*Journal of Survey Statistics and Methodology (2024) **12**, 814–839*

THREE APPROACHES TO IMPROVE INFERENCES BASED ON SURVEY DATA COLLECTED WITH MIXED-MODE DESIGNS

WENSHAN YU*
MICHAEL R. ELLIOTT
TRIVELLORE E. RAGHUNATHAN

Mixed-mode designs have become increasingly common in survey data collection. Although different modes often have different measurement properties, the standard practice is to treat mixed-mode data as if they had been collected with a single mode, neglecting the potential impact of mode effects. To account for potential mode effects when making inferences for mixed-mode samples, we propose (i) a Testimator approach, (ii) a Bayesian approach, and (iii) a model averaging approach. In the Testimator approach, we test whether the means and the variances of mixed-mode samples are the same. If the means are the same, we take the average of mode-specific estimates. If the means are different, we take the average when we have no prior information about preferred modes and take the smaller (or larger) estimate when we have prior information about preferred modes (e.g., a smaller estimate is better). In the Bayesian approach, we assume some prior information. We use a data-driven method to determine whether there are mode effects. If there are no mode effects, we draw inferences using a common mean model. If there are mode effects, we draw inferences using the data collected with the mode that produces smaller estimates. In the model averaging approach, we combine estimates of different models (characterized by whether assume same means and variances across modes) using marginal posteriors as weights. We

WENSHAN YU is a PhD candidate in the Program of Survey and Data Science at the University of Michigan, Ann Arbor, MI, USA. MICHAEL R. ELLIOTT is Professor of Biostatistics and Research Professor of Survey Methodology at the University of Michigan, Ann Arbor, MI, USA. TRIVELLORE E. RAGHUNATHAN is Professor of Biostatistics and Research Professor of Survey Methodology at the University of Michigan, Ann Arbor, MI, USA.

We gratefully acknowledge the financial support provided by the Daniel Katz Dissertation Fellowship for this research. We thank Dr. James Wagner and Dr. Tuba Suzer Gurtekin for their valuable comments, as well as Dr. Brady West and Dr. Katharine Abraham for their efforts in editing a portion of the manuscript. We also thank the Editor, the Associate Editor, and the anonymous referees for their helpful comments and suggestions.

*Address correspondence to Wenshan Yu, Survey Research Center, Institute for Social Research, University of Michigan–Ann Arbor, 426 Thompson Street, Ann Arbor, MI 48104-2321, USA; E-mail: yuwens@umich.edu.

https://doi.org/10.1093/jssam/smae012 Advance access publication 29 April 2024

evaluate the approaches in simulation studies and find that they achieve robust inferences compared to the standard approach. We apply the methods to the Arab Barometer study, which employs a randomized mixed-mode design.

KEY WORDS: Bayesian; Inference; Mixed-mode; Model averaging; Testimator.

Statement of Significance

As mixed-mode surveys become increasingly popular, analytical strategies that account for different measurement properties associated with each mode are needed. Our proposed approaches tease out potential mode effects compared to the standard approach that simply pools the mixed-mode data. We consider both cases where prior information about the direction of mode effects is known and where it is not and provide practitioners with ready-to-implement procedures to analye their mixed-mode data that are susceptible to measurement error.

1. INTRODUCTION

The use of mixed-mode designs in survey practice is a response to an increasingly difficult survey climate. Compared to the 1970s and 1980s, the general population is much less interested in taking a survey, and it is much harder to contact a potential sample (Massey and Tourangeau 2013). The use of mixed-mode designs can be helpful because each mode is associated with different nonresponse and cost properties (Voogt and Saris 2005). For example, a mail survey is usually cheaper than a face-to-face (FTF) interview, but the latter is more effective for recruiting reluctant participants. In responsive and adaptive designs, by tailoring modes for different participants, survey organizations have a better chance to achieve a well-balanced sample at a lower cost. Despite these benefits, there is a potential threat introduced by using multiple modes—the measurement properties of different modes can be different. For example, when using interviewer-administered modes as a follow-up to a web survey, responses to sensitive questions can be more prone to social desirability bias. This calls into question how best to analyze the pooled data collected with mixed-mode designs. This article aims to propose novel approaches to address this issue.

To begin, we first need to understand how mode effects threaten the validity of mixed-mode surveys. A large volume of research has been devoted to studying mode effects, that is, any influence on survey responses that is due to modes (Lavrakas 2008). For example, the item nonresponse rate is often

higher in self-administered modes than in interviewer-administered modes (de Leeuw and Desiree 1992). Telephone (TEL) surveys may produce recency effects, in which respondents tend to choose the last response option heard. In contrast, primacy effects are more common in visually-presented modes (such as mail surveys), in which respondents are more inclined to choose the first option they see (Schwarz et al. 1985). One of the most consistent findings from previous studies is that responses collected from interviewer-administered modes are more prone to social desirability bias than responses collected with self-administered modes (Tourangeau and Smith 1996; Dillman and Christian 2005; Kreuter et al. 2008). Presser and Stinson find that compared to interviewer-administered modes, claims of weekly religious attendance reduce by one-third in the self-reported mode (Presser and Stinson 1998). Holbrook and Krosnick show that compared to TEL surveys, social desirability bias in voter turnout reports is minimal in Internet surveys (Holbrook and Krosnick 2010). These findings suggest that the accuracy of responses across modes can be different and self-administered modes consistently provide better responses to sensitive questions. We refer to the phenomenon that different modes produce different measurement errors as mode measurement effects (de Leeuw et al. 2018).

模式量測效應

模式選擇效應

On the other hand, mode selection effects refer to the phenomenon that different modes produce different nonresponse errors (Vannieuwenhuyze et al. 2010; Vannieuwenhuyze and Revilla 2013; de Leeuw et al. 2018). Mode selection effects occur when respondent characteristics differ across modes in ways that are correlated with the variable(s) of interest. Researchers do not equivalently value the two types of mode effects mentioned above. Generally, mode selection effects are a wanted property of mixed-mode designs because researchers can achieve better sample balance by making use of the selection effects (de Leeuw 2018). Mode measurement effects are unwanted because they can result in inconsistent response quality and thus need to be accounted for (de Leeuw 2018). However, as researchers only observe responses from each participant with one mode, selection effects and measurement effects are confounded. Caution needs to be taken to account for the selection effects when adjusting for the mode measurement effects.

1.1 Literature Review

To combine data collected with mixed-mode designs, one line of research focuses on developing estimate-level weights such that the resulting final estimate has some desirable properties.

Suzer-Gurtekin et al. (2012) multiply impute the potential values of what would have been observed with another mode(s) so that each case has an observed value of a variable of interest collected with one mode and multiple impute values of the variable of interest that would have been collected with

other modes. They use weights to combine these mode-specific estimates. Buelens and van den Brakel (2015) propose to fix the weights of mode-specific estimates in longitudinal or cross-sectional surveys such that mode-related measurement error remains comparable across waves. Brick et al. (2021) develop an adaptive mode adjustment to address the differential non-response properties of mixed-modes.

Another line of research aims to calibrate mixed-mode data so that it approximates what would have been collected with a single mode. To do that, researchers need to specify a reference mode, which in many cases will be the mode that theoretically contributes to better data quality based on previous findings. It also might be the more prevalent mode in the survey such that consistency of the results with this mode is of interest to researchers (Kolenikov and Kennedy 2014). After determining a reference mode, researchers need to derive potential outcomes for cases in the nonreference mode. Examples include Powers et al. (2005), Elliott et al. (2009), Kolenikov and Kennedy (2014), and Park et al. (2016). One limitation of this approach is that the accuracy of the resulting estimate strongly relies on the choice of the reference mode, which has to be predetermined. When there is no prior knowledge of the reference mode, it is unclear what to do with the mode-specific estimates.

1.2 Context of This Study

Mixed-mode inference has become a pressing necessity since coronavirus disease 2019 (COVID-19), as many large survey projects have been forced to shift data collection modes due to restricted social contact. The Arab Barometer study, which is the application considered in this article, is one of them. It is the largest repository of public opinion data in the Middle East and North Africa (MENA) region. In wave 6 (2020), they shifted from FTF alone to mixed-mode designs (FTF and TEL). The research team applied a mixed-mode experiment in Jordan, and they intend to estimate the population quantities using the data collected in the experiment while teasing out potential mode effects. On one hand, literature reports that FTF can reduce socially desirable reporting relative to TEL, possibly because of enhanced rapport built between interviewers and respondents during FTF interviews (Holbrook et al. 2003). Other credible literature finds no clear differences between FTF and TEL on sensitive items (Klausch et al. 2013). We aim to address the research question of how to make inferences to incorporate the potential mode effects by proposing three new approaches in this article. Although mode selection effects are a central concern when adjusting for mode measurement effects, they can be dealt with common approaches like propensity score adjustments. Therefore, the three approaches proposed in this study focus on accounting for the different measurement properties when combining mode-specific estimates.

2. PROPOSED METHODS

This article proposes three approaches to combine mode-specific estimates: (i) a Testimator approach, (ii) a Bayesian approach, and (iii) a model averaging approach. We compare the approaches with two Naïve approaches. For illustration purposes, we consider continuous outcomes that follow normal distributions in this article. However, the proposed approaches are not limited to normal outcomes; they can be extended for various types of variables.

