Improving Inferences Based on Survey Data Collected Using Mixed-mode Designs

by

Wenshan Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Survey and Data Science)
in the University of Michigan
2024

Doctoral Committee:

Professor Trivellore E. Raghunathan, co-Chair
Professor Michael R. Elliott, co-Chair
Assistant Research Scientist Tuba Suzer Gurtekin
Professor Jacqui Smith
Research Professor James R. Wagner

Wenshan Yu

yuwens@umich.edu

ORCID iD: 0000-0001-8603-4946

ACKNOWLEDGEMENTS

The process of writing this dissertation has been a challenging but rewarding journey for me. This journey would not have been possible without the constant and generous help from the MPSM faculties, peers, colleagues in ISR, my friends, and my family.

First, I deeply appreciate my advisors, Trivellore Raghunathan and Michael Elliott, for guiding me over the past three years. Raghu has always been available to offer his wise suggestions and feedback whenever I needed them. Mike has consistently provided useful and insightful comments on every version of my dissertation drafts. They have not only guided me in my research but also taught me how to mentor students by setting perfect examples. I appreciate all the time and patience they devoted to me, and I will always consider them to be my academic role models in the field of survey statistics and methodology.

I also cannot express enough gratitude to my GSRA supervisor and my committee member, Jacqui Smith. She inspires me with many interesting social science and methodological ideas and enlightens me about the real concerns faced in data collection. I am fortunate to have James Wagner and Tuba Suzer Gurtekin on my committee as well. I enjoyed talking with James since I was a master's student in the program. Tuba Suzer Gurtekin has generously shared much of her experience in mixed-mode design and inferences with me.

In addition, I appreciate the writing training I received from participating in the doctoral seminar instructed by Brady West and Katharine Abraham. The research assistant work I did with Sunghee Lee during the master's program equipped me with a solid foundation for future research. I am grateful to Zeina Mneimneh and Fred Conrad for their mentoring during the master's program.

I will treasure my memories at MPSM, ISR, and Umich, where I met many great coworkers and friends. Your generous help, warm personalities, and wise suggestions have helped me become a better person while pursuing a PhD degree. Last, I thank my families for their ongoing emotional support, especially during the Covid pandemic.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS ii

LIST OF FIGURES v

LIST OF TABLES vi

ABSTRACT viii

CHAPTER

1 Introduction 1

1.1 Background 1
1.2 Dissertation Structure 4

2 Three Approaches to Adjust for Mode Effects 6

2.1 Introduction 6
2.1.1 Literature Review 7
2.2 Proposed Methods 8
2.2.1 The Testimator Approach 10
2.2.2 The Bayesian Approach 12
2.2.3 The Model Averaging Approach 15
2.2.4 Naïve Approach 1: Use the Estimate in the Preferred Direction (Naïve Preferred) 18
2.2.5 Naïve Approach 2: Pool the Data (Naïve Pooled) 18
2.3 Simulation Study 19
2.4 Application: Arab Barometer Wave 6 Jordan Experiment 28
2.5 Discussion 32
2.6 Appendix 36
2.6.1 Appendix A: Derivation of Weights in the Model Averaging Approach 36
2.6.2 Appendix B: Additional Simulation Results 38

3 Investigating Mode Effects in Interviewer Variances Using Two Representative Multi-mode Surveys 59

3.1 Introduction 59
3.2 The Arab Barometer Study 62
3.2.1 Study Description 62
3.2.2 Analytical Strategy 63

3.2.3 Results 66
3.3 Health and Retirement Study 2016 75
3.3.1 Study Description 75
3.3.2 Analytical Strategy 76
3.3.3 Results 77
3.4 Simulation Study 87
3.4.1 Arab Barometer Study 87
3.4.2 Health and Retirement Study 90
3.5 Discussion 92
3.6 Appendix 95

4 Using Principal Stratification to Detect Mode Effects across Waves in a Longitudinal Study 112

4.1 Introduction 112
4.2 Methods 116
4.2.1 Framework 116
4.2.2 Data 117
4.2.3 Cross-sectional Analysis 120
4.2.4 Longitudinal Analysis 127
4.3 Results 130
4.3.1 The HRS 2016 130
4.3.2 The HRS 2018 132
4.3.3 The HRS 2016-2018 140
4.4 Discussion 145
4.5 Appendix 148
4.5.1 Longitudinal Analysis Supplementary Tables 148

5 Conclusion and Future Work 152

BIBLIOGRAPHY 157

LIST OF FIGURES

FIGURE

2.1 Flowchart of the Testimator Approach when some Information about the Preferred Direction is Available 12
2.2 Bias and RMSE in the Simulation Study when Information about the Preferred Direction is Available 22
2.3 Coverage Rates and Interval Length in the Simulation Study when Information about the Preferred Direction is Available 23
2.4 Bias and RMSE in the Simulation Study when no Information is Available . . . 26
2.5 Coverage Rate and Length in the Simulation Study when no Information is Available 27
2.6 Distribution of the Outcome Variable in the Arab Barometer Application 30

3.1 Interviewer Workloads Per Mode in the Arab Barometer Study 66
3.2 Interviewer Workloads Per Mode in the Health and Retirement Study 78

4.1 Visualization of the HRS 2016-2018 Mixed-mode Designs 114
4.2 Observed Data Structure of HRS 2016 and 2018 120

LIST OF TABLES

TABLE

2.1 Simulation Scenarios 19

2.2 Probability of the Model Being Selected in the Testimator Approach (ρ = 0.95 and n = 500) 24

2.3 Average Weights in the Model Averaging Approach (ρ = 0.95 and n = 500) . . . 25

2.4 Results when Smaller Estimates are Preferred 32

2.5 Results when There is no Prior Information 32

2.6 Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and ρ = 0.95) 38

2.7 Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions (n = 500 and ρ = 0.95) 41

2.8 Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and ρ = 0.75) 43

2.9 Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions (n = 500 and ρ = 0.75) 46

2.10 Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and ρ = 0.5) 48

2.11 Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions (n = 500 and ρ = 0.5) 51

2.12 Probabilities of the Model Being Selected in the Testimator Approach (n = 5000 and ρ = 0.95) 54

2.13 Average Weight of Model Averaging Approach (n = 5000 and ρ = 0.95) 55

2.14 Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 5000 and ρ = 0.95) 56

3.1 Unweighted Distribution of Outcome variables in the Arab Barometer Study across Interviewers by Modes 68

3.2 Unweighted Distribution of Sample Characteristics of the Arab Barometer Study across Interviewers by Modes 70

3.3 Interviewer Variances Per Mode for Selected Items in the Arab Barometer Study Adjusting for Covariates Using Bayesian Estimation 73

3.4 Distribution of Sample Characteristics in the Health and Retirement Study across Interviewers by Modes 79

3.5 Distribution of Outcome variables in the Health and Retirement Study across Interviewers by Modes 81

3.6 Interviewer Variances Per Mode for Selected Items in Health and Retirement Study Adjusting for Covariates Using Bayesian Estimation 85
3.7 Simulation study using the Arab Barometer Setup 89
3.8 Simulation study using the HRS Setup 91
3.9 Outcome Variables Used in the Arab Barometer Study 95
3.10 Interviewer Variances Per Mode for Selected Items in the Arab Barometer Study Adjusting for Covariates 98
3.11 Outcome Variables Used in the Health and Retirement Study 101
3.12 Interviewer Variances Per Mode for Selected Items in Health and Retirement Study Adjusting for Covariates 104
3.13 Interviewer Variances Per Mode for Depression Items in Health and Retirement Study Adjusting for Covariates Using Likelihood Estimation 109

4.1 Descriptive Statistics of the HRS sample (Unweighted) 118
4.2 Models Per Principal Stratum for the HRS 2016 121
4.3 Principal Strata considered for HRS 2018 126
4.4 Principal Strata considered for HRS 2016-2018 129
4.5 Mode Effect Estimates for the HRS 2016 131
4.6 Fraction of Missing Information and Sample Size for the Principal Strata in the HRS 2016 134
4.7 Stratum-specific Mode Effect Estimates for the HRS 2018 135
4.8 Fraction of Missing Information and Sample Size for the Principal Strata in the HRS 2018 137
4.9 Combined Mode Effect Estimates for the HRS 2018 140
4.10 Stratum-specific Mode Effect Estimates for the HRS 16-18 142
4.11 Combined Mode Effect Estimates for the HRS 16-18 145
4.12 Information about the Principal Strata for the HRS 16-18 149
4.13 Fraction of Missing Information the Combined Principal Strata for the HRS 16-18151

ABSTRACT

Mixed-mode designs have become increasingly common in survey data collection. However, different modes may have different measurement properties, which need to be accounted for when analyzing mixed-mode data. This dissertation investigates the presence of mode effects in means and interviewer variances in both cross-sectional and longitudinal studies, and it develops methods to incorporate mode effects when making inferences. Specifically, Study 1 proposes three approaches to detect and address potential mode effects in cross-sectional data collected with randomized mixed-mode designs. We applied this work to assess face-to-face (FTF) versus telephone (TEL) mode effects in a randomized mixed-mode experiment conducted in Wave 6 of the Arab Barometer Study (ABS). The methods developed in this study can offer tools for data collection agencies and researchers to analyze mixed-mode data. Study 2 examines whether interviewer variances remain consistent across different modes (e.g., FTF versus TEL) in two mixed-mode studies (the ABS and the Health and Retirement Study [HRS] 2016), representing different interviewer assignment schemes. The results can help inform interviewer training strategies and mixed-mode designs. Study 3 investigates mode effects in a longitudinal study when different mixed-mode designs are used across waves. Here, we considered the 2016 and 2018 waves of the HRS, since the HRS 2018 first introduced a sequential WEB-TEL mixed-mode design, alongside the typical FTF and TEL modes. Given that not all respondents would participate in the survey regardless of the mode used, we turned to the causal inference literature—specifically, principal stratification—to account for mode choice as a post-treatment observed variable. This study illustrates the application of principal stratification for mixed-mode inference, with the findings potentially guiding future mode assignment strategies. We examine mode effects cross-sectionally among respondents estimated to be able to complete the study in any of the compared modes; and we consider time effects within modes, again among respondents estimated to be capable of completing the survey in a given mode across both waves.

CHAPTER 1

Introduction

1.1 Background

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 [1]. Given declining response rates, survey organizations have made greater efforts to maintain sample sizes and response rates. Achieving higher response rates by increasing the level of effort (e.g., making more calls) generally does not seem to lead to large differences in estimates [2, 3]. However, increasing the number of different types of survey efforts (e.g., varying modes) does seem to lead to differences [4]. Note that mixed-mode designs may refer to 1) using multiple modes of contact to communicate with respondents or 2) using multiple modes by which participants can respond. Within the scope of this dissertation, mixed-mode designs refer to the latter.

The use of mixed-mode designs can be helpful because each mode is associated with different nonresponse and cost properties [5]. 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 (RAD), 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, a potential threat is 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.

To answer the above question, 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, i.e., any influence on survey responses that is due to modes [6]. For example, the item nonresponse rate is often higher in self-administered modes than in interviewer-administered modes [7]. 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 [8]. 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 [9, 10, 11]. Presser and Stinson find that compared to interviewer-administered modes, claims of weekly religious attendance reduce by one-third in the self-reported mode [12]. Holbrook and Krosnick show that compared to TEL surveys, social desirability bias in voter turnout reports is minimal in Internet surveys [13]. 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 [14].

On the other hand, mode selection effects refer to the phenomenon that different modes produce different nonresponse errors [15, 16, 14]. Mode selection effects occur when respondent characteristics differ across modes in ways that are correlated with the variable(s) of interest. Suzer-Gurtekin, Heeringa, and Valliant [17] use "mode choice" to indicate the mode that a participant uses to respond to surveys, which can be different from the mode assigned if a participant is offered with more than one mode. This dissertation also uses the term to refer to the mode that a participant uses to respond. In this context, mode selection effects exist when participants' mode choices are not independent of the outcome variable.

Consider a hypothetical scenario where we are interested in estimating the mean income of a target population. We have decided to collect survey data using a mobile text survey and FTF interviews. Compared to the mobile text survey, the FTF mode is more likely to recruit and retain an older sample. As older people generally have a higher level of income, it is reasonable to suspect there are mode selection effects in this case.

Mode selection effects can be present in various types of mixed-mode studies. To date, there are three main types of mixed-mode designs: randomized, sequential, and concurrent [18]. Randomized designs are used mainly for methodological studies, where the focus is on the estimation of the mode effect itself; in contrast to sequential and concurrent designs, where varying modes are is used to improve response rates or sampling frame coverage. Although sample members are randomly assigned to different modes initially, they may not respond to the assigned mode. If the nonresponse is differential across modes, respondents in one mode can systematically differ from respondents in other modes in key survey outcome variables. However, the randomization allows the effect of mode assignment to be estimated.

Sequential designs use cost-effective modes for initial data collection and follow-up nonrespondents with more expensive interview modes. Thus, participants with higher response propensity will receive cost-effective self-administered modes, while hard-to-recruit participants will receive more expensive interview modes [18]. In this case, if response propensities are correlated with the outcome variable, the mode choices are not independent of the outcome, which leads to mode selection effects.

Concurrent designs can be used to increase the coverage of a survey. For example, survey organizations may provide the offline population in a web survey the option of a mail survey [18]. In adaptive designs, survey practitioners use multiple modes to improve recruitment efficiency by tailoring modes to sample members' response propensities [18]. In concurrent mixed-mode designs that use different data collection approaches for different subpopulations, participants in one mode are different from participants in another mode on one or more variables by design. If these variables are associated with the outcome variable, the mode choices are again not independent of the outcome.

The two types of mode effects mentioned above are not equivalently valued by researchers. 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 [18]. Mode measurement effects are unwanted because they can result in inconsistent response quality and thus need to be accounted for [18]. 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.

1.2 Dissertation Structure

This dissertation revolves around improving inferences for data collected with mixed-mode designs. In Chapter 2, we consider how to analyze data collected with a randomized mixed-mode design while accounting for potential mode effects on means. Chapter 3 explores another source of mode effects: interviewer variances. Chapter 4 discusses the analytical procedures in longitudinal settings. The topics covered in these chapters are essential for developing comprehensive and versatile inference tools to account for different mixed-mode designs under various settings.

Specifically, Chapter 2 proposes three approaches to account for potential mode effects when making the inferences: 1) a "testimator" approach, 2) a Bayesian approach, and 3) a model averaging extension of the Bayesian approach. We evaluate the approaches in a simulation study and Arab Barometer study data, where FTF is the benchmark mode, and TEL is the comparison mode.

While studies about mode comparisons focused on bias properties; few of them have investigated variance across modes in mixed-mode designs. While many factors (such as interviewer, measurement error including primacy or recenecy effects, and respondent heterogeneity) may affect the mode-specific variances, Chapter 3 investigates whether interviewer variances are equal across modes in mixed-mode studies. We use data collected with two designs to answer the research question. In the first design, when interviewers are responsible for either FTF or TEL mode, we examine whether there are mode differences in interviewer variance using the Arab Barometer wave 6 Jordan data. In the second design, we draw on Health and Retirement Study (HRS) 2016 core survey data to examine the question on three topics when interviewers are responsible for both modes.

While the previous two chapters focused on cross-sectional studies, Chapter 4 examines mode effects in a longitudinal setting where different mixed-mode designs are used across waves. We treat the mode of data collection as the treatment, employing a potential outcome framework to multiply impute the potential response status of cases if assigned to another mode and the associated potential outcomes. After imputation, we construct principal strata based on the observed and the predicted response status of each case and estimate mode effects within each principal stratum. Last, we make inference by combining mode effect estimates across the principal strata and the imputed datasets. We apply this analytical strategy to the HRS 2016 and 2018 core surveys.

Finally, in Chapter 5, we discuss the findings and limitations of the previous chapters and suggest future directions for extending this work.

CHAPTER 2

Three Approaches to Adjust for Mode Effects

2.1 Introduction

Mixed-mode inference has become a pressing necessity since COVID-19, as many large survey projects have been forced to shift data collection modes due to restricted social contact [19, 20, 21]. The Arab Barometer study, which is the application considered in this paper, 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 have shifted from face-to-face (FTF) alone to mixed-mode designs (FTF and telephone [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 [22]. On the other hand, some literature finds no clear differences between FTF and TEL on sensitive items [23]. 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 paper. 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.

2.1.1 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. [17] 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 [24] 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. [25] develop an adaptive mode adjustment to address the differential nonresponse 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 [17, 26]. 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 [26]. After determining a reference mode, researchers need to derive potential outcomes for cases in the non-reference mode. Examples include Powers, Michra, and Young [27], Elliott et al. [28], Kolenikov and Kennedy [26], and Park, Kim, and Park [29]. 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.

2.2 Proposed Methods

This paper proposes three approaches to combine mode-specific estimates: 1) a Testimator approach, 2) a Bayesian approach, and 3) 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 paper. 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, 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 paper:

1. At least one mode provides an unbiased estimate of the population mean (either $\delta_a = 0$ or $\delta_b = 0$, but we don't 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.

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.

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 non-sensitive 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. Since expressing dissatisfaction with government performance on COVID 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 paper.

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 towards 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.

2.2.1 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.

2.2.1.1 When there is some information about the preferred direction

1. We first 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 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 in $[-\infty, F(\frac{\alpha_1}{2}, n_a - 1, n_b - 1)]$ 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 $\frac{(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b})^2}{\frac{(\frac{s_a^2}{n_a})^2}{n_a - 1} + \frac{(\frac{s_b^2}{n_b})^2}{n_b - 1}}$.

   (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 $[\hat{\theta} - t_{\frac{\alpha_2}{2}, v}(\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2})^{-\frac{1}{2}}, \hat{\theta} + t_{\frac{\alpha_2}{2}, v}(\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2})^{-\frac{1}{2}}]$.

   (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 2.1.

2.2.1.2 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 $100(1 - \gamma)\%$ confidence intervals for u_a and u_b. The confidence intervals are computed as

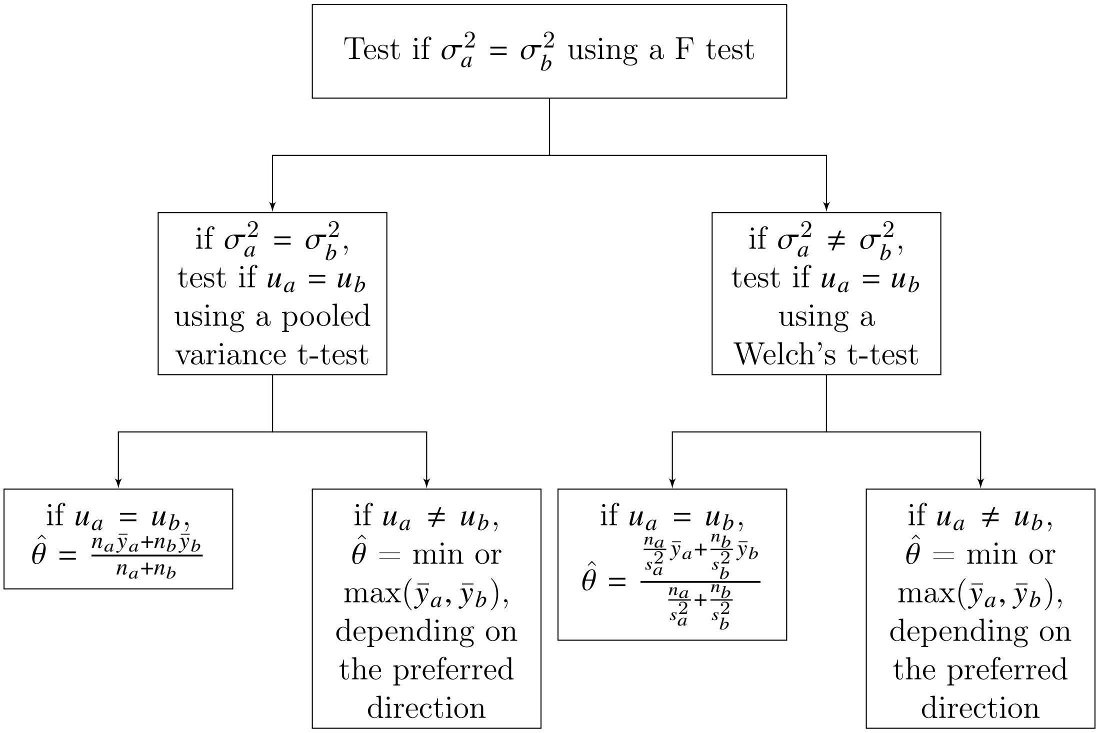

Test if $\sigma_a^2 = \sigma_b^2$ using a F test
if $\sigma_a^2 = \sigma_b^2$, test if $u_a = u_b$ using a pooled variance t-test
if $\sigma_a^2 \neq \sigma_b^2$, test if $u_a = u_b$ using a Welch's t-test
if $u_a = u_b$, $\hat{\theta} = \frac{n_a\bar{y}_a + n_b\bar{y}_b}{n_a+n_b}$
if $u_a \neq u_b$, $\hat{\theta}$ = min or max$(\bar{y}_a, \bar{y}_b)$, depending on the preferred direction
if $u_a = u_b$, $\hat{\theta} = \frac{\frac{n_a}{s_a^2}\bar{y}_a + \frac{n_b}{s_b^2}\bar{y}_b}{\frac{n_a}{s_a^2} + \frac{n_b}{s_b^2}}$
if $u_a \neq u_b$, $\hat{\theta}$ = min or max$(\bar{y}_a, \bar{y}_b)$, depending on the preferred direction

Figure 2.1: Flowchart of the Testimator Approach when some Information about the Preferred Direction is Available

$[L_a, U_a] = \bar{y}_a \mp t_{n_a-1}\frac{S_a}{\sqrt{n_a}}$ 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.

