 2, 3$, with $\boldsymbol{x}_{0\ell}$ and $\tilde{\boldsymbol{x}}_{0\ell}$ taken to be the covariates of non-bilingual enumerators on the ℓth attempt who have and have not received the training, respectively. The particular codings of $\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell}$ are shown in Table 4. The six defined odds ratios can be estimated by maximum likelihood using $\hat{\boldsymbol{\beta}}$ obtained from the model fit, and an approximate $1-\alpha$ level confidence interval is given for each using $\hat{\vartheta}_\ell \pm z_{\alpha/2} \cdot \widehat{\mathrm{SE}}(\hat{\vartheta}_\ell)$, with $z_{\alpha/2}$ the $1-\alpha/2$ quantile of the standard normal distribution, estimated standard error

$$\widehat{\mathrm{SE}}(\hat{\vartheta}_\ell) = \left\{(\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell})^\top [\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})](\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell})\right\}^{1/2},$$

Table 3: Estimates obtained from fitting Model IV. Note to reviewer: rows pertaining to ACO effects have been relabeled and reorded in accordance with Table 1.

Variable	Estimate	SE	z-value	p-value
Intercept	-0.4033	0.0417	-9.6649	$<$ 2e-16
Bil	2.4940	0.0198	125.9200	$<$ 2e-16
Train	-0.6882	0.4901	-1.4041	0.1603
$\mathrm{I}(\ell = 2)$	1.5552	0.0280	55.4500	$<$ 2e-16
$\mathrm{I}(\ell = 3)$	1.5742	0.0477	33.0238	$<$ 2e-16
aco01	-0.2038	0.0489	-4.1654	3.11e-05
aco02	-0.2317	0.0434	-5.3408	9.25e-08
aco03	0.0123	0.0422	0.2926	0.7698
aco04	-0.5654	0.0550	-10.2879	$<$ 2e-16
aco05	-0.6002	0.0391	-15.3444	$<$ 2e-16
aco06	-0.3660	0.0411	-8.9121	$<$ 2e-16
aco07	-0.4213	0.0423	-9.9482	$<$ 2e-16
aco08	-0.4964	0.0481	-10.3171	$<$ 2e-16
aco09	-0.5690	0.0425	-13.3801	$<$ 2e-16
aco10	-0.4593	0.0464	-9.8917	$<$ 2e-16
aco11	-0.0795	0.0441	-1.8028	0.0714
aco12	-0.1042	0.0574	-1.8169	0.0692
aco13	0.3785	0.0782	4.8408	1.29e-06
aco14	-0.2769	0.0442	-6.2647	3.74e-10
aco15	-0.3300	0.0449	-7.3472	2.02e-13
aco16	-0.2198	0.0487	-4.5171	6.27e-06
aco17	0.1547	0.0532	2.9054	0.0037
aco18	-0.8093	0.0400	-20.2360	$<$ 2e-16
aco19	-0.7171	0.0401	-17.8744	$<$ 2e-16
$\mathrm{Bil} \cdot \mathrm{Train}$	0.7899	0.4893	1.6142	0.1065
$\mathrm{Bil} \cdot \mathrm{I}(\ell = 2)$	-2.9982	0.0312	-96.1384	$<$ 2e-16
$\mathrm{Bil} \cdot \mathrm{I}(\ell = 3)$	-2.8545	0.0593	-48.1610	$<$ 2e-16
$\mathrm{Train} \cdot \mathrm{I}(\ell = 2)$	1.2543	0.5907	2.1234	0.0337
$\mathrm{Train} \cdot \mathrm{I}(\ell = 3)$	0.7758	0.7847	0.9887	0.3228
$\mathrm{Bil} \cdot \mathrm{Train} \cdot \mathrm{I}(\ell = 2)$	-1.4085	0.5938	-2.3721	0.0177
$\mathrm{Bil} \cdot \mathrm{Train} \cdot \mathrm{I}(\ell = 3)$	-0.0944	0.8003	-0.1180	0.9061

Table 4: Coding for difference in characteristics between trained and untrained enumerators, $\boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell}$, for the six log-odds ratios. The variables involving aco01, ..., aco19 are omitted from the display and are coded as zeros in all cases.