We assume two modes (mode A and mode B) are used. We consider a normally distributed outcome y observed in mode A as drawn from an identically and independent normal distribution $y_{ai} \sim N(u_a, \sigma_a^2)$, where $u_a = \theta + \delta_a$, u_a is the population mean when data is collected via mode A, θ reflects the true population mean, δ_a represents the bias occurred due to mode A, and σ_a^2 includes both the unit level population variance and random measurement error associated with mode A. Similarly, for mode B, y_{bi} follows $N(u_b, \sigma_b^2)$, where $u_b = \theta + \delta_b$, u_b is the population mean when data is collected via mode B, δ_b indicates the bias occurred due to mode B and σ_b^2 equals the sum of random measurement error associated with mode B and unit level population variance. In an analytical sample, we assume mode A is used on n_a subjects, and mode B is used on n_b subjects. We denote the sample mean and the standard derivation of data collected with mode A as $\bar{y}_a$ and S_a. Similarly, we denote the sample mean and the standard deviation derived using data collected with mode B as $\bar{y}_b$ and S_b.

From the setup, the only estimable quantities are u_a, u_b, σ_a^2, and σ_b^2. To infer the population mean θ, we need additional information and assumptions to make inferences. The following assumptions are made in the article:

(1) At least one mode provides an unbiased estimate of the population mean (either $\delta_a = 0$ or $\delta_b = 0$, but we do not know which is 0). This assumption guarantees that the population mean is estimable, despite the presence of mode effects.

(2) y_{ai} and y_{bi} are never jointly observed. 不存在一個人同時填A與B的資料

(3) Mode selection (denoted as M_i) is independent from the potential outcomes (y_{ai} and y_{bi}). In the real data application, we relax the assumption to conditional independence given covariates. This assumption guarantees the identification of mode measurement effects.

受訪者分到A或B不能與他原本會回答的內容有關
ex.
X
受訪者自己選 -> 不支持政府者害怕電話監控 -> 選擇FTF

O
調查員抽籤分配，避免模式選擇效應

given covariates: 將某些背景變數控制住，模式選擇會跟潛在回答獨立 ex. 直接各拉50個面訪與電訪的老年人比較

Drawing on observed data, we cannot know which mode leads to unbiased estimates; thus, we use external information (such as preferred directions) to help make inferences. For example, for sensitive questions, which are more subject to mode effects than nonsensitive questions, researchers may know which direction of estimates indicates more honest reports based on substantive knowledge. As an illustration, the Arab Barometer survey asked how satisfied Jordan participants were with the government's performance in responding to COVID-19. Since expressing dissatisfaction with government

performance on COVID-19 may pose risks to respondents, researchers may anticipate that a lower mode-specific estimate of the satisfaction is likely to represent more truthful answers.

We consider the following settings and inference strategies in this article.

存在mode effect，且知道偏好的估計方法

(1) When mode effects exist and we know a preferred direction of the estimates (setting 1), we take the estimate in the preferred direction to estimate the population mean.
(2) When mode effects do not exist ($\delta_a = \delta_b = 0$, setting 2), we estimate the population mean to be the same as the estimated mode-specific means ($\hat{\theta} = \hat{u}_a = \hat{u}_b$).
(3) When mode effects exist but the preferred direction is unknown (setting 3), we estimate the population mean as the average of the estimated mode-specific means ($\hat{\theta} = \frac{\hat{u}_a + \hat{u}_b}{2}$) and propagate the uncertainty associated with the setting.

We develop three approaches in these settings inspired by two different philosophies. The first assesses whether data can be pooled or not by testing if mode-specific means are the same using some cutoff values. Approaches based on this philosophy provide a clear-cut answer to whether mode effects exist in the data and then develop inferences accordingly. The second considers all possible models generating the mixed-mode data and then averages across models using weights that reflect the likelihood of the model. Inference based on this philosophy can accommodate more than one view toward mode effects and thus makes a distinction with inference strategies based on the first philosophy. Built on the first philosophy, we propose the Testimator and the Bayesian approaches; in the spirit of the second philosophy, we develop a Bayesian model averaging approach. We compare the proposed methods to two Naïve approaches, one that simply uses the mode providing the preferred direction while dropping the other cases, and one that pools the data.

2.1 The Testimator Approach

In this approach, we consider a two-step testing procedure. First, we use the F test to test whether sample variances of the modes are the same. Depending on the result, we use a corresponding t-test to evaluate if the means are the same. Based on the results, we make inferences accordingly.

2.1.1 When there is some information about the preferred direction.

(1) We test if $\sigma_a^2 = \sigma_b^2$ using a two-tailed F test. We calculate the F statistic as $\frac{S_a^2}{S_b^2}$ and refer it to $F(n_a - 1, n_b - 1)$. Users may determine the significance level of the F test (denoted as α_1).

(2) If the F statistic is within the interval $[F(\frac{\alpha_1}{2}, n_a - 1, n_b - 1), F(1 - \frac{\alpha_1}{2}, n_a - 1, n_b - 1)]$, then we cannot reject the null hypothesis that $\sigma_a^2 = \sigma_b^2$. Next, we use a two-tailed pooled variance t-test to test whether there are differences between u_a and u_b assuming a common variance $\sigma_a^2 = \sigma_b^2 = \sigma^2$. We denote the significance level of the t-test as α_2. The t-statistic is constructed as $\frac{\bar{y}_a - \bar{y}_b}{s\sqrt{(n_a^{-1} + n_b^{-1})}}$, where $s = \sqrt{\frac{(n_a - 1)s_a^2 + (n_b - 1)s_b^2}{n_a + n_b - 2}}$.

(a) If the t-statistic falls within $[-t_{\frac{\alpha_2}{2}, n_a + n_b - 2}, t_{\frac{\alpha_2}{2}, n_a + n_b - 2}]$, we compute an estimate of the population mean as $\hat{\theta} = \frac{n_a \bar{y}_a + n_b \bar{y}_b}{n_a + n_b}$, assuming $u_a = u_b = \theta$. We construct the confidence interval (CI) as $[\theta - t_{\frac{\alpha_2}{2}, n_a + n_b - 2} \frac{s}{\sqrt{n_a + n_b - 2}}, \theta + t_{\frac{\alpha_2}{2}, n_a + n_b - 2} \frac{s}{\sqrt{n_a + n_b - 2}})]$.

(b) If the t-statistic falls outside of the interval, we estimate θ using the smaller (or larger) value of $\bar{y}_a$ and $\bar{y}_b$ and construct the CI using $\bar{y}_a \pm t_{n_a - 1, \frac{\gamma}{2}} \frac{s_a}{\sqrt{n_a}}$ if $\bar{y}_a \leq \bar{y}_b$ or $\bar{y}_b \pm t_{n_b - 1, \frac{\gamma}{2}} \frac{s_b}{\sqrt{n_b}}$ if $\bar{y}_a > \bar{y}_b$, depending on whether we are assuming that the smaller or the larger estimate is better. Otherwise, we use $\bar{y}_b \pm t_{n_b - 1, \frac{\gamma}{2}} \frac{s_b}{\sqrt{n_b}}$ to construct the CI. Note that $1 - \gamma$ represents the confidence level of the CI, and γ can be interpreted as type I error rate. This parameter γ can be set to values different from α_1 and α_2, which determine the significance levels of the F test and the t-tests, respectively.

(3) If the F statistic falls within $[-\infty, F(\frac{\alpha_1}{2}, n_a - 1, n_b - 1)]$ or $[F(1 - \frac{\alpha_1}{2}, n_a - 1, n_b - 1), \infty]$, we construct $t = \frac{\bar{y}_a - \bar{y}_b}{\sqrt{s_a^2 n_a^{-1} + s_b^2 n_b^{-1}}}$ assuming unequal variances $\sigma_a^2 \neq \sigma_b^2$, with degrees of freedom (v) as $\frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{\left(\frac{s_a^2}{n_a}\right)^2}{n_a - 1} + \frac{\left(\frac{s_b^2}{n_b}\right)^2}{n_b - 1}}$.

(a) If t falls within $[-t_{1 - \frac{\alpha_2}{2}, v}, t_{1 - \frac{\alpha_2}{2}, v}]$, we estimate $\hat{\theta} = \frac{\frac{n_a}{s_a^2} \bar{y}_a + \frac{n_b}{s_b^2} \bar{y}_b}{\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2}}$ assuming $u_a = u_b = \theta$ and with CI given by $\left[\hat{\theta} - t_{\frac{\alpha_2}{2}, v} \left(\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2}\right)^{-\frac{1}{2}}, \hat{\theta} + t_{\frac{\alpha_2}{2}, v} \left(\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2}\right)^{-\frac{1}{2}}\right]$.

(b) If t falls out of the interval and we know the preferred mode, we can make inferences similar to 2(b). Specifically, we estimate θ based on the prior information about the preferred mode (e.g., the smaller value of $\bar{y}_a$ and $\bar{y}_b$) and construct CI accordingly.

We summarize the steps of the Testimator approach in figure 1.

