Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment

by Michael R. Elliott, Brady T. West, Xinyu Zhang and Stephanie Coffey

Release date: June 21, 2022

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2022
Vol. 48, No. 1, pp. 25-48
Statistics Canada, Catalogue No. 12-001-X

The anchoring method: Estimation of interviewer effects in the absence of interpenetrated sample assignment

Michael R. Elliott, Brady T. West, Xinyu Zhang and Stephanie Coffey[1]

Abstract

Methodological studies of the effects that human interviewers have on the quality of survey data have long been limited by a critical assumption: that interviewers in a given survey are assigned random subsets of the larger overall sample (also known as interpenetrated assignment). Absent this type of study design, estimates of interviewer effects on survey measures of interest may reflect differences between interviewers in the characteristics of their assigned sample members, rather than recruitment or measurement effects specifically introduced by the interviewers. Previous attempts to approximate interpenetrated assignment have typically used regression models to condition on factors that might be related to interviewer assignment. We introduce a new approach for overcoming this lack of interpenetrated assignment when estimating interviewer effects. This approach, which we refer to as the "anchoring" method, leverages correlations between observed variables that are unlikely to be affected by interviewers ("anchors") and variables that may be prone to interviewer effects to remove components of within-interviewer correlations that lack of interpenetrated assignment may introduce. We consider both frequentist and Bayesian approaches, where the latter can make use of information about interviewer effect variances in previous waves of a study, if available. We evaluate this new methodology empirically using a simulation study, and then illustrate its application using real survey data from the Behavioral Risk Factor Surveillance System (BRFSS), where interviewer IDs are provided on public-use data files. While our proposed method shares some of the limitations of the traditional approach – namely the need for variables associated with the outcome of interest that are also free of measurement error – it avoids the need for conditional inference and thus has improved inferential qualities when the focus is on marginal estimates, and it shows evidence of further reducing overestimation of larger interviewer effects relative to the traditional approach.

Key Words: Clustering; Intraclass correlation; Design effects; Behavioral Risk Factor Surveillance System.

1. Introduction

Despite the best efforts of survey organizations to standardize the training of both face-to-face and telephone survey interviewers (Fowler and Mangione, 1989), numerous researchers have shown that estimates of key population parameters tend to vary between interviewers (e.g., Groves, 2004; Schnell and Kreuter, 2005; West and Olson, 2010; West and Blom, 2017). This variability may be due to verbal or nonverbal signals sent (likely unintentionally) by different interviewers, or by demographic features of the interviewer that reveal interviewer preferences and expectations (West and Blom, 2017). Even simpler factual items and self-administered items have been found to show variation across interviewers, despite the random assignment of respondents to interviewers (e.g., Kish, 1962; Groves and Magilavy, 1986; O'Muircheartaigh and Campanelli, 1998).

This intra-interviewer correlation, generally referred to as an *interviewer effect*, reduces the efficiency of survey estimates and decreases effective sample sizes given fixed survey costs in a manner similar to

1. Michael R. Elliott, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, MI 41809, University of Michigan Department of Biostatistics, School of Public Health, 1415 Washington Heights, Ann Arbor, MI 48109. E-mail: mrelliot@umich.edu; Brady T. West and Xinyu Zhang, University of Michigan Institute for Social Research, 426 Thompson St., Ann Arbor, MI 41809; Stephanie Coffey, US Census Bureau, 4600 Silver Hill Rd, Suitland-Silver Hill, MD 20746.

cluster sampling, due to the presence of a common effect across subjects that induces correlation. It can be conceptualized in statistical terms as a random effect common to all observations obtained by a given interviewer, whose variance is termed "interviewer variance". Accounting for this variance is critical to get correct statistical inference. In addition, as part of data collection monitoring, survey managers can use unbiased estimates of interviewer effects to identify interviewers that are having extreme effects on particular survey outcomes in real time and may need additional training to curb inappropriate behaviors.

A key assumption in the estimation of interviewer variance – whether via random effects models, or indirectly through use of generalized estimation equation/Taylor Series approaches – is interpenetrated sampling, or the random assignment of sampled cases to interviewers. Thus Schnell and Kreuter (2005) estimate interviewer effects in a face-to-face survey where interviewers are nested within PSUs and respondents within a PSU are randomly assigned to an interviewer, while O'Muircheartaigh and Campanelli (1998) use a cross-classified model in a design where respondents are randomly assigned to interviewers who worked in multiple PSUs. Interpenetrated sampling helps to ensure unbiased estimation of interviewer variance by ensuring there is no "spurious" variance introduced by certain types of respondents being more likely to be assigned to a given interviewer (e.g., older respondents being associated with interviewers working during the day), just as randomization ensures unbiased estimation of treatment effects in clinical trials. Unfortunately, interpenetrated sampling is logistically infeasible in many sample designs.

Recent studies of interviewer variance have adopted ad-hoc analytic approaches to "adjusting" for the effects of selected covariates that may introduce spurious correlation within interviewers based on non-interpenetrated sample designs (e.g., covariates describing features of sampling areas), claiming that any remaining variance in survey estimates across interviewers is mostly attributable to the interviewers (West and Blom, 2017). While this approach may in principle work to reduce spurious correlations between interviewers and outcomes if such covariates are available, it comes at the price of requiring conditional inference for the substantive variable of interest. This is particularly problematic if our goal is inference that properly accounts for interviewer effects in variance estimation without inappropriately adjusting for covariates that are not of interest. For example, if our interest is in the mean of a survey variable Y, $E(Y)=\mu,$ while appropriately accounting for the additional variance introduced by "clustering" from multiple interviewers conducted by a single interviewer, adjusting for multiple covariates $\left(X_1,\ldots,X_p\right)$ yields an estimator of β_0 under the model $E(Y)=\beta_0+\sum_{k=1}^{p}\beta_k x_k$. It is clear that $\mu\neq\beta_0$ unless either $\beta_1=\ldots=\beta_p=0$ (in which case there cannot be adjustment for spurious correlations between interviewers and outcome), $E\left(X_1\right)=\ldots=E\left(X_p\right)=0,$ or there is some extremely unlikely cancellation of regression components. (For readers familiar with causal inference, this is somewhat analogous to marginal structural models (MSMs), which avoid using confounders in a regression model while still accounting for confounding, Joffe, Ten Have, Feldman and Kimmel (2004), although our approach is fully model-based rather than model-assisted as in MSMs.) While centering the covariates can guarantee the second condition in the absence of interactions, this is not always desirable or noted, and even if doable may not

leave the remaining residuals with the desired distributional characteristics. With the present study, we aim to provide survey researchers with a means to estimate interviewer variance (either to improve the quality of estimates or inform survey operations) in the absence of interpenetration without conditioning on covariates in the traditional manner.

Our approach, which we refer to as the "anchoring" method, leverages correlations between observed variables that are unlikely to be affected by interviewers ("anchors") and variables that may be prone to interviewer effects (e.g., sensitive or complex factual questions) to statistically remove components of within-interviewer correlations that a lack of interpenetrated assignment may introduce. The improved estimates of interviewer effects on survey measures will increase the ability of survey analysts to correct estimates of interest for interviewer effects, and enable survey managers to adaptively manage a data collection in real time and intervene when particular interviewers are producing survey outcomes that vary substantially from expectations.

In Section 2, we provide some background on the important problem of interviewer variance, as well as a discussion of its estimation and impact on inference. In Section 3, we introduce the anchoring method and its development in a frequentist and Bayesian framework, as well as the heuristic interpretation and issues related to choice of variables. In Section 4 we empirically evaluate the properties of this new method using a simulation study, and in Section 5 we illustrate the method using real data from the Behavioral Risk Factor Surveillance System (BRFSS). In Section 6 we provide concluding remarks as well as some discussion of implementation and monitoring of the method in practise.

2. Background

2.1 Interviewer variance

Between-interviewer variance affects survey estimates in a manner similar to the design effects introduced by cluster sampling. One can estimate the multiplicative increase in the total variance of an estimated mean as $\text{deff} = 1 + \rho_{\text{int}}(m-1)$, where m is the average number of interviews conducted by individual interviewers and ρ_{int} is the within-interviewer correlation in answers elicited to a particular survey question (Kish, 1965). Typical values of 35 respondents per interviewer and 0.03 for ρ_{int} would therefore *double* the estimated variance of the mean, relative to the variance with ρ_{int} equal to zero. Failure to account for the within-interviewer correlation introduced by interviewer effects leads to *misspecification effects* (Skinner, Holt and Smith, 1989), resulting in anti-conservative inference due to underestimation of standard errors.