	Bilingual			Non-Bilingual		
Variable	$\ell = 1$	$\ell = 2$	$\ell = 3$	$\ell = 1$	$\ell = 2$	$\ell = 3$
Intercept	0	0	0	0	0	0
Bil	0	0	0	0	0	0
Train	1	1	1	1	1	1
$\mathrm{I}(\ell = 2)$	0	0	0	0	0	0
$\mathrm{I}(\ell = 3)$	0	0	0	0	0	0
Bil $\cdot$ Train	1	1	1	0	0	0
Bil $\cdot \mathrm{I}(\ell = 2)$	0	0	0	0	0	0
Bil $\cdot \mathrm{I}(\ell = 3)$	0	0	0	0	0	0
Train $\cdot \mathrm{I}(\ell = 2)$	0	1	0	0	1	0
Train $\cdot \mathrm{I}(\ell = 3)$	0	0	1	0	0	1
Bil $\cdot$ Train $\cdot \mathrm{I}(\ell = 2)$	0	1	0	0	0	0
Bil $\cdot$ Train $\cdot \mathrm{I}(\ell = 3)$	0	0	1	0	0	0

Table 5: Log-odds ratios and associated 90% confidence intervals under Model IV.

		90% CI	
	Estimate	Lo	Hi
$\vartheta_1^{\mathrm{Bil}}$	0.102	0.031	0.172
$\vartheta_2^{\mathrm{Bil}}$	-0.053	-0.162	0.057
$\vartheta_3^{\mathrm{Bil}}$	0.783	0.520	1.046
$\vartheta_1^{\neg\mathrm{Bil}}$	-0.688	-1.494	0.118
$\vartheta_2^{\neg\mathrm{Bil}}$	0.566	0.017	1.115
$\vartheta_3^{\neg\mathrm{Bil}}$	0.088	-0.924	1.099

and where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\beta}})$ is an estimate of the variance of $\hat{\boldsymbol{\beta}}$ which is obtained as `vcov(glm4_out)` using the `glm` fit.

Let us first ignore dependence between the log-odds ratios. A log-odds ratio larger than zero indicates that the odds of response are increased with enumerator training (compared to without training) for the given attempt number and bilingual status. The finding is significant with type I error $\alpha = 0.10$ when both endpoints of the associated confidence interval are larger than zero. Similarly, a log-odds ratio smaller than zero indicates that the odds of response are decreased with enumerator training, and the finding is significant when both endpoints are smaller than zero. The finding is not statistically significant when zero is contained in the interval.

Table 5 displays the estimates of the log-odds ratios and associated confidence intervals. The results appear to confirm the intuition that training should not have a negative effect on response rates, with none of the intervals being entirely below zero. However, for attempts 1 and 3 with bilingual enumerators, log-odds ratios are significantly larger than zero.

We now consider a more formal test to interpret log-odds ratios jointly, acknowledging their dependence. Here we are specifically concerned with bilingual enumerators; the training was intended for bilingual enumerators but was assigned to nonbilingual enumerators as well due to logistical difficulties. The union-

Table 6: Holm procedure with $m = 3$ tests and $\alpha = 0.10$.

Test	Estimate	SE	z-value	p-value	Adjusted Significance	Decision
H_{02}	-0.053	0.067	-0.790	0.215	0.034	Do not reject
H_{01}	0.102	0.043	2.364	0.991	0.050	Do not reject
H_{03}	0.783	0.160	4.899	> 0.999	0.100	Do not reject

intersection test (Casella and Berger, 2002, Section 8.2)

$$H_0 : \bigcap_{\ell=1}^{L} [\vartheta_\ell^{\text{Bil}} \geq 0] \quad \text{versus} \quad H_1 : \bigcup_{\ell=1}^{L} [\vartheta_\ell^{\text{Bil}} < 0] \tag{2}$$

assumes the null hypothesis that response probabilities remain the same or increase in all $L = 3$ attempts in cases where enumerators receive the training. The null hypothesis is rejected when there is evidence of reduced response probability in any of the three attempts, which would be an unexpected finding.