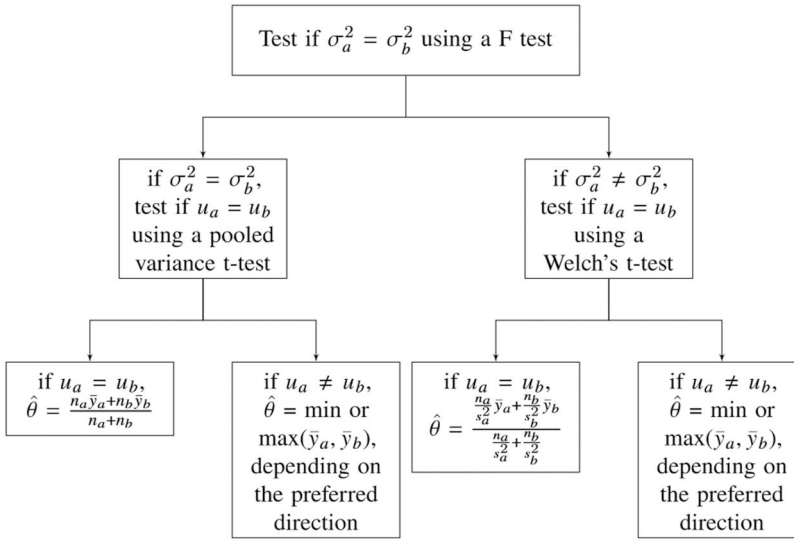


Figure 1. Flowchart of the Testimator Approach When Information about the Preferred Direction Is Available.

2.1.2 When there is no information about preferred directions.

We follow the same steps to test whether variances and means are the same across modes as in the previous setting. However, when we detect differences in the means, we compute two $100(1-\gamma)\%$ CIs for u_a and u_b. The CIs are computed as $[L_a, U_a] = \bar{y}_a \mp t_{n_a-1}\frac{S_a}{\sqrt{n_a}}$ for u_a and $[L_b, U_b] = \bar{y}_b \mp t_{n_b-1}\frac{S_b}{\sqrt{n_b}}$ for u_b. Then we compare L_a and L_b and take the smaller one as the lower bound of the population mean ($L = \min(L_a, L_b)$). We compute the upper bound by taking the larger one in U_a and U_b ($U = \max(U_a, U_b)$). The final interval we use for inference is $[L, U]$. We consider the midpoint of the interval ($\frac{L+U}{2}$) as an estimate of the population mean. We do not expect the point estimator to be unbiased or outperform the Naïve approach in bias reduction. Due to the lack of external information, we recommend focusing on interval estimation of the population mean.

F檢定變異數 - t檢定平均值 - 建構CI - 將下界min與上界max合併成一個區間 - 取區間中點

2.2 The Bayesian Approach

In this approach, we distinguish between testing and inference phases. During testing, we assume mixed-mode data has different means and variances and use posterior draws to compute effect size estimates. The effect size informs model selection and subsequent inferences.

2.2.1 When there is some information about the preferred direction.

(1) We consider a model that assumes different means and variances for data collected with modes A and B, from which we obtain the posterior draws of u_a and u_b:

$$y_{ai} \sim N(u_a, \sigma_a^2), y_{bi} \sim N(u_b, \sigma_b^2)\,.$$

why 條件機率? 可代表對平均數的信心程度

We use conjugate priors for u_a, u_b, σ_a^2, and σ_b^2: $u_a|\sigma_a^{-2} \sim N(\beta_a, \frac{\sigma_a^2}{k_a})$, $u_b|\sigma_b^{-2} \sim N(\beta_b, \frac{\sigma_b^2}{k_b})$, $\sigma_a^{-2} \sim G(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2)$, $\sigma_b^{-2} \sim G(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2)$. The hyperparameters β_a and β_b reflect the prior belief about u_a and u_b. The hyperparameters k_a and k_b control the contribution of prior information to the posterior population mean (u_a and u_b), with larger values of k increasing the contribution of the prior mean to the posterior. The hyperparameters τ_a^2 and τ_b^2 are the prior estimates of precision for modes A and B, while v_a and v_b allow for different levels of confidence in τ_a^2 and τ_b^2, respectively. Users can determine these hyperparameters. Diffuse priors can be considered when there are no external resources (e.g., expert opinion or historical data) for prior information, as illustrated in the simulation study in this article. The parameterization of the gamma distribution in this article is suggested by scaled inverse chi-square distributions, such that if $\sigma_a^2 \sim \text{Scale-inv} - \chi^2(v_a, \tau_a^2)$, then $\sigma_a^2 \sim \text{Inv-Gamma}(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2)$ and $\sigma_a^{-2} \sim G(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2)$.

(2) We compute estimates of effect size using the posterior draws of u_a, u_b, σ_a^2, and σ_b^2: $\hat{\eta} = \dfrac{\hat{u}_a - \hat{u}_b}{\sqrt{\dfrac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{2}}}$.

兩mode平均差距

兩Mode平均標準差

衡量mode之間是否有顯著的mode effe

(3) We create cutoff values $(\epsilon_{lw}, \epsilon_{up})$ for $\hat{\eta}$ by computing the 50, 75, 90, and 95 percent quantile-based credible interval of $\hat{\eta}$ and check if the interval includes 0.

(a) If $\epsilon_{lw} \le 0 \le \epsilon_{up}$, draw estimates of the population mean $\hat{\theta}$ from the following model that assumes a common mean u for both modes ($\hat{\theta} = u$). We assume a normal prior on u, with mean β_0 and variance ψ^2.

$$y_{ai} \sim N(u, \sigma_a^2), y_{bi} \sim N(u, \sigma_b^2), u \sim N(\beta_0, \psi^2),$$
$$\sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2\right), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2\right).$$

(b) If $\epsilon_{lw} > 0$, using only data collected by mode B to estimate the population mean ($\hat{\theta} = u_b$), as we illustrate the approaches by considering a

smaller estimate is preferred. Similarly, we assume a normal prior on u_b, with mean β_b and variance ψ_b^2.

$$y_{bi} \sim N(u_b, \sigma_b^2), u_b \sim N(\beta_b, \psi_b^2), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\lambda_b^2\right).$$

(c) If $\epsilon_{up} < 0$, using only data collected by mode A to estimate the population mean ($\hat{\theta} = u_a$). Again, we illustrate the approaches assuming a smaller estimate is preferred. We assume a normal prior on u_a, with mean β_a and variance ψ_a^2.

$$y_{ai} \sim N(u_a, \sigma_a^2), u_a \sim N(\beta_a, \psi_a^2), \sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\lambda_a^2\right)$$

(4) We compute the posterior mean of $\hat{\theta}$ as the estimate of the population mean. We use quantile-based intervals to quantify the uncertainty.

2.2.2 When there is no information about preferred directions.
In this scenario, we first compute the estimate of effect sizes as previously suggested. If $\epsilon_{lw} \leq 0 \leq \epsilon_{up}$, we generate draws of the population mean from the posterior distribution of the common mean, u, using the subsequent model:

$$y_{ai} \sim N(u, \sigma_a^2), y_{bi} \sim N(u, \sigma_b^2), u \sim N(\beta_0, \psi^2),$$
$$\sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2\right), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2\right).$$

If $\epsilon_{lw} > 0$ or $\epsilon_{up} < 0$, we generate draws of the population mean from the pooled draws of $\hat{u}_a$ and $\hat{u}_b$ from the different mean model (obtained when computing the effect size):

$$y_{ai} \sim N(u_a, \sigma_a^2), y_{bi} \sim N(u_b, \sigma_b^2),$$
$$u_a|\sigma_a^{-2} \sim N\left(\beta_a, \frac{\sigma_a^2}{k_a}\right), u_b|\sigma_b^{-2} \sim N\left(\beta_b, \frac{\sigma_b^2}{k_b}\right),$$
$$\sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2\right), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2\right).$$

Specifically, we generate R numbers of Bernoulli random variates, each with a probability equal to 0.5. we take a draw from $\hat{u}_a$ when the random variate is 0 and take a draw from $\hat{u}_b$ when the random variate is 1. Depending on whether the interval includes 0, we use the draws either from the common mean model or the different mean model to compute a posterior mean and a credible interval.

2.3 The Model Averaging Approach

This approach accounts for the uncertainty in four proposed models through Bayesian model averaging. Bayesian model averaging is a statistical method

that accounts for uncertainties in model selection in a principled manner and thus avoids the risks of making over-confident inferences (Hoeting et al. 1999). Unlike the Testimator and Bayesian approaches, this approach does not aim to find a single model that best describes the data, instead averaging over all possible models with weights proportional to the probability that one of the models is correct.

2.3.1 When there is some information about the preferred direction.

(1) We assume four models that differ in specifying same or different means and same or different variances on data collected with two modes. We use similar priors and notation in this approach as those used in the Bayesian approach.