2.2.2 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.

2.2.2.1 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:

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

   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 paper. The parameterization of the Gamma distribution in this paper is suggested by scaled inverse chi-square distributions, such that if $\sigma_a^2 \sim$ Scale-inv-$\chi^2(v_a, \tau_a^2)$, then $\sigma_a^2 \sim$ 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: $\hat{\eta} = \frac{\hat{u}_a - \hat{u}_b}{\sqrt{\frac{\hat{\sigma}_a^2 + \hat{\sigma}_b^2}{2}}}$.

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

   (a) If $\epsilon_{lw} \leq 0 \leq \epsilon_{up}$, 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.

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

(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.

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

(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.

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

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.

2.2.2.2 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:

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

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):

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

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

2.2.3 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 [30]. Unlike the Testimator and Bayesian approaches, this approach does not aim to find a single model that best describes the data; instead, it averages over all possible models with weights proportional to the probability that one of the models is correct.

2.2.3.1 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 (2.1). 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 hyper-parameters for priors of u_a, u_b, σ_a^{-2}, and

σ_b^{-2}.

$$y_{ai} \sim N(u_a, \sigma_a^2), y_{bi} \sim N(u_b, \sigma_b^2)$$
$$u_a|\sigma_a^{-2} \sim N(\beta_a, \frac{\sigma_a^2}{k_a}), u_b|\sigma_b^{-2} \sim N(\beta_b, \frac{\sigma_b^2}{k_b}), \sigma_a^{-2} \sim G(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2), \sigma_b^{-2} \sim G(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2) \tag{2.1}$$

Model 2 assumes a common mean (u) but different variances for mixed-mode data (2.2). This model accounts for the scenario when the use of modes leads to shifts in the variances but not the means.

$$y_{ai} \sim N(u, \sigma_a^2), y_{bi} \sim N(u, \sigma_b^2), u \sim N(\beta_0, \psi^2), \sigma_a^{-2} \sim G(\frac{v_a}{2}, \frac{v_a}{2}\tau_a^2), \sigma_b^{-2} \sim G(\frac{v_b}{2}, \frac{v_b}{2}\tau_b^2). \tag{2.2}$$

Model 3 assumes different means but a common variance (σ^2, 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}).

$$y_{ai} \sim N(u_a, \sigma^2), y_{bi} \sim N(u_b, \sigma^2)$$
$$u_a|\sigma^{-2} \sim N(\beta_a, \frac{\sigma^2}{k_a}), u_b|\sigma^2 \sim N(\beta_b, \frac{\sigma^2}{k_b}), \sigma^{-2} \sim G(\frac{v}{2}, \frac{v}{2}\lambda^2) \tag{2.3}$$

Model 4 assumes a common mean and variance (2.4). 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 $\frac{\sigma^2}{k}$.

$$y_{ai} \sim N(u, \sigma^2), y_{bi} \sim N(u, \sigma^2), u \sim N(\beta_0, \frac{\sigma^2}{k}), \sigma^{-2} \sim G(\frac{v}{2}, \frac{v}{2}\lambda^2) \tag{2.4}$$

2. We calculate the marginal posterior $\pi(M|y_a, y_b)$ for each model using analytical in-

tegration (Appendix A). 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 (MCMC) algorithm to derive the weights. This may be particularly useful when dealing with non-conjugate priors or when working with small sample sizes.

3. We then draw R times 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 $(1 - \gamma) \times 100\%$ credible interval.

2.2.3.2 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 (2.2.2.2). This is as opposed to 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.

2.2.4 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 $E(\hat{\theta}) = u_a \Phi(\frac{u_b - u_a}{s}) + u_b \Phi(\frac{u_a - u_b}{s}) - s\phi(\frac{u_b - u_a}{s})$, 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 [31]. Using the same notation, we can write the second moment as $E(\hat{\theta}^2) = (u_a + \frac{\sigma_a^2}{n_a})\Phi(\frac{u_b - u_a}{s}) + (u_b + \frac{\sigma_b^2}{n_b})\Phi(\frac{u_a - u_b}{s}) - (u_a + u_b)s\phi(\frac{u_b - u_a}{s})$ [31]. Then, the variance of the estimator can be expressed as

$$\begin{aligned} Var(\hat{\theta}) =& E(\theta^2) - E(\theta)^2 \\ =& (u_a + \frac{\sigma_a^2}{n_a})\Phi(\frac{u_b - u_a}{s}) + (u_b + \frac{\sigma_b^2}{n_b})\Phi(\frac{u_a - u_b}{s}) - \\ & (u_a + u_b)s\phi(\frac{u_b - u_a}{s}) - (u_a\Phi(\frac{u_b - u_a}{s}) + u_b\Phi(\frac{u_a - u_b}{s}) - s\phi(\frac{u_b - u_a}{s}))^2. \end{aligned} \tag{2.5}$$

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 $(1 - \gamma) \times 100\%$ confidence interval using a Z distribution.

2.2.5 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 $(1 - \gamma) \times 100\%$ confidence interval.

2.3 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 simulate nine scenarios to mimic varying levels of mode effects by adjusting u_b and σ_b^2 (See Table 2.1 for the scenarios).

Table 2.1: Simulation Scenarios

Scenarios	u_b	σ_b^2
1	0	1
2	0	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

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

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 to 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% ($\gamma = 0.05$) confidence interval for $\hat{\theta}$ in this approach. In the Bayesian and the model averaging approach, we consider the following hyper-parameters 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 5000 draws from the posterior distribution.

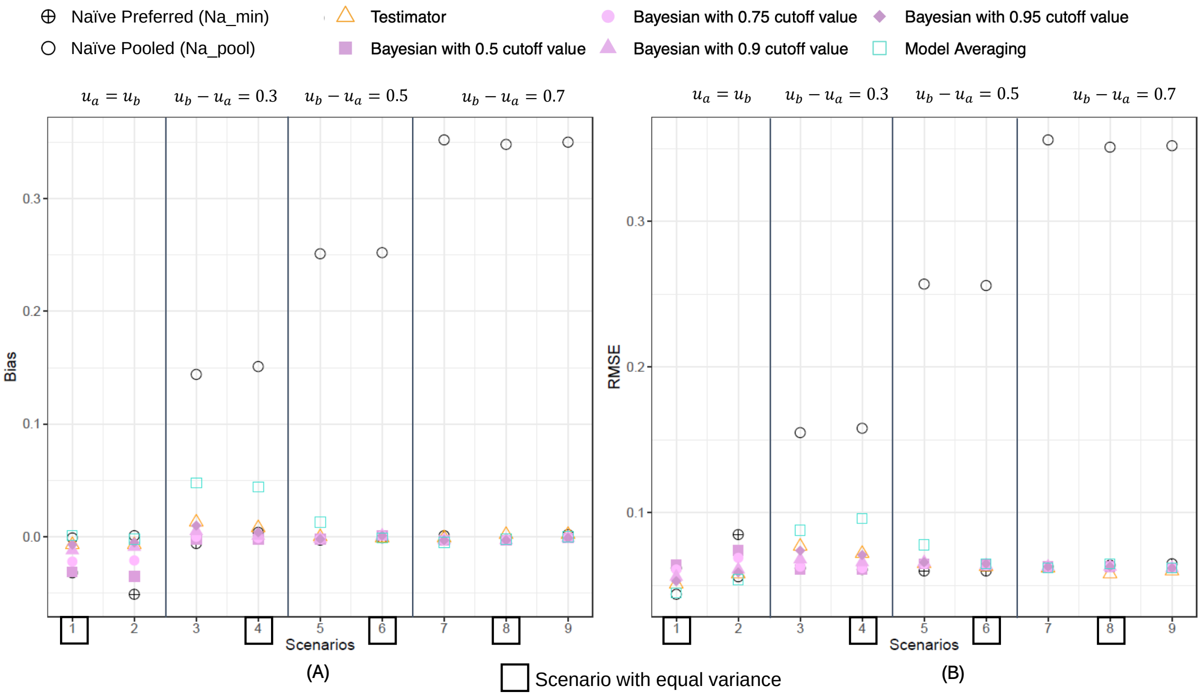
We compare the approaches on their performances in bias, RMSE, coverage rate, actual interval length, and expected mean interval length. Bias is given by bias$(\hat{\theta}) = \frac{\sum \hat{\theta}}{500} - \theta$. RMSE is given by RMSE$(\hat{\theta}) = \sqrt{\frac{\sum(\hat{\theta}-\theta)^2}{500}}$. Coverage rate is the percentage of derived intervals including the true population mean (θ). Actual interval length is the average interval 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 1) this might best reflect the high correlation between responses collected with different modes, and 2) simulation results don't show significant variations across different ρ values. Results for $\rho = 0.75$ and $\rho = 0.5$ can be found in the Appendix (Tables A3, A4, A5, and A6).

Figures 2.2 and 2.3 show the performances of the point estimates and uncertainty measures respectively when prior information is available and $\rho = 0.95$ (Appendix Table A1). 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.

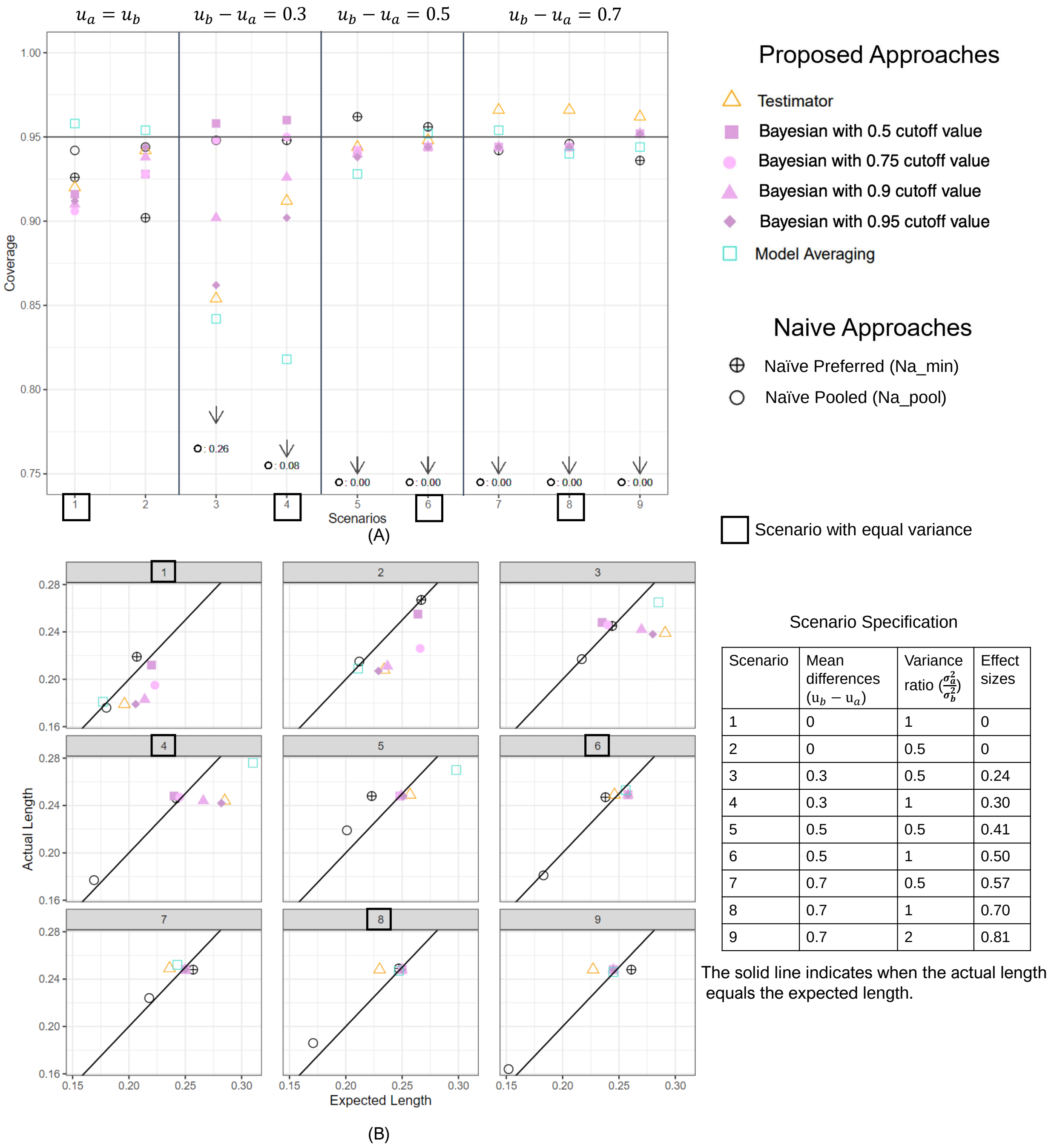
To compare the three approaches, we note that when mode effects do not exist, a wider interval (e.g., 95%) in the Bayesian approach performs better. This is expected because a wider interval is associated with fewer false positive errors, thus providing more accurate estimates of the population mean 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 to 9). When mode effects do not exist (Scenarios 1 and 2), the coverage rates of the Testimator approach are lower than 0.95 (Figure 2.3). 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 Bayesian approach that applies more sensitive cutoff values (Figure 2.2, i.e., the ones computed from 50%, 75%, and 90% intervals).

To further illustrate the performances of the Testimator and the model averaging approaches, we show the probabilities of selecting four models in Tables 2.2 and 2.3. 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 2.2.3. We note that when the effect size is less than 0.50, both the Testimator approach (the probabilities $p = 0.764$ and 0.874 in Scenarios 3 and 4 respectively) and the model averaging approach (0.345 , 0.595, and 0.830 respectively for Scenarios 3 to 5) have a lower probability of picking the



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 2.2: Bias and RMSE in the Simulation Study when Information about the Preferred Direction is Available



Scenario Specification

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

The solid line indicates when the actual length equals the expected length.

Figure 2.3: Coverage Rates and Interval Length in the Simulation Study when Information about the Preferred Direction is Available

right model. In these scenarios, besides the correct model, they are most likely to choose the model with equal means (Models 2 and 4).

Table 2.2: 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

Notes: "D" stands for different and "S" stands for same. The correct model of a scenario are 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 size is computed as $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}}$.

Figures 2.4 and 2.5 present the simulation results when we have no information about preferred modes or directions. Detailed results can be found in Table A2 in the Appendix. 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% coverage rate. Thus, 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

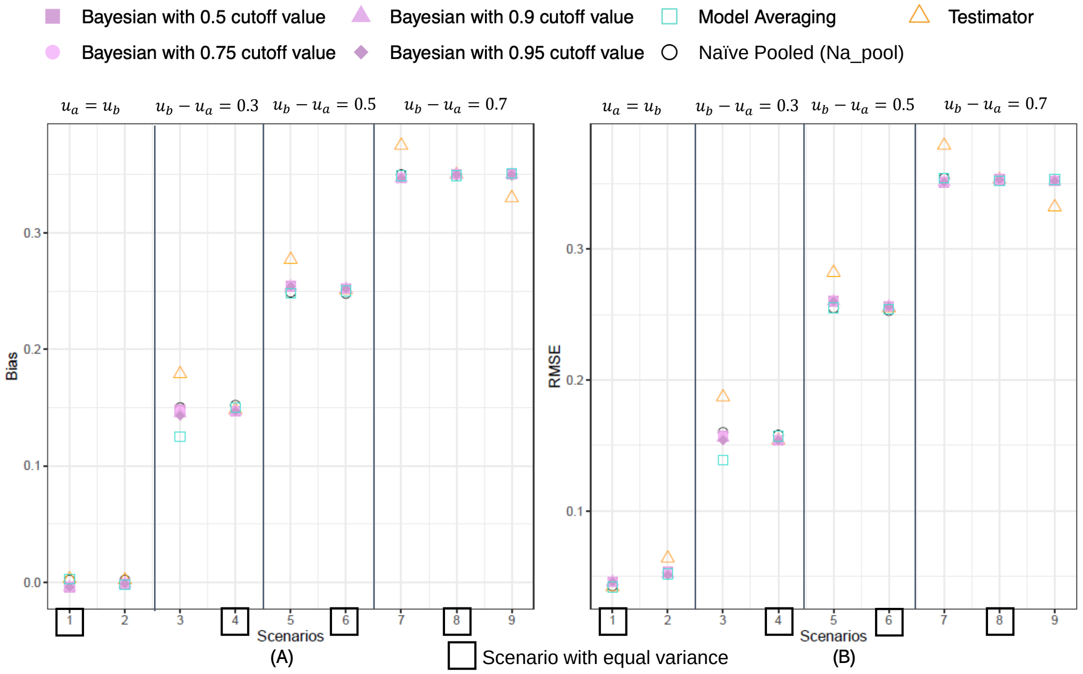
Table 2.3: Average Weights in the Model Averaging Approach (ρ = 0.95 and n = 500)

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

Notes: "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 $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2+\sigma_b^2}{2}}}$.

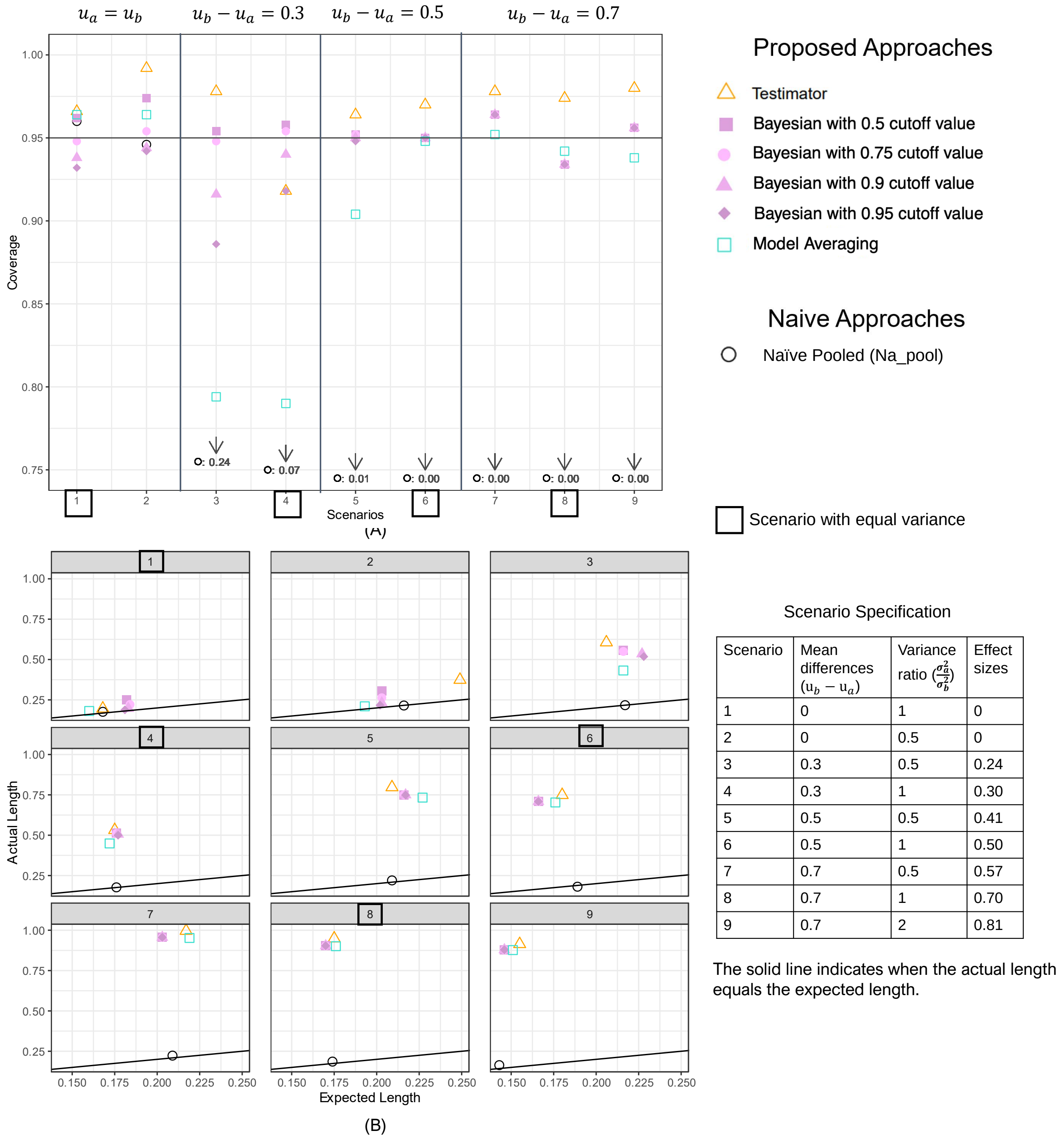
estimated effective size [$\hat{\eta}$] in the testing phase), the narrower the interval width. This makes sense as when we apply more strict criteria (i.e., wider "intervals") to detect mode effects, we are more likely to find no mode effects and thus make inferences from the common mean model. This 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%) does not make a difference as they all lead to the same conclusion. Lastly, the model averaging approach shows poorer coverage in Scenarios 3 to 5. This is again related to the diffuse weight distribution in these scenarios (Table 2.3), 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



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 2.4: Bias and RMSE in the Simulation Study when no Information is Available



Scenario Specification

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

The solid line indicates when the actual length equals the expected length.