To test (2), we compute one-sided tests for ϑ_1^{Bil}, ϑ_2^{Bil}, and ϑ_3^{Bil} and use the Holm procedure (Holm, 1979) for multiple comparisons to test each of $H_{0\ell} : \vartheta_\ell^{\text{Bil}} \geq 0$ at an adjusted significiance level; see Algorithm 1. Our decision will be not to reject H_0 if none of the $H_{0\ell}$ are rejected. To justify the Holm procedure as an appropriate test for (2), note that Holm is based on a family of m tests with corresponding null hypotheses $H_{01}, \ldots, H_{0m}$ where each test is of size α. The familywise error rate is defined as $\text{FWER} = \text{P}(V \geq 1)$, where V is the number of falsely rejected hypotheses among $H_{01}, \ldots, H_{0m}$. The Holm procedure has the property that it controls FWER in the strong sense, with $\text{FWER} \leq \alpha$ under any configuration of $H_{01}, \ldots, H_{0m}$ being true (Shaffer, 1995). A special case is the global null hypothesis where all $H_{01}, \ldots, H_{0m}$ are true so that $V \geq 1$ is equivalent to a Type I error occurring. Therefore, a decision rule checking whether any hypothesis is rejected by Holm serves as a size α test of (2). Table 6 shows that there is indeed insufficient evidence to reject any of the hypotheses $H_{0\ell} : \vartheta_\ell \geq 0$ in favor of $H_{1\ell} : \vartheta_\ell < 0$, and thus to not reject H_0.

Algorithm 1 The Holm step-down procedure adjusts for m simultaneous comparisons to achieve an overall ("family-wise") error rate of α.

1. Let $P_{(1)} \leq \ldots \leq P_{(m)}$ be the ordered p-values and $H_{0(1)}, \ldots, H_{0(m)}$ be their associated null hypotheses.
2. Let $\alpha_{(j)} = \alpha/(m + 1 - j)$, $j = 1, \ldots, m$, be adjusted significance levels.
3. Let k be the smallest j where $P_{(j)} > \alpha_{(j)}$.
4. Hypotheses $H_{0(1)}, \ldots, H_{0(k-1)}$ are rejected.

5 Revisiting Power

To justify that there was an adequate number of households included in the original design of the study, Raim et al. (2023) proposed a method based on the power of a test of the generalized linear hypothesis $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{C}\boldsymbol{\beta} \neq \boldsymbol{0}$. With the benefit of hindsight, we can reconsider the power to better reflect how the experiment was actually carried out in the field and alterations which have been made to the statistical procedures in Sections 4.

Recall that the power of a test is the probability of a test procedure (correctly) rejecting the null hypothesis when the alternative hypothesis is true. Let $\vartheta = \boldsymbol{c}^\top \boldsymbol{\beta}$ denote one of the log-odds ratios defined in Section 4 with $\boldsymbol{c} = \boldsymbol{x}_{0\ell} - \tilde{\boldsymbol{x}}_{0\ell}$ as the corresponding fixed covariate value. Consider a test of the one-sided hypothesis $H_0 : \vartheta \geq 0$ versus $H_1 : \vartheta < 0$ at significance level $\alpha = 0.10$. Using a standard large sample normal approximation, we may reject H_0 if $\mathcal{Z} < z_\alpha$, where $\mathcal{Z} = \boldsymbol{c}^\top \hat{\boldsymbol{\beta}} / \{\boldsymbol{c}^\top \mathcal{I}^{-1}(\hat{\boldsymbol{\beta}}) \boldsymbol{c}\}^{1/2}$ and $z_\alpha \approx -1.282$. Here,

$\mathcal{I}(\boldsymbol{\beta}) = \boldsymbol{X}^\top \boldsymbol{D}(\boldsymbol{\beta})\boldsymbol{X}$ is the information matrix of the continuation-ratio logit likelihood with $\boldsymbol{D}(\boldsymbol{\beta})$ a diagonal matrix whose elements are $g(\boldsymbol{x}_{i\ell}^\top\boldsymbol{\beta})\prod_{b=1}^{\ell-1}[1-G(\boldsymbol{x}_{ib}^\top\boldsymbol{\beta})]$ for $\ell = 1,\ldots,w_i$ and $i = 1,\ldots,n$, where G and g are the CDF and density of the standard logistic distribution, respectively. When the actual effect size in the population is $\Delta = \boldsymbol{c}^\top\boldsymbol{\beta}$, the power of the test is

$$\begin{aligned}
\varpi(\boldsymbol{\beta}) &= \mathrm{P}\left(\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}^\top\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}}} < z_\alpha\right) \\
&= \mathrm{P}\left(\frac{\boldsymbol{c}^\top\hat{\boldsymbol{\beta}}}{\sqrt{\boldsymbol{c}^\top\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c}}} - \frac{\Delta}{\sqrt{\boldsymbol{c}^\top\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{c}}} < z_\alpha - \frac{\Delta}{\sqrt{\boldsymbol{c}^\top\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{c}}}\right) \\
&\approx \Phi\left(z_\alpha - \frac{\Delta}{\sqrt{\boldsymbol{c}^\top\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{c}}}\right).
\end{aligned}$$