Model 1 assumes different means and different variances for data collected with modes A and B (1). This model fits the scenario when modes used in data collection lead to shifts in both means and variances. We write model 1 as follows, where $\beta_a, k_a, \beta_b, k_b, v_a, v_b, \tau_a^2$, and τ_b^2 are the hyperparameters for priors of u_a, u_b, σ_a^{-2}, and σ_b^{-2}.

$$\begin{aligned} &y_{ai} \sim N\left(u_a, \sigma_a^2\right), y_{bi} \sim N\left(u_b, \sigma_b^2\right), u_a|\sigma_a^{-2} \sim N\left(\beta_a, \frac{\sigma_a^2}{k_a}\right), \\ &u_b|\sigma_b^{-2} \sim N\left(\beta_b, \frac{\sigma_b^2}{k_b}\right), \sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2\right), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2\right). \end{aligned} \tag{1}$$

Model 2 assumes a common mean (u) but different variances for mixed-mode data (2). This model accounts for the scenario when the use of modes leads to shifts in the variances but not the means.

$$\begin{aligned} &y_{ai} \sim N(u, \sigma_a^2), y_{bi} \sim N(u, \sigma_b^2), u \sim N(\beta_0, \psi^2), \\ &\sigma_a^{-2} \sim G\left(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2\right), \sigma_b^{-2} \sim G\left(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2\right). \end{aligned} \tag{2}$$

Model 3 assumes different means but a common variance, denoted σ^2 (3). This model considers the scenario when the use of modes leads to shifts in the means but not the variances. We use $\frac{v}{2}$ and $\frac{v}{2}\lambda^2$ as the shape and rate parameters, respectively, of the gamma distribution used as priors for the common precision (σ^{-2}).

$$\begin{aligned} &y_{ai} \sim N\left(u_a, \sigma^2\right), y_{bi} \sim N\left(u_b, \sigma^2\right), \\ &u_a|\sigma^{-2} \sim N\left(\beta_a, \frac{\sigma^2}{k_a}\right), u_b|\sigma^2 \sim N\left(\beta_b, \frac{\sigma^2}{k_b}\right), \sigma^{-2} \sim G\left(\frac{v}{2}, \frac{v}{2}\lambda^2\right). \end{aligned} \tag{3}$$

Model 4 assumes a common mean and variance (4). This model is appropriate when there are no shifts across mixed-mode data in either means or variances. We consider a conjugate normal prior on the common mean u with mean β_0 and variance $\frac{\sigma^2}{k}$.

$$y_{ai} \sim N(u, \sigma^2), y_{bi} \sim N(u, \sigma^2), u \sim N\left(\beta_0, \frac{\sigma^2}{k}\right), \sigma^{-2} \sim G\left(\frac{\nu}{2}, \frac{\nu}{2}\lambda^2\right). \quad (4)$$

(2) We calculate the marginal posterior $\pi(M|y_a, y_b)$ for each model using analytical integration (appendix A in the supplementary data online). After computing the marginal posteriors for each model (M), we compute the weight of each model as $W_M = \frac{\pi(M|y_a,y_b)}{\sum \pi(M|y_a,y_b)}$, where $\pi(M|y_a, y_b)$ is the marginal posterior computed in this step. Note that it is possible to use a Markov Chain Monte Carlo algorithm to derive the weights. This may be particularly useful when dealing with nonconjugate priors or when working with small sample sizes.

(3) We take R draws from a multinomial distribution using the weights (W_M) as the probabilities. The multinomial random variates indicate from which model we should draw $\hat{\theta}$. We draw $\hat{\theta}$ differently across models. In models 1 and 3, when we have some information about preferred directions (e.g., smaller the better), we do a pairwise comparison between $\hat{u}_a$ and $\hat{u}_b$ and take the smaller one as $\hat{\theta}$. In models 2 and 4, we directly draw $\hat{\theta}$ from the posterior distributions of u.

(4) Using the R draws of $\hat{\theta}$ obtained from the previous step, we compute the posterior mean, posterior variance, and a $(1-\gamma) \times 100$ percent credible interval.

2.3.2 When there is no information about preferred directions.
In this scenario, we consider the same four models proposed earlier. We generate draws of $\hat{u}_a$, $\hat{u}_b$, or $\hat{u}$ based on the models and compute the marginal posterior ($\pi(M|y_a, y_b)$) as in the previous scenario. However, for models 1 and 3, which assume different means, we generate draws of $\hat{\theta}$ from the pooled draws of $\hat{u}_a$ and $\hat{u}_b$ in a manner consistent with the Bayesian approach (section 2.2.2), instead of leveraging information about preferred directions to decide between $\hat{u}_a$ and $\hat{u}_b$. We follow the same procedure of using multinomial random variates to determine from which model we should draw θ, with the weights ($W_M = \frac{\pi(M|y_a,y_b)}{\sum \pi(M|y_a,y_b)}$) dictating the drawing probabilities.

2.4 Naïve Approach 1: Use the Estimate in the Preferred Direction (Naïve Preferred)

In this naïve approach, we consider the estimator of the population mean as $\hat{\theta} = \min$ or $\max(\bar{y}_a, \bar{y}_b)$, depending on the preferred direction. We illustrate the approach by considering smaller estimates are preferred. In this case, the estimated population mean is given by $\hat{\theta} = \bar{y}_a I(\bar{y}_a \leq \bar{y}_b) + \bar{y}_b I(\bar{y}_b \leq \bar{y}_a)$.

The expectation of the estimator is given by $E(\hat{\theta}) = u_a\Phi\left(\frac{u_b-u_a}{s}\right) + u_b\Phi\left(\frac{u_a-u_b}{s}\right) - s\phi\left(\frac{u_b-u_a}{s}\right)$, where $s = \sqrt{\frac{\sigma_a^2}{n_a} + \frac{\sigma_b^2}{n_b}}$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and the cdf of the standard normal distribution (Nadarajah and Kotz 2008). Using the same notation, we can write the second moment as $E(\hat{\theta}^2) = (u_a + \frac{\sigma_a^2}{n_a})\Phi\left(\frac{u_b-u_a}{s}\right) + (u_b + \frac{\sigma_b^2}{n_b})\Phi\left(\frac{u_a-u_b}{s}\right) - (u_a + u_b)s\phi\left(\frac{u_b-u_a}{s}\right)$ (Nadarajah and Kotz 2008). Then, the variance of the estimator can be expressed as (5).

$$
\begin{aligned}
Var\left(\hat{\theta}\right) &= E\left(\theta^2\right) - E(\theta)^2 \\
&= \left(u_a + \frac{\sigma_a^2}{n_a}\right)\Phi\left(\frac{u_b-u_a}{s}\right) + \left(u_b + \frac{\sigma_b^2}{n_b}\right)\Phi\left(\frac{u_a-u_b}{s}\right) - \\
&\quad (u_a+u_b)s\phi\left(\frac{u_b-u_a}{s}\right) - \left(u_a\Phi\left(\frac{u_b-u_a}{s}\right) + u_b\Phi\left(\frac{u_a-u_b}{s}\right) - s\phi\left(\frac{u_b-u_a}{s}\right)\right)^2.
\end{aligned} \tag{5}
$$

We then use sample means ($\bar{y}_a$ and $\bar{y}_b$) to estimate population means (u_a and u_b) and sample variances (s_a^2 and s_b^2) to estimate population variances (σ_a^2 and σ_b^2). We compute a $(1-\gamma)\times 100$ percent CI using a Z distribution.

2.5 Naïve Approach 2: Pool the Data (Naïve Pooled)

The second naïve approach ignores mode effects and pools the mixed-mode data as if they had been collected with a single mode. In the approach, we estimate the population mean using a sample mean ($\bar{y}$) and estimate its standard error by dividing the sample standard deviation by the square root of the sample size ($n_a + n_b$). We use a t distribution with degrees of freedom as $(n_a + n_b - 1)$ to compute a $(1-\gamma)\times 100$ percent CI.

3. SIMULATION STUDY

We consider nine scenarios for normal outcomes assuming simple random sampling from an infinite superpopulation. The data generation model is as follows:

$$
\begin{pmatrix} y_{ai} \\ y_{bi} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ u_b \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_b \\ \rho\sigma_b & \sigma_b^2 \end{pmatrix}\right).
$$

We assume mode A collects unbiased data, with a true superpopulation mean of 0 ($\theta = u_a = 0$). We set the variance of mode A to be 1 ($\sigma_a^2 = 1$). We

Table 1. Simulation Scenarios

Scenarios	u_b	σ_b^2
1	0	1
2	0.3	2
3	0.3	2
4	0.3	1
5	0.5	2
6	0.5	1
7	0.7	2
8	0.7	1
9	0.7	0.5

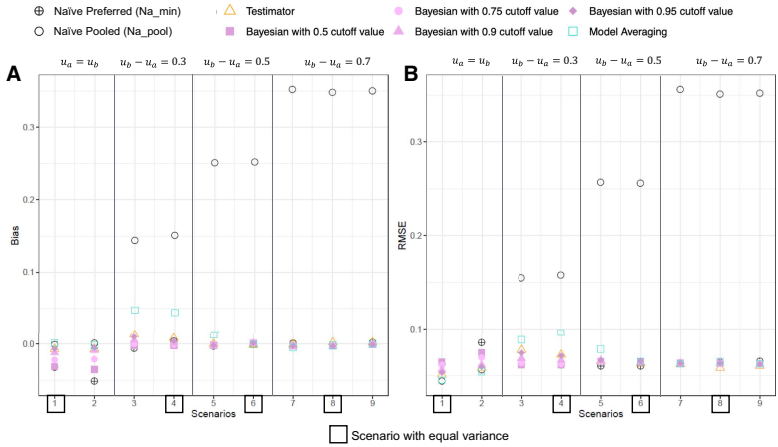
NOTE.—In all simulated scenarios, $u_a = 0$ and $\sigma_a^2 = 1$.

simulate nine scenarios to mimic varying levels of mode effects by adjusting u_b and σ_b^2 (see table 1 for the scenarios).