Figure 2.5: Coverage Rate and Length in the Simulation Study when no Information is Available

large samples, we increase the sample size to 5,000 and re-run 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 Appendix). 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 to 5, see Table A9 in the Appendix). 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, the increased sample size eliminates the sensitivity of simulation results to the choice of interval length in the Bayesian approach (Scenarios 3 to 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.

2.4 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 = 2531, n_{FTF} = 1193, n_{TEL} = 1338).

As mentioned before, the research team applied a randomized mixed-mode experiment in Jordan, where 1/3 of the households were interviewed only 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 telephone 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 telephone 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 paper, we construct a measure of satisfaction towards 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: 1) Government's performance on security, 2) Government's performance on keeping price down, 3) Government's performance on responding to COVID, 4) Government's performance overall, 5) Government's performance on education system, and 6) 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. Figure 2.6 shows the distributions of the outcome variable in pooled data, data collected via FTF, and data collected via TEL; they can be seen as approximately following normal distributions.

We consider two settings in this application: 1) when we are aware that a smaller estimate is closer to the truth and 2) 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 [32]. We consider the second scenario as a sensitivity analysis.

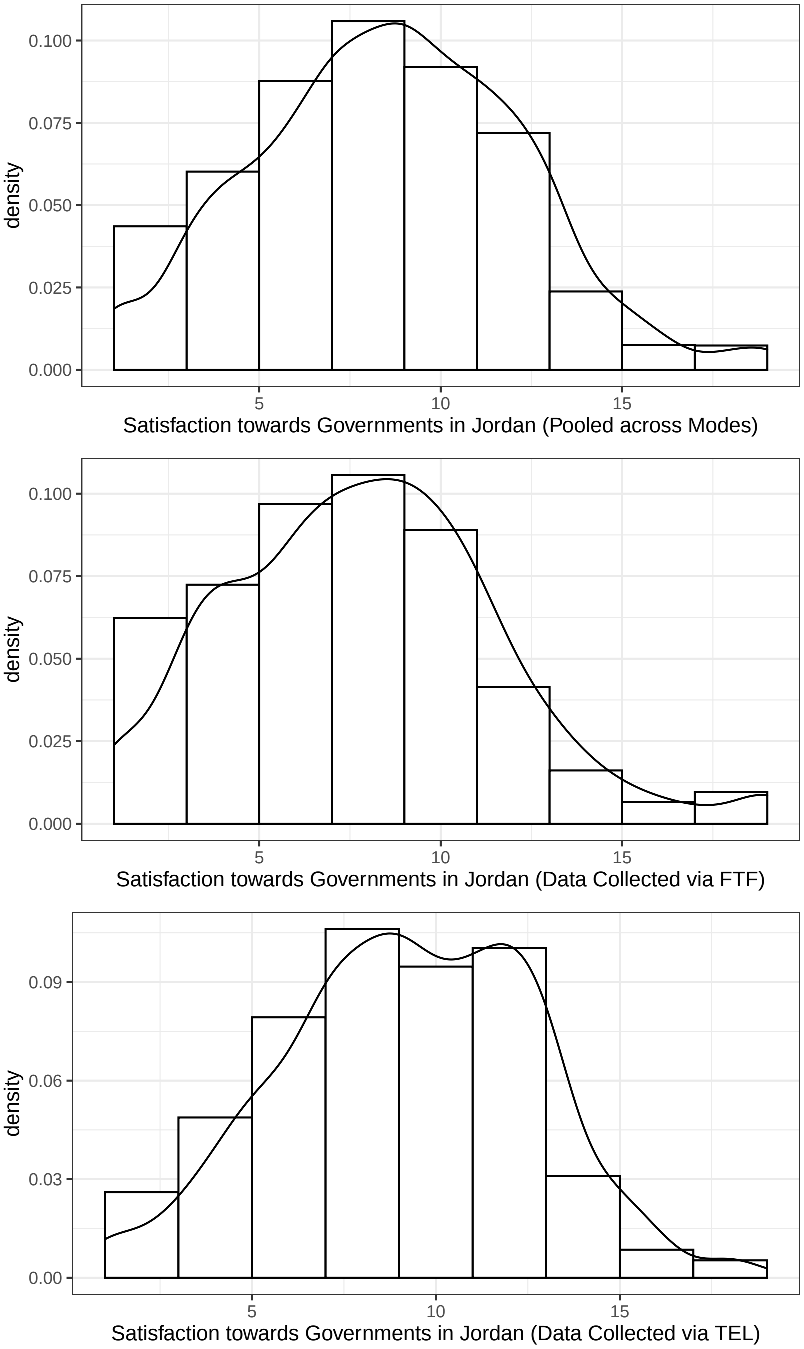


Figure 2.6: Distribution of the Outcome Variable in the Arab Barometer Application

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 2.4 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 reporting relative to TEL [22]. 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 2.5. 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.

Table 2.4: 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

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

Table 2.5: 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

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

2.5 Discussion

This paper 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), 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%). 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% cutoff value in the Bayesian approach and they can consider a narrower interval (such as 75% or 90%) 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 paper 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 paper.

This paper 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 [33]. Nevertheless, this paper 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 paper 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 paper 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 paper provides novel approaches to connect the testing and the inferences of mode effects. Kolenikov and Kennedy [26] classify mode effects literature into three aims: 1) determining the magnitude of mode effects, 2) providing population estimates, and 3) obtaining case-level estimates. This paper 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 paper addresses the important research gap by proposing procedures for different scenarios depending on whether we have prior information about preferred directions or not.

This paper 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 mode 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 paper 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 confidence intervals in this particular setting.

2.6 Appendix

2.6.1 Appendix A: Derivation of Weights in the Model Averaging Approach

For Model 1, we first integrate u_a and u_b with respect to conjugate normal priors with known means (β_a and β_b) and known hyperparameters (k_a and k_b), then integrate σ_a^{-2} and σ_b^{-2} with respect to gamma priors with known shapes ($\frac{v_a}{2}$ and $\frac{v_b}{2}$) and rates ($\frac{v_a}{2}\tau_a^2$ and $\frac{v_a}{2}\tau_b^2$). The marginal posterior of Model 1 ($\pi(M = 1|y_a, y_b)$) is computed as follows:

$$\frac{\Gamma(\frac{v_a+n_a}{2})}{(\frac{v_a\tau_a^2+(n_a-1)s_a^2+k_a n_a(\beta_a-\bar{y}_a)^2(k_a+n_a)^{-1}}{2})^{\frac{v_a+n_a}{2}}}\sqrt{\frac{k_a}{k_a+n_a}}\frac{(\frac{v_a}{2}\tau_a^2)^{\frac{v_a}{2}}}{\Gamma(\frac{v_a}{2})}\times$$
$$\frac{\Gamma(\frac{v_b+n_b}{2})}{(\frac{v_b\tau_b^2+(n_b-1)s_b^2+k_b n_b(\beta_b-\bar{y}_b)^2(k_b+n_b)^{-1}}{2})^{\frac{v_b+n_b}{2}}}\sqrt{\frac{k_b}{k_b+n_b}}\frac{(\frac{v_b}{2}\tau_b^2)^{\frac{v_b}{2}}}{\Gamma(\frac{v_b}{2})} \tag{2.6}$$

Note that we omitted $((\sqrt{2\pi})^{n_a+n_b}\pi(y_a)\pi(y_b))^{-1}$ in weights for each model as the component is the same across the weights and can be cancelled out in the end.

For Model 2, we don't have an analytically closed form for the weight; thus, we provide a workaround by substituting t densities with normal densities, which achieves good approximation when degrees of freedom are large.

We first integrate σ_a^{-2} and σ_b^{-2} with respect to gamma priors with known shapes ($\frac{v_a}{2}$ and $\frac{v_b}{2}$) and rates ($\frac{v_a}{2}\tau_a^2$ and $\frac{v_a}{2}\tau_b^2$), to obtain a function of u:

$$\frac{1}{\sqrt{2\pi}\psi}\exp(-\frac{(u-\beta_0)^2}{2\psi^2})\frac{(\frac{v_a}{2}\tau_a^2)^{\frac{v_a}{2}}}{\Gamma(\frac{v_a}{2})}\frac{(\frac{v_b}{2}\tau_b^2)^{\frac{v_b}{2}}}{\Gamma(\frac{v_b}{2})}\times$$
$$\frac{\Gamma(\frac{v_a+n_a}{2})}{(\frac{v_a\tau_a^2+(n_a-1)s_a^2+n_a(u-\bar{y}_a)^2}{2})^{\frac{v_a+n_a}{2}}}\frac{\Gamma(\frac{v_b+n_b}{2})}{(\frac{v_b\tau_b^2+(n_b-1)s_b^2+n_b(u-\bar{y}_b)^2)}{2})^{\frac{v_b+n_b}{2}}} \tag{2.7}$$

Next, we write part of Formula 2.7 as the product of the kernel of t densities (degrees of freedom: $v_a + n_a - 1$) and a constant, then approximate the t densities with normal densities as the degrees of freedom are usually large.

$$
\begin{aligned}
&(\frac{v_a\tau_a^2 + (n_a - 1)s_a^2 + n_a(u - \bar{y}_a)^2}{2})^{-\frac{v_a+n_a}{2}} \\
&= (\frac{v_a\tau_a^2 + (n_a - 1)s_a^2}{2})^{-\frac{v_a+n_a}{2}}(1 + \frac{n_a(u - \bar{y}_a)^2}{v_a\tau_a^2 + (n_a - 1)s_a^2})^{-\frac{v_a+n_a}{2}} \\
&= (\frac{v_a\tau_a^2 + (n_a - 1)s_a^2}{2})^{-\frac{v_a+n_a}{2}}(1 + \frac{1}{v_a + n_a - 1}(\frac{n_a(u - \bar{y}_a)}{\sqrt{\frac{v_a\tau_a^2+(n_a-1)s_a^2}{(v_a+n_a-1)n_a}}})^2)^{-\frac{v_a+n_a}{2}} \\
&\approx (\frac{v_a\tau_a^2 + (n_a - 1)s_a^2}{2})^{-\frac{v_a+n_a}{2}}\frac{\Gamma(\frac{v_a+n_a-1}{2})\sqrt{\pi(v_a + n_a - 1)}\sqrt{\frac{v_a\tau_a^2+(n_a-1)s_a^2}{(v_a+n_a-1)n_a}}}{\Gamma(\frac{v_a+n_a}{2})}\times \\
&\frac{1}{\sqrt{2\pi}\sqrt{\frac{v_a\tau_a^2+(n_a-1)s_a^2}{(v_a+n_a-1)n_a}}}exp(-\frac{(u - \bar{y}_a)^2}{2\frac{v_a\tau_a^2+(n_a-1)s_a^2}{(v_a+n_a-1)n_a}})
\end{aligned}
\tag{2.8}
$$

We write the other part involving $\bar{y}_b$ in Formula 2.7 in a similar fashion.

$$
\begin{aligned}
&(\frac{v_b\tau_b^2 + (n_b - 1)s_b^2 + n_b(u - \bar{y}_b)^2}{2})^{-\frac{v_b+n_b}{2}} \\
&\approx (\frac{v_b\tau_b^2 + (n_b - 1)s_b^2}{2})^{-\frac{v_b+n_b}{2}}\frac{\Gamma(\frac{v_b+n_b-1}{2})\sqrt{\pi(v_b + n_b - 1)}\sqrt{\frac{v_b\tau_b^2+(n_b-1)s_b^2}{(v_b+n_b-1)n_b}}}{\Gamma(\frac{v_b+n_b}{2})}\times \\
&\frac{1}{\sqrt{2\pi}\sqrt{\frac{v_b\tau_b^2+(n_b-1)s_b^2}{(v_b+n_b-1)n_b}}}exp(-\frac{(u - \bar{y}_b)^2}{2\frac{v_b\tau_b^2+(n_b-1)s_b^2}{(v_b+n_b-1)n_b}})
\end{aligned}
\tag{2.9}
$$

Based on Formula 2.8 and 2.9, we can write Formula 2.7 as the kernel of normal densities and integrate u accordingly. We obtain the final marginal posterior for Model 2 as follows:

$$
\begin{aligned}
&\frac{(\frac{v_a}{2}\tau_a^2)^{\frac{v_a}{2}}}{\Gamma(\frac{v_a}{2})}\frac{(\frac{v_b}{2}\tau_b^2)^{\frac{v_b}{2}}}{\Gamma(\frac{v_b}{2})}\frac{\Gamma(\frac{v_a+n_a-1}{2})\Gamma(\frac{v_b+n_b-1}{2})}{2}\sqrt{v_a + n_a - 1}\sqrt{v_b + n_b - 1}\times \\
&\quad\sqrt{\frac{AB}{2\psi^2B + 2\psi^2A + AB}}[\frac{v_a\tau_a^2 + s_a^2(n_a - 1)}{2}]^{-\frac{v_an_a}{2}}[\frac{v_b\tau_b^2 + s_b^2(n_b - 1)}{2}]^{-\frac{v_b+n_b}{2}}\times \\
&\qquad exp[(\frac{1}{A} + \frac{1}{B} + \frac{1}{2\psi^2})(\frac{\frac{\bar{y}_a}{A} + \frac{\bar{y}_b}{B} + \frac{\beta_0}{2\psi^2}}{\frac{1}{A} + \frac{1}{B} + \frac{1}{2\psi^2}})^2 - (\frac{\bar{y}_a^2}{A} + \frac{\bar{y}_b^2}{B} + \frac{\beta_0^2}{2\psi^2})],
\end{aligned}
\tag{2.10}
$$

where

$$
A = 2\frac{v_a\tau_a^2 + (n_a - 1)s_a^2}{(v_a + n_a - 1)n_a}, B = 2\frac{v_b\tau_b^2 + (n_b - 1)s_b^2}{(v_b + n_b - 1)n_b}.
$$

For Model 3, we first integrate u_a and u_b separately using normal distributions as in Model 1, then integrate σ^{-2} using a gamma distribution with respect to a gamma prior with known shape parameter $\frac{v}{2}$ and rate parameter $\frac{v}{2}\lambda^2$.

$$\frac{\Gamma(\frac{v+n_a+n_b}{2})}{(\frac{v\lambda^2+(n_a-1)s_a^2+(n_b-1)s_b^2+k_a n_a(\beta_a-\bar{y}_a)^2(k_a+n_a)^{-1}+k_b n_b(\beta_b-\bar{y}_b)^2(k_b+n_b)^{-1}}{2})^{\frac{v+n_a+n_b}{2}}}\times \sqrt{\frac{k_a k_b}{(k_a+n_a)(k_b+n_b)}}\frac{(\frac{v}{2}\lambda^2)^{\frac{v}{2}}}{\Gamma(\frac{v}{2})} \tag{2.11}$$

For Model 4, we similarly first integrate u with respect to a conjugate normal prior with known mean β_0 and variance $\frac{\sigma^2}{k}$, then integrate σ^{-2} with respect to a gamma prior with known shape $\frac{v}{2}$ and rate $\frac{v}{2}\lambda^2$.

$$\frac{\Gamma(\frac{v+n_a+n_b}{2})}{(\frac{v\lambda^2+(n_a-1)s_a^2+(n_b-1)s_b^2+(kn_a(\beta_a-\bar{y}_a)^2+kn_b(\beta_b-\bar{y}_b)^2+n_a n_b(\bar{y}_a-\bar{y}_b)^2(k+n_a+n_b)^{-1}}{2})^{\frac{v+n_a+n_b}{2}}}\times \sqrt{\frac{k}{(k+n_a+n_b)}}\frac{(\frac{v}{2}\lambda^2)^{\frac{v}{2}}}{\Gamma(\frac{v}{2})} \tag{2.12}$$

2.6.2 Appendix B: Additional Simulation Results

Table 2.6: Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and $\rho = 0.95$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Na_min	-0.032	0.062	0.926	0.207	0.219
Na_pool	-0.001	0.044	0.942	0.180	0.176
Te	-0.007	0.051	0.920	0.196	0.179
B50	-0.031	0.064	0.916	0.220	0.212
B75	-0.022	0.061	0.906	0.223	0.195
B90	-0.012	0.056	0.910	0.214	0.183
B95	-0.007	0.053	0.912	0.206	0.179
MA	0.001	0.045	0.958	0.177	0.181

Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na_min	-0.051	0.085	0.902	0.267	0.267
Na_pool	0.001	0.056	0.944	0.212	0.215
Te	-0.007	0.058	0.942	0.234	0.208
B50	-0.035	0.074	0.928	0.264	0.255
B75	-0.021	0.069	0.928	0.266	0.226
B90	-0.009	0.061	0.938	0.237	0.211
B95	-0.005	0.059	0.944	0.229	0.207
MA	-0.002	0.054	0.954	0.211	0.209
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Na_min	-0.006	0.062	0.948	0.244	0.245
Na_pool	0.144	0.155	0.264	0.217	0.217
Te	0.013	0.077	0.854	0.291	0.239
B50	-0.002	0.061	0.958	0.235	0.248
B75	0.000	0.063	0.948	0.240	0.246
B90	0.005	0.068	0.902	0.270	0.242
B95	0.010	0.074	0.862	0.280	0.238
MA	0.048	0.088	0.842	0.285	0.265
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na_min	0.004	0.061	0.948	0.242	0.246
Na_pool	0.151	0.158	0.084	0.169	0.177
Te	0.008	0.072	0.912	0.285	0.244
B50	-0.002	0.061	0.960	0.240	0.248
B75	-0.001	0.062	0.950	0.245	0.247
B90	0.001	0.066	0.926	0.266	0.244
B95	0.004	0.071	0.902	0.282	0.242
MA	0.044	0.096	0.818	0.310	0.276
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Na_min	-0.003	0.060	0.962	0.223	0.248
Na_pool	0.251	0.257	0.004	0.201	0.219
Te	0.000	0.065	0.944	0.257	0.249
B50	-0.002	0.065	0.942	0.248	0.248
B75	-0.002	0.065	0.942	0.248	0.248
B90	-0.002	0.066	0.940	0.249	0.248
B95	-0.002	0.066	0.938	0.251	0.248