2.2 Estimation of interviewer variance

Researchers may wish to estimate interviewer variance for correct statistical inference (Elliott and West, 2015), to identify interviewers having unusual effects on data collection outcomes for purposes of responsive survey design, or as the focus of a methodological study designed to reduce its impact by

understanding its causes (e.g., Brunton-Smith, Sturgis and Williams, 2012; Sakshaug, Tutz and Kreuter, 2013). Interpenetrated designs, which assign sampled cases to interviewers at random, allow for interviewer variance to be accounted for using standard methods that account for clustering in the observed data: generalized estimating equations (Liang and Zeger, 1986) or mixed-effects models (Laird and Ware, 1982; Stiratelli, Laird and Ware, 1984). Temporarily ignoring sampling weights, a simple model for a normally-distributed variable of interest that accounts for interviewer variance is

$$Y_{ijk} = \mu + a_i + b_{ij} + \varepsilon_{ijk},\ a_i \sim N\left(0, \sigma_a^2\right),\ b_{ij} \sim N\left(0, \sigma_b^2\right),\ \varepsilon_{ijk} \sim N\left(0, \sigma^2\right), \tag{2.1}$$

where i indexes a primary sampling unit (PSU), j indexes the interviewer within the i^{th} PSU, and k the respondent associated with the j^{th} interviewer in the i^{th} PSU. Assuming that all of the error terms are independent, that there are an average of J interviewers in each of the I PSUs, and that there are an average of K interviews per interviewer, the variance of the mean estimator $\hat{\mu} = \bar{y}$ is approximately inflated by a factor of $1 + \rho_a (JK - 1) + \rho_b (K - 1)$, where $\rho_a = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma^2}$ and $\rho_b = \frac{\sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma^2}$. As a practical matter, when the variance of $\hat{\mu}$ is the only quantity of interest, the second stage of clustering due to an interviewer can be ignored, as in an "ultimate cluster" design (Kalton, 1983). Treating the random effect of the PSU as $\tilde{a}_i = a_i + \sum_{j=1}^{J} b_{ij}$ with variance $\sigma_{\tilde{a}}^2 = \sigma_a^2 + J\sigma_b^2$, the variance of the mean estimator $\hat{\mu}$ is inflated by a factor of $1 + \rho_{\tilde{a}} (JK - 1)$, where $\rho_{\tilde{a}} = \frac{\sigma_{\tilde{a}}^2}{\sigma_{\tilde{a}}^2 + \sigma^2}$.

If multiple interviewers are nested within a single PSU as assumed in (2.1), interviewer variances can still be estimated for methodological purposes using multistage hierarchical linear models. However, for reasons of cost efficiency, many area probability samples require a given interviewer to restrict their efforts to a single sampling area (e.g., the U.S. National Survey of Family Growth; see Lepkowski, Mosher, Groves, West, Wagner and Gu, 2013), which completely aliases the components of variance due to interviewers and areas. Such designs preclude any type of direct estimation of interviewer variance, although from a purely analytic perspective, accounting for clustering using the PSU IDs in analysis will account for the additional interviewer variance introduced.

For other types of surveys – and in particular telephone surveys – this "automatic" accommodation of interviewer effects at the variance estimation stage afforded by "ultimate cluster" approaches does not occur. A spectacular example of this is the Behavioral Risk Factor Surveillance System (BRFSS; Centers for Disease Control, 2013), a massive annual telephone survey sponsored by the Centers for Disease Control that is the only Federal health survey designed to provide state-level estimates of key health factors such as smoking rates, obesity measures, and cancer screening. Elliott and West (2015) found no evidence that any substantial proportion of the 1,000 + manuscripts published using BRFSS data accounted for interviewer effects when conducting variance estimation based on these data, despite variance inflation factors of 10 or more at the state level for estimates such as mean self-rated health. These authors found evidence of substantial interviewer effects for selected survey items, and variability in the variance of these effects themselves across states, when applying both model-based and design-based approaches to estimate the variance (although this analysis used naïve estimators in contrast to

either the standard regression or the anchoring methods discussed here, and so may have overestimated this variance).

Importantly, secondary analysts still do not know for sure if these components of variance are arising due to sampling variability, true measurement error introduced by the interviewers, or differential non-response among the interviewers. Because of the design effect definition noted above, their impact on inference can still be large even if the intra-class correlation (ICC) is small or moderate, since interviewers typically conduct many interviews. Thus when Groves and Magilavy (1986) found mean ICCs between 0.002 and 0.02 among 25 to 55 variables across each of nine telephone surveys of political, health, and economic issues, the design effect would range between 1.04 and 1.38 for studies in which interviewers average 20 interviews each, and between 1.10 and 1.98 if interviewers average 50 interviews each. Some outcomes can have much higher ICCs – Cernat and Sakshaug (2021) found ICCs on the order of 0.30 for biometric measures, which would yield design effects on the order of 15 if 50 interviews were conducted per interviewer. Although interviewer variance studies for face-to-face data collections tend to be rare because interpenetrated sample designs are more difficult to implement in such settings, Schnell and Kreuter (2005) found a median overall design effect of 2.0 in a multi-stage sample survey on fear of crime, which was mostly attributable to interviewer effects rather than spatial clustering. Thus the need for analysts to accommodate interviewer effects is clear.

2.3 Accounting for interviewer variance in inference in the absence of interpenetration

As noted in Section 2.2, when interviewers are nested within PSUs, standard methods of variance estimation based on "ultimate clusters" (Kalton, 1983) that account for the dependence of observations within a PSU will "automatically" absorb measurement error due to interviewers into the within-PSU correlation. However, whenever interviewers are not nested within PSUs – as can occur in some area probability samples where interviewers cross sampling unit segments (e.g., O'Muircheartaigh and Campanelli, 1998; Vassallo, Durrant and Smith, 2017) – clustering induced by interviewer effects must be accounted for directly. In such situations, cross-classified random effects models (Rasbash and Goldstein, 1994) of the form

$$E\left(Y_{hij}\right)=\theta+a_h+b_i,\ a_h\sim N\left(0,\tau_a^2\right),\ b_i\sim N\left(0,\tau_b^2\right) \tag{2.2}$$

may be employed, where h indexes PSUs, i indexes interviewers, and j indexes interviews conducted by the i^{th} interviewer (e.g., O'Muircheartaigh and Campanelli, 1998; Schnell and Kreuter, 2005; Biemer, 2010; Durrant, Groves, Staetsky and Steele, 2010). Extensions of these models are also possible for non-linear link functions using generalized linear mixed models (e.g., Vassallo et al., 2017).

Unfortunately, interpenetration can fail, either due to differential non-response error among interviewers (West and Olson, 2010; West, Kreuter and Jaenichen, 2013), non-random shift assignment (e.g., with daytime interviewers more likely to interview non-working respondents), or other common

practices used to increase response rates, such as assigning experienced interviewers to more difficult respondents (Brunton-Smith et al., 2012). In the absence of interpenetration, standard methods to account for interviewer variance may lead to "spurious" correlations within interviewers that have nothing to do with interviewer-induced measurement error.

The literature is not completely devoid of approaches for estimating (and accommodating) interviewer variance in non-interpenetrated sample designs. Fellegi (1974), Biemer and Stokes (1985), Kleffe, Prasad and Rao (1991), and Gao and Smith (1998) developed statistical methods for area probability samples that assumed interpenetration for a random subset of PSUs, and a single interviewer in each of the remaining PSUs. More recent work has considered methods for estimation of interviewer variance in binary survey variables in related settings of *partial interpenetration* (von Sanden and Steel, 2008). Rohm, Carstensen, Fischer and Gnambs (2021) used a two-parameter item response theory model to separate area and interviewer effects under this assumption, which de-confounds interviewer and area effects to the extent that each interviewer recruits in multiple areas and vice versa (although lack of random assignment within an area can still yield some degree of variance component bias). These methods are useful for obtaining estimates of interviewer variance separate from area homogeneity for purposes of assessing the independent impact of such variance. However, they are not relevant for our more general setting of interest, where interviewers may not cross PSUs and are not working random subsamples of the full sample (i.e., no interpenetration).