The first two scenarios mimic the situation when there are no mode effects. Scenarios 3 and 4 are designed to reflect small mode effects, scenarios 5 and 6 to represent medium mode effects, and scenarios 7–9 to capture large mode effects. In this simulation study, we consider correlations ρ of 0.5, 0.75, and 0.95. We draw 500 samples from the superpopulation, with a sample size of 500 ($n_a = 250, n_b = 250$) for each. We first explore the setting where we know the preferred direction (setting 1), incorporating the preference for smaller estimates when comparing the three approaches to the two Naïve approaches. We then examine the setting without information on preferred modes or directions (setting 2). In setting 2, the population mean remains 0, and we compare the three approaches to Naïve approach 2 (Naïve Pooled).

In the Testimator approach, we set the significance level for both the F-test (α_1) and the t-test (α_2) at 0.05. Additionally, we consider a 95 percent ($\gamma = 0.05$) CI for $\hat{\theta}$ in this approach. In the Bayesian and the model averaging approach, we consider the following hyperparameters when specifying the priors: $\psi^2 = 100$, $\beta_a = \beta_b = \beta_0 = 0, k_a = k_b = k = 0.01$, $\frac{v_a}{2} = \frac{v}{2} = \frac{v_b}{2} = 0.001$, and $\frac{v_a}{2}\tau_a^2 = \frac{v}{2}\lambda^2 = \frac{v_b}{2}\tau_b^2 = 0.001$. We choose these values so that the resulting priors are weakly informative, thereby allowing the data to predominantly influence the posterior distribution. In the Bayesian and the model averaging approaches, we obtain 5,000 draws from the posterior distribution.

We compare the approaches on their performances in bias, root mean square error (RMSE), coverage rate, actual interval length, and expected mean interval length. Bias is given by $\text{bias}(\hat{\theta}) = \frac{\sum \hat{\theta}}{500} - \theta$. RMSE is given by $\text{RMSE}(\hat{\theta}) = \sqrt{\frac{\sum (\hat{\theta} - \theta)^2}{500}}$. Coverage rate is the percentage of derived intervals including the true population mean (θ). Actual interval length is the average interval



Scenario 1: $u_b = 0$ and $\sigma_b^2 = 1$. Effect size: 0. Scenario 2: $u_b = 0$ and $\sigma_b^2 = 2$. Effect size: 0. Scenario 3: $u_b = 0.3$ and $\sigma_b^2 = 2$. Effect size: 0.24. Scenario 4: $u_b = 0.3$ and $\sigma_b^2 = 1$. Effect size: 0.30. Scenario 5: $u_b = 0.5$ and $\sigma_b^2 = 2$. Effect size: 0.41. Scenario 6: $u_b = 0.5$ and $\sigma_b^2 = 1$. Effect size: 0.50. Scenario 7: $u_b = 0.7$ and $\sigma_b^2 = 2$. Effect size: 0.57. Scenario 8: $u_b = 0.7$ and $\sigma_b^2 = 1$. Effect size: 0.7. Scenario 9: $u_b = 0.7$ and $\sigma_b^2 = 0.5$. Effect size: 0.81. In all scenarios, $u_a = 0$ and $\sigma_a^2 = 1$.

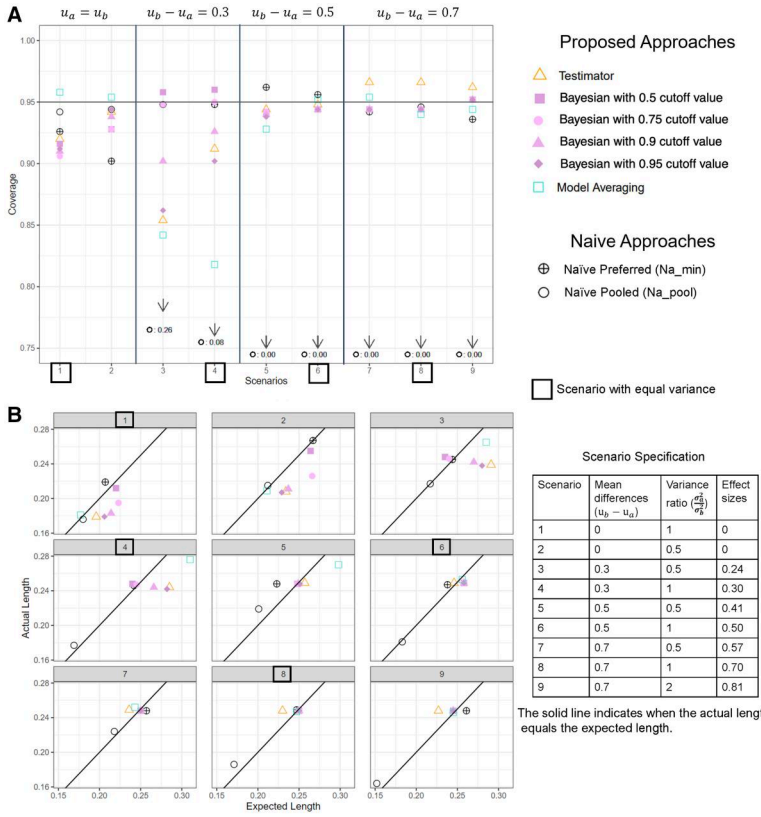
Figure 2. Bias and RMSE in the Simulation Study When Information about the Preferred Direction Is Available.

length computed using derived intervals (upper bound—lower bound) across all simulation samples. Expected interval length is the theoretical interval length computed using empirical distributions computed from all simulation samples. Effect size in the simulation study is defined as $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}}$.

We focus on presenting the results of $\rho = 0.95$, as (i) this might best reflect the high correlation between responses collected with different modes and (ii) simulation results do not show significant variations across different ρ values. Results for $\rho = 0.75$ and $\rho = 0.5$ can be found in tables A3, A4, A5, and A6 in the supplementary data online.

Figures 2 and 3 show the performances of the point estimates and uncertainty measures respectively when prior information is available and $\rho = 0.95$ (table A1 in the supplementary data online). When mode effects do not exist (scenarios 1 and 2), the Naïve Preferred (`Na_min`) is largely biased and the Naïve Pooled (`Na_pool`) is almost unbiased. In contrast, when mode effects are present, the Naïve Preferred is unbiased and the Naïve Pooled is very biased. Our proposed approaches largely reduce biases regardless of whether mode effects are present; thus providing more robust inferences than the Naïve approaches.

To compare the three approaches, we note that when mode effects do not exist, a wider interval (e.g., 95 percent) in the Bayesian approach performs better. This is expected because a wider interval is associated with fewer false positive errors, thus providing more accurate estimates of the population mean



Scenario	Mean differences $(u_b - u_a)$	Variance ratio $(\frac{\sigma_b^2}{\sigma_a^2})$	Effect sizes
1	0	1	0
2	0	0.5	0
3	0.3	0.5	0.24
4	0.3	1	0.30
5	0.5	0.5	0.41
6	0.5	1	0.50
7	0.7	0.5	0.57
8	0.7	1	0.70
9	0.7	2	0.81

Figure 3. Coverage Rates and Interval Length in the Simulation Study When Information about the Preferred Direction Is Available.

in that case. When mode effects are present, a narrower interval works better as it is more sensitive in detecting mode effects. Therefore, in determining the cutoff values in the Bayesian approach, there is a tradeoff between increasing power and the risk of making a type I error. We note that the model averaging approach has a large bias in scenarios 3–5, where effect sizes are smaller than 0.5. This is related to the diffuse weight distribution in the scenarios, which will be discussed further in the next paragraph. Except in these scenarios, the model averaging approach shows minor bias and good coverage properties. The Testimator approach performs well in moderate and large mode effects (scenarios 5–9). When mode effects do not exist (scenarios 1 and 2), the coverage rates of the Testimator approach are lower than 0.95 (figure 3). When mode effects are small (effect sizes smaller than 0.5, scenarios 3 and 4), the approach has slightly larger biases, compared to the Naïve Preferred and the

Table 2. Probability of the Model Being Selected in the Testimator Approach ($\rho = 0.95$ and $n = 500$)

Scenarios	Effect sizes	Models (mean–variance)			
		M1 (D–D)	M2 (S–D)	M3 (D–S)	M4 (S–S)
1. $u_b = 0,\ \sigma_b^2 = 1$	0.00	0.000	0.044	0.046	**0.910**
2. $u_b = 0,\ \sigma_b^2 = 2$	0.00	0.058	**0.942**	0.000	0.000
3. $u_b = 0.3,\ \sigma_b^2 = 2$	0.24	**0.764**	0.236	0.000	0.000
4. $u_b = 0.3, \sigma_b^2 = 1$	0.30	0.052	0.004	**0.874**	0.070
5. $u_b = 0.5, \sigma_b^2 = 2$	0.41	**0.994**	0.006	0.000	0.000
6. $u_b = 0.5, \sigma_b^2 = 1$	0.50	0.054	0.000	**0.946**	0.000
7. $u_b = 0.7, \sigma_b^2 = 2$	0.57	**1.000**	0.000	0.000	0.000
8. $u_b = 0.7, \sigma_b^2 = 1$	0.70	0.034	0.000	**0.966**	0.000
9. $u_b = 0.7, \sigma_b^2 = 0.5$	0.81	**1.000**	0.000	0.000	0.000

NOTE.—"D" stands for different and "S" stands for same. The correct model of a scenario is marked in bold in the table. Model 1 corresponds to the different mean different variance model, model 2 is the same mean different variance model, model 3 is the different mean and same variance model, and model 4 is the same mean and same variance model. Effect sizes are computed as $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}}$.