Here we have informally used large sample properties, including convergence in distribution $\hat{\boldsymbol{\beta}} \xrightarrow{d} \mathrm{N}\left(\boldsymbol{\beta}, \mathcal{I}^{-1}(\boldsymbol{\beta})\right)$ and convergence in probability $\boldsymbol{c}^\top\mathcal{I}^{-1}(\hat{\boldsymbol{\beta}})\boldsymbol{c} \xrightarrow{p} \boldsymbol{c}^\top\mathcal{I}^{-1}(\boldsymbol{\beta})\boldsymbol{c}$, as $n \to \infty$. Formal investigations of consistency and asymptotic normality in generalized linear models may be found in the work of Fahrmeir and Kaufmann (1985) and related references.

To study power as a function of Δ, note that $\boldsymbol{\beta}$ may take on many values under the restriction $\boldsymbol{c}^\top\boldsymbol{\beta} = \Delta$. Raim et al. (2023) consider a method of choosing a representative value of $\boldsymbol{\beta}$ for a given Δ, based on minimizing power over a set of restrictions. We now consider a variation of this idea based on a one-sided test and making use of the data at hand. We compute the power at the solution $\tilde{\boldsymbol{\beta}}$ of linear equation $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{a}$ with $\boldsymbol{A}$ and $\boldsymbol{a}$ constructed as follows. The vector $\boldsymbol{a} = (\boldsymbol{b}_0, \Delta)$ represents a set of restrictions with Δ the size of the effect under consideration in the hypothesis. Let $\boldsymbol{A} = [\boldsymbol{B}^\top \ \boldsymbol{c}]^\top \in \mathbb{R}^{d\times d}$ be non-singular with $\boldsymbol{B} \in \mathbb{R}^{(d-1)\times d}$. We take $\boldsymbol{b}_0 \in \mathbb{R}^{d-1}$ to be $\boldsymbol{B}\hat{\boldsymbol{\beta}}$ so that the equation $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}_0$ anchors $\tilde{\boldsymbol{\beta}}$ to the MLE $\hat{\boldsymbol{\beta}}$ from the observed data. This choice of $\boldsymbol{B}$ and $\boldsymbol{b}_0$ yields a nearest $\boldsymbol{\beta}$ to $\hat{\boldsymbol{\beta}}$ where $\boldsymbol{c}^\top\tilde{\boldsymbol{\beta}} = \Delta$, in the sense that

$$\begin{aligned}
\|\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top\boldsymbol{A}^\top\boldsymbol{A}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\boldsymbol{B}^\top\boldsymbol{B} + \boldsymbol{c}\boldsymbol{c}^\top)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \|\boldsymbol{B}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 + \|\boldsymbol{c}^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\|^2 \\
&= \|\boldsymbol{B}\hat{\boldsymbol{\beta}} - \boldsymbol{b}_0\|^2 + \|\boldsymbol{c}^\top\hat{\boldsymbol{\beta}} - \Delta\|^2
\end{aligned}$$

for any $\boldsymbol{\beta}$ with $\boldsymbol{c}^\top\boldsymbol{\beta} = \Delta$ and $\boldsymbol{B}\boldsymbol{\beta} = \boldsymbol{b}_0$. The term $\|\boldsymbol{B}\hat{\boldsymbol{\beta}} - \boldsymbol{b}_0\|^2$ is nonnegative for any $\boldsymbol{b}_0$ and vanishes with $\boldsymbol{b}_0 = \boldsymbol{B}\hat{\boldsymbol{\beta}}$. The matrix $\boldsymbol{B}$ is taken to be $\boldsymbol{Q}_2^\top$ in the QR decomposition

$$\boldsymbol{c} = [\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]\begin{bmatrix} r_1 \\ \boldsymbol{0} \end{bmatrix} = \boldsymbol{Q}_1 r_1,$$

where the columns of $[\boldsymbol{Q}_1 \ \boldsymbol{Q}_2]$ form an orthogonal basis and r_1 is a scalar. Note that the columns of $\boldsymbol{Q}_2$ are a basis for the null space of the matrix $\boldsymbol{c}^\top$ because $\boldsymbol{c}^\top(\boldsymbol{Q}_2\boldsymbol{x}) = r_1\boldsymbol{Q}_1^\top\boldsymbol{Q}_2\boldsymbol{x} = 0$ for any $\boldsymbol{x} \in \mathbb{R}^{d-1}$. Therefore,