MA	0.013	0.078	0.928	0.298	0.270
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na_min	-0.001	0.060	0.956	0.238	0.247
Na_pool	0.252	0.256	0.000	0.183	0.181
Te	-0.001	0.063	0.948	0.246	0.249
B50	0.001	0.064	0.944	0.258	0.249
B75	0.001	0.064	0.944	0.258	0.249
B90	0.001	0.064	0.944	0.258	0.249
B95	0.001	0.065	0.944	0.258	0.249
MA	-0.001	0.065	0.952	0.256	0.253
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na_min	0.001	0.063	0.942	0.257	0.248
Na_pool	0.352	0.356	0.000	0.218	0.224
Te	-0.001	0.062	0.966	0.236	0.249
B50	-0.003	0.063	0.944	0.250	0.248
B75	-0.003	0.063	0.944	0.250	0.248
B90	-0.003	0.063	0.944	0.250	0.248
B95	-0.003	0.063	0.944	0.250	0.248
MA	-0.005	0.062	0.954	0.243	0.252
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na_min	-0.003	0.064	0.946	0.247	0.249
Na_pool	0.348	0.351	0.000	0.171	0.186
Te	0.002	0.058	0.966	0.230	0.248
B50	-0.003	0.063	0.944	0.250	0.248
B75	-0.003	0.063	0.944	0.250	0.248
B90	-0.003	0.063	0.944	0.250	0.248
B95	-0.003	0.063	0.944	0.250	0.248
MA	-0.002	0.065	0.940	0.247	0.247
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na_min	0.002	0.065	0.936	0.261	0.248
Na_pool	0.350	0.352	0.000	0.152	0.164
Te	0.002	0.060	0.962	0.227	0.248
B50	-0.001	0.062	0.952	0.245	0.248
B75	-0.001	0.062	0.952	0.245	0.248
B90	-0.001	0.062	0.952	0.245	0.248

B95	-0.001	0.062	0.952	0.245	0.248
MA	0.000	0.062	0.944	0.245	0.246

Notes: "Na_min" refers to the Naïve approach that takes the smaller estimate, "Na_pool" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.7: Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions ($n = 500$ and $\rho = 0.95$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Na	0.002	0.043	0.960	0.168	0.176
Te	0.003	0.042	0.966	0.168	0.196
B50	-0.004	0.046	0.962	0.182	0.251
B75	-0.004	0.046	0.948	0.184	0.221
B90	-0.004	0.046	0.938	0.183	0.197
B95	-0.004	0.046	0.932	0.181	0.188
MA	0.003	0.042	0.964	0.160	0.181
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na	0.002	0.053	0.946	0.216	0.215
Te	0.002	0.064	0.992	0.249	0.374
B50	-0.001	0.054	0.974	0.203	0.306
B75	-0.001	0.053	0.954	0.203	0.263
B90	-0.001	0.052	0.944	0.203	0.231
B95	-0.001	0.051	0.942	0.202	0.217
MA	-0.002	0.052	0.964	0.193	0.210
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Na	0.150	0.160	0.240	0.217	0.217
Te	0.179	0.187	0.978	0.206	0.606
B50	0.148	0.157	0.954	0.216	0.556

B75	0.147	0.157	0.948	0.216	0.550
B90	0.145	0.156	0.916	0.227	0.535
B95	0.143	0.154	0.886	0.228	0.518
MA	0.125	0.139	0.794	0.216	0.432
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na	0.152	0.158	0.068	0.176	0.177
Te	0.148	0.154	0.918	0.175	0.532
B50	0.147	0.154	0.958	0.176	0.512
B75	0.147	0.154	0.954	0.176	0.511
B90	0.147	0.154	0.940	0.177	0.506
B95	0.147	0.154	0.918	0.177	0.500
MA	0.150	0.157	0.790	0.172	0.449
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Na	0.249	0.255	0.006	0.209	0.220
Te	0.277	0.282	0.964	0.209	0.797
B50	0.254	0.260	0.952	0.216	0.750
B75	0.254	0.260	0.952	0.216	0.750
B90	0.254	0.260	0.950	0.217	0.750
B95	0.254	0.260	0.948	0.217	0.749
MA	0.248	0.255	0.904	0.227	0.733
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na	0.248	0.253	0.000	0.189	0.181
Te	0.251	0.255	0.970	0.180	0.749
B50	0.252	0.256	0.950	0.166	0.709
B75	0.252	0.256	0.950	0.166	0.709
B90	0.252	0.256	0.950	0.166	0.709
B95	0.252	0.256	0.950	0.166	0.709
MA	0.250	0.254	0.948	0.176	0.703
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na	0.350	0.354	0.000	0.209	0.223
Te	0.375	0.379	0.978	0.217	0.996
B50	0.347	0.351	0.964	0.203	0.956
B75	0.347	0.351	0.964	0.203	0.956
B90	0.347	0.351	0.964	0.203	0.956
B95	0.347	0.351	0.964	0.203	0.956

MA	0.349	0.354	0.952	0.219	0.952
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na	0.350	0.353	0.000	0.174	0.186
Te	0.350	0.353	0.974	0.175	0.949
B75	0.350	0.353	0.934	0.170	0.904
B90	0.350	0.353	0.934	0.170	0.904
B95	0.350	0.353	0.934	0.170	0.904
MA	0.349	0.352	0.942	0.176	0.901
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na	0.350	0.352	0.000	0.143	0.164
Te	0.330	0.332	0.980	0.155	0.914
B50	0.350	0.352	0.956	0.146	0.879
B75	0.350	0.352	0.956	0.146	0.879
B90	0.350	0.352	0.956	0.146	0.879
B95	0.350	0.352	0.956	0.146	0.879
MA	0.351	0.353	0.938	0.151	0.877

Notes: "Na" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.8: Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and ρ = 0.75)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Na_min	-0.032	0.062	0.922	0.206	0.219
Na_pool	0.001	0.047	0.946	0.180	0.176
Te	-0.005	0.051	0.934	0.200	0.178
B50	-0.036	0.065	0.934	0.217	0.214
B75	-0.026	0.060	0.930	0.215	0.195

B90	-0.015	0.053	0.938	0.205	0.182
B95	-0.011	0.051	0.942	0.203	0.179
MA	-0.002	0.045	0.942	0.186	0.181
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na_min	-0.052	0.086	0.910	0.270	0.267
Na_pool	0.001	0.056	0.932	0.221	0.216
Te	-0.009	0.062	0.936	0.249	0.210
B50	-0.038	0.082	0.920	0.292	0.252
B75	-0.027	0.077	0.916	0.270	0.231
B90	-0.014	0.069	0.922	0.274	0.215
B95	-0.009	0.064	0.930	0.257	0.210
MA	-0.002	0.054	0.932	0.215	0.208
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Na_min	-0.006	0.062	0.956	0.235	0.245
Na_pool	0.152	0.161	0.204	0.211	0.217
Te	0.010	0.073	0.866	0.273	0.239
B50	-0.004	0.065	0.938	0.254	0.246
B75	-0.003	0.067	0.928	0.265	0.245
B90	0.001	0.070	0.904	0.278	0.242
B95	0.006	0.074	0.876	0.301	0.238
MA	0.047	0.094	0.790	0.309	0.260
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na_min	0.004	0.061	0.942	0.251	0.246
Na_pool	0.153	0.158	0.052	0.163	0.178
Te	0.004	0.071	0.914	0.289	0.244
B50	0.001	0.068	0.918	0.275	0.248
B75	0.001	0.069	0.914	0.279	0.247
B90	0.003	0.072	0.898	0.293	0.245
B95	0.006	0.078	0.868	0.323	0.242
MA	0.045	0.100	0.820	0.317	0.276
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Na_min	-0.003	0.059	0.954	0.239	0.247
Na_pool	0.247	0.253	0.008	0.213	0.219
Te	-0.002	0.061	0.958	0.238	0.249
B50	-0.006	0.064	0.942	0.244	0.248

B75	-0.006	0.064	0.942	0.244	0.248
B90	-0.006	0.064	0.942	0.244	0.248
B95	-0.006	0.065	0.940	0.245	0.247
MA	0.011	0.078	0.918	0.302	0.269
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na_min	0.000	0.061	0.950	0.246	0.247
Na_pool	0.249	0.253	0.000	0.177	0.181
Te	0.001	0.063	0.952	0.244	0.249
B50	-0.001	0.064	0.954	0.246	0.248
B75	-0.001	0.064	0.954	0.246	0.248
B90	-0.001	0.064	0.954	0.246	0.248
B95	-0.001	0.064	0.954	0.246	0.248
MA	0.008	0.068	0.948	0.264	0.255
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na_min	0.003	0.064	0.950	0.243	0.248
Na_pool	0.349	0.353	0.000	0.221	0.224
Te	0.002	0.065	0.946	0.252	0.249
B50	0.005	0.063	0.948	0.245	0.248
B75	0.005	0.063	0.948	0.245	0.248
B90	0.005	0.063	0.948	0.245	0.248
B95	0.005	0.063	0.948	0.245	0.248
MA	-0.005	0.063	0.952	0.247	0.252
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na_min	-0.003	0.064	0.946	0.251	0.249
Na_pool	0.351	0.354	0.000	0.173	0.186
Te	0.003	0.064	0.956	0.243	0.249
B50	0.002	0.062	0.956	0.234	0.249
B75	0.002	0.062	0.956	0.234	0.249
B90	0.002	0.062	0.956	0.234	0.249
B95	0.002	0.062	0.956	0.234	0.249
MA	-0.001	0.064	0.942	0.259	0.248
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na_min	0.002	0.065	0.936	0.262	0.248
Na_pool	0.347	0.349	0.000	0.146	0.164
Te	0.000	0.062	0.936	0.254	0.249

B50	0.000	0.061	0.954	0.243	0.248
B75	0.000	0.061	0.954	0.243	0.248
B90	0.000	0.061	0.954	0.243	0.248
B95	0.000	0.061	0.954	0.243	0.248
MA	0.002	0.063	0.950	0.238	0.246

Notes: "`Na_min`" refers to the Naïve approach that takes the smaller estimate, "`Na_pool`" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.9: Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions ($n = 500$ and $\rho = 0.75$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Na	-0.001	0.043	0.946	0.175	0.176
Te	-0.004	0.045	0.948	0.175	0.193
B50	0.004	0.045	0.982	0.178	0.250
B75	0.004	0.045	0.964	0.178	0.224
B90	0.004	0.045	0.956	0.178	0.201
B95	0.004	0.045	0.954	0.178	0.191
MA	0.001	0.047	0.940	0.189	0.183
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na	-0.007	0.052	0.970	0.191	0.215
Te	0.002	0.054	0.958	0.214	0.220
B50	-0.002	0.055	0.974	0.211	0.302
B75	-0.002	0.055	0.958	0.212	0.261
B90	-0.002	0.053	0.950	0.207	0.232
B95	-0.002	0.052	0.950	0.203	0.217
MA	0.002	0.054	0.958	0.196	0.211
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					

Na	0.149	0.158	0.234	0.209	0.217
Te	0.162	0.175	0.880	0.254	0.535
B50	0.149	0.159	0.952	0.218	0.553
B75	0.149	0.159	0.948	0.222	0.549
B90	0.146	0.157	0.900	0.228	0.528
B95	0.143	0.155	0.860	0.233	0.505
MA	0.129	0.142	0.768	0.228	0.420
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na	0.152	0.159	0.074	0.177	0.177
Te	0.149	0.156	0.904	0.173	0.527
B50	0.151	0.157	0.960	0.167	0.510
B75	0.151	0.157	0.954	0.167	0.509
B90	0.151	0.157	0.942	0.167	0.505
B95	0.151	0.157	0.900	0.169	0.496
MA	0.149	0.155	0.794	0.168	0.452
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Na	0.252	0.257	0.004	0.200	0.220
Te	0.273	0.279	0.984	0.214	0.803
B50	0.252	0.258	0.944	0.232	0.747
B75	0.252	0.258	0.944	0.232	0.746
B90	0.252	0.258	0.944	0.232	0.746
B95	0.252	0.258	0.942	0.232	0.745
MA	0.243	0.249	0.888	0.207	0.726
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na	0.250	0.255	0.000	0.191	0.181
Te	0.248	0.252	0.978	0.175	0.749
B50	0.252	0.256	0.936	0.172	0.710
B75	0.252	0.256	0.936	0.172	0.710
B90	0.252	0.256	0.936	0.172	0.710
B95	0.252	0.256	0.936	0.172	0.710
MA	0.245	0.249	0.940	0.174	0.700
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na	0.354	0.357	0.000	0.196	0.224
Te	0.376	0.380	0.978	0.199	1.007
B50	0.351	0.355	0.960	0.212	0.961

B75	0.351	0.355	0.960	0.212	0.961
B90	0.351	0.355	0.960	0.212	0.961
B95	0.351	0.355	0.960	0.212	0.961
MA	0.353	0.357	0.944	0.215	0.959
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na	0.346	0.349	0.000	0.170	0.186
Te	0.353	0.356	0.982	0.176	0.946
B50	0.352	0.355	0.952	0.172	0.910
B75	0.352	0.355	0.952	0.172	0.910
B90	0.352	0.355	0.952	0.172	0.910
B95	0.352	0.355	0.952	0.172	0.910
MA	0.349	0.352	0.952	0.173	0.904
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na	0.350	0.352	0.000	0.152	0.164
Te	0.334	0.336	0.970	0.150	0.909
B50	0.350	0.352	0.944	0.152	0.882
B75	0.350	0.352	0.944	0.152	0.882
B90	0.350	0.352	0.944	0.152	0.882
B95	0.350	0.352	0.944	0.152	0.882
MA	0.345	0.347	0.950	0.151	0.880

Notes: "Na" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.10: Performances in Normal Outcomes When Smaller Estimates Are Preferred (n = 500 and $\rho = 0.5$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
NNa_min	-0.032	0.063	0.924	0.214	0.218

Na_pool	0.004	0.044	0.962	0.164	0.176
Te	-0.006	0.048	0.946	0.193	0.179
B50	-0.030	0.062	0.922	0.211	0.213
B75	-0.020	0.060	0.916	0.212	0.195
B90	-0.011	0.054	0.924	0.213	0.184
B95	-0.006	0.051	0.932	0.205	0.180
MA	-0.001	0.048	0.942	0.183	0.181
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na_min	-0.052	0.086	0.912	0.266	0.267
Na_pool	0.004	0.055	0.960	0.209	0.216
Te	-0.011	0.066	0.928	0.249	0.210
B50	-0.037	0.078	0.920	0.260	0.253
B75	-0.026	0.073	0.918	0.264	0.229
B90	-0.014	0.065	0.926	0.261	0.214
B95	-0.010	0.062	0.932	0.242	0.209
MA	-0.001	0.051	0.962	0.198	0.210
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Na_min	-0.006	0.062	0.954	0.236	0.245
Na_pool	0.151	0.161	0.216	0.204	0.217
Te	0.013	0.077	0.860	0.284	0.237
B50	0.003	0.066	0.936	0.254	0.247
B75	0.004	0.068	0.920	0.256	0.245
B90	0.009	0.073	0.886	0.289	0.241
B95	0.014	0.078	0.852	0.304	0.238
MA	0.047	0.091	0.838	0.280	0.263
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na_min	0.004	0.062	0.948	0.247	0.246
Na_pool	0.152	0.159	0.070	0.180	0.178
Te	0.006	0.069	0.900	0.283	0.243
B50	-0.004	0.066	0.936	0.260	0.247
B75	-0.004	0.067	0.930	0.267	0.247
B90	-0.002	0.070	0.918	0.283	0.245
B95	0.001	0.076	0.890	0.324	0.242
MA	0.042	0.099	0.830	0.319	0.278
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					

Na_min	-0.003	0.060	0.948	0.243	0.247
Na_pool	0.255	0.260	0.004	0.206	0.220
Te	0.001	0.067	0.930	0.264	0.248
B50	-0.001	0.063	0.948	0.241	0.248
B75	-0.001	0.063	0.948	0.241	0.248
B90	-0.001	0.063	0.948	0.241	0.248
B95	-0.001	0.063	0.946	0.245	0.248
MA	0.010	0.075	0.922	0.288	0.268
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na_min	0.000	0.061	0.954	0.242	0.247
Na_pool	0.248	0.252	0.000	0.176	0.181
Te	0.003	0.062	0.956	0.237	0.248
B50	0.000	0.060	0.962	0.238	0.248
B75	0.000	0.060	0.962	0.238	0.248
B90	0.000	0.060	0.962	0.238	0.248
B95	0.000	0.061	0.962	0.238	0.248
MA	-0.001	0.065	0.960	0.253	0.253
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na_min	0.004	0.064	0.952	0.241	0.249
Na_pool	0.352	0.356	0.000	0.206	0.224
Te	0.002	0.063	0.960	0.238	0.250
B50	-0.001	0.067	0.938	0.266	0.248
B75	-0.001	0.067	0.938	0.266	0.248
B90	-0.001	0.067	0.938	0.266	0.248
B95	-0.001	0.067	0.938	0.266	0.248
MA	0.001	0.062	0.952	0.243	0.252
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na_min	-0.003	0.064	0.946	0.252	0.249
Na_pool	0.353	0.355	0.000	0.181	0.186
Te	-0.001	0.060	0.966	0.227	0.249
B50	0.002	0.062	0.950	0.243	0.249
B75	0.002	0.062	0.950	0.243	0.249
B90	0.002	0.062	0.950	0.243	0.249
B95	0.002	0.062	0.950	0.243	0.249

MA	-0.001	0.058	0.972	0.227	0.247
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na_min	0.002	0.065	0.938	0.259	0.248
Na_pool	0.348	0.350	0.000	0.157	0.165
Te	-0.002	0.063	0.948	0.251	0.250
B50	-0.002	0.065	0.940	0.252	0.248
B75	-0.002	0.065	0.940	0.252	0.248
B90	-0.002	0.065	0.940	0.252	0.248
B95	-0.002	0.065	0.940	0.252	0.248
MA	-0.001	0.064	0.948	0.248	0.247

Notes: "Na_min" refers to the Naïve approach that takes the smaller estimate, "Na_pool" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.11: Performances in Normal Outcomes When There is no Information about Preferred Modes or Directions ($n = 500$ and $\rho = 0.5$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Na	-0.002	0.043	0.956	0.170	0.176
Te	-0.001	0.045	0.944	0.179	0.194
B50	0.000	0.044	0.974	0.175	0.251
B75	0.000	0.044	0.960	0.175	0.222
B90	0.000	0.044	0.958	0.175	0.193
B95	0.000	0.044	0.956	0.175	0.186
MA	-0.001	0.043	0.966	0.166	0.181
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Na	-0.002	0.056	0.948	0.223	0.215
Te	0.000	0.056	0.934	0.219	0.221
B50	-0.003	0.053	0.978	0.211	0.300

B75	-0.003	0.052	0.964	0.210	0.261
B90	-0.003	0.052	0.956	0.201	0.228
B95	-0.003	0.051	0.956	0.199	0.218
MA	-0.001	0.048	0.966	0.187	0.213
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Na	0.150	0.160	0.236	0.212	0.217
Te	0.169	0.180	0.878	0.244	0.550
B50	0.148	0.157	0.938	0.213	0.547
B75	0.147	0.157	0.926	0.214	0.541
B90	0.145	0.156	0.902	0.220	0.526
B95	0.142	0.153	0.862	0.227	0.505
MA	0.128	0.141	0.802	0.239	0.434
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Na	0.149	0.156	0.086	0.166	0.177
Te	0.154	0.160	0.900	0.176	0.534
B50	0.152	0.158	0.948	0.178	0.509
B75	0.152	0.158	0.948	0.178	0.508
B90	0.152	0.158	0.930	0.178	0.502
B95	0.152	0.158	0.902	0.178	0.495
MA	0.151	0.158	0.808	0.179	0.453
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Na	0.252	0.258	0.008	0.222	0.220
Te	0.274	0.280	0.966	0.213	0.793
B50	0.250	0.256	0.946	0.218	0.754
B75	0.250	0.256	0.946	0.218	0.754
B90	0.250	0.256	0.946	0.218	0.754
B95	0.250	0.256	0.944	0.218	0.753
MA	0.247	0.254	0.924	0.231	0.735
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Na	0.247	0.251	0.000	0.184	0.181
Te	0.254	0.258	0.966	0.181	0.746
B50	0.254	0.258	0.940	0.173	0.716
B75	0.254	0.258	0.940	0.173	0.716
B90	0.254	0.258	0.940	0.173	0.716
B95	0.254	0.258	0.940	0.173	0.716

MA	0.252	0.256	0.934	0.174	0.699
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Na	0.350	0.354	0.000	0.214	0.224
Te	0.377	0.380	0.968	0.204	1.006
B50	0.350	0.355	0.924	0.213	0.941
B75	0.350	0.355	0.924	0.213	0.941
B90	0.350	0.355	0.924	0.213	0.941
B95	0.350	0.355	0.924	0.213	0.941
MA	0.350	0.354	0.956	0.199	0.951
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Na	0.352	0.354	0.000	0.167	0.187
Te	0.351	0.354	0.974	0.175	0.955
B50	0.350	0.353	0.952	0.174	0.908
B75	0.350	0.353	0.952	0.174	0.908
B90	0.350	0.353	0.952	0.174	0.908
B95	0.350	0.353	0.952	0.174	0.908
MA	0.351	0.354	0.952	0.180	0.906
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Na	0.349	0.351	0.000	0.163	0.164
Te	0.331	0.333	0.970	0.155	0.915
B50	0.353	0.355	0.962	0.154	0.882
B75	0.353	0.355	0.962	0.154	0.882
B90	0.353	0.355	0.962	0.154	0.882
B95	0.353	0.355	0.962	0.154	0.882
MA	0.351	0.353	0.958	0.149	0.879

Notes: "Na" refers to the Naïve approach that pools the mixed-mode data, "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

Table 2.12: Probabilities of the Model Being Selected in the Testimator Approach ($n = 5000$ and $\rho = 0.95$)

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	0.002	0.036	0.048	**0.914**
2.$u_b = 0$, $\sigma_b^2 = 2$	0.00	0.040	**0.960**	0.000	0.000
3.$u_b = 0.3$, $\sigma_b^2 = 2$	0.24	**1.000**	0.000	0.000	0.000
4.$u_b = 0.3$, $\sigma_b^2 = 1$	0.30	0.052	0.000	**0.948**	0.000
5.$u_b = 0.5$, $\sigma_b^2 = 2$	0.41	**1.000**	0.000	0.000	0.000
6.$u_b = 0.5$, $\sigma_b^2 = 1$	0.50	0.062	0.000	**0.938**	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	0.042	0.000	**0.958**	0.000
9.$u_b = 0.7$, $\sigma_b^2 = 0.5$	0.81	**1.000**	0.000	0.000	0.000

Notes: "D" stands for different and "S" stands for same. The correct model of a scenario are 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 $\frac{u_a - u_b}{\sqrt{\frac{\sigma_a^2 + \sigma_b^2}{2}}}$.