Bayesian approach that applies more sensitive cutoff values (figure 2, i.e., the ones computed from 50, 75, and 90 percent intervals).

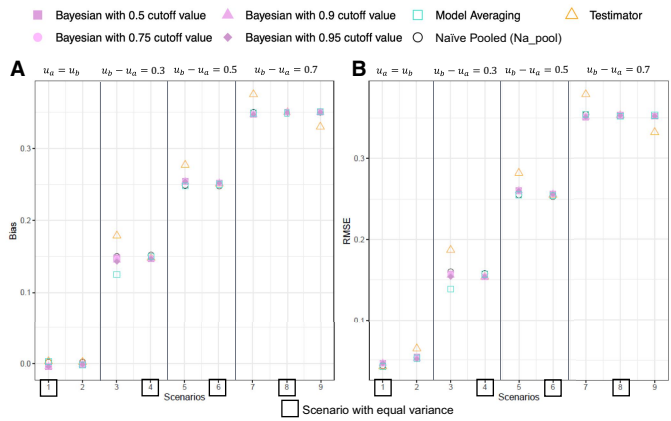
To further illustrate the performances of the Testimator and the model averaging approaches, we show the probabilities of selecting four models in tables 2 and 3. We compute the probabilities as the frequency of a model being selected divided by the number of simulations in the Testimator approach. In the model averaging approach, we compute the probabilities using the marginal posteriors as introduced in section 2.3. We note that when the effect size is less than 0.50, both the Testimator approach (the probabilities $p = .764$ and .874 in scenarios 3 and 4, respectively) and the model averaging approach (0.345, 0.595, and 0.830, respectively, for scenarios 3–5) have a lower probability of selecting the right model. In these scenarios, besides the correct model, they are most likely to choose the model with equal means (models 2 and 4).

Figures 4 and 5 present the simulation results when we have no information about preferred modes or directions. Detailed results can be found in table A2 in the supplementary data online. The three proposed methods and the Naïve Pooled lead to a similar level of bias in the point estimate of the population mean. Moreover, the bigger the mode effects, the larger the bias in the point estimates of all approaches. However, the three approaches achieve a much better coverage rate than the Naïve approach. When mode effects are present, the interval computed with the Naïve rarely includes the true population mean, while the three approaches mostly achieve a 95 percent coverage rate. Thus,

Table 3. Average Weights in the Model Averaging Approach ($\rho = 0.95$ and $n = 500$)

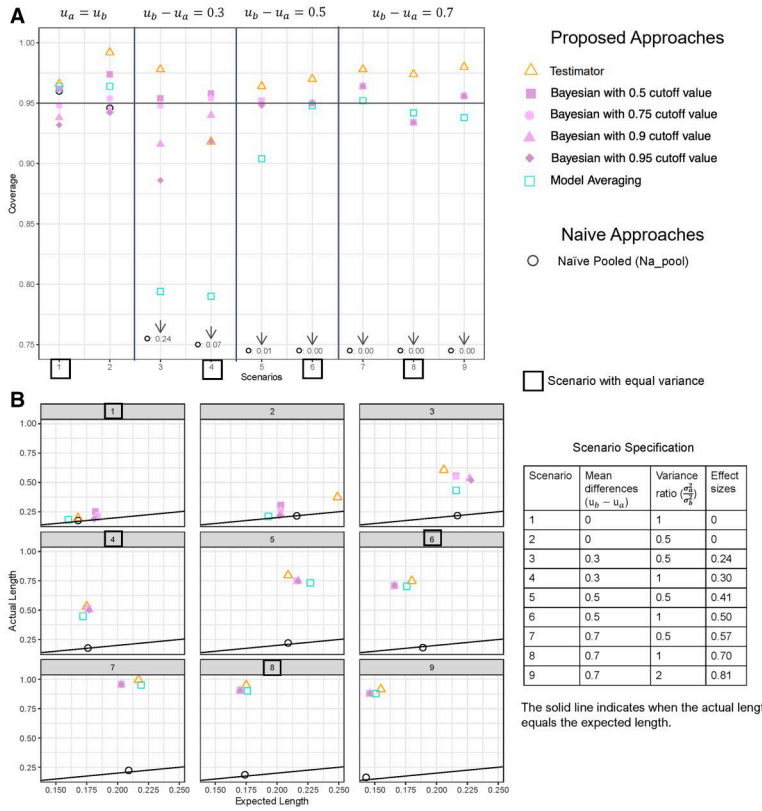
Scenarios	Effect sizes	Models (mean–variance)			
		M1 (D–D)	M2 (S–D)	M3 (D–S)	M4 (S–S)
1. $u_b = 0, \sigma_b^2 = 1$	0.00	0.000	0.001	0.022	**0.977**
2. $u_b = 0, \ \sigma_b^2 = 2$	0.00	0.018	**0.911**	0.002	0.070
3. $u_b = 0.3, \ \sigma_b^2 = 2$	0.24	**0.345**	0.576	0.028	0.051
4. $u_b = 0.3, \sigma_b^2 = 1$	0.30	0.001	0.000	**0.595**	0.404
5. $u_b = 0.5, \ \sigma_b^2 = 2$	0.41	**0.830**	0.079	0.081	0.010
6. $u_b = 0.5, \ \sigma_b^2 = 1$	0.50	0.002	0.000	0.989	0.009
7. $u_b = 0.7, \ \sigma_b^2 = 2$	0.57	**0.900**	0.003	0.098	0.000
8. $u_b = 0.7, \ \sigma_b^2 = 1$	0.70	0.001	0.000	**0.999**	0.000
9. $u_b = 0.7, \ \sigma_b^2 = 0.5$	0.81	**0.896**	0.000	0.104	0.000

NOTE.—"D" stands for different and "S" stands for same. The correct model of a scenario is marked in bold in the table. Model 1 corresponds to the different mean different variance model, model 2 is the same mean different variance model, model 3 is the different mean and same variance model, and model 4 is the same mean and same variance model. Effect sizes are computed as $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}}$.



Scenario 1: $u_b = 0$ and $\sigma_b^2 = 1$. Effect size: 0. Scenario 2: $u_b = 0$ and $\sigma_b^2 = 2$. Effect size: 0. Scenario 3: $u_b = 0.3$ and $\sigma_b^2 = 2$. Effect size: 0.24. Scenario 4: $u_b = 0.3$ and $\sigma_b^2 = 1$. Effect size: 0.30. Scenario 5: $u_b = 0.5$ and $\sigma_b^2 = 2$. Effect size: 0.41. Scenario 6: $u_b = 0.5$ and $\sigma_b^2 = 1$. Effect size: 0.50. Scenario 7: $u_b = 0.7$ and $\sigma_b^2 = 2$. Effect size: 0.57. Scenario 8: $u_b = 0.7$ and $\sigma_b^2 = 1$. Effect size: 0.7. Scenario 9: $u_b = 0.7$ and $\sigma_b^2 = 0.5$. Effect size: 0.81. In all scenarios, $u_a = 0$ and $\sigma_a^2 = 1$.

Figure 4. Bias and RMSE in the Simulation Study When No Information Is Available.



Scenario	Mean differences ($u_b - u_a$)	Variance ratio ($\frac{\sigma_a^2}{\sigma_b^2}$)	Effect sizes
1	0	1	0
2	0	0.5	0
3	0.3	0.5	0.24
4	0.3	1	0.30
5	0.5	0.5	0.41
6	0.5	1	0.50
7	0.7	0.5	0.57
8	0.7	1	0.70
9	0.7	2	0.81

Figure 5. Coverage Rates and Interval Length in the Simulation Study When No Information Is Available.

although the proposed methods rarely reduce bias over the Naïve approach, they do have better coverage.

To compare the three approaches in this setting, we note that the Testimator approach provides slightly more conservative inferences than the other two approaches (except in scenario 4). In the Bayesian approach, the wider the "interval" chosen (the interval of the estimated effective size [$\hat{\eta}$] in the testing phase), the narrower the interval width. This makes sense as when we apply more strict criteria (i.e., wider "intervals") to detect mode effects, we are more likely to find no mode effects and thus make inferences from the common mean model. This leads to smaller variability than pooling draws from different mean models. Yet, if mode effects are very large, the choice of the width of the "interval" (i.e., 50, 75, 90, or 95 percent) does not make a difference as they all lead to the same conclusion. Lastly, the model averaging approach shows poorer coverage in scenarios 3–5. This is again related to the diffuse weight distribution in these scenarios (table 3), as the approach often picks the

same mean models (models 2 and 4) when the different mean models (models 1 and 3) are correct.