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{B} \\ \boldsymbol{c}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & r_1 \end{bmatrix}\begin{bmatrix} \boldsymbol{Q}_2^\top \\ \boldsymbol{Q}_1^\top \end{bmatrix}$$

is nonsingular. The power of the test with effect size Δ is therefore computed as $\varpi(\tilde{\boldsymbol{\beta}})$ with $\tilde{\boldsymbol{\beta}}$ the solution of $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{a}$.

Remark 1. An appealing property of this construction is that $\tilde{\boldsymbol{\beta}}$ differs from $\hat{\boldsymbol{\beta}}$ only in coordinates where $\boldsymbol{c}$ is non-zero. Such coordinates are displayed in Table 4. To see that this is the case, first note that we may

write any $\boldsymbol{\beta}$ as $\boldsymbol{\beta} = \boldsymbol{Q}\boldsymbol{Q}^\top\boldsymbol{\beta} = \boldsymbol{Q}_1\boldsymbol{Q}_1^\top\boldsymbol{\beta} + \boldsymbol{Q}_2\boldsymbol{Q}_2^\top\boldsymbol{\beta}$ with $\boldsymbol{Q} = [\boldsymbol{Q}_1\ \boldsymbol{Q}_2]$. The solution $\tilde{\boldsymbol{\beta}}$ to $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{a}$ is

$$\boldsymbol{A}^{-1}\boldsymbol{a} = \begin{bmatrix}\boldsymbol{Q}_2 & \boldsymbol{Q}_1\end{bmatrix}\begin{bmatrix}\boldsymbol{I} & \boldsymbol{0}\\ \boldsymbol{0} & r_1^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{Q}_2^\top\hat{\boldsymbol{\beta}}\\ \Delta\end{bmatrix} = \boldsymbol{Q}_2\boldsymbol{Q}_2^\top\hat{\boldsymbol{\beta}} + r_1^{-1}\Delta\boldsymbol{Q}_1.$$

Denoting $\boldsymbol{e}_j$ as the jth column of a $d \times d$ identity matrix, we have

$$\begin{aligned}\hat{\beta}_j - \tilde{\beta}_j &= \boldsymbol{e}_j^\top\left\{\hat{\boldsymbol{\beta}} - \boldsymbol{A}^{-1}\boldsymbol{a}\right\}\\ &= \boldsymbol{e}_j^\top\left\{[\boldsymbol{Q}_1\boldsymbol{Q}_1^\top\hat{\boldsymbol{\beta}} + \boldsymbol{Q}_2\boldsymbol{Q}_2^\top\hat{\boldsymbol{\beta}}] - [\boldsymbol{Q}_2\boldsymbol{Q}_2^\top\hat{\boldsymbol{\beta}} + r_1^{-1}\Delta\boldsymbol{Q}_1]\right\}\\ &= \boldsymbol{e}_j^\top\boldsymbol{Q}_1\left\{\boldsymbol{Q}_1^\top\hat{\boldsymbol{\beta}} - r_1^{-1}\Delta\right\}. \qquad (3)\end{aligned}$$

Note that $\boldsymbol{Q}_1$ is a single-vector basis for the column space $\{\alpha\boldsymbol{c} : \alpha \in \mathbb{R}\}$ so that (3) is zero when j corresponds to an element of $\boldsymbol{c}$ which is zero.