Table 2.13: Average Weight of Model Averaging Approach (n = 5000 and ρ = 0.95)

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	0.000	0.002	**0.998**
2.$u_b = 0$, $\sigma_b^2 = 2$	0.00	0.005	**0.995**	0.000	0.000
3.$u_b = 0.3$, $\sigma_b^2 = 2$	0.24	**1.000**	0.000	0.000	0.000
4.$u_b = 0.3$, $\sigma_b^2 = 1$	0.30	0.000	0.000	**1.000**	0.000
5.$u_b = 0.5$, $\sigma_b^2 = 2$	0.41	**1.000**	0.000	0.000	0.000
6.$u_b = 0.5$, $\sigma_b^2 = 1$	0.50	0.002	0.000	**0.998**	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	0.000	0.000	**1.000**	0.000
9.$u_b = 0.7$, $\sigma_b^2 = 0.5$	0.81	**1.000**	0.000	0.000	0.000

Notes: Correct models are marked in bold in the table. We differentiate each model by whether it assumes the same mean and same variances across modes. When specifying a scenario, "Same" means no shift in mean/variances, "Small" and "Large" means small and large shifts in means/variances respectively, and "Medium" means medium shifts in means.

Table 2.14: Performances in Normal Outcomes When Smaller Estimates Are Preferred ($n = 5000$ and $\rho = 0.95$)

Method	Bias	RMSE	Coverage	Expected Length	Actual Length
Scenario 1: $u_b = 0$, $\sigma_b^2 = 1$. Effect size: 0					
Te	-0.001	0.016	0.946	0.062	0.056
B50	-0.008	0.019	0.924	0.065	0.066
B75	-0.005	0.018	0.916	0.069	0.061
B90	-0.002	0.017	0.922	0.065	0.057
B95	-0.001	0.016	0.934	0.059	0.056
MA	-0.001	0.014	0.954	0.054	0.056
Scenario 2: $u_b = 0$, $\sigma_b^2 = 2$. Effect size: 0					
Te	-0.002	0.019	0.940	0.075	0.065
B50	-0.011	0.023	0.948	0.076	0.077
B75	-0.007	0.021	0.948	0.077	0.070
B90	-0.004	0.019	0.954	0.072	0.066
B95	-0.003	0.018	0.954	0.067	0.065
MA	-0.001	0.016	0.952	0.063	0.065
Scenario 3: $u_b = 0.3$, $\sigma_b^2 = 2$. Effect size: 0.24					
Te	-0.001	0.020	0.942	0.081	0.079
B50	-0.001	0.021	0.958	0.077	0.078
B75	-0.001	0.021	0.958	0.077	0.078
B90	-0.001	0.021	0.958	0.077	0.078
B95	-0.001	0.021	0.958	0.077	0.078
MA	-0.001	0.020	0.948	0.079	0.078
Scenario 4: $u_b = 0.3$, $\sigma_b^2 = 1$, Effect size: 0.30					
Te	0.000	0.019	0.968	0.072	0.079
B50	0.001	0.020	0.936	0.082	0.078
B75	0.001	0.020	0.936	0.082	0.078
B90	0.001	0.020	0.936	0.082	0.078
B95	0.001	0.020	0.936	0.082	0.078
MA	0.001	0.020	0.954	0.076	0.078
Scenario 5: $u_b = 0.5$, $\sigma_b^2 = 2$. Effect size: 0.41					
Te	-0.001	0.019	0.968	0.073	0.079
B50	-0.001	0.020	0.958	0.076	0.078

B75	-0.001	0.020	0.958	0.076	0.078
B90	-0.001	0.020	0.958	0.076	0.078
B95	-0.001	0.020	0.958	0.076	0.078
MA	0.000	0.020	0.948	0.078	0.078
Scenario 6: $u_b = 0.5$, $\sigma_b^2 = 1$. Effect size: 0.50					
Te	0.001	0.020	0.952	0.077	0.079
B50	-0.001	0.020	0.954	0.077	0.078
B75	-0.001	0.020	0.954	0.077	0.078
B90	-0.001	0.020	0.954	0.077	0.078
B95	-0.001	0.020	0.954	0.077	0.078
MA	0.001	0.020	0.950	0.077	0.078
Scenario 7: $u_b = 0.7$, $\sigma_b^2 = 2$. Effect size: 0.57					
Te	-0.002	0.020	0.956	0.076	0.079
B50	0.001	0.021	0.946	0.079	0.078
B75	0.001	0.021	0.946	0.079	0.078
B90	0.001	0.021	0.946	0.079	0.078
B95	0.001	0.021	0.946	0.079	0.078
MA	0.000	0.019	0.972	0.070	0.078
Scenario 8: $u_b = 0.7$, $\sigma_b^2 = 1$. Effect size: 0.70					
Te	0.000	0.020	0.958	0.073	0.079
B50	0.000	0.020	0.960	0.075	0.078
B75	0.000	0.020	0.960	0.075	0.078
B90	0.000	0.020	0.960	0.075	0.078
B95	0.000	0.020	0.960	0.075	0.078
MA	0.000	0.019	0.948	0.078	0.078
Scenario 9: $u_b = 0.7$, $\sigma_b^2 = 0.5$. Effect size: 0.81					
Te	0.000	0.020	0.968	0.074	0.079
B50	0.000	0.021	0.936	0.080	0.078
B75	0.000	0.021	0.936	0.080	0.078
B90	0.000	0.021	0.936	0.080	0.078
B95	0.000	0.021	0.936	0.080	0.078
MA	0.000	0.019	0.956	0.073	0.078

Notes: "Te" refers to the testimator approach, "B50" refers to the Bayesian approach with 50% interval, "B75" refers to the Bayesian approach with 75% interval, "B90" is the Bayesian approach with 90% interval, "B95" is the Bayesian approach with 95% interval, and "MA" refers to the Model Averaging approach.

CHAPTER 3

Investigating Mode Effects in Interviewer Variances Using Two Representative Multi-mode Surveys

3.1 Introduction

Interviewers play a central role in survey data collection. Depending on the mode and sampling design of data collection, they may need to list addresses to generate sampling frames, recruit respondents, ask survey questions, and record participants' responses. Therefore, from a total survey error framework, interviewers can affect survey data quality by generating or reducing coverage error, nonresponse error, measurement error, and processing error [34]. Most research examining interviewers' effects focuses on measurement error [35, 36, 37], which can be further decomposed into a systematic part, the bias due to interviewers (when respondents alter answers either because of the presence of interviewers or their observable traits), and a random component, interviewer variance. This interviewer variance inflates the uncertainty of the estimates, sometimes to an even greater degree than the correlation induced by geographical clustering [38]. This study focuses on determining the effect of different modes of data collection – specifically telephone (TEL) versus face-to-face (FTF) – on interviewer variances in mixed-mode surveys.

Interviewer variances were first studied in the context of face-to-face interviews [39]. When telephone surveys became an alternative to FTF interviews, researchers evaluated interviewer variances in telephone surveys and generally found that they were less substantial than those in personal surveys [40, 41, 42]. Specifically, the intraclass correlation ρ_{int}, a common measure used to assess interviewer effects and defined by the ratio of interviewer variances to the total variance, ranged from 0.005 to 0.102 in FTF surveys, whereas those computed in centralized TEL surveys ranged from 0.0018 to 0.0184 [40, 42]. The finding is aligned with theoretical expectations, as interviewers in the centralized TEL setting are more closely monitored and supervised than field interviewers are [43]. Since then, the research domain has received little scholarly attention. However, as mixed-mode designs become increasingly used, the subject of study calls for more research. There is a lack of first-hand evidence as the prior findings are mostly based on different surveys that employ one mode (FTF or TEL). Besides, mixed-mode surveys naturally provide an opportunity where the survey context and the questionnaires are highly comparable (if not the same) when comparing interviewer variances in both modes. Furthermore, depending on whether interviewers are responsible for both modes in mixed-mode surveys, interviewers can potentially carry their influence from one mode to another. These factors can lead to different results in comparing interviewer variances between modes.

Investigating mode effects in interviewer variances is also useful to facilitate mixed-mode designs and serve as an indicator of data quality. First, quantifying mode-specific interviewer variance can help researchers to determine and choose the mode with low interviewer variance in a multimode design. The current state-of-the-art mixed-mode inference strategy focuses on the bias property of modes [28, 26], but little was done to incorporate the potential heterogeneous variance structure [44]. Part of the reason is that little literature sheds light on the variance properties of mixed-mode designs [45], especially what goes into the variances. Second, identifying the questions associated with large interviewer variance mode effects can inform how interviewer variance is generated and thus might be reduced. For example, researchers show that attitudinal, sensitive, ambiguous, complex, and open-ended questions are generally more vulnerable to interviewer effects [43], as those questions introduce more opportunities for the interviewer to help the respondents [34]. If sensitive questions only present a large interviewer effect in FTF but not in TEL, that may suggest the questions bring a burden to field interviewers. To address that, survey organizations can provide additional training to standardize how to ask the question or use other approaches [such as audio computer-assisted self-interviewing [ACASI] or the item count technique [13]] to collect information for sensitive items. Third, in mixed-mode designs where interviewers are responsible for both modes, we can potentially find specific interviewers that have a large effect on responses in both modes or only in one mode, which provide the basis for real-time intervention and interviewer training at a more granular level.

In this paper, we consider two representative multi-mode studies: 1) the Arab Barometer Study (ABS) Wave 6 Jordan experiment and 2) the Health and Retirement Study (HRS) 2016. Drawing on both data sources, we consider mode effects in interviewer variances for interviewers in different countries, for different target populations, and for a variety of outcome variables. Additionally, the use of the two studies offers distinct perspectives for examining our research question. The ABS interviewer design is commonly used in surveys where different modes are managed by separate data collection agencies, resulting in different interviewers across modes. On the other hand, the HRS interviewer design, where the same interviewers are utilized in both modes, facilitates a more precise estimation of the differences in interviewer variances solely due to modes, by eliminating the portion of interviewer variances that result from using different interviewers across modes.

The remainder of this paper is organized as follows. In Section 2, we describe the study design and analytical strategy, and present the results using our first data source – ABS. Section 3 introduces the second data source – HRS, along with the corresponding analytical approach and the results pertaining to interviewer variance associated with the HRS data. In Section 4, we conduct a simulation study to illustrate the power to detect mode effects in interviewer variances using both the ABS and the HRS setup. Finally, in Section 5, we discuss the implications of our study.

3.2 The Arab Barometer Study

3.2.1 Study Description

The ABS is the largest repository of public opinion data in the Middle East and North Africa (MENA) region. In wave 6, it embedded a mode experiment in Jordan between March and April 2021, where participants were randomly assigned to either a personal interview or a TEL recontact interview. Center for Strategic Studies in Jordan conducted the field work using the 2015 Population and Housing Census as the sampling frame. They implemented an area probability sample stratified on governorate and urban-rural cleavages. Separate interviewers were used in the FTF and TEL interviews. 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 telephone follow-up. In the FTF mode, 31 interviewers collected data from 1,193 respondents, while 13 interviewers interviewed 1,212 participants via phone.

We focus on three types of outcome variables (Y): 1) sensitive political questions (6 items), 2) less sensitive international questions (3 items), and 3) whether reported do not know or refused to answer international relationship questions (3 items). Except for the item missing indicators, the other outcome variables were initially measured by four ordinal categories; we collapsed them into binary outcomes by setting the cutoff point in the middle (see the original and the collapsed categories in the Appendix A).

Outcome variables (Y) can be subject to two types of mode effects: 1) mode effects that lead to a shift in the means of outcome variables (referred to as mode effects in means) and 2) mode effects in interviewer variances. We consider, in total, q interviewers collect information in only one of two modes (FTF and TEL) from n sample units from a finite population. Interviewers also collect respondent-level covariates (X) that are predictive of the outcome variables (Y). The covariates (X) are assumed to be independent of any mode effects. We consider covariates (X) including respondents' age, gender, marital status, household size, and regions in this paper.

3.2.2 Analytical Strategy

First, to illustrate the descriptive statistics of interviewer variation in the collected responses, we compute the between-interviewer standard deviation (SD) and the average within-interviewer SD. Specifically, we calculate the average proportions for each variable and interviewer ($\bar{y}_{(m)j}$). In the ABS setup, where interviewers are nested within each mode, these statistics are inherently mode-specific; therefore, we enclose m in parentheses to emphasize this point. We then calculate the SD of these average proportions across interviewers, termed the between-interviewer SD. The within-interviewer SD (v_j^m) is derived from the responses collected by each interviewer. The average within-interviewer SD is computed as the mean of the within-interviewer SDs across all interviewers for each mode. We show the formula to compute the relevant statistics in 3.1, where i indexes respondents, j indexes interviewers, m indexes modes, $n_{(m)j}$ reflects the number of interviews conducted by interviewer j using mode m, n_m represents the number of respondents in mode m, n_j^m indicates the number of interviewers using mode m, and $y_{i(m)j}$ indicates the responses provided by respondent i interviewed by interviewer j using mode m. From the perspective of survey data collection agencies, a small SD between interviewers and a large average within-interviewer SD are desirable, as this may indicate an interviewer assignment that is close to random and minimal effects from interviewers on the collected responses. We report the statistics for both the covariates and the outcomes of interest. The statistics for the covariates can suggest interviewer selection effects, thereby highlighting the importance of considering the covariates in the final analytical model. The statistics for the outcome variables may provide initial evidence of the presence of interviewer effects and justify further investigation.

$$
\begin{aligned}
&\text{Average proportion per interviewer } \bar{y}_{(m)j} = \frac{\sum_i^{n_(m)j} y_{i(m)j}}{n_{(m)j}} \\
&\text{Average proportion per mode } \bar{y}_m = \frac{\sum_i^{n_m} y_{i(m)j}}{n_m} \\
&\text{Between-interviewer SD } = \sqrt{\frac{\sum_j^{n_j^m} (\bar{y}_{(m)j} - \bar{y}_m)^2}{n_j^m}} \\
&\text{Within-interviewer SD } v_j^m = \sqrt{\frac{\sum_i^{n_{(m)j}} (\bar{y}_{i(m)j} - \bar{y}_{(m)j})^2}{n_{(m)j}}} \\
&\text{Average within-interviewer SD } = \frac{\sum_j^{n_j^m} v_j^m}{n_j^m}
\end{aligned}
\tag{3.1}
$$

To test whether interviewer variances are equal across modes, since all the outcome variables are binary, we fit the following probit model to each of the variables, where m indexes modes (f for FTF and t for TEL), M and $J_{j,j=1,\dots,q-1}$ are dummy variables (length of n) to indicate modes ($M = 1$ for the FTF mode and $M = 0$ for the TEL mode) and interviewers:

$$
\begin{aligned}
&Y^*_{ij(m)} = \beta_0 + \beta_1 M_i + b_{j(m)} + \epsilon_{ij(m)}, \\
&Y_{ij(m)} = 1 \text{ if } Y^*_{ij(m)} > 0 \text{ and } Y_{ij(m)} = 0 \text{ if } Y^*_{ij(m)} \leq 0, \\
&b_{j(m)} \sim N(0, \sigma_m^2), \\
&\epsilon_{ij(m)} \sim N(0, 1), \\
&\sigma_f, \sigma_t \sim half - T(3, 1) \text{ (for Bayesian modeling)}, \\
&\gamma, \beta_0, \beta_1 \sim N(0, 10^6) \text{ (for Bayesian modeling)}
\end{aligned}
\tag{3.2}
$$

In Model 3.2, the interviewer random effects are represented as $b_{j(m)}$ as interviewers are nested within the modes. Our research question, "Are interviewer variances equal between modes in a randomized mixed-mode design?" is addressed by evaluating if $\alpha = log(\sigma_f) - log(\sigma_t)$ is equal to zero for each variable in Model 3.2. To determine this, we examine if the

95% confidence or HPD credible intervals of α include zero. If the intervals do not include zero for some variables, it suggests that the interviewer variances are not equal between modes for those variables.

By fitting 3.2, we can also obtain estimates of mode effects (β_1) for each variable by computing and testing if the quantity differs from 0. Note that the estimates may include some mode selection effects; despite the random mode assignment, differential nonresponse can happen across the modes [46].