In general, the Testimator, Bayesian, and model averaging approaches are useful across all scenarios, as they achieve robust inferences and improve coverage rates compared to the Naïve approaches. Yet, they can be suboptimal in certain scenarios, especially when the mode effects are small or nonexistent. To explore whether the issues may be alleviated in large samples, we increase the sample size to 5,000 and rerun the simulation when a smaller estimate is preferred (setting 1). We illustrate the probability distribution of the Testimator and model averaging approach over four models with $n = 5{,}000$ in tables A7 and A8 in the supplementary data online. In the model averaging approach, the correct model of each scenario always has a weight approximating 1, largely reducing bias when the true effect size is smaller than 0.5 (scenarios 3–5, see table A9 in the supplementary data online). For the Testimator approach, the probabilities of picking the correct model when mode effects are small (scenarios 3 and 4) increase a great deal. However, when there is no shift in variances in the population (scenarios 1, 4, 6, and 8), the probabilities of the Testimator approach selecting different mean models do not converge to 0 as sample size increases. In table A9 in the supplementary data online, the increased sample size eliminates the sensitivity of simulation results to the choice of interval length in the Bayesian approach (scenarios 3–9 when mode effects are present). These results suggest that large sample sizes can largely improve the performances of the three approaches, especially in scenarios of small effect sizes.

In sum, when prior information is available, the simulation results show that our proposed methods provide robust inferences, compared to pooling the mixed-mode data or always taking the estimate in the preferred direction. When there is no prior information, the proposed methods provide intervals with generally good coverage properties, outperforming the approach that pools the data.

4. APPLICATION: ARAB BAROMETER WAVE 6 JORDAN EXPERIMENT

To illustrate how the proposed methods can be applied to mixed-mode surveys with complex sample designs, as well as to provide an example application of the proposed methods, we consider the Arab Barometer wave 6 Jordan mixed-mode experiment data ($n = 2,531$, $n_{\text{FTF}} = 1,193$, $n_{\text{TEL}} = 1,338$).

As mentioned before, the research team applied a randomized mixed-mode experiment in Jordan, where 1/3 of the households were interviewed via FTF, and the remaining households were assigned to the TEL mode. The TEL-assigned households were initially recruited via FTF for a short 5-minute survey, and the majority of the survey items were asked approximately a week

later in a TEL follow-up. Since the mode assignment was randomized and both mode groups were recruited via FTF, selection effects would be attributable to attrition to the TEL follow-up for participants assigned to TEL. We assume the attrition from the FTF screening is missing at random. To account for it, we apply attrition weights to the TEL group so that final TEL respondents resemble the initial TEL group who received the quick FTF interview on key demographic variables (i.e., education, age, gender, number of people in a household, marital status, and region).

In this article, we construct a measure of satisfaction toward government ("`gov_sat`") as our outcome variable (range: 1–24). The measure is constructed as 25 minus the sum of six 4-point ordinal variables: (i) government's performance on security, (ii) government's performance on keeping price down, (iii) government's performance on responding to COVID-19, (iv) government's performance overall, (v) government's performance on education system, and (vi) government's performance on healthcare system. The first three variables are coded as 1 = "Completely satisfied," 2 = "Satisfied," 3 = "Dissatisfied," and 4 = "Completely dissatisfied." The last three variables are coded as 1 = "Very good," 2 = "Good," 3 = "Bad," and 4 = "very bad." We reverse code the variable so that the higher the outcome, the more satisfied a participant is with the government.

We consider two settings in this application: (i) when we are aware that a smaller estimate is closer to the truth and (ii) when we cannot determine the preferred direction. The first setting is convincing in this context because regime support can be subject to self-censorship in authoritarian countries (Robinson and Tannenberg 2019). We consider the second scenario as a sensitivity analysis.

To account for the complex survey designs employed in this study, we compute sample variances (S_w^2), sampling variances (v_w^2), and sample means ($\bar{y}_w$), accounting for weights, stratification, and clustering separately for FTF and TEL modes. Using these quantities, we then compute design effects and the effective sample size for each mode. Finally, we incorporate the effective sample size and design-based sufficient statistics (sample means and sample variances) into the proposed approaches to account for complex survey designs. We use linearization to compute the sampling variances respectively for the two modes, implemented by the svymean function in the survey package in R. Table 4 shows the results when we know that smaller estimates are preferred. Because all four cutoff values in the Bayesian approaches lead to the same results, we only show one set of estimates for the Bayesian approach. The Testimator, Bayesian, and model averaging approaches give a similar point estimate as using data collected via FTF alone. This suggests that the proposed approaches detect substantial mode effects between FTF and TEL and mostly rely on FTF data to make population inferences. The point estimate computed using FTF data is much lower than the estimate computed using TEL data. This is in line with previous findings that FTF can reduce socially desirable

Table 4. Results When Smaller Estimates Are Preferred

	Proposed approaches			Naïve approaches		
	Testimator	Bayesian	Model averaging	Pool data	FTF	TEL
Estimate	7.521	7.519	7.521	8.295	7.521	9.088
Interval	7.197, 7.844	7.193, 7.849	7.197, 7.851	8.058, 8.531	7.200, 7.841	8.825, 9.351
Interval length	0.647	0.656	0.654	0.473	0.642	0.526

NOTE.—In the Testimator approach, the interval corresponds to a 95 percent confidence interval. In the Bayesian and the model averaging approach, the interval refers to a 95 percent credible interval. Other choices of the cutoff values (50, 75, and 90 percent) in the Bayesian approach lead to the same results as 95 percent in this application.

Table 5. Results When There Is No Prior Information

	Proposed approaches			Naïve approaches		
	Testimator	Bayesian	Model averaging	Pool data	FTF	TEL
Estimate	8.275	8.293	8.297	8.295	7.521	9.088
Interval	7.197, 9.352	7.254, 9.307	7.246, 9.304	8.058, 8.531	7.200, 7.841	8.825, 9.351
Interval length	2.155	2.053	2.058	0.473	0.642	0.526

NOTE.—In the Testimator approach, the interval corresponds to a 95 percent confidence interval. In the Bayesian and the model averaging approach, the interval refers to a 95 percent credible interval. Other choices of the cutoff values (50, 75, and 90 percent) in the Bayesian approach lead to the same results as 95 percent in this application.

reporting relative to TEL (Holbrook et al. 2003). In Jordan's context, respondents may fear being eavesdropped on or be suspicious about the identity of interviewers during phone interviews. Meanwhile, in FTF contacts, respondents have more visual clues to determine the identity of interviewers. Consequently, participants may have a higher trust in interviewers in FTF than in TEL and thus tend to report more honest answers in FTF.

We present the results when having no information about preferred directions in table 5. In this case, the proposed approaches provide point estimates similar to the pooled estimate. Yet, the intervals created by the proposed

approaches are much wider than the Naïve estimates, which reflect the additional uncertainty associated with the scenario. Therefore, the intervals computed from the proposed approaches have a higher chance of including the true population mean.

5. DISCUSSION

This article proposes three procedures to account for potential mode effects when dealing with mixed-mode data. By embedding testing procedures and incorporating available information about mode effects, we achieve robust inferences compared to standard approaches such as picking a single mode, or ignoring potential mode effects. All three approaches are proposed to achieve the same purpose; however, the ideas behind the approaches are different. The Testimator approach can be seen as frequentist model selection. In the approach, we follow a sequential testing procedure, where the use of the *t*-test depends on the *F* test results. The Bayesian approach is a Bayesian version of model selection. In the Bayesian model averaging approach, the equality of means and variances is evaluated concurrently. In addition, the model averaging approach combines estimates across different models, while the other two approaches apply explicit testing procedures to select the most plausible model.

Conceptually, we expect the model averaging approach to be the most robust method as it incorporates the uncertainty in the models. However, simulation results show a clear advantage for any of the three methods. Compared to the Testimator and the Bayesian, the model averaging approach is less sensitive in detecting small mode effects, especially when the sample size is limited. This observation may be due to the weights used in this approach, which are linked to the marginal posteriors of the model being correct given data. For instance, examining the marginal posterior of model 1 (refer to appendix A in the supplementary data online), we observe that the contributions of $\bar{y}_a$ and $\bar{y}_b$ are very small relative to s_a^2 and s_b^2 when we use diffuse priors (i.e., small k_a and k_b). This can result in suboptimal performance of the model averaging approach when effect sizes are less than 0.5. The study also suggests special caution is needed when mode effects are small, since all the proposed approaches show larger bias in the scenarios except for the Bayesian approach with a small cutoff value (e.g., 50 percent). This issue will be alleviated in practical settings when the sample size for each mode is larger. Researchers may use smaller critical values to enlarge the rejection region. However, this option comes with the price of increasing the type 1 error rate when mode effects do not exist in reality. We recommend researchers start with a 95 percent cutoff value in the Bayesian approach. They can consider a narrower interval (such as 75 or 90 percent) when additional sensitivity is needed to detect small mode effects. The idea also applies to the Testimator and the model averaging approaches, where the significance levels can be modified.