We now write $\tilde{\boldsymbol{\beta}}(\Delta)$ to emphasize dependence of $\tilde{\boldsymbol{\beta}}$ on Δ. We compute the quantity $\varpi(\tilde{\boldsymbol{\beta}}(\Delta))$ over a range of $\Delta \in [-2, -0.01]$ for each of the six log-odds ratios considered in Section 4. The results are plotted in Figure 1. To interpret a point $(\Delta, \varpi(\tilde{\boldsymbol{\beta}}(\Delta)))$ on the curve, the reader can consider the statement "had the true value for this particular log-odds ratio been Δ, there would be power $\varpi(\tilde{\boldsymbol{\beta}}(\Delta))$ to correctly reject the associated null hypothesis." Values of Δ in the curve may be compared to estimates in Table 5 which were observed in the data, as a reference. Note that power curves originate at the nominal 0.10 significance level at $\Delta = 0$ and increase to one as Δ decreases; i.e., effects with larger magnitudes are easier to detect than ones closer to zero. The curves for bilingual enumerators have a steeper increase than those for non-bilingual enumerators because a larger portion of the former received the training. It is interesting to note that the curves for attempts 1 and 2 cross in the non-bilingual case, with attempt 2 starting below attempt 1 but eventually overtaking it.

To support interpretation of the power, Figure 2 plots the elements of $\tilde{\boldsymbol{\beta}}(\Delta)$ varying within each power curve; i.e., Figures 2a–2f correspond to the six log-odds ratios plotted in Figure 1. Recall that each $\tilde{\boldsymbol{\beta}}(\Delta)$ represents a modification of $\hat{\boldsymbol{\beta}}$ to achieve effect size $\boldsymbol{c}^\top\tilde{\boldsymbol{\beta}} = \Delta$, and Δ is varied from -0.01 to -2. By Remark 1, only coordinates that correspond to nonzero entries in $\boldsymbol{c}$ are modified from $\hat{\boldsymbol{\beta}}$. Let us consider, as a concrete example, the log-odds ratio corresponding to the effect of bilingual enumerators on attempt $\ell = 1$. Figure 2a shows that the coefficients for Train and Bil $\cdot$ Train in $\tilde{\boldsymbol{\beta}}(\Delta)$ decrease as the effect size is decreased from $\Delta = 0$. The value for Train decreases from -0.744 to -1.739 while the value for Bil $\cdot$ Train decreases from 0.734 to -0.261. Combining Figure 2a with Figure 1, we notice that there is power 0.773 to detect effect size $\Delta = -0.09$, which is represented in the power study by $\tilde{\boldsymbol{\beta}}(\Delta)$ with Train and Bil $\cdot$ Train coefficients -0.784 and 0.694, respectively. Figures 2b–2f show paths of the coefficients for varying Δ under the five remaining log-odds ratios.