Suppose evidence suggests that $\alpha \neq 0$, we then consider whether the mode-specific interviewer variance is spurious due to the lack of interpenetrated designs by adding respondent-level covariates (x_{si}, where s denotes covariate s) to Model 3.2:

$$\begin{aligned}
&Y^*_{ij(m)} = \beta_0 + \beta_1 M_i + b_{j(m)} + \sum_s^S \gamma_s x_{si} + \epsilon_{ij(m)}, \\
&Y_{ij(m)} = 1 \text{ if } Y^*_{ij(m)} > 0 \text{ and } Y_{ij(m)} = 0 \text{ if } Y^*_{ij(m)} \leq 0, \\
&b_{j(m)} \sim N(0, \sigma^2_m), \\
&\epsilon_{ij(m)} \sim N(0, 1), \\
&\sigma_f, \sigma_t \sim half - T(3, 1) \text{ (for Bayesian modeling)}, \\
&\boldsymbol{\gamma}, \beta_0, \beta_1 \sim N(0, 10^6) \text{ (for Bayesian modeling)}
\end{aligned} \tag{3.3}$$

We implement the models using both likelihood (Proc Nlmixed) and Bayesian approaches (Proc MCMC) in the SAS programming language. In the likelihood approach, we take log transformation on σ^2_f and σ^2_t to stabilize the variance of the parameters and improve the coverage property. We compute the variance of the estimated α using the delta method, given by $var(\alpha) = \frac{1}{4}var(log(\sigma^2_f)) + \frac{1}{4}var(log(\sigma^2_t))$ (see the derivations in the Appendix B), then use a normal distribution to estimate the 95% confidence interval. In the Bayesian approach, we use one chain with 200,000-300,000 draws, depending on the autocorrelation and effective sample size, and select every 100th value as the thinning rate. For the ease

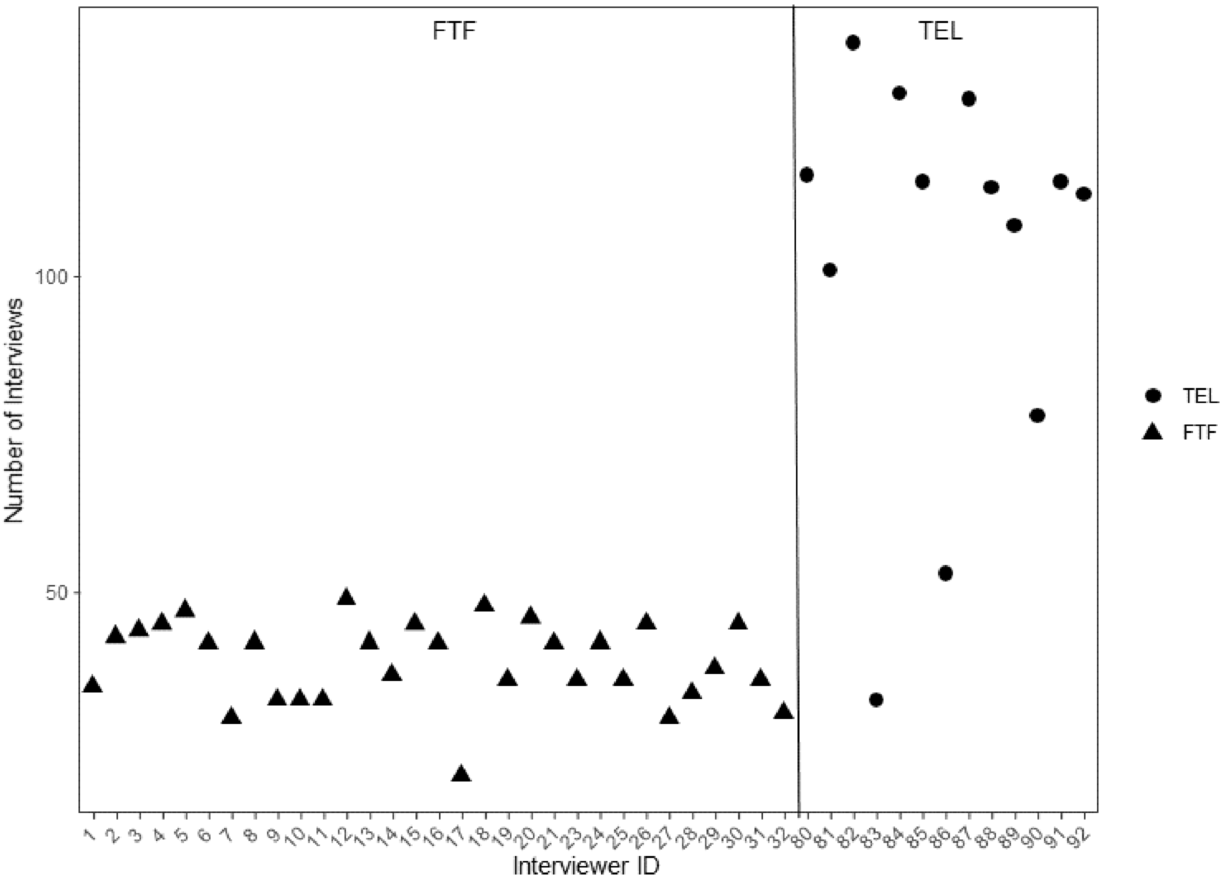


Figure 3.1: Interviewer Workloads Per Mode in the Arab Barometer Study

of illustration, we only report the results of the model with covariates added and estimated using Bayesian modeling (Model 3.3) in the later section.

3.2.3 Results

3.2.3.1 Descriptive Statistics

We assume interviewers are interchangeable in this paper. To partly evaluate this assumption, we present the interviewer workloads in the FTF and TEL modes in the ABS in Figure 3.1. In Figure 3.1, we note that in the FTF mode, each interviewer conducts a similar number of interviews. In contrast, both the mean and the variation in the number of interviews per interviewer are larger and more variable in the TEL mode.

We report unweighted mode-specific sample means, between-interviewer standard deviations (SDs), and average within-interviewer SDs in Table 3.1. From Table 3.1. First, we observe that for sensitive political questions, the average proportions reported via telephone (TEL) are generally higher than those reported in face-to-face interviews (FTF), suggesting that TEL may be associated with more positive reporting. Second, between-interviewer SDs in FTF are generally larger than those in TEL for most outcomes, while the average within-interviewer SD is larger in TEL than in FTF for sensitive political questions and missing indicators. This provides some initial evidence that interviewers seem to have a larger effect in FTF than in TEL. We provide the distribution of the outcome variables per interviewer in Appendix C.

Table 3.1: Unweighted Distribution of Outcome variables in the Arab Barometer Study across Interviewers by Modes

Questions	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average Within interviewer SD (FTF)	Average Within interviewer SD (TEL)
Sensitive political questions						
1. Freedom of the media	0.403	0.588	0.191	0.117	0.455	0.480
2. trust in government	0.356	0.533	0.165	0.122	0.455	0.487
3. trust in courts	0.594	0.770	0.139	0.123	0.477	0.398
4. satisfied with healthcare	0.491	0.592	0.155	0.071	0.482	0.489
5. performance on inflation	0.140	0.243	0.146	0.142	0.291	0.406
6. performance during COVID-19	0.402	0.576	0.171	0.161	0.464	0.470
International Questions						
7. favorable of the United States	0.394	0.415	0.187	0.189	0.467	0.459
8. favorable of Germany	0.488	0.560	0.224	0.186	0.464	0.464
9. favorable of China	0.468	0.507	0.207	0.203	0.470	0.463
Whether missing on international questions (constructed)						
10. missing on favorable of the United States	0.253	0.297	0.235	0.158	0.341	0.425
11. missing on favorable of Germany	0.320	0.381	0.247	0.199	0.384	0.442
12. missing on favorable of China	0.283	0.329	0.252	0.180	0.359	0.431

We show unweighted sample characteristics in the FTF and the TEL modes in Table 3.2. Under the randomized mixed-mode design, the Jordan sample is roughly balanced on key demographic and socioeconomic variables (age, gender, education, marital status, household size, and region) across modes. However, there are slightly more males (0.55 vs 0.50) respondents in the TEL mode relative to the FTF mode, possibly due to differential nonresponse. We note that for these covariates, the between-interviewer SD in FTF is usually much larger than that in TEL, suggesting potentially larger selection effects in FTF, since we assume the covariates are not susceptible to measurement error.

Table 3.2: Unweighted Distribution of Sample Characteristics of the Arab Barometer Study across Interviewers by Modes

Respondent Variables	Mean (FTF)	Mean (TEL)	Between interviewer SD (FTF)	Between interviewer SD (TEL)	Average Within interviewer SD (FTF)	Average Within interviewer SD (TEL)
Age 18-24	0.166	0.164	0.085	0.039	0.361	0.369
Age 25-34	0.226	0.203	0.072	0.038	0.415	0.402
Age 35-44	0.227	0.215	0.088	0.052	0.412	0.408
Age 45-54	0.199	0.219	0.069	0.031	0.394	0.414
Age 55+	0.183	0.198	0.071	0.032	0.381	0.399
Male	0.497	0.549	0.291	0.041	0.369	0.499
Less than secondary education	0.345	0.337	0.125	0.106	0.463	0.463
Secondary education	0.365	0.357	0.098	0.082	0.477	0.474
Higher than secondary education	0.290	0.307	0.101	0.051	0.445	0.461
Unmarried	0.238	0.264	0.106	0.063	0.412	0.438
Married	0.693	0.684	0.082	0.062	0.459	0.463
Divorced, widows, separated	0.069	0.053	0.044	0.024	0.230	0.219
Household size: Less than 3	0.208	0.222	0.083	0.040	0.399	0.416
Household size: 4-5	0.345	0.349	0.081	0.061	0.475	0.475
Household size: 6-7	0.281	0.288	0.079	0.072	0.447	0.449
Household size: 8+	0.165	0.141	0.089	0.056	0.353	0.330
Region: Central	0.523	0.509	0.154	0.255	0.482	0.429
Region: North	0.261	0.282	0.101	0.188	0.424	0.388
Region: South	0.216	0.209	0.175	0.119	0.333	0.367

3.2.3.2 Mode effects in Means and Interviewer Variances

This section reports the modeling results that incorporate respondent information (Model 3.3) using Bayesian estimation in Table 3.3. With respect to the mode effects in means, we observe negative estimates for all sensitive items. For example, the probability of an unmarried male participant aged 18-24, with higher than secondary education, living in a household with fewer than three individuals, and residing in the North region of Jordan, reporting that media freedom is guaranteed to a great or medium extent, decreases by 17.9% when interviewed via face-to-face (FTF) methods compared to telephone (TEL) interviews. The 17.9% is calculated using $\phi(\beta_0+\beta_1+\sum_s^S \gamma_s x_{si})\beta_1$, where ϕ is the pdf of a standard normal distribution and S is the number of covariates (x). The estimates of γ_s are not provided in the paper but can be provided upon request. The negative mode effects in means suggest that respondents expressed lower opinions of the government when answering FTF interviews, which could be more honest responses given Jordan's authoritarian regime. Table 3.3 also indicates that missing rates for international questions are lower in FTF interviews compared to TEL interviews (though this is not statistically significant at the 0.05 level). We did not incorporate sample weights in the analysis as our focus of inference is repeated sampling under the same survey design.

Next, we turn our attention to the interviewer variances. Firstly, the magnitude of interviewer variances is generally large in the ABS. For sensitive political questions, the interviewer variances range from 0.018 to 0.393 (Table 3.3). Previous literature examining interviewer effects usually reported interviewer intraclass correlation (ρ_{int}) to reflect the proportion of variance due to interviewers. To compute mode-specific $\rho_{m,int}$, we can use the formula $\rho_{m,int} = \frac{var_{m,int}}{1+var_{m,int}}$, since the residual variance in the probit model is 1. Consequently, the previously mentioned results correspond to ρ_{int} ranging from 0.018 to 0.282. As a reference, based on the literature, a value of ρ_{int} below 0.01 is considered small, while a value higher than 0.12 is regarded as large [47]. In Table 3.3, we observe that $\rho_{f,int}$ and $\rho_{t,int}$ can vary substantially for the same outcome. For example, for satisfaction with healthcare, $\rho_{f,int}$

is 0.125, while $\rho_{t,int}$ is 0.029. It is important to consider these differences when using the $\rho_{m,int}$ values to calculate the effective sample sizes associated with a specific data collection mode.

For one sensitive item, performance in the healthcare system, we observe marginally significant difference in interviewer variances in Table 3.3 using Bayesian estimation (and significant using likelihood estimation, see Appendix D). In this item, the estimates of interviewer variances are considerably larger in the FTF mode. For 5 out of 6 sensitive items, FTF interviewer variances are somewhat larger than TEL interviewer variances. The differences are not statistically significant, possibly due to the limited power determined by the small number of interviewers in this study. The larger interviewer variances in FTF are consistent with theoretical expectations, as interviewers may exhibit greater heterogeneity in administering sensitive questions and establishing rapport with respondents during in-person interviews.

Counterintuitively, for substantive responses to nonsensitive international attitude questions (items 7-9), the interviewer variance estimates are generally larger in TEL compared to FTF (not significantly). The interviewer variances of whether reporting “don’t know” or refusing to answer the nonsensitive international questions are larger in FTF than in TEL (significant on the first item). This finding may be because interviewers assigned to FTF mode tried to persuade respondents to give substantive answers, and whether the persuasion happens or is successful can differ by interviewers.

Table 3.3: Interviewer Variances Per Mode for Selected Items in the Arab Barometer Study Adjusting for Covariates Using Bayesian Estimation

Questions	σ_f^2	σ_t^2	$\rho_{f,int}$	$\rho_{t,int}$	α	β_1
			Sensitive political questions			
1. Freedom of the media	0.252 [0.122, 0.428]	0.135 [0.036, 0.284]	0.201 [0.109, 0.3]	0.119 [0.035, 0.221]	0.355 [-0.223, 0.898]	**-0.526** **[-0.795, -0.222]**
2. trust in government	0.188 [0.083, 0.322]	0.127 [0.029, 0.275]	0.158 [0.077, 0.244]	0.113 [0.028, 0.216]	0.239 [-0.382, 0.838]	**-0.504** **[-0.768, -0.238]**
3. trust in courts	0.113 [0.038, 0.201]	0.214 [0.05, 0.445]	0.102 [0.037, 0.167]	0.176 [0.048, 0.308]	-0.291 [-0.94, 0.318]	**-0.555** **[-0.881, -0.273]**
4. satisfied with healthcare	0.143 [0.051, 0.251]	0.03 [0, 0.075]	0.125 [0.049, 0.201]	0.029 [0, 0.07]	0.906 [-0.054, 1.758]	**-0.278** **[-0.475, -0.085]**
5. performance on inflation	0.393 [0.153, 0.672]	0.204 [0.051, 0.435]	0.282 [0.133, 0.402]	0.169 [0.049, 0.303]	0.361 [-0.275, 0.927]	**-0.523[-0.861, -0.153]**
6. performance during COVID-19	0.202 [0.084, 0.34]	0.224 [0.07, 0.443]	0.168 [0.077, 0.254]	0.183 [0.065, 0.307]	-0.026 [-0.602, 0.508]	**-0.51** **[-0.841, -0.205]**
			International Questions			

7. favor US	0.198 [0.074, 0.34]	0.362 [0.104, 0.719]	0.165 [0.069, 0.254]	0.266 [0.094, 0.418]	-0.278 [-0.841, 0.282]	-0.057 [-0.45, 0.318]
8. favor Germany	0.292 [0.12, 0.514]	0.33 [0.092, 0.663]	0.226 [0.107, 0.339]	0.248 [0.084, 0.399]	-0.037 [-0.603, 0.548]	-0.147 [-0.551, 0.236]
9. favor China	0.205 [0.083, 0.361]	0.378 [0.116, 0.787]	0.17 [0.077, 0.265]	0.274 [0.104, 0.44]	-0.282 [-0.869, 0.245]	-0.15 [-0.549, 0.19]
Whether missing on international questions (constructed)						
10. missing on favor US	0.995 [0.48, 1.71]	0.343 [0.104, 0.668]	0.499 [0.324, 0.631]	0.255 [0.094, 0.4]	**0.557** **[0.014,** **1.121]**	-0.298 [-0.805, 0.172]
11. missing on favor Germany	0.844 [0.404, 1.324]	0.464 [0.16, 0.857]	0.458 [0.288, 0.57]	0.317 [0.138, 0.461]	0.324 [-0.169, 0.839]	-0.287 (0.24) [-0.765, 0.149]
12. missing on favor China	0.936 [0.434, 1.552]	0.452 [0.118, 0.933]	0.483 [0.303, 0.608]	0.311 [0.106, 0.483]	0.398 [-0.134, 0.949]	-0.244 [-0.73, 0.229]

Notes: Significant results are marked in bold. β_1 refers to the mode effect estimates in means. α_1 refers to the mode effect estimates in interviewer variances. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,int}$ and $\rho_{t,int}$ are interviewer intraclass correlation in FTF and TEL, respectively.

3.3 Health and Retirement Study 2016

3.3.1 Study Description

The HRS is a longitudinal panel study that surveys people over age 50 (and their spouses) in the United States. It is conducted biennially, started in 1992, and has studied more than 43,000 people [48]. The HRS is sponsored by the National Institute on Aging (grant number NIA U01AG009740) and is conducted by the University of Michigan. The HRS sample was drawn using a multistage, national area-clustered probability sample frame [49]. Since 2006, the HRS has initiated the rotation of enhanced FTF interviews (during which physical and biological measures and a psychosocial questionnaire are collected, in addition to the regular information collection) and TEL interviews across waves for participants, unless they are older than 80 years, newly recruited into the sample, or spouses of another HRS participant. In these cases, they rotate between regular FTF and enhanced FTF interviews. In this study, we are interested in analyzing the HRS 2016 data, when the Late Baby Boomers (LBB) cohort was added to replenish the HRS sample. Although not every interviewer collects data in both modes, under the HRS design, interviewers are responsible for data collection in both FTF and TEL modes. The HRS 2016 was fielded from April 2016 to April 2018, with a sample size of 20,912 [response rate: 82.8%, [50]]. In our analytical sample, we excluded respondents who were missing data on mode indicators, interviewer IDs, and covariates, resulting in a sample size of 20,868.

We consider four types of outcome variables in the HRS study, including 1) nine items of the Center for Epidemiologic Studies Depression Scale (CESD), 2) six items of interviewer observations, and 3) a three-item physical activity scale (see the Appendix E for the question wordings, the original response categories and categories used in the study). We consider nine respondent-level covariates (X), including age, sex, race / ethnicity, interview language, education, whether respondents are coupled and working. All participants are included in our sample, unless they are missing data in either the outcome or predictor variables. Missing rates for predictor variables are minor, and those for outcome variables are less than 0.05.

3.3.2 Analytical Strategy

Similar to the descriptive statistics reported in the ABS, we report the between-interviewer SD and the average within-interviewer SD to gain an intuitive understanding of the interviewer effects in the outcome variables examined in the HRS.

Next, we fit multilevel models to each of the outcome variables using the same notation as in Model 3.2. Unlike the ABS, interviewers are not nested in model hence a single interviewer can interview in both modes, and thus interviewer effects can be correlated across modes. Therefore we posit a bivariate normal model for the interviewer effects:

$$
\begin{aligned}
&Y^*_{ijm} = \beta_0 + \beta_1 M_i + b_{jm} + \sum_s^S \gamma_s x_{si} + \epsilon_{ijm},\\
&Y_{ijm} = 1 \text{ if } Y^*_{ijm} > 0 \text{ and } Y_{ijm} = 0 \text{ if } Y^*_{ijm} \leq 0,\\
&\begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix}\right),\\
&\epsilon_{ijm} \sim N(0,1),\\
&\sigma_f, \sigma_t \sim half-T(3,1) \text{ (for Bayesian modeling)},\\
&\rho \sim U(-1,1) \text{ (for Bayesian modeling)},\\
&\boldsymbol{\gamma}, \beta_0, \beta_1 \sim N(0, 10^6) \text{ (for Bayesian modeling)}.
\end{aligned}
\tag{3.4}
$$

Similarly, we use $\alpha = log(\sigma_f) - log(\sigma_t)$ as a metric to answer our research question. To test if α is equal to zero for each variable, we assess if the 95% credible intervals or confidence intervals include zero. Additionally, to control for interviewer selection effects, we include respondent-level covariates as fixed effects in the model.

We apply the Fisher Z transformation ($z = \frac{1}{2}ln(\frac{1+\rho}{1-\rho})$) when constructing the 95% confidence interval for ρ in the likelihood approach. We calculate the variance of α using the delta method, given by $var(\alpha) = \frac{1}{4}var(log(\sigma_f^2)) + \frac{1}{4}var(log(\sigma_t^2)) - \frac{1}{2}cov(log(\sigma_f^2), log(\sigma_t^2))$, which is slightly different from the ABS (see the derivations in Appendix B).

3.3.3 Results

3.3.3.1 Descriptive Statistics

First, we illustrate the interviewer load in Figure 3.2. In HRS 2016, 382 interviewers were employed for data collection. The number of interviews conducted in FTF and TEL is very different across interviewers. Eighty-two (21.5%) interviewers exclusively conducted telephone interviews, while thirty-seven (9.7%) solely conducted in-person interviews. The remaining 263 (68.9%) interviewers conducted both types of interviews. All interviews are included in the analysis, although estimation of the covariances between the FTF and TEL effects within interviewer are limited to the subsample of interviewers who conducted both types of interviews.

Second, we present unweighted sample characteristics for both FTF and TEL modes in Table 3.4. Compared to TEL respondents, a higher proportion of FTF respondents are under 60 or over 80 years old, belong to minority groups, are not in a relationship, have not completed high school, and are currently employed. This unbalanced sample distribution underscores the importance of including demographic and socioeconomic status variables in the analytical model when analyzing interviewer effects. Comparing the statistics from HRS to those from ABS, we note that in the HRS, the between-interviewer SDs are generally higher and the average within-interviewer SDs are generally lower. This suggests that interviewer selection effects are potentially a larger threat when analyzing interviewer variance in the HRS. This is consistent with our expectations, as randomized mode assignment is applied in ABS but not in HRS.