Another common method found in the literature for grappling with the problem of non-interpenetrated sample designs when estimating interviewer variance is adjustment for the effects of respondent- and area- or interviewer-level covariates in multilevel models (Hox, 1994; Schaeffer, Dykema and Maynard, 2010; West, Kreuter and Jaenichen, 2013). These methods are largely ad-hoc, and rely on the assumption that the included covariates adequately account for all sources of variability that arise from the areas (and would thus be attributed to the interviewers if the covariates were not accounted for). This approach suffers from two major shortcomings. First, many studies, and especially those relying on publicly available data, may not contain sufficient area- or interviewer-level covariate information to adequately account for the lack of randomization in interviewer assignment. Second, the resulting estimators condition on these covariates, and these conditional estimators are typically not of interest, with the focus being on either marginal estimates of descriptive parameters, such as means or totals, or parameters in models that typically do not condition on (or include) covariates.

3. The anchoring method

As noted in Section 2.3, existing methods adjust for possible interviewer effects introduced at the recruitment and measurement stages of data collection by including respondent- and area- or interviewer-level covariates in multilevel models (Hox, 1994), but such adjustment may be erroneous if part of the interviewer variance is simply arising due to non-interpenetrated sampling. As noted by Elliott and West

(2015), if subjects with similar values on a variable of interest are assigned to interviewers in a non-random fashion – for example, if a telephone interviewer working day shifts tends to interview older respondents, where age may be correlated with main variables of interest – these variables will be correlated with specific interviewers. However, we are just re-ordering the random sample, not introducing measurement error in the manner described in Section 1, e.g., West and Blom (2017). Thus the actual data are not being altered, and there are no true interviewer effects: we term the resulting within-interviewer correlation "spurious" from a variance inflation perspective. Thus estimating interviewer effects while failing to account for differential sample assignment can lead to conservative inferences, resulting in misleadingly large estimates of interviewer variance, p-values and confidence intervals that are too wide, and incorrect operational decisions based on predicted effects for individual interviewers.

To address this important gap in the literature, we describe an "anchoring" method that analysts can use to estimate the unique components of variance due to interviewer effects on selection and measurement. The method aims to leverage correlations between variables where interviewer measurement error is of concern and variables known – or reasonably believed – to be free of measurement error to remove the fraction of the within-interviewer correlation that is due to non-interpenetrated sample assignment. In the simplest case, if we have two variables, one (Y_1) treated as measurement error-free (the "anchor") and one (Y_2) treated as possibly having interviewer-induced measurement error, and our objective is to estimate a mean of Y_2, we fit a multilevel model to the observed data for the two variables that includes a random interviewer effect only for the variable subject to measurement error:

$$y_{ijk} = \mu_k + I(k=2)\, b_i + \varepsilon_{ijk}. \tag{3.1}$$

In (3.1), $i = 1, \ldots, I$ indexes interviewers, $j = 1, \ldots, J_i$ indexes respondents within interviewers, $k = 1, 2$ indexes the variable (1 = anchor, 2 = variable of interest), $b_i \sim N(0, \sigma_b^2)$ is the interviewer effect, and

$$\begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix} \right).$$

Our focus of inference in this manuscript is μ_2, although σ_b^2 or b_i may also be of interest if the focus is on interviewer variance or determining individual interviewers who are contributing to that variance.

To provide a heuristic explanation of why this proposed "anchoring" approach works, assume that y_{ij1} and y_{ij2} net of b_i are almost perfectly correlated. Since y_{ij1} lacks measurement error, it can serve as a proxy for the non-measurement error component of y_{ij2}, absorbing artificial error in y_{ij2} induced in the ordering of the data. Lack of interpenetration means that estimating a linear mixed model using y_{ij2} only will yield an upwardly biased estimate of σ_b^2. If $\sigma_{12} > 0$, information will be available to reduce the bias in $\hat{\sigma}_b^2$, with large samples and high correlations between ε_{ij1} and ε_{ij2} yielding increasingly accurate estimates of σ_b^2 and thus of the true impact of the interviewer-induced measurement error on the variance of $\hat{\mu}_2$.

This approach easily extends to the setting where $K-1 \geq 2$ "anchoring" variables free from measurement error are available:

$$y_{ijk} = \mu_k + I\left(k = K\right) b_{iK0} + \varepsilon_{ijk},\ i = 1, \ldots, I,\ j = 1, \ldots, J_i, k = 1, \ldots, K. \tag{3.2}$$

In this case, the first $K-1$ variables are assumed to be free of interviewer measurement error and the K^{th} variable is the variable of interest, $b_{iK0} \sim N\left(0, \tau^2\right)$, and $\left(\varepsilon_{ij1} \ldots \varepsilon_{ijK}\right)^T \sim N_K\left(0, \Sigma\right)$, where Σ is an unstructured $K \times K$ covariance matrix. Alternatively, instead of using (3.2) directly, we can reduce (3.2) back to the bivariate setting in (3.1) by replacing Y_{1i} with the best linear predictor of Y_{Ki} using the anchoring variables: $\hat{Y}_{Ki} = E\left(Y_{Ki} \middle| Y_{1i}, \ldots, Y_{K-1,i}\right) = \hat{\boldsymbol{\beta}}^T \mathbf{X}_i$ where $\mathbf{X}_i = \left(Y_{1i}, \ldots, Y_{K-1,i}\right)^T$ and $\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y}_K$.

3.1 Estimation remarks

One can use standard linear mixed model software (e.g., SAS PROC MIXED) to fit the models in (3.1) or (3.2) and obtain a restricted maximum likelihood (REML) point estimate $\hat{\mu}_2$ together with an associated variance estimate. We have provided an annotated example of such code in the supplemental materials. Weights used to account for unequal probabilities of selection, non-response adjustment, and calibration to known population values can be incorporated using pseudo-maximum likelihood estimation (PML; Pfeffermann, Skinner, Holmes, Goldstein and Rasbash, 1998; Rabe-Hesketh and Skrondal, 2006) when fitting the models in (3.1) or (3.2). We would generally recommend that interviewers be assigned a weight of 1 when fitting weighted multilevel models of these forms, to mimic the notion of simple random sampling of interviewers from a hypothetical population of interviewers. The weights for respondents should be rescaled to sum to the final respondent count for each interviewer (Carle, 2009), and extensions of the PML method outlined by Veiga, Smith and Brown (2014) and Heeringa, West and Berglund (2017, Chapter 11) can be used to incorporate the rescaled weights in estimation of the residual covariance structure in (3.1) or (3.2). In multistage samples where interviewers cross geographic areas, cross-classified random effects models (Rasbash and Goldstein, 1994) can also be utilized.

3.2 The Bayesian anchoring method

In the presence of prior information on the parameters of interest in this model (e.g., in a repeated cross-sectional survey using interviewer administration), the models in (3.1) or (3.2) can also be fitted using a Bayesian approach to incorporate the prior information. In repeated surveys that carefully monitor interviewer performance, good predictions of individual interviewer effects based on the estimated variance component are important. Given historical data from a survey with the same essential design conditions, one can estimate the parameters of interest in (3.1) using this historical data, and then define informative prior distributions for these parameters. (Examples of these types of surveys would include high-quality government sponsored surveys with repeated cross-sectional data collections, such as the National Health Interview Survey or, for the example considered in this paper, the Behavioral Risk Factor

Surveillance System.) Specifically, we consider a prior distribution on the interview effect standard deviation σ_b that follows a half t distribution (Gelman, 2006) with degrees of freedom ν and standard deviation s:

$$p\left(\sigma_b \mid \nu, s\right) = \frac{2\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi s^2}}\left(1+\frac{\sigma_b^2}{\nu s^2}\right)^{-\frac{\nu+1}{2}}, \quad \tau \geq 0. \tag{3.3}$$

Following Gelman, we assume $\nu = 3$, and we estimate s based on prior estimates of interviewer effects. We consider standard weak priors for the fixed effect means: $p\left(\mu_k\right) \overset{\text{ind}}{\sim} N\left(0, 10^6\right)$ and for the residual variance:

$$p\begin{pmatrix}\sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2\end{pmatrix} \sim \text{INV} - \text{WISHART}\left(2, I\right).$$

This approach offers advantages relative to likelihood ratio testing approaches that rely on asymptotic theory, particularly for smaller samples. By using prior information to constrain the resulting posterior distribution for the interviewer variance components, it generally prevents extremely large draws of the variance component while not constraining the means or residual variances. It also constrains posterior draws of variance components to be greater than zero, enabling inference based on small components of variance, while frequentist model-fitting procedures generally fix such estimates of variance components to be exactly zero (which equates to a rather unreasonable assumption that each interviewer produces exactly the same survey estimate; West and Elliott, 2014). In such cases, the effects of interviewers (even if they are small) would be ignored completely; the Bayesian approach would still enable small effects to be integrated into the inference. The Bayesian approach also yields credible intervals for the interviewer variance components based on posterior draws.