As for the choice between the Bayesian and the Testimator approaches, the Testimator may be easier to implement, while the Bayesian approach provides a solution for researchers who favor Bayesian methods over frequentist approaches.

This article considers one type of prior information: the preferred directions. Depending on the mode used in data collection and findings in the literature, other prior information may be useful in inferences. For example, instead of a preferred direction, researchers may have a preferred mode when they have reasons to believe one mode gives less biased estimates than another mode. In this case, the question becomes whether the other mode provides comparable estimates as the preferred mode. The three approaches can be easily adapted to address the question by always taking the estimate provided by the preferred mode if differences are detected between the mode-specific estimates. When differences are not detected, we can compute the estimate of the population as some average of mode-specific estimates in a similar fashion as this article.

This article uses very weakly informative normal priors for means u_a, u_b and inverse-gamma priors for variances σ_a^2, σ_b^2. It is noted that half-t priors for Gaussian standard deviation parameters perform better than inverse-gamma family priors in hierarchical models (Gelman 2006). Nevertheless, this article still uses the inverse-gamma priors for their conjugate properties, which greatly simplify the process of computing weights in the model averaging approach. To remain consistent across the approaches, we also use inverse-gamma family priors in the Bayesian approach.

This article applies the proposed methods to a relatively simple mixed-mode scenario: a randomized mixed-mode experiment. If randomization is achieved, participants assigned to each mode should be homogeneous and any selection effects are attributable to nonresponse or attrition, depending on sample designs. This article computes attrition weights to account for the selection effects. However, if sequential mixed-mode or concurrent mixed-mode designs are used, the sample composition across modes may differ by design. We recognize that this is a more realistic scenario for multimode designs. Under that circumstance, selection effects can be a bigger caveat for population inferences and thus necessitate more advanced tools (such as propensity score adjustments) in causal inference to account for them. For example, propensity score stratification can be used in these scenarios to achieve balance in the distributions of covariates across modes. Specifically, we can apply the proposed approaches in each propensity stratum and then combine estimates across strata using relative sample sizes as weights. The relative samples sizes are computed as $\frac{n_{mh}}{n_m}$, where n_{mh} stands for the sample size in propensity stratum h in mode m and n_m means the total sample size with mode m. In this case, users do not rely solely on data collected with one mode but may utilize information gathered with different modes in various strata.

This article provides novel approaches to connect the testing and the inferences of mode effects. Kolenikov and Kennedy (2014) classify mode effects

literature into three aims: (i) determining the magnitude of mode effects, (ii) providing population estimates, and (iii) obtaining case-level estimates. This article connects the first two types of studies. Despite the copious findings made by aim 1 literature, no prior study provides principled approaches to incorporate such information to adjust for mode effects when making population inferences. This article addresses the important research gap by proposing procedures for different scenarios depending on whether we have prior information about preferred directions or not.

This article provides a useful framework for combining mode-specific estimates produced from multimode designs. We illustrate the proposed approaches using normal outcomes. However, these approaches can be adapted for other types of variables. For example, the Bayesian and the model averaging approaches can easily account for binary variables using a latent probit framework. In the Testimator approach, we can test whether $p_a = p_b$, where p_a and p_b represent population proportions measured by modes A and B, using a pooled Z test of proportions. Semiparametric methods like bootstrap can be used for other types of variables. Furthermore, the approaches developed in this article assume two modes for data collection, but they can be adapted for scenarios with three modes (e.g., Web, TEL, and FTF). Additionally, the simulation study suggests that the Testimator approach might result in overly conservative inferences when the preferred direction is unknown. Future work could explore alternative methods for constructing robust and well-calibrated CIs in this particular setting.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

REFERENCES

Brick, J. M., Kennedy, C., Cervantes-Flores, I., and Mercer, A. W. (2021), "An Adaptive Mode Adjustment for Multimode Household Surveys," *Journal of Survey Statistics and Methodology*, 10, 1024–1047.

Buelens, B., and van den Brakel, J. A. (2015), "Measurement Error Calibration in Mixed-Mode Sample Surveys," *Sociological Methods & Research*, 44, 391–426.

de Leeuw, D., and Desiree, E. (1992), *Data Quality in Mail, Telephone and Face to Face Surveys*, Amsterdam: TT Publikaties.

de Leeuw, E. D., Suzer-Gurtekin, Z. T., and Hox, J. J. (2018), "The Design and Implementation of Mixed-Mode Surveys," in *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*, eds. T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, and B. Dorer, Hoboken, NJ: John Wiley & Sons, Inc, pp. 387–408.

de Leeuw, E. D. (2018), "Mixed-Mode: Past, Present, and Future," *Survey Research Methods*, 12, 75–89.

Dillman, D. A., and Christian, L. M. (2005), "Survey Mode as a Source of Instability in Responses across Surveys," *Field Methods*, 17, 30–52.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K., and Giordano, L. (2009), "Effects of Survey Mode, Patient Mix, and Nonresponse on CAHPS® Hospital Survey Scores," *Health Services Research*, 44, 501–518.

Gelman, A. (2006), "Prior Distributions for Variance Parameters in Hierarchical Models (Comment on Article by Browne and Draper)," *Bayesian Analysis*, 1, 515–534.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999), "Bayesian Model Averaging: A Tutorial (with Comments by M. Clyde, David Draper and EI George, and a Rejoinder by the Authors)," *Statistical Science*, 14, 382–417.

Holbrook, A. L., Green, M. C., and Krosnick, J. A. (2003), "Telephone versus Face-to-Face Interviewing of National Probability Samples with Long Questionnaires: Comparisons of Respondent Satisficing and Social Desirability Response Bias," *Public Opinion Quarterly*, 67, 79–125.

Holbrook, A. L., and Krosnick, J. A. (2010), "Social Desirability Bias in Voter Turnout Reports: Tests Using the Item Count Technique," *Public Opinion Quarterly*, 74, 37–67.

Klausch, T., Hox, J. J., and Schouten, B. (2013), "Measurement Effects of Survey Mode on the Equivalence of Attitudinal Rating Scale Questions," *Sociological Methods & Research*, 42, 227–263.

Kolenikov, S., and Kennedy, C. (2014), "Evaluating Three Approaches to Statistically Adjust for Mode Effects," *Journal of Survey Statistics and Methodology*, 2, 126–158.

Kreuter, F., Presser, S., and Tourangeau, R. (2008), "Social Desirability Bias in Cati, Ivr, and Web Surveysthe Effects of Mode and Question Sensitivity," *Public Opinion Quarterly*, 72, 847–865.

Lavrakas, P. J. (2008), *Encyclopedia of Survey Research Methods*, Thousand Oaks, CA: Sage publications.

Massey, D. S., and Tourangeau, R. (2013), "Where Do we Go from Here? Nonresponse and Social Measurement," *The Annals of the American Academy of Political and Social Science*, 645, 222–236.

Nadarajah, S., and Kotz, S. (2008), "Exact Distribution of the Max/Min of Two Gaussian Random Variables," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16, 210–212.

Park, S., Kim, J. K., and Park, S. (2016), "An Imputation Approach for Handling Mixed-Mode Surveys," *The Annals of Applied Statistics*, 10, 1063–1085.

Powers, J. R., Mishra, G., and Young, A. F. (2005), "Differences in Mail and Telephone Responses to Self-Rated Health: Use of Multiple Imputation in Correcting for Response Bias," *Australian and New Zealand Journal of Public Health*, 29, 149–154.

Presser, S., and Stinson, L. (1998), "Data Collection Mode and Social Desirability Bias in Self-Reported Religious Attendance," *American Sociological Review*, 63, 137–145.

Robinson, D., and Tannenberg, M. (2019), "Self-Censorship of Regime Support in Authoritarian States: Evidence from List Experiments in China," *Research & Politics*, 6, 205316801985644.

Schwarz, N., Hippler, H.-J., Deutsch, B., and Strack, F. (1985), "Response Scales: Effects of Category Range on Reported Behavior and Comparative Judgments," *Public Opinion Quarterly*, 49, 388–395.

Suzer-Gurtekin, Z. T., Heeringa, S., and Vaillant, R. (2012), "Investigating the Bias of Alternative Statistical Inference Methods in Sequential Mixed-Mode Surveys," in *Proceedings of the JSM, Section on Survey Research Methods*, Alexandria, VA: American Statistical Association, pp. 4711–2.

Tourangeau, R., and Smith, T. W. (1996), "Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format, and Question Context," *Public Opinion Quarterly*, 60, 275–304.

Vannieuwenhuyze, J., Loosveldt, G., and Molenberghs, G. (2010), "A Method for Evaluating Mode Effects in Mixed-Mode Surveys," *Public Opinion Quarterly*, 74, 1027–1045.

Vannieuwenhuyze, J. T. A., and Revilla, M. (2013), "Relative Mode Effects on Data Quality in Mixed-Mode Surveys by an Instrumental Variable," *Survey Research Methods*, 7, 157–168.

Voogt, R., and Saris, W. (2005), "Mixed Mode Designs: Finding the Balance between Nonresponse Bias and Mode Effects," *Journal of Official Statistics*, 21, 367.

Journal of Survey Statistics and Methodology, 2024, 12, 814–839
https://doi.org/10.1093/jssam/smae012
Article