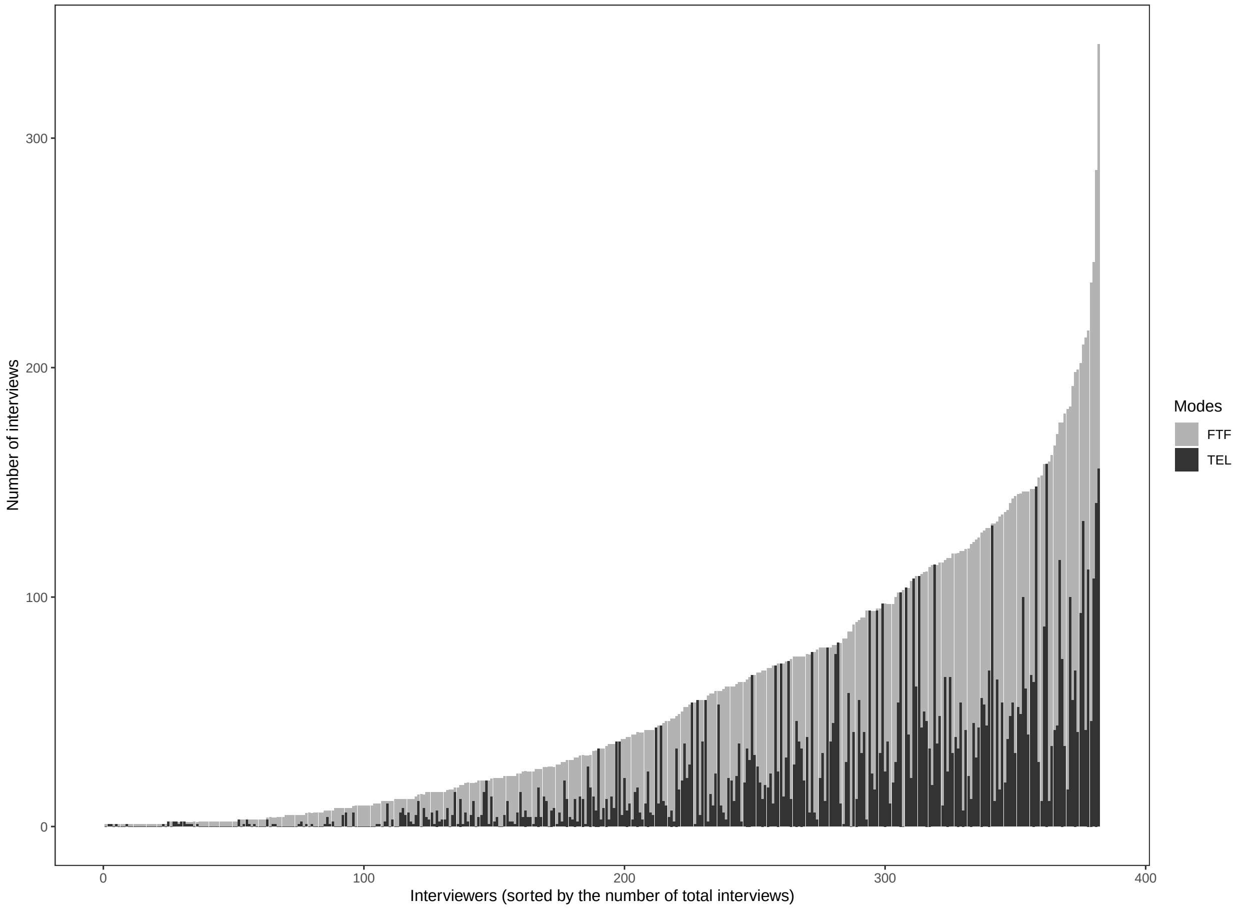


Figure 3.2: Interviewer Workloads Per Mode in the Health and Retirement Study

Table 3.4: Distribution of Sample Characteristics in the Health and Retirement Study across Interviewers by Modes

Respondent Characteristics	Mean (FTF)	Mean (TEL)	Between-interviewer SD (FTF)	Between-interviewer SD (TEL)	Average Within-interviewer SD (FTF)	Average Within-interviewer SD (TEL)
Age: less than 60	0.449	0.305	0.359	0.315	0.294	0.363
Age: 60-69	0.188	0.343	0.156	0.239	0.260	0.398
Age: 70-79	0.181	0.287	0.150	0.255	0.240	0.342
Age: 80+	0.182	0.066	0.185	0.151	0.232	0.163
Currently Working	0.368	0.329	0.280	0.265	0.426	0.427
Male	0.414	0.415	0.177	0.198	0.503	0.503
Spanish-speaking Hispanic	0.091	0.085	0.194	0.212	0.087	0.072
English-speaking Hispanic	0.077	0.074	0.158	0.146	0.198	0.189
Black	0.219	0.200	0.274	0.246	0.315	0.344
White	0.613	0.641	0.315	0.300	0.376	0.397
Coupled	0.601	0.632	0.260	0.275	0.431	0.428
Education:less than 12 years	0.203	0.188	0.190	0.225	0.348	0.324
Education:12 years	0.303	0.290	0.188	0.218	0.420	0.416
Education:13-15 years	0.259	0.268	0.196	0.200	0.406	0.421
Education:16 years +	0.249	0.262	0.214	0.220	0.391	0.374

Next, we present the descriptive statistics of the HRS, including mode-specific sample means, between-interviewer standard deviation (SD), and average within-interviewer SD in Table 3.5. First, for the CESD scale, the prevalence rates are generally higher in face-to-face (FTF) interviews than in telephone (TEL) interviews, suggesting that FTF may be associated with more honest reporting. Second, interviewers report FTF respondents as more attentive, understanding questions better, less cooperative, having less difficulty remembering but more difficulty hearing things, and interviewed with higher quality. Third, the magnitude of the between-interviewer SD appears larger in the interviewer observation and physical activity items compared to the CESD items, indicating potentially different levels of interviewer effects in different outcomes.

Table 3.5: Distribution of Outcome variables in the Health and Retirement Study across Interviewers by Modes

Questions	Mean (FTF)	Mean (TEL)	Between-interviewer SD (FTF)	Between-interviewer SD (TEL)	Average Within-interviewer SD (FTF)	Average Within-interviewer SD (TEL)
CESD questions						
1. you felt depressed.	0.155	0.117	0.178	0.151	0.303	0.264
2. you felt that everything you did was an effort.	0.329	0.251	0.227	0.212	0.428	0.388
3. your sleep was restless.	0.347	0.301	0.217	0.220	0.443	0.424
4. you were happy (REVERSED CODE).	0.175	0.143	0.193	0.174	0.324	0.294
5. you felt lonely.	0.203	0.153	0.186	0.159	0.364	0.319
6. you enjoyed life (REVERSED CODE).	0.112	0.077	0.153	0.116	0.258	0.212
7. you felt sad.	0.243	0.192	0.216	0.186	0.378	0.351
8. you could not get going.	0.210	0.171	0.179	0.171	0.373	0.332
9. Depressed ($\geq$ 4 symptoms)	0.182	0.119	0.188	0.141	0.335	0.277
Interviewer Observations						
10. attentive to the questions	0.799	0.793	0.209	0.235	0.336	0.319
11. understanding of the questions	0.459	0.469	0.270	0.304	0.440	0.403
12. cooperation	0.718	0.663	0.258	0.281	0.376	0.396
13. difficulty remembering things	0.540	0.591	0.289	0.313	0.419	0.385
14. difficulty hearing you	0.807	0.741	0.198	0.252	0.322	0.361
15. quality of this interview	0.592	0.624	0.322	0.327	0.380	0.363
Physical activity						
16. vigorous sports or activities	0.353	0.347	0.219	0.221	0.441	0.451
17. moderately energetic sports or activities	0.679	0.657	0.213	0.232	0.423	0.439
18. mildly energetic sports or activities	0.809	0.779	0.167	0.201	0.357	0.374

3.3.3.2 Mode Effects in Means and Interviewer Variances

Last, we discuss the modeling results presented in Table 3.6 using Bayesian estimation. Positive mode effects in means are found in four of the nine depression items. These items are “felt depressed,” “everything was an effort,” “sleep was restless,” and “overall indicator for depression.” For example, for a female under 60 years old, who is an English-speaking Hispanic, not in a relationship, not currently employed, and with less than a high school education, participating in a FTF interview increases the probability of being classified as depressive by 8.01%, compared to a TEL interview. Similarly, we compute 8.01% using $\phi(\beta_0 + \beta_1 + \sum_s^S \gamma_s x_{si})\beta_1$, where ϕ is the pdf of a standard normal distribution and S is the number of covariates (x). Since depressive symptoms constitute sensitive information, and admitting to them might cause embarrassment for respondents, we believe that a higher level of reported depressive symptoms is closer to the truth. For the interviewer observation items, positive mode effects in means are present in three out of six items. In the FTF mode, interviewers rated respondents as more cooperative, with better hearing and overall quality of the interview, compared to the TEL mode (Table 3.6). Lastly, in the physical activity items, respondents tend to report engaging in mildly energetic sports more often when responding via FTF, compared to TEL.

We observe smaller interviewer variances in the substantive responses in HRS (Table 3.6) compared to the ABS. For depression items, the interviewer variances in FTF and TEL range from 0.002 to 0.032, corresponding to ICCs between 0.002 and 0.031. In the physical activity items, the interviewer variances range from 0.007 (ICC: 0.007) to 0.031 (ICC: 0.030). When comparing the magnitude of interviewer variances across variables, we notice larger interviewer variances for the interviewer observation items (ranging from 0.271 [ICC: 0.273] to 0.881 [ICC: 0.788]).

In terms of mode effects in interviewer variances, we find significant differences for three out of the eighteen questions examined in the HRS study, specifically one in the depression scale and two in the interviewer observation questions (Table 3.6). When asking participants if they felt sad, the results reveal that FTF is associated with larger interviewer variances. Additionally, interviewer variance in the FTF mode is marginally larger than in the TEL mode for the item “everything was an effort”. Generally, for the depression items, the interviewer variances in the FTF mode are larger than those in the TEL mode for seven out of nine items, though not always significantly. This outcome aligns with the Arab Barometer findings and may be due to interviewers approaching sensitive items differently in FTF compared to the TEL mode.

In assessing whether respondents have any difficulty remembering and hearing things, the results suggest that TEL interviewer variances are larger than FTF interviewer variances. This finding may be attributed to interviewers having fewer cues to evaluate interview quality in TEL, as opposed to FTF, where interviewers can rely on respondents’ facial expressions or body language to infer participants’ ability to hear questions. This might lead to responses being primarily determined by interviewers’ subjective judgments and thus causing larger variances. Regarding the physical activity items, there is no evidence to reject the null hypothesis that interviewer variances are equal between modes.

It is not surprising to find higher correlations ($\rho > 0.8$) between the random interviewer effects across modes for interviewer observation variables, which interviewers directly answer. In contrast, for the other two scales (CESD and physical activity scales), the effects of interviewers on responses are mediated through respondents, resulting in a smaller and less stable correlation between the FTF and TEL modes.

Although we focus on reporting the Bayesian results, we provide the inferences from both the likelihood and the Bayesian procedures in Appendix F. We note that, in general, the estimates from the two procedures are similar, except when estimating the correlation (ρ). The correlations are associated with wide intervals in the CESD scales and the physical activity items. Moreover, the point estimates of the correlation are sometimes quite different between the two procedures, especially for the two types of items mentioned above. On two items, “happy” and “felt sad”, the correlation cannot be estimated using the likelihood approach. This might be due to the small interviewer variances in the scale, making the estimation of the covariance numerically challenging and thus unstable. Additionally, this might be attributed to the unbalanced interviewer burden between modes. Approximately 30% of interviewers only conduct interviews in one mode, and 51% of interviewers carry out fewer than five interviews in either FTF or TEL. This imbalance may result in insufficient information for estimating ρ.

To address the numerical challenges and evaluate whether the estimation of other parameters (e.g., σ_f^2, σ_t^2, and α) is sensitive to ρ, we set ρ to 0 and to the posterior mean obtained with the Bayesian procedure, and rerun Model 3.4 for the CESD items. We find that the estimates of the interviewer variances remain nearly unchanged when specifying ρ to different values or estimating ρ (see details in Appendix G). Thus, we conclude that there is little sensitivity in the inferences provided by the likelihood estimation to ρ.

Table 3.6: Interviewer Variances Per Mode for Selected Items in Health and Retirement Study Adjusting for Covariates Using Bayesian Estimation

Questions	σ_f^2	σ_t^2	$\rho_{f,int}$	$\rho_{t,int}$	α	β_1	ρ
				CESD questions			
felt depressed	0.011	0.013	0.011	0.013	0.044	**0.056**	0.07
	[0, 0.022]	[0, 0.03]	[0.000, 0.022]	[0.000, 0.029]	[-1.148, 1.533]	[0.005, 0.114]	[-0.551, 0.874]
everything was an effort	0.025	0.007	0.024	0.007	0.746	**0.118**	-0.128
	[0.013, 0.037]	[0.001, 0.016]	[0.013, 0.036]	[0.001, 0.016]	[-0.002, 1.496]	[0.071, 0.175]	[-0.56, 0.254]
restless sleep	0.002	0.005	0.002	0.005	-0.486	**0.053**	0.337
	[0, 0.007]	[0, 0.012]	[0.000, 0.007]	[0.000, 0.012]	[-1.89, 0.925]	[0.011, 0.095]	[-0.162, 0.849]
happy	0.011	0.011	0.011	0.011	0.128	0.032	**-0.518**
	[0.003, 0.021]	[0, 0.022]	[0.003, 0.021]	[0.000, 0.022]	[-0.889, 1.333]	[-0.024, 0.083]	[-0.989, -0.006]
lonely	0.006	0.006	0.006	0.006	0.178	0.048	0.055
	[0, 0.014]	[0, 0.016]	[0.000, 0.014]	[0.000, 0.016]	[-1.455, 1.846]	[-0.005, 0.099]	[-0.108, 0.218]
enjoyed life	0.01	0.007	0.010	0.007	0.551	0.061	**0.56**
	[0.001, 0.021]	[0, 0.025]	[0.001, 0.021]	[0.000, 0.024]	[-1.096, 2.148]	[-0.005, 0.134]	[0.223, 0.921]
felt sad	0.032	0.003	0.031	0.003	**1.694**	0.046	**0.296**
	[0.018, 0.048]	[0, 0.009]	[0.018, 0.046]	[0.000, 0.009]	[0.463, 3.775]	[-0.01, 0.097]	[0.037, 0.577]
could not get going	0.02	0.02	0.020	0.020	-0.051	0.051	0.274
	[0.007, 0.029]	[0.006, 0.035]	[0.007, 0.028]	[0.006, 0.034]	[-0.732, 0.515]	[-0.01, 0.109]	[-0.346, 0.797]
overall indicator	0.016	0.012	0.016	0.012	0.093	**0.15**	0.244
	[0.002, 0.024]	[0.001, 0.027]	[0.002, 0.023]	[0.001, 0.026]	[-0.901, 1.075]	[0.102, 0.207]	[-0.18, 0.573]
				Interviewer Observations			
attentive	0.298	0.351	0.230	0.260	-0.081	0.018	**0.878**
	[0.233, 0.356]	[0.262, 0.431]	[0.189, 0.263]	[0.208, 0.301]	[-0.197, 0.038]	[-0.049, 0.088]	[0.803, 0.955]
understanding	0.413	0.465	0.292	0.317	-0.058	0 [-0.064,	**0.91**
	[0.341, 0.493]	[0.366, 0.56]	[0.254, 0.330]	[0.268, 0.359]	[-0.149, 0.043]	0.061]	[0.861, 0.958]
cooperation	0.459	0.41	0.315	0.291	0.057	**0.178**	**0.931**
	[0.378, 0.556]	[0.321, 0.51]	[0.274, 0.357]	[0.243, 0.338]	[-0.039, 0.138]	[0.108, 0.236]	[0.881, 0.971]

remembering	0.483	0.605	0.326	0.377	**-0.112**	-0.062	**0.931**
	[0.392, 0.574]	[0.489, 0.721]	[0.282, 0.365]	[0.328, 0.419]	[-0.205, -0.028]]	[-0.124, 0.002]	[0.885, 0.972]
hearing	0.271	0.375	0.213	0.273	**-0.161**	**0.151**	**0.87**
	[0.212, 0.335]	[0.274, 0.462]	[0.175, 0.251]	[0.215, 0.316]	[-0.284, -0.037]	[0.084, 0.229]	[0.795, 0.947]]
Overall quality	0.881	0.788	0.468	0.441	0.057	**0.086**	**0.94**
	[0.749, 1.04]	[0.641, 0.949]	[0.428, 0.510]	[0.391, 0.487]	[-0.032, 0.14]	[0.014, 0.158]	[0.913, 0.983]
				Physical activity			
vigorous sports	0.017	0.007	0.017	0.007	0.523	-0.037	0.36
	[0.007, 0.026]	[0, 0.015]	[0.007, 0.025]	[0.000, 0.015]	[-0.209, 1.45]	[-0.081, 0.014]	[-0.446, 0.827]
moderate sport	0.015	0.019	0.015	0.019	-0.086	0.031	0.233
	[0.006, 0.024]	[0.004, 0.033]	[0.006, 0.023]	[0.004, 0.032]	[-0.655, 0.464]	[-0.019, 0.078]	[-0.351, 0.698]
mild sport	0.02	0.031	0.020	0.030	-0.355	**0.134**	0.144
	[0.002, 0.03]	[0.014, 0.052]	[0.002, 0.029]	[0.014, 0.049]	[-1.097, 0.324]	[0.073, 0.184]	[-0.264, 0.962]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,int}$ is the interviewer intraclass correlation associated with the FTF mode. $\rho_{t,int}$ is the interviewer intraclass correlation associated with the TEL mode. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects.

3.4 Simulation Study

To understand the repeated sampling properties of our proposed method, including the power to detect mode effects in the typically modest interviewer sample sizes available, we conducted simulation studies using the ABS and the HRS setup.

3.4.1 Arab Barometer Study

This simulation study is designed such that the number of respondents ($n = 2521$) and interviewers (13 in the TEL mode and 31 in the FTF mode) are the same as the ABS, as well as how respondents are matched to interviewers. We consider four scenarios, 1) no difference scenario where the FTF interviewer variance is equal to the TEL interviewer variance ($\sigma_f^2 = \sigma_t^2 = 0.14, \alpha_0 = -0.98$ and $\alpha = 0$), 2) small differences where $\sigma_f^2 = 0.20, \sigma_t^2 = 0.14, \alpha_0 = -0.98$ and $\alpha = 0.18$, 3) medium differences where $\sigma_f^2 = 0.24, \sigma_t^2 = 0.14, \alpha_0 = -0.98$ and $\alpha = 0.27$, and 4) large differences where $\sigma_f^2 = 0.50, \sigma_t^2 = 0.14, \alpha_0 = -0.98$ and $\alpha = 0.64$. We consider the true data generation model as follows:

$$\eta_i = \Phi(\beta_0 + \beta_1 M_{ij} + b_{j(m)}),$$

$$b_{j(m)} \sim N(0, \sigma_m^2),$$

$$y_i \sim Bernoulli(\eta_i),$$

where i indexes respondents, j indexes interviewers, m indicates modes (f or t), $\Phi()$ is the cumulative distribution function of the standard normal distribution, and M is a $n \times 1$ vector of the mode that each participant used to participate in the survey.

We fit the same analytical model (3.2) to the simulated data, implemented separately using Proc Nlmixed and Proc MCMC in the SAS programming language. The simulation is repeated $K = 200$ times, where for each iteration, the point estimates, standard errors, and 95% confidence intervals or credible intervals of β_1, σ_f^2, σ_t^2, and α are computed and saved. Based on these statistics, we report the bias, coverage rate, SE ratio, and power in each scenario for the parameters.

$$
\begin{aligned}
&Bias(\hat{\delta}) = \frac{1}{K}\sum_{k}^{K}\hat{\delta_k} - \delta,\\
&Coverage\ Rate(\hat{\delta}) = \frac{1}{K}\sum_{k}^{K} I(\hat{\delta}_{k,lw} < \delta \,\&\, \hat{\delta}_{k,up} > \delta),\\
&SE\ Ratio(\hat{\delta}) = \frac{1}{K}\sum_{k}^{K}\sqrt{v\hat{a}r(\hat{\delta_k})}/\sqrt{\frac{1}{K-1}\sum_{k}^{K}(\hat{\delta_k} - \bar{\hat{\delta_k}})^2},\\
&Power(\hat{\delta}) = 1 - \frac{1}{K}\sum_{k}^{K} I(\hat{\delta}_{k,lw} < 0 \,\&\, \hat{\delta}_{k,up} > 0)\ when\ \delta \neq 0,
\end{aligned}
$$

where δ refers to the parameters that we are interested in estimating (i.e., σ_f^2, σ_t^2, β_1, and α), $\hat{\delta_k}$ is the estimated point estimate of δ obtained in iteration K, $\hat{\delta}_{k,lw}$ and $\hat{\delta}_{k,up}$ is the lower bound and upper bound of the estimated parameter.