3.3 Choosing anchoring variables

A key assumption underlying both the standard regression-based approach and the "anchoring" method is that selected variables are free from interviewer-induced error. Like the "missing at random" assumption in the missing data literature, we do not expect that there will often be cases where we can be certain of this, but that approximations may be available based on simple demographic measures (e.g., age) or other factual questions with simple response options (e.g., current employment) and little room for the introduction of interviewer error. The identification of error-free covariates in advance of data collection is an important substantive and methodological component of this approach, and prior methodological literature on the variables most prone to interviewer effects (West and Blom, 2017) can be consulted for this component of the approach.

As we note above, if we have multiple error-free covariates measured on the respondents, we can preserve their predictive power (and thus the correlation of the anchor's residuals with the residuals of the variable of interest) by computing a linear predictor of the variable of interest from a linear model that includes fixed effects of all of the error-free covariates. We consider such an approach in our simulation

studies and applications, and compare it with the "standard" approach of simply adjusting for these covariates in a multilevel model in an effort to improve the estimate of the interviewer variance component (Hox, 1994).

Finally, the anchoring approach employs mixed-effects models that should yield correct estimates with a sufficient amount of data. However, these models may be more difficult to fit, especially for smaller samples, and we therefore also consider alternative Bayesian approaches when evaluating the anchoring approach.

4. Simulation study

We first consider an empirical simulation study of the proposed "anchoring" approach. We repeatedly simulated samples of data from a quadrivariate normal distribution, $\left(Y_{1ij}^{*}\ Y_{2ij}^{*}\ Y_{3ij}^{*}\ Z_{ij}\right) \sim N_4\left(\boldsymbol{\mu}, \Sigma\right)$, where $j = 1, \ldots, J = 30$ indexes hypothetical respondents nested within $i = 1, \ldots, I = 30$ interviewers, $Y_{kij(z)} = Y_{kij(z)}^{*} + I\left(k = 3\right) b_i$ for $b_i \sim N\left(0, \sigma_b^2\right)$, and $Y_{kij(z)}^{*}$ is ordered by the values of Z_{ij} prior to assignment of respondents to the 30 interviewers. $Y_{1ij(z)}$ and $Y_{2ij(z)}$ are the observed values without interviewer-induced measurement error, while $Y_{3ij(z)}$ is observed with interviewer-induced measurement error, and Z_{ij} is a (nuisance and unobserved) covariate that induces extraneous variability when the design is treated as interpenetrated. (One might think of Y_1 and Y_2 as measurement-error free demographic variables and Y_3 as a continuous self-reported overall health measure, which is potentially prone to interviewer effects, and Z as amount of time spent at home, which is associated with interviewer scheduling by shift.)

Given this data generating model, we note that a higher correlation of Z with the other measurements will introduce what appears to be interviewer variance because of the ordering of $Y_{kij(z)}^{*}$ by the values of Z_{ij} above and beyond the true random interviewer effects on Y_2 (given by b_i). This is the lack of interpenetrated assignment that we wish to adjust for with the proposed anchoring method, which aims to isolate the unique interviewer variance σ_b^2 that does not arise from simple assignment of cases to interviewers. For simplicity, we assume that $\mu_{Y_1} = \mu_{Y_2} = \mu_{Y_3} = \mu_z = \mu$, $\sigma_{Y_1}^2 = \sigma_{Y_2}^2 = \sigma_{Y_3}^2 = \sigma_Z^2 = 1$ and $\rho_{Y_1Y_2} = \rho_{Y_1Y_3} = \rho_{Y_1Z} = \rho_{Y_2Y_3} = \rho_{Y_2Z} = \rho_{Y_3Z} = \rho$.

We consider four models used to estimate the mean of Y_3 and the associated interviewer effect variance:

Unadjusted: $Y_{3ij} \sim N\left(\mu_3 + b_i, \sigma_3^2\right)$

Adjusted: $Y_{3ij} \sim N\left(\mu_3 + \beta_1 y_{1ij} + \beta_2 y_{2ij} + b_i, \sigma_3^2\right)$

Anchoring: $\begin{pmatrix} Y_{1ij} \\ Y_{2ij} \\ Y_{3ij} \end{pmatrix} \sim N_3\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 + b_i \end{bmatrix}, \Sigma\right)$

Anchoring-Linear Predictor: $\begin{pmatrix} \hat{Y}_{3ij} \\ Y_{3ij} \end{pmatrix} \sim N_2 \left(\begin{bmatrix} \mu_1 \\ \mu_2 + b_i \end{bmatrix}, \Sigma \right)$

where y_{kij} is the observed realization of Y_{kij}, $b_i \sim N\left(0, \sigma_b^2\right)$, and, in the anchoring-linear predictor model, $\hat{Y}_{3ij} = \hat{\beta}_0 + \hat{\beta}_1 y_{1ij} + \hat{\beta}_2 y_{2ij}$ where $\hat{\boldsymbol{\beta}}$ is obtained from the linear regression of Y_3 on Y_1 and Y_2. We estimate the mean of Y_3 as the REML estimator of μ_3 and similarly the associated interviewer effect variance as the REML estimator of σ_b^2.

We consider the power to reject the null hypothesis that the mean of the observed variables is zero (at the 0.05 level) and the empirical bias in the estimation of the variance of the random interviewer effects, σ_b^2. We evaluated the empirical bias by computing the difference between the mean of the simulated estimates of the variance component and the true value of the variance component specific to a given simulation scenario. Our simulation study design considers a full factorial design where $\mu = \{0, 0.5\}$, $\rho = \{0.25, 0.5, 0.75\}$, and $\sigma_b^2 = \{0.1, 0.5, 0.9\}$. We generated 200 independent simulations for each of the 18 cross-classifications of values on these parameters. Table 4.1 presents the results of the simulation study.

Table 4.1
Results of the empirical simulation study. Best performing method italicized (note that when $\mu = 0$, ideal power if 0.05)

True values of model parameters			Power: $H_0 : \mu = 0$ *vs.* $H_A : \mu \neq 0$				Empirical Bias of $\hat{\sigma}_b^2$			
μ	ρ	σ_b^2	Unadjusted	Adjusted	Anchor	Anchor-Linear Predictor	Unadjusted	Adjusted	Anchor	Anchor-Linear Predictor
0	0.25	0.1	0.03	*0.04*	*0.04*	*0.04*	0.063	0.029	*0.027*	*0.027*
0	0.25	0.5	0.03	0.08	*0.04*	*0.04*	0.070	*0.022*	0.033	0.032
0	0.25	0.9	0.07	*0.04*	*0.06*	*0.06*	0.078	*0.037*	0.044	0.043
0	0.5	0.1	0.00	*0.03*	0.02	0.02	0.255	0.061	*0.056*	*0.056*
0	0.5	0.5	0.01	*0.04*	0.03	0.03	0.247	0.058	*0.054*	*0.053*
0	0.5	0.9	0.02	*0.04*	*0.04*	*0.04*	0.251	*0.049*	0.061	0.061
0	0.75	0.1	0.00	*0.02*	0.01	0.01	0.568	*0.074*	0.078	0.076
0	0.75	0.5	0.00	0.04	*0.05*	*0.05*	0.555	0.098	*0.084*	*0.084*
0	0.75	0.9	*0.04*	*0.04*	*0.06*	*0.06*	0.602	*0.099*	0.103	0.103
0.5	0.25	0.1	*1.00*	*1.00*	*1.00*	*1.00*	0.069	*0.025*	0.032	0.032
0.5	0.25	0.5	*0.96*	0.68	*0.96*	*0.96*	0.072	0.044	*0.034*	*0.034*
0.5	0.25	0.9	*0.76*	0.48	0.75	0.75	0.075	0.040	*0.039*	*0.037*
0.5	0.5	0.1	*1.00*	0.87	*1.00*	*1.00*	0.261	0.062	0.062	*0.061*
0.5	0.5	0.5	0.92	0.44	*0.96*	*0.96*	0.269	*0.062*	0.067	0.067
0.5	0.5	0.9	0.75	0.24	*0.80*	*0.80*	0.248	0.068	0.064	*0.063*
0.5	0.75	0.1	*1.00*	0.62	*1.00*	*1.00*	0.567	0.079	0.078	*0.077*
0.5	0.75	0.5	0.81	0.27	*0.96*	*0.96*	0.507	0.103	*0.082*	*0.082*
0.5	0.75	0.9	0.58	0.22	*0.70*	*0.70*	0.598	*0.100*	0.106	0.106