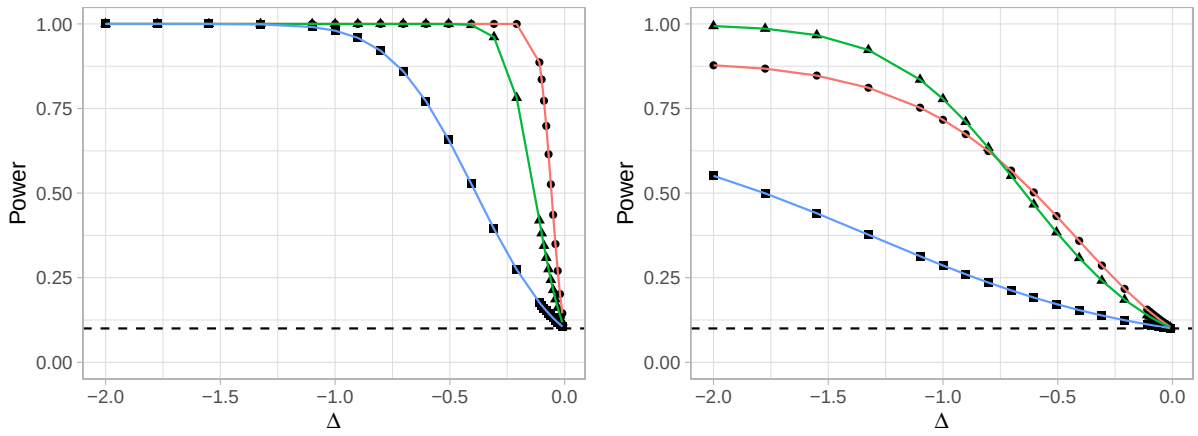
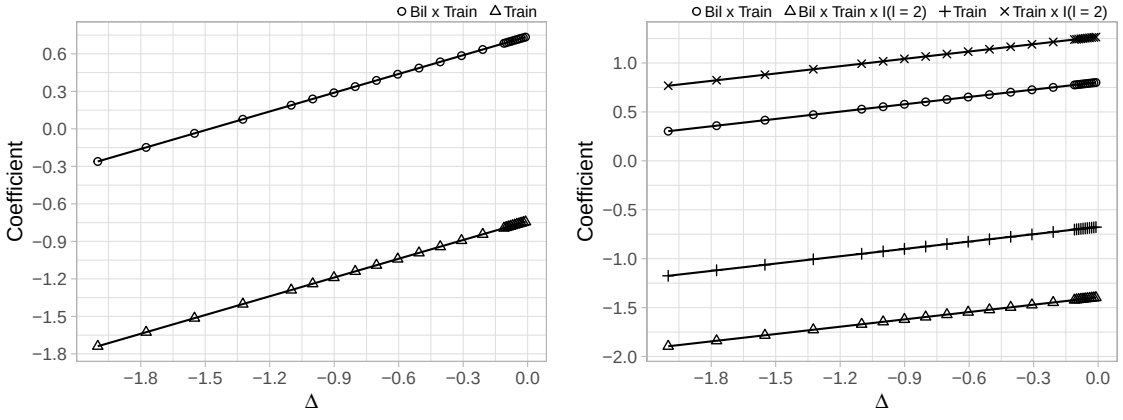
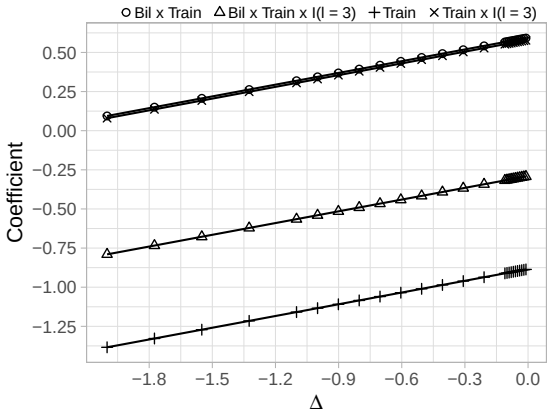


Figure 1: Effect size Δ versus power $\varpi(\tilde{\beta}(\Delta))$ for tests associated with the six log-odds ratios. Attempts 1 (circles), 2 (triangles), and 3 (squares) are displayed for (a) bilingual enumerators and (b) non-bilingual enumerators. The dashed line is the nominal significance of 0.10.

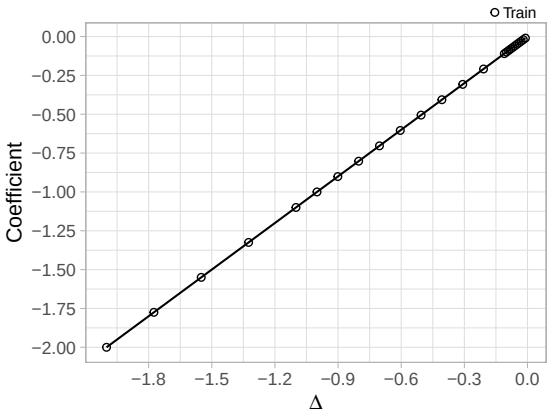


(a) Bilingual, $\ell = 1$.

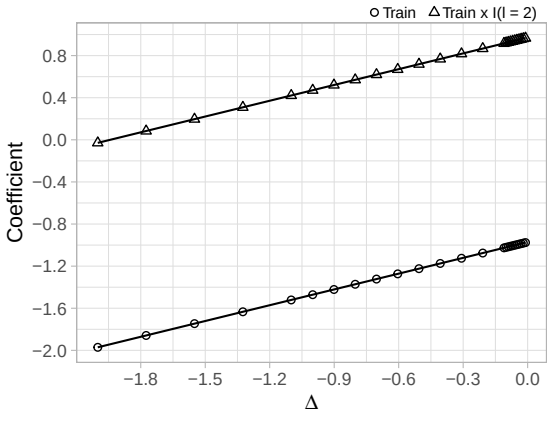
(b) Non-bilingual, $\ell = 1$.