Several notable patterns emerge from the simulation results in Table 4.1. First, as the values of ρ increase, the anchoring method produces larger reductions in the overestimation of interviewer variance relative to the unadjusted model. Recall that this was expected by design, given the initial ordering of the observations by Z prior to assignment to interviewers, which introduces artificial variance among the interviewers. Similarly, as anticipated, estimation of the interviewer variance using covariate adjustment is similar to the anchoring method when this variance is not large, although there is evidence of a somewhat larger reduction in bias when the variance is large.

In addition, for the non-zero values of μ, higher values of ρ yield larger improvements in power when using the anchoring method when compared with the unadjusted estimator, since more of the extraneous variance is correctly allocated. Both the unadjusted and an anchoring method yield higher power than the adjusted estimator, since the adjusted estimator is biased for non-zero means of Y_{1ij} and Y_{2ij} when they are correlated with Y_{3ij}. Smaller values of ρ approximate an interpenetrated design, and as a result, the unadjusted estimation approach does not produce substantially different results from the adjusted or anchoring approach. The empirical bias in the estimation of σ_b^2 is unrelated to the value of σ_b^2 but is entirely a function of ρ, since that drives the spurious within-interviewer correlation due to the unobserved Z. Finally, we note that replacing the actual values of Y_{1ij} and Y_{2ij} with a summary measure based on their linear prediction of Y_{3ij} yields virtually identical results to their direct use in the anchoring method. This is partly a function of their common normality; we discuss this limitation in the Discussion section below.

5. Application to the Behavioral risk factor surveillance system

To further illustrate the implementation of our proposed approach, we analyze data from the 2011 and 2012 Behavioral Risk Factor Surveillance System (BRFSS; https://www.cdc.gov/brfss/index.html). The BRFSS is a major national health survey in the U.S. that employs interviewer administration via the telephone, and is one of the few national surveys that provides data users with interviewer identification variables in the public-use versions of its data sets (Elliott and West, 2015). This enables the estimation of interviewer variance components for any BRFSS measures. We only use data from the publicly available data files for these two years in this study.

For illustration purposes, we consider the case where the variable of interest (Y_2) is perceived health status (1 = poor, …, 5 = excellent). We define an "anchoring" variable (Y_1) as the linear predictor of perceived health status from a linear regression model fitted using OLS that includes age, an indicator of obtaining a college degree, an indicator of being a female, and an indicator of white race/ethnicity as covariates. We chose these respondent-level covariates for this application for three reasons: 1) we believe that they are likely to be reported with minimal differential measurement error among interviewers (West and Blom, 2017); 2) they are associated with interviewer assignment, as telephone interviewers tend to work calling shifts at different times of the day, and interview time of day is associated with age and

education (e.g., older respondents and respondents with lower levels of education may be more likely to be interviewed during the day); and 3) they also tend to be correlated with perceived health status (Franks, Gold and Fiscella, 2003).

As part of the application, we also compare the ability of the anchoring method based on this linear predictor to reduce estimates of variance components to that of the more "standard" method that is often used in practice: simply adjusting for these respondent-level covariates in a multilevel model, in an effort to adjust for the fixed effects of these covariates when evaluating the interviewer variance component (Hox, 1994). We make two remarks about this approach, specifically with respect to this application:

1. Centering of the covariates at their means (whether they are binary or continuous) is critical to this approach if inference is focused on the mean of Y_2, as a failure to do this will lead to biased "conditional" estimates of the mean on that variable that depend on the values of the covariates (rather than the overall mean). This is not relevant for the anchoring method.
2. In some cases interviewer-level covariates could be expected to explain more of the artificial interviewer variance due to non-interpenetrated assignment than respondent-level covariates (e.g., area-level socio-demographic information; Hox, 1994; West and Blom, 2017). However, the BRFSS does not provide any interviewer-level covariates.

5.1 Frequentist approach

We considered both frequentist and Bayesian approaches in our analysis, and performed separate analyses of the BRFSS data from each of the 50 states and the District of Columbia for each approach. We only retained cases with complete data on all analysis variables of interest to ensure a common case base no matter the type of analysis being performed. First, in the frequentist approach, we started by estimating means of self-reported health from a given state that assumed independent and identically distributed (i.i.d.) data (i.e., ignoring random interviewer effects):

$$Y_{ij2} = \mu_2 + \varepsilon_{ij2}, \quad \varepsilon_{ij2} \sim N\left(0, \sigma_2^2\right). \tag{5.1}$$

We then fit a "naïve" mixed-effects model including random interviewer effects (of the form in (2.1) but without random PSU effects, given the absence of PSUs in the BRFSS design) to the self-reported health data (ignoring the other covariates), assuming interpenetrated sample assignment within each state:

$$Y_{ij2} = \mu_2 + b_i + \varepsilon_{ij2}, \quad b_i \sim N\left(0, \sigma_b^2\right). \tag{5.2}$$

We estimated the interviewer variance component based on this model and tested the variance component for significance using a mixture-based likelihood ratio test (West and Olson, 2010). We also evaluated the ratio of the estimated variance of mean self-reported health when naively accounting for the interviewer effects to the variance of the mean assuming simple random sampling (i.e., i.i.d. data). The

literature generally refers to this ratio, shown in (5.3), as an "interviewer effect" on a particular descriptive estimate:

$$\text{IntEff}_{\text{naive}} = \frac{\text{var}_{\text{naive}}\left(\hat{\mu}_2\right)}{\text{var}_{\text{iid}}\left(\hat{\mu}_2\right)}. \tag{5.3}$$

Next, after fitting a linear regression model to the perceived health status variable and computing the linear predictor of perceived health status based on the estimated coefficients (denoted in (5.4) by y_{ij1}), we fit the model in (3.1) to implement the anchoring approach:

$$\begin{aligned} y_{ij1} &= \mu_1 + \varepsilon_{ij1} \\ y_{ij2} &= \mu_2 + b_i + \varepsilon_{ij2} \\ b_i &\sim N\left(0, \sigma_b^2\right) \\ \begin{pmatrix} \varepsilon_{ij1} \\ \varepsilon_{ij2} \end{pmatrix} &\sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}\right). \end{aligned} \tag{5.4}$$

Here $Y_{ij1} = \hat{\beta}_0 + \sum_p \hat{\beta}_p x_{ip}$, where $\hat{\boldsymbol{\beta}}$ is obtained from the linear regression of the p anchoring covariates (of which there are four in this application). We then computed the same ratio in (5.3) based on the anchoring approach, where anchoring would be expected to reduce the bias in the estimate of the interviewer effect that would be arising from the naïve approach.

Next, we fitted a model representing the "standard" adjustment approach (Hox, 1994) as follows:

$$Y_{ij2} = \mu_2 + \sum_p \beta_p x_p + b_i + \varepsilon_{ij2},\ b_i \sim N\left(0, \sigma_b^2\right). \tag{5.5}$$

In (5.5), the x_p represent the centered respondent-level covariates indexed by p (the same four anchoring covariates as in (5.4)), with corresponding fixed effects. We once again computed the ratio in (5.3) representing the estimated interviewer effect for comparison with the other approaches. To keep the focus on the potential reduction in bias in the estimation of the interviewer effect, we ignored sampling weights in these analyses.

5.2 Bayesian approach

Next, in the Bayesian approach, we applied the same types of comparative analyses to evaluate the anchoring method, varying whether prior information about the interviewer variance component from the 2011 BRFSS was used (yes or no). This prior information came from implementing the anchoring approach with the same linear predictor in 2011 to determine a prior estimate of the interviewer variance component. In all cases, we assumed non-informative prior distributions for the fixed effects (which recall from (3.1) define the means of the two variables) and the residual variances and covariances in the models.

We defined an informative prior distribution for the standard deviation of the random interviewer effects using (3.3), where the standard deviation s is given by the estimated standard deviation of the random interviewer effects for the same state in 2011, and used the weak priors on μ and Σ defined in Section 2.3. We implemented the Bayesian approach using PROC MCMC in the SAS software, and annotated examples of the code used are available in the supplemental materials.

5.3 Results

Figure 5.1 presents four scatter plots, enabling comparisons of the naïve estimates of the interviewer effects on the mean of perceived health status for each of the 50 states and the District of Columbia with the adjusted estimates based on the anchoring method, the "standard" adjustment method, and the two alternative Bayesian approaches to implementing the anchoring method. All estimates of interviewer effects were computed using (5.3).

The plots vary in terms of the methods used to implement the estimation approaches. We first consider a plot of the adjusted estimates of the interviewer effects based on the anchoring method against the naïve estimates of the interviewer effects from (5.3), using the frequentist approach described above (Figure 5.1a). The next plot (Figure 5.1b) presents the adjusted estimates based on the "standard" adjustment approach of including the covariates in a multilevel model. The third plot (Figure 5.1c) considers the first Bayesian anchoring method with a non-informative prior. Finally, the fourth plot (Figure 5.1d) once again considers the Bayesian anchoring method, only this time with the aforementioned informative prior based on analyses of the 2011 BRFSS data.

In general, we see that the anchoring method has a tendency to reduce estimates of the interviewer effects, regardless of the approach used. Data points below the 45-degree lines in each plot indicate states where a particular adjustment method reduced the estimates of the interviewer effects. In particular, the "standard" adjustment method will more often *increase* estimates of the interviewer effects in a non-trivial fashion relative to the naïve approach (Figure 5.1b).

Table 5.1 presents mean estimates and ranges of the interviewer effects across the 50 states and D.C. under the different methods. The anchoring method tended to reduce the estimates relative to the naïve method more often than the adjustment method, with 88.2% and 72.5% of states seeing a reduction in the estimated interviewer effects when using the frequentist and informative Bayesian anchoring methods, respectively (compared to only 60.8% of states when using the adjustment method). There is evidence in Table 5.1 that the use of prior information helps when applying the Bayesian anchoring method, but the frequentist version of the anchoring method still has the best performance overall. In some cases these reductions in the interviewer effect relative to the naïve approach were substantial: five of the states had reductions in the estimated interviewer effect of at least 33% regardless of the type of anchoring method used. In some cases, the anchoring approach did lead to slight increases in the estimated interviewer effects. These were predominantly cases where the interviewer effects were very small (suggesting that

the proposed adjustment would not be necessary, and that any resulting increases in the estimates were simply noise).

Figure 5.1 Scatter plots comparing the anchoring and naïve estimates of the interviewer effects for the 50 states and the District of Columbia, by estimation approach (NI = non-informative prior; Inf = weakly informative prior, based on analyses of the 2011 BRFSS data). Points belowthe 45-degree lines in each plot indicate states where a particular adjustment method reduced the estimates of the interviewer effects below that of the naïve estimate.

Figure 5.1a: Frequentist - Anchoring

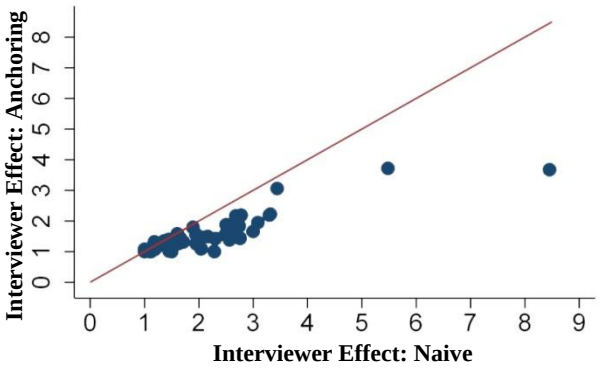


Figure 5.1b: Frequentist - Adjustment

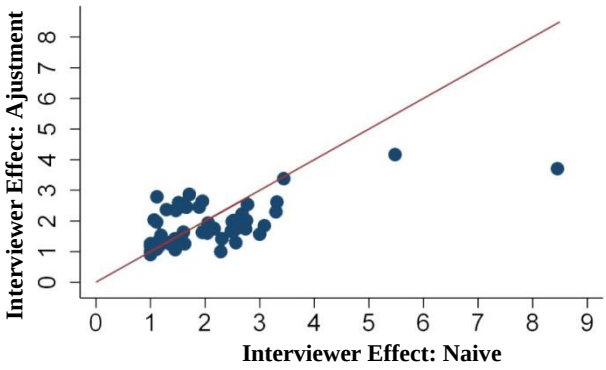


Figure 5.1c: Bayesian - Anchoring (NI)

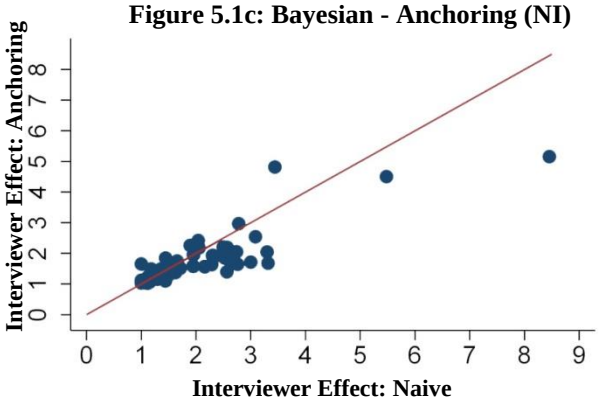


Figure 5.1d: Bayesian - Anchoring (Inf)

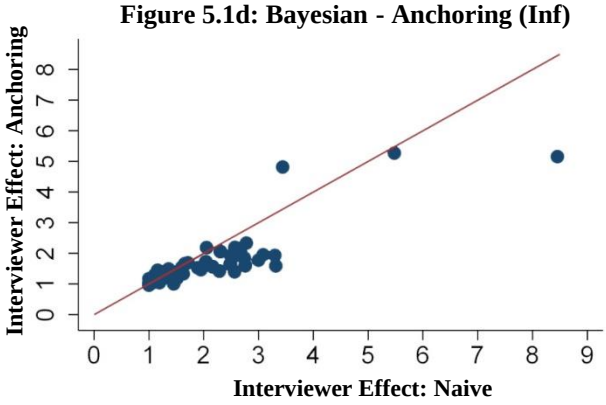


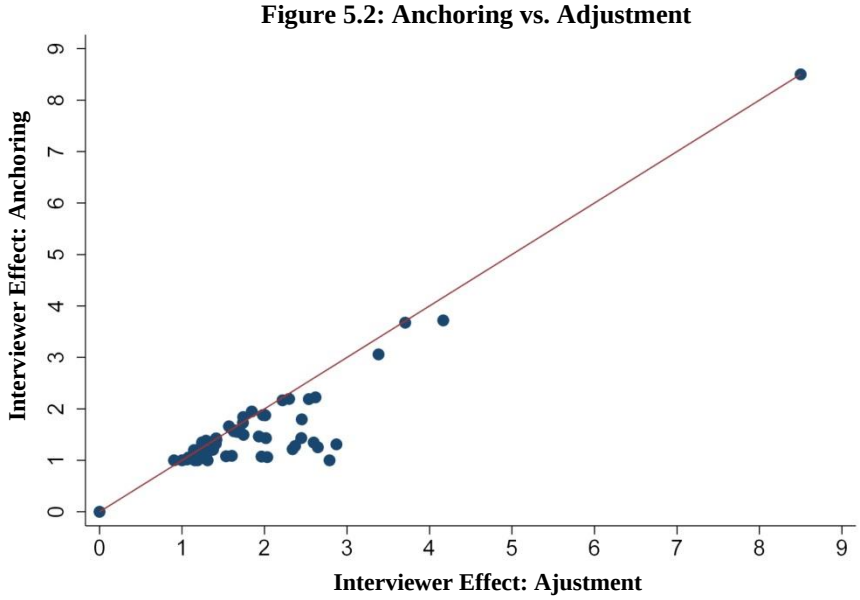
Table 5.1
Means and ranges of interviewer effects across the 50 States and the District of Columbia under the competing approaches