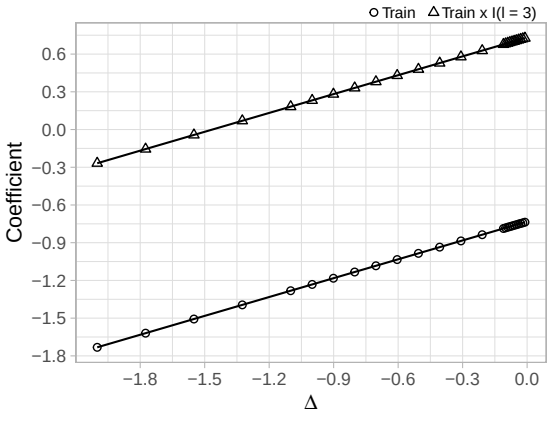
(c) Bilingual, $\ell = 2$.



(d) Non-bilingual, $\ell = 2$.



(e) Bilingual, $\ell = 3$.



(f) Non-bilingual, $\ell = 3$.

Figure 2: Variations in elements of $\tilde{\boldsymbol{\beta}}(\Delta)$ with Δ, corresponding to Figure 1. Labels at the top of each plot indicate which coefficients are plotted; labels not plotted remain fixed at their value in $\hat{\boldsymbol{\beta}}$.

6 Conclusions

We have studied the relationship between an enumerator training module for Spanish-language interviews and response rates for applicable households. Data from an experiment in the 2020 Census were analyzed using continuation-ratio logit multinomial regression. Six log-odds ratios were considered to express the training effect for both bilingual and non-bilingual enumerators at contact attempts 1, 2, and 3. Several of the estimates— corresponding to attempts 1 and 3 for bilingual enumerators and attempt 2 for non-bilingual enumerators—were significantly larger than zero, providing evidence of a positive association with response rate. Evidence of a negative association was not seen at a 0.10 significance level. We do not consider these findings to be surprising, as it had been anticipated that inclusion of the training should not harm response rate. We further suspect that the training may have a strong positive influence on response rate in certain contact situations; however, exploration of this may require controlling more carefully for variability—in the characteristics of enumerators and respondents and the circumstances of their interactions—than is possible using the current study and its resulting data.

Several methodological considerations arose in the course of this study which may be of interest in future work. We considered the use of random effects to adjust for variability in response rate due to enumerators. Data mapping each contact attempt to a particular enumerator were indeed available. However, with a large number of enumerators in the data, attempts to fit generalized linear mixed models proved to be computationally burdensome. Another consideration is that the ability of CRL model to capture a sequence of contacts is limited. Ideally, a maximum number of attempts—three in our case—should not need to be prespecified. Furthermore, it would be beneficial to account for contacts and responses via alternative modes such as internet, proxies, and administrative records, which are commonly used in modern official statistics operations.

Acknowledgements

This work has been reviewed for disclosure risk and approved with DRB clearance number CBDRB–FY24–CBSM003–004. We are grateful to a number of Census Bureau staff who helped to support this analysis. Thanks to Amy Fischer, Rhonda Cleveland, and Nelson Ur—as well as field staff across the country—for their support in implementing the experiment within the 2020 Census. We thank Sabin Lakhe, Valeria Trigo Vasconcellos, Stephen Jack Fisher, John Wilen, Venkatsubramaniam Chandrasekharan, and Christine Yun for support in assembling the analysis dataset and to access computational resources. Thanks to Thomas Mathew and Kimberly Sellers for discussions during planning of the experiment which helped to influence the methodology. Thanks to Luke Larsen, Thomas Mathew, and Tommy Wright for providing reviews of the manuscript.

References

George Casella and Roger L. Berger. *Statistical Inference*. Brooks/Cole, Cengage Learning, 2nd edition, 2002.

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

Ludwig Fahrmeir and Heinz Kaufmann. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. *The Annals of Statistics*, 13(1):342–368, 1985.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Mikelyn Meyers, Renee Ellis, Andrew Raim, Patricia Goerman, Kathleen Kephart, Kim Aspinwall, Patricia LeBaron, and Emilia Peytcheva. Evaluation of spanish-speaking enumerator training experiment, 2024+. (In preparation).

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Andrew M. Raim, Thomas Mathew, Kimberly F. Sellers, Renee Ellis, and Mikelyn Meyers. Design and sample size determination for experiments on nonresponse followup using a sequential regression model. *Journal of Official Statistics*, 39(2):173–202, 2023.

Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

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

Gerhard Tutz. Ordinal regression: A review and a taxonomy of models. *WIREs Computational Statistics*, 14(2), 2022.