Estimation Approach	Interviewer Effects: Mean (Range)	Percentage of States with a Reduction
Frequentist – Naive	2.06 (1.00 – 8.45)	-
Frequentist – Adjustment	1.85 (0.90 – 4.17)	60.8%
Frequentist – Anchoring	1.51 (1.00 – 3.72)	88.2%
Bayesian – Anchoring, Non-Informative	1.79 (1.03 – 5.16)	58.8%
Bayesian – Anchoring, Informative	1.70 (0.96 – 5.27)	72.5%

When comparing the anchoring method with the "standard" adjustment method, we found consistent evidence of the anchoring method producing larger reductions in the estimated interviewer effects.

Figure 5.2 compares the estimated interviewer effects for the 50 states and D.C. when using the anchoring method and the adjustment method, considering the frequentist results only. We see that the interviewer estimates based on the adjustment method tend to be larger than the estimates based on the anchoring approach.

Figure 5.2 Scatter plot comparing the anchoring and adjusted estimates of the interviewer effects for the 50 states and the District of Columbia.



In general, we did not find significant benefits of using a Bayesian approach to implement the anchoring method in this application. We did find that for 92.5% of the states, the 95% credible interval for the interviewer variance component was smaller in width when using the informative prior than the credible interval based on the non-informative prior, as would be expected. However, the posterior medians of the interviewer variance components tended to be similar based on both Bayesian anchoring methods (Pearson correlation = 0.73).

6. Discussion

We have developed and evaluated a new method for estimating interviewer effects in the absence of interpenetrated assignment of sampled units to interviewers. Via a simulation study and applications using real survey data from the BRFSS, we have demonstrated the ability of the proposed anchoring method to improve estimates of interviewer effects in situations where interpenetrated assignment may not be feasible and interviewer variance may be arising from the underlying sample assignments. The anchoring method can also easily be applied in a Bayesian framework, leveraging prior information to improve the quality of predictions and inferences related to interviewer components of variance.

In interviewer-administered survey data collections, interviewer effects should generally be monitored as part of an ongoing data collection to prevent excessive problems with interviewer variance in survey

outcomes at the conclusion of the data collection. Survey managers responsible for this type of monitoring will likely benefit from the anchoring method, improving any real-time intervention decisions made for individual interviewers in a responsive survey design framework. Real-time interventions/re-trainings for interviewers who are found to have extreme effects on production outcomes or variables of scientific interest that in reality only reflect the features of the areas in which they are working and not actual interviewer performance will be at best inefficient and at worst could cause interviewers who are otherwise performing well to be inappropriately criticized and perhaps to leave a given study.

When using the anchoring method in practice, we would suggest that it be described as a method that "adjusts estimates of interviewer variance components for spurious within-interviewer correlation in survey measures of interest that can arise due to non-random assignment of sampled units to interviewers." We emphasize the importance of a sound theoretical selection of an anchoring variable (or variables) that ideally has the optimal properties described in this paper. In the absence of an anchoring variable with these optimal properties, we argue that "clean" estimation of interviewer variance in a non-interpenetrated sample design may simply not be possible, and that analysts 1) adjust for as many respondent-, interviewer-, and area-level covariates as possible when attempting to estimate the interviewer variance, and 2) report estimates of uncertainty associated with the estimated variance components, preferably using Bayesian approaches. This will prevent over-estimation of interviewer variance components and possible attribution of lower data quality to interviewers that are already performing extremely challenging tasks in the field.

There are several limitations to our proposed method. Perhaps the largest is the requirement for "anchoring" variables to not be subject to interviewer error and still be highly correlated with the substantive variable of interest. In our BRFSS example, we considered age, gender, race, and education as "anchoring" self-rated health. Although age is self-reported and thus possibly subject to some degree of measurement error (for example, reporting younger ages or rounding ages), we do not see an obvious mechanism by which this would be induced by the interviewer, although of course that possibility remains. A similar argument can be made for the other three factors, although the possibility of interviewer induced measurement error is slightly stronger due to issues such as "liking" between interviewers and respondents (West and Blom, 2017). In addition, the normality assumption that we make in the paper is highly restrictive. To deal with this in our application, we replaced the multivariate anchoring model (3.2) with a model that summarized the multiple anchors into a linear predictor that we then used in the bivariate anchoring model (3.1). While this linear predictor is effectively a sufficient statistic in the case where all of the anchoring variables are normal, as shown in the simulation study, it is more of an ad-hoc solution when some or all of the anchoring variables are non-normal, as was the setting in our application.

A more principled solution when one or more of the components of Y are dichotomous variable would be to consider extensions such as probit random effects models, replacing y_{ijk} in (3.1) with a latent y_{ijk}^{*}, where the observed $y_{ijk} = I\left(y_{ijk}^{*} > 0\right)$ and the variance $\sigma_k^2 = 1$ for identifiability, for all values of k where

y_{ijk} is dichotomous. More ambitiously, we could use a Gaussian random effects copula model (Wu and de Leon, 2014) for arbitrary distributions for Y^*. Standard software will not accommodate such models, although methods that integrate over random effects or use fully Bayesian approaches could be considered. Next, while other sources of measurement error are potentially important to address in inference, our focus here is on measurement error variance introduced by interviewer effects and its estimation in the absence of interpenetration. Finally, we note that our approach, like its competitors, relies on observed data and is thus not a replacement for true interpenetration, which ensures that all forms of non-random assignment (observed and unobserved) are eliminated.

In addition to extending the anchoring method to the case of regression coefficients and non-normal variables, future applications also need to consider contexts where the correlations of the anchoring variables with the survey variables of interest that may be prone to interviewer effects are modest at best. Our simulation study suggests that good anchoring variables having strong associations with the survey variables of interest are important for the effectiveness of this method, and future studies should also focus on the identification of sound anchoring variables (like age, education, etc.) that are unlikely to be affected by interviewers and could serve as useful anchors in other applications.

Acknowledgements

Funding for this research was provided by NIH Grant #1R01AG058599-01. The authors would like to thank the editor, associate editor, and two reviewers for their guidance which has improved the manuscript.

Supplemental materials

SAS code implementing the different approaches

The SAS code below can be used to implement the anchoring method using a standard frequentist approach. Implementing this approach requires the data to be in a "long" structure with two observations per subject (corresponding to the two variables), where the variable X2 is an indicator variable for the anchoring variable (1 = the observation on Y is the anchor, 0 = the observation on Y is the variable of interest), the variable X1 is an indicator for the variable of interest (1 = the observation on Y is the variable of interest, 0 = the observation on Y is the anchor), INTVID is the interviewer ID, and OBS is a subject ID:

```
proc mixed data=yourlongdata;
  class INTVID;
  model y = x2 / solution;
  random x1 / sub=INTVID;
  repeated / sub=obs type=un r rcorr;
run;
```

The SAS code below can be used to fit the naïve model using a Bayesian approach with a weakly informative prior. This approach requires the data to be in the same "long" format:

```
proc mcmc data=yourlongdata seed=41279 nmc=20000 thin=25;
  where x1 = 1; /* only fit model to variable of interest */
  parms B0 S2;
  parms Sigma 1;
  prior B: ~ normal(0, var=1e6); /* prior for means */
  prior S2 ~ igamma(0.01, scale = 0.01); /* NI prior for resid. var. */
  prior Sigma ~ t(0, sd=0.045, df=3, lower=0); /* informative prior for SD of
  interviewer effects, per Gelman (2006); SD of distribution is estimated SD
  of random interviewer effects from 2011, constrains posterior */
  random Gamma ~ normal(0,sd=Sigma) subject=INTVID;
  Mu = B0 + Gamma; /* model with interviewer effects only for variable of
  interest */
  model y ~ normal(Mu,var=S2);
run;
```

Finally, the SAS code below can be used to implement the anchoring method using a Bayesian approach with a weakly informative prior. Implementing this approach requires the data to be in a wide structure, with one row per case and interviewer IDs (INTVID):

```
proc mcmc data=yourwidedata seed=41279 nmc=20000 thin=25;
  array y[2] genhlthmdd age10; /* var1=variable of interest, var2=anchor */
  array Mu[2]; /* vector of two observations for each case */
  array Cov[2,2]; /* residual covariance matrix */
  array S[2,2]; /* for defining prior of COV */
  array H[2] 0 H1; /* H1 = fixed effect for change in mean for anchor */
  parms B0 Cov; /* intercept (mean of variable of interest) and residual
  covariance matrix */
  parms H1 0; /* change in mean for anchor */
  parms Sigma 1;
  prior B: H: ~ normal(0, var=1e6); /* normal prior for fixed effects */
  prior Cov ~ iwish(2,S); /* prior for 2x2 residual covariance matrix */
  prior Sigma ~ t(0, sd=0.045, df=3, lower=0); /* informative prior for SD of
  interviewer effects, per Gelman (2006); SD of distribution is estimated SD
  of random interviewer effects from 2011, constrains posterior */
  begincnst;
    call identity(S); /* use identity matrix in defining prior for residual
    covariance matrix (non-informative) */
  endcnst;
  random Gamma ~ normal(0,sd=Sigma) subject=INTVID;
  Mu[1] = B0 + Gamma; /* interviewer effect only applies to variable of
  interest */
  Mu[2] = B0 + H1; /* mean for anchor (note: this parameterization used to
  ensure easy calculation of posterior SD of B0 for interviewer effects */
  model y ~ mvn(Mu, Cov);
run;
```

References

Biemer, P.P. (2010). Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5), 817-848.

Biemer, P.P., and Stokes, S.L. (1985). Optimal design of interviewer variance experiments in complex surveys. *Journal of the American Statistical Association*, 80(389), 158-166.

Brunton-Smith, I., Sturgis, P. and Williams, J. (2012). Is success in obtaining contact and cooperation correlated with the magnitude of the interviewer variance? *Public Opinion Quarterly*, 76, 265-286.

Carle, A.C. (2009). Fitting multilevel models in complex survey data with design weights: Recommendations. *BMC Medical Research Methodology*, 9, 49-62.

Centers for Disease Control (2013). Behavioral Risk Factor Surveillance System: OVERVIEW: BRFSS 2012. Accessed at http://www.cdc.gov/brfss/annual_data/2012/pdf/Overview_2012.pdf.

Cernat, A., and Sakshaug, J.W. (2021). Interviewer effects in biosocial survey measurements. *Field Methods*, 33, 236-252.

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

Elliott, M.R., and West, B.T. (2015). "Clustering by interviewer": A source of variance that is unaccounted for in single-stage health surveys. *American Journal of Epidemiology*, 182, 118-126.

Fellegi, I.P. (1974). An improved method of estimating the correlated response variance. *Journal of the American Statistical Association*, 69,496-501.

Fowler, F.J., and Mangione, T.W. (1989). *Standardized Survey Interviewing: Minimizing Interviewer-Related Error*. Newbury Park: Sage.

Franks, P., Gold, M.R. and Fiscella, K. (2003). Sociodemographics, self-rated health, and mortality in the US. *Social Science & Medicine*, 56, 2505-2514.

Gao, S., and Smith, T.M.F. (1998). A constrained MINQU estimator of correlated response variance from unbalanced data in complex surveys. *Statistica Sinica*, 8, 1175-1188.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, 1, 515-533.

Groves, R.M. (2004). Chapter 8: The interviewer as a source of survey measurement error. *Survey Errors and Survey Costs (2nd Edition)*. New York: Wiley-Interscience.

Groves, R.M., and Magilavy, L.J. (1986). Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50, 251-266.

Heeringa, S.G., West, B.T. and Berglund, P.A. (2017). *Applied Survey Data Analysis, Second Edition*. Boca Raton, FL: Chapman Hall/CRC Press.

Hox, J.J. (1994). Hierarchical regression models for interviewer and respondent effects. *Sociological Methods and Research*, 22, 300-318.

Joffe, M.M., Ten Have, T.R., Feldman, H.I. and Kimmel, S.E. (2004). Model selection, confounder control, and marginal structural models: Review and new applications. *The American Statistician*, 58, 272-279.

Kalton, G. (1983). *Introduction to Survey Sampling*, Sage Publications: London, UK.

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57, 92-115.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Kleffe, J., Prasad, N.G.N. and Rao, J.N.K. (1991). "Optimal" estimation of correlated response variance under additive models. *Journal of the American Statistical Association*, 86, 144-150.

Laird, N.M., and Ware, J.H. (1982). Random-effects models for longitudinal data. *Biometrics*, 38, 963-974.

Lepkowski, J.M., Mosher, W.D., Groves, R.M., West, B.T., Wagner, J. and Gu, H. (2013). Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth. National Center for Health Statistics. *Vital Health Stat*, 2(158).

Liang, K.-Y., and Zeger, S.L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73, 13-22.

O'Muircheartaigh, C.A., and Campanelli, P. (1998). The relative impact of interviewer effects and sample design effects on survey precision. *Journal of the Royal Statistical Society, Series A*, 161, 63-77.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.

Rabe-Hesketh, S., and Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society-A*, 169, 805-827.

Rasbash, J., and Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross-classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics*, 19, 337-350.

Rohm, T., Carstensen, C.H., Fischer, L. and Gnambs, T. (2021). Disentangling interviewer and area effects in large-scale educational assessments using cross-classified multilevel item response models. *Journal of Survey Statistics and Methodology*, 9, 722-744.

Sakshaug, J.W., Tutz, V. and Kreuter, F. (2013). Placement, wording, and interviews: identifying correlates of consent to link survey and administrative data. *Survey Research Methods*, 7, 133-144.

Schaeffer, N.C., Dykema, J. and Maynard, D.W. (2010). Interviewers and Interviewing. In *Handbook of Survey Research, Second Edition* (Eds., J.D. Wright and P.V. Marsden), Bingley, U.K.: Emerald Group Publishing Limited.

Schnell, R., and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21, 389-410.

Skinner, C.J., Holt, D. and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons, Inc.

Stiratelli, R., Laird, N. and Ware, J. (1984). Random effects models for serial observations with binary responses. *Biometrics*, 40, 961-971.

Vassallo, R., Durrant, G. and Smith, P. (2017). Separating interviewer and area effects by using a cross-classified multilevel logistic model: Simulation findings and implications for survey designs. *Journal of the Royal Statistical Society, Series A*, 180, 531-550.

Veiga, A., Smith, P.W.F. and Brown, J.J. (2014). The use of sample weights in multivariate multilevel models with an application to income data collected by using a rotating panel survey. *Journal of the Royal Statistical Society (Series C)*, 63, 65-84.

von Sanden, N., and Steel, D. (2008). Optimal estimation of interviewer effects for binary response variables through partial interpenetration. Centre for Statistical and Survey Methodology, University of Wollongong, Working Paper 04-08.

West, B.T., and Blom, A.G. (2017). Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5, 175-211.

West, B.T., and Elliott, M.R. (2014). Frequentist and Bayesian approaches for comparing interviewer variance components in two groups of survey interviewers. *Survey Methodology*, 40, 2, 163-188. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2014002/article/14092-eng.pdf.

West, B.T., and Olson, K. (2010). How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74, 1004-1026.

West, B.T., Kreuter, F. and Jaenichen, U. (2013). "Interviewer" effects in face-to-face surveys: A function of sampling, measurement error or nonresponse? *Journal of Official Statistics*, 29, 277-297.

Wu, B., and de Leon, A.R. (2014). Gaussian copula mixed models for clustered mixed outcomes, with application in developmental toxicology. *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 39-56.