Table 3.7 displays the simulation results using the Arab Barometer setup. When $\alpha = 0$, the power reported in Table 3.7 represents the Type 1 error rate. We observe that the power to reject the null hypothesis stating that interviewer variances are equal ($\alpha = 0$) is limited across the scenarios. However, as the differences grow larger (0.18-0.64), the power does increase from 0.075 to 0.520 in the Bayesian procedure and from 0.110 to 0.633 in the frequentist approach. There are some differences in the power provided by the likelihood and Bayesian approaches. This is because the likelihood procedures do not offer nominal coverage rates in Scenarios 1 to 3; as a result, the power obtained from the likelihood and Bayesian procedures is based on different significance levels. The small power of α is primarily due to the very limited number of interviewers in both FTF and TEL modes. Conversely, the power of rejecting the null hypothesis that there are no mode effects in means (β_1) when the alternative hypothesis is true is considerably higher (around 0.90). However, as α becomes larger and the interviewer variances increase simultaneously, we observe a declining power of β_1, due to the decline in effective sample size from the increased ICC.

Table 3.7: Simulation study using the Arab Barometer Setup

Parameters	Likelihood results				Bayesian results			
	Bias	Coverage rate	SE ratio	Power	Bias	Coverage rate	SE ratio	Power
Scenario 1: No differences								
$\sigma_f^2 = 0.14$	-0.002	0.950	1.000	N/A	0.017	0.940	1.059	N/A
$\sigma_t^2 = 0.14$	-0.001	0.955	1.014	N/A	0.049	0.975	1.346	N/A
$\beta_1 = 0.5$	-0.003	0.965	1.023	0.935	0.006	0.955	1.121	0.930
$\alpha = 0$	0.028	0.930	0.888	0.070	-0.033	0.985	1.107	0.015
Scenario 2: Small differences								
$\sigma_f^2 = 0.20$	-0.012	0.960	0.948	N/A	0.028	0.975	1.105	N/A
$\sigma_t^2 = 0.14$	-0.007	0.935	0.974	N/A	0.059	0.955	1.161	N/A
$\beta_1 = 0.5$	-0.002	0.940	0.926	0.950	-0.001	0.950	1.078	0.900
$\alpha = 0.18$	0.042	0.920	0.928	0.110	-0.020	0.950	0.955	0.075
Scenario 3: Medium differences								
$\sigma_f^2 = 0.24$	-0.002	0.920	0.947	N/A	0.039	0.920	0.980	N/A
$\sigma_t^2 = 0.14$	-0.013	0.955	1.009	N/A	0.061	0.980	1.311	N/A
$\beta_1 = 0.5$	0.004	0.935	0.940	0.920	-0.010	0.960	1.184	0.860
$\alpha = 0.27$	0.079	0.905	0.922	0.230	-0.042	0.960	1.075	0.085
Scenario 4: Large differences								
$\sigma_f^2 = 0.50$	-0.007	0.970	1.058	N/A	0.078	0.950	1.093	N/A
$\sigma_t^2 = 0.14$	-0.009	0.960	1.055	N/A	0.054	0.935	1.231	N/A
$\beta_1 = 0.5$	0.022	0.935	0.965	0.824	-0.016	0.980	1.097	0.690
$\alpha = 0.64$	0.079	0.945	0.906	0.633	0.012	0.955	0.882	0.520

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances.

3.4.2 Health and Retirement Study

In the simulation study using the HRS setup, we consider the following data generation model using the same notations as in the ABS simulation study. We use b_{jf} to represent random interviewer effects in the FTF mode and b_{jt} to represent random interviewer effects in the TEL mode:

$$\eta_i = \Phi(\beta_0 + \beta_1 M_{ij} + b_{jf} M_{ij} + b_{jt}(1 - M_{ij})),$$

$$\begin{pmatrix} b_{jf} \\ b_{jt} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_f^2 & \rho\sigma_f\sigma_t \\ \rho\sigma_f\sigma_t & \sigma_t^2 \end{pmatrix}\right).$$

$$y_i \sim Bernoulli(\eta_i),$$

We consider four scenarios: 1) $\sigma_f^2 = \sigma_t^2 = 0.03, \alpha_0 = -1.75$, and $\alpha = 0$; 2) $\sigma_f^2 = 0.05, \sigma_t^2 = 0.03, \alpha_0 = -1.75$, and $\alpha = 0.26$; 3) $\sigma_f^2 = 0.06, \sigma_t^2 = 0.03, \alpha_0 = -1.75$, and $\alpha = 0.35$; 4) $\sigma_f^2 = 0.09, \sigma_t^2 = 0.03, \alpha_0 = -1.75$, and $\alpha = 0.55$. Across all scenarios, $\beta_1 = 0.5$ and $\rho = 0.5$. We report bias, coverage rate, SE ratio, power for these parameters and the logarithmic differences of interviewer variances between FTF and TEL (α) in Table 3.4.2.

Table 3.4.2 illustrates that as α rises from 0 to 0.55, the power correspondingly increases from 0.035 to 0.990 using the Bayesian procedure, and from 0.035 to 0.935 employing the likelihood approach. The findings suggest that When α is large enough, we can achieve a reasonably high power using the HRS setup. Upon comparing Table 3.7 and Table 3.4.2, we observe that the power to reject the null hypothesis asserting equal interviewer variances, when the alternative hypothesis holds true, surpasses that in the ABS simulation. This outcome aligns with expectations, given the larger number of interviewers involved in the HRS. In addition, we note that the likelihood approach may not always reach the 95% nominal coverage rates (in Scenarios 3 and 4), thus the power computed using the likelihood and the Bayesian procedures are based on different significance levels.

Table 3.8: Simulation study using the HRS Setup

Parameters	Likelihood results				Bayesian results			
	Bias	Coverage rate	SE ratio	Power	Bias	Coverage rate	SE ratio	Power
Scenario 1: No differences								
$\sigma_f^2 =$ 0.03	-0.000	0.980	1.085	N/A	0.003	0.965	1.704	N/A
$\sigma_t^2 =$ 0.03	-0.001	0.975	1.049	N/A	0.002	0.935	1.469	N/A
$\beta_1 = 0.5$	-0.002	0.940	1.029	1.000	-0.000	0.960	1.128	1.000
$\rho = 0.5$	0.012	0.965	1.009	0.470	-0.020	0.925	1.061	0.690
$\alpha = 0$	0.022	0.965	1.019	0.035	0.047	0.965	0.928	0.035
Scenario 2: Small differences								
$\sigma_f^2 =$ 0.05	0.000	0.940	0.999	N/A	0.001	0.955	1.507	N/A
$\sigma_t^2 =$ 0.03	-0.000	0.975	1.125	N/A	0.002	0.945	1.249	N/A
$\beta_1 = 0.5$	0.003	0.960	0.996	1.000	0.001	0.950	0.983	1.000
$\rho = 0.5$	0.020	0.980	1.084	0.695	-0.021	0.925	1.032	0.755
$\alpha = 0.26$	0.018	0.940	0.978	0.270	0.008	0.940	0.934	0.295
Scenario 3: Medium differences								
$\sigma_f^2 =$ 0.06	-0.001	0.945	0.999	N/A	0.001	0.950	1.268	N/A
$\sigma_t^2 =$ 0.03	-0.001	0.975	1.045	N/A	0.002	0.940	1.103	N/A
$\beta_1 = 0.5$	-0.001	0.920	0.993	1.000	0.001	0.965	1.007	1.000
$\rho = 0.5$	0.011	0.970	1.030	0.665	-0.009	0.930	1.014	0.815
$\alpha = 0.35$	0.024	0.910	0.919	0.510	0.008	0.945	0.949	0.530
Scenario 4: Large differences								
$\sigma_f^2 =$ 0.09	0.000	0.930	0.983	N/A	0.002	0.950	1.201	N/A
$\sigma_t^2 =$ 0.03	-0.001	0.955	1.054	N/A	-0.001	0.915	1.089	N/A
$\beta_1 = 0.5$	0.004	0.950	1.009	1.000	-0.002	0.955	1.031	1.000

$\rho = 0.5$	0.009	0.985	1.085	0.750	0.004	0.970	1.121	0.860
$\alpha = 0.55$	0.029	0.915	0.977	0.935	0.070	0.950	0.955	0.990

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects.

3.5 Discussion

This paper explores the presence of mode effects in interviewer variances across multiple items in two national surveys. In the ABS, we find statistical evidence for differing interviewer effects between the FTF and TEL modes in one (marginally) out of six sensitive items and one out of three item missing indicators. Besides, for sensitive items and missing indicators in the ABS, interviewer variances from the FTF mode are generally larger than those from the TEL mode. Utilizing the 2016 HRS data, we observe significant mode effects in interviewer variances on two depression items (one marginally) and two interviewer observation item. For sensitive depression items, a similar pattern emerges, with larger interviewer variances in FTF than in TEL. These findings indicate that sensitive questions and item missing items are crucial challenges when stabilizing interviewer variances between modes. In addition, the magnitude of interviewer variances are much larger on interviewer observation items than substantive responses. Evidence suggests that TEL interviewer variances are larger than FTF interviewer variances on these items. This could be because these questions involve more subjective evaluations and may offer greater opportunities to reduce interviewer variances by standardizing interviewer protocols for such items, especially in the TEL mode.

Simulation studies suggest that it is possible to achieve reasonable power with either the ABS or HRS setup if there are substantial mode effects in interviewer variances. However, with small mode effects, the power is limited, especially in the ABS setup. The observation of significant mode effects in interviewer variances in both the ABS and HRS data highlights the importance of considering the role of modes on interviewer effects, particularly when addressing sensitive topics and item nonresponse. Given the typically limited number of interviewers employed in most surveys, a null finding may not necessarily indicate equal interviewer variance. However, it is still useful for survey agencies to consider such investigation as a positive finding is valid and should capture the attention of researchers.

The literature has extensively documented whether modes affect measurement errors at the respondent level [10, 11]. However, few studies have investigated whether and how modes influence interviewer-related measurement errors, particularly following the widespread adoption of mixed-mode designs. This paper addresses this gap by analyzing two national surveys with distinct mixed-mode design features, such as the number of interviewers and whether the interviewers are nested under modes. When interviewers are nested under modes, it is hard to determine if the observed differences are attributable to modes or interviewers. The current modeling approach presumes that all systematic differences between responses collected in TEL and FTF are a consequence of modes, not interviewers. If survey organizations possess information on interviewer characteristics, they can evaluate this assumption by comparing the characteristics of interviewers between modes. Such an analysis would help disentangle the effects of modes from those of interviewers, providing valuable insights for survey data quality.

For designs that allow interviewers to collect data in both modes, the models presented in this paper enable the estimation of individual interviewer effects in each mode. This is useful for detecting interviewers with a substantial impact on responses in one or both modes. Utilizing these estimated interviewer effects, we can further identify if specific interviewers consistently exhibit large effects across variables, potentially signaling the need for intervention by interviewer supervisors. If particular variables are associated with significant interviewer variances in a certain mode, this may warrant improved interviewer training for those items. For instance, based on this study's findings, a more standardized interview protocol could be considered for sensitive items and when respondents answer "don't know" to questions in FTF mode. As such, we recommend that survey agencies incorporate these analyses into their routine data quality assessments. Future research could investigate whether interviewer characteristics can explain the differential interviewer effects observed across modes, potentially shedding light on the underlying mechanisms at play.

When determining which mode to use for generating population estimates in mixed-mode studies, it is desirable to have smaller bias and lower interviewer variances, which might result in smaller mean squared error. However, in reality, the mode with smaller bias and lower interviewer variance may not always be the same, as shown in this paper. For instance, FTF interviews may be linked with less bias but larger interviewer variance. How to balance the trade-offs between bias and variance in a formal method will be a topic for future research. This study showcases two survey examples to evaluate mode effects both in means and interviewer variances. If such analyses are routinely adopted by researchers who design and implement mixed-mode studies, more evidence can be accumulated about whether and how interviewers could have performed differently in different modes of data collection. This can become the basis for developing future mixed-mode protocols. When reporting the results of the analysis, we recommend that survey agencies explain how their interviewers are assigned to or self-select different modes and clarify whether the resultant mode effects in interviewer variances are consistent with their expectations.

In this paper, we observe some discrepancies between the results obtained from the maximum likelihood procedure and the Bayesian procedure implemented in the SAS programming language. When interviewer variances are small, fitting the analytical model with correlated interviewer random effects across modes using the likelihood approach can be challenging. In this situation, the Bayesian approach can be particularly useful, as employing proper and informative priors helps ensure that we draw inferences from proper posterior distributions.

This study has two main limitations. First, like other similar studies [51, 40], it faces the issue of limited statistical power, as demonstrated in the simulation study. Second, both surveys lack randomization in the interviewer assignment scheme. Ideally, when estimating interviewer variances, interpenetrated designs should be used to ensure that the variability is solely due to the interviewer measurement process, rather than differences among respondents. As a workaround for the absence of randomization, we included respondent characteristics in the analysis model. However, interviewer variances might still be overestimated due to unobserved covariates not accounted for in the models.

3.6 Appendix

Appendix A: Outcome Variables Used in the Arab Barometer Study

Table 3.9: Outcome Variables Used in the Arab Barometer Study

Questions	Original response categories	Collapsed response categories
Sensitive political questions		
Freedom of the media to criticize the things government does?	1. Guaranteed to a great extent 2. Guaranteed to a medium extent 3. Guaranteed to a limited extent 4. Not guaranteed at all	1. Guaranteed to a great or medium extent 0. Guaranteed to a limited extent or not guaranteed at all
How much trust do you have in government?	1. A great deal of trust 2. Quite a lot of trust 3. Not a lot of trust 4. No trust at all	1. A great deal of or quite a lot of trust 0. Not a lot of trust or no trust at all
How much trust do you have in courts and the legal system?	1. A great deal of trust 2. Quite a lot of trust 3. Not a lot of trust 4. No trust at all	1. A great deal of or quite a lot of trust 0. Not a lot of trust or no trust at all
How satisfied are you with the healthcare system in our country?	1. Completely satisfied 2. Satisfied 3. Dissatisfied 4. Completely dissatisfied	1. Completely satisfied or satisfied 0. Dissatisfied or completely dissatisfied
How would you evaluate the current government's performance on keeping prices down?	1. Very good 2. Good 3. Bad 4. Very bad	1. Very good or good 0. Bad or very bad

How would you evaluate the current government's performance on responding to the COVID-19 outbreak?	1. Very good 2. Good 3. Bad 4. Very bad	1. Very good or good 0. Bad or very bad
Less Sensitive International Questions		
Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of the United States.	1. Very favorable 2. Somewhat favorable 3. Somewhat unfavorable 4. Very unfavorable	1. Very or somewhat favorable 0. Somewhat or very unfavorable
Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of Germany.	1. Very favorable 2. Somewhat favorable 3. Somewhat unfavorable 4. Very unfavorable	1. Very or somewhat favorable 0. Somewhat or very unfavorable
Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of China.	1. Very favorable 2. Somewhat favorable 3. Somewhat unfavorable 4. Very unfavorable	1. Very or somewhat favorable 0. Somewhat or very unfavorable
Whether missing on international questions (constructed)		
Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of the United States.	Don't know or refused to answer (Interviewer: do not read)	1. Don't know or refused to answer 0. Answered
Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of Germany.	Don't know or refused to answer (Interviewer: do not read)	1. Don't know or refused to answer 0. Answered

Please tell me if you have a very favorable, somewhat favorable, somewhat unfavorable, or very unfavorable opinion of China.	Don't know or refused to answer (Interviewer: do not read)	1. Don't know or refused to answer 0. Answered

Appendix B: Derivations of the Variance of α Using Delta Method

$$
\begin{aligned}
& var(\alpha) = var(log(\sigma_f) - log(\sigma_t)) \\
& = var(log(\sigma_f)) + var(log(\sigma_t)) - 2cov(log(\sigma_f), log(\sigma_t)) \\
& = \frac{1}{4}var(2log(\sigma_f)) + \frac{1}{4}var(2log(\sigma_t)) - 2cov(log(\sigma_f), log(\sigma_t)) \\
& = \frac{1}{4}var(log(\sigma_f^2)) + \frac{1}{4}var(log(\sigma_t^2)) - \frac{1}{4} \times 2cov(log(\sigma_f^2), log(\sigma_t^2)) \\
& = \frac{1}{4}var(log(\sigma_f^2)) + \frac{1}{4}var(log(\sigma_t^2)) - \frac{1}{2}cov(log(\sigma_f^2), log(\sigma_t^2))
\end{aligned}
$$

We express $var(\alpha)$ as a function of $var(log(\sigma_f^2))$, $var(log(\sigma_t^2))$, and $cov(log(\sigma_f^2), log(\sigma_t^2))$, as we apply a log transformation to σ_t^2 and σ_f^2 to stabilize their variances. The covariance between $log(\sigma_f^2)$ and $log(\sigma_t^2)$ can be assumed to be 0 when the random interviewer effects of FTF and TEL are not correlated, as is the case in the ABS. In contrast, in the HRS, when the random interviewer effects are correlated across modes, the covariance between the two estimates should be considered when calculating $var(\alpha)$.

Appendix C: Full Results on Interviewer Variances in the Arab Barometer Study

Table 3.10: Interviewer Variances Per Mode for Selected Items in the Arab Barometer Study Adjusting for Covariates

Questions	Likelihood Results				Bayesian Results			
	σ_f^2	σ_t^2	α	β_1	σ_f^2	σ_t^2	α	β_1
				Sensitive political questions				
1. Freedom								
of the	0.222	0.092	0.443	**-0.529**	0.252	0.135	0.355	**-0.526**
media	(0.071)	(0.044)	(0.291)	**(0.132)**	(0.082)	(0.076)	(0.298)	**(0.146)**
	[0.116,	[0.034,	[-0.127,	**[-0.794,**	[0.122,	[0.036,	[-0.223,	**[-0.795,**
	0.425]	0.244]	1.013]	**-0.263]**	0.428]	0.284]	0.898]	**-0.222]**
2. trust								
in	0.159	0.085	0.311	**-0.5**	0.188	0.127	0.239	**-0.504**
government	(0.054)	(0.041)	(0.293)	**(0.121)**	(0.066)	(0.079)	(0.311)	**(0.134)**
	[0.081,	[0.032,	[-0.263,	**[-0.745,**	[0.083,	[0.029,	[-0.382,	**[-0.768,**
	0.313]	0.225]	0.886]	**-0.255]**	0.322]	0.275]	0.838]	**-0.238]**
3. trust	0.093	0.145	-0.225	**-0.553**	0.113	0.214	-0.291	**-0.555**
in courts	(0.036)	(0.069)	(0.306)	**(0.133)**	(0.046)	(0.119)	(0.327)	**(0.154)**
	[0.042,	[0.056,	[-0.825,	**[-0.822,**	[0.038,	[0.05,	[-0.94,	**[-0.881,**
	0.202]	0.378]	0.374]	**-0.285]**	0.201]	0.445]	0.318]	**-0.273]**
4. satisfied								
with	0.118	0.018	**0.955**	**-0.285**	0.143	0.03	0.906	**-0.278**
healthcare	(0.042)	(0.013)	**(0.411)**	**(0.089)**	(0.056)	(0.024)	(0.467)	**(0.1)**
	[0.057,	[0.004,	**[0.149,**	**[-0.465,**	[0.051,	[0,	[-0.054,	**[-0.475,**
	0.244]	0.078]	**1.761]**	**-0.105]**	0.251]	0.075]	1.758]	**-0.085]**

5. performance on inflation	0.327 (0.112) [0.164, 0.653]	0.139 (0.064) [0.055, 0.351]	0.428 (0.286) [-0.134, 0.989]	**-0.515** **(0.163)** **[-0.843,** **-0.186]**	0.393 (0.146) [0.153, 0.672]	0.204 (0.112) [0.051, 0.435]	0.361 (0.305) [-0.275, 0.927]	**-0.523** **(0.182)** **[-0.861,** **-0.153]**
6. performance during COVID-19	0.172 (0.057) [0.088, 0.336]	0.165 (0.074) [0.067, 0.406]	0.019 (0.278) [-0.526, 0.563]	**-0.495** **(0.146)** **[-0.79,** **-0.201]**	0.202 (0.069) [0.084, 0.34]	0.224 (0.112) [0.07, 0.443]	-0.026 (0.285) [-0.602, 0.508]	**-0.51** **(0.166)** **[-0.841,** **-0.205]**
International Questions								
7. favor US	0.163 (0.058) [0.08, 0.333]	0.262 (0.115) [0.108, 0.634]	-0.238 (0.282) [-0.79, 0.314]	-0.05 (0.173) [-0.399, 0.299]	0.198 (0.073) [0.074, 0.34]	0.362 (0.184) [0.104, 0.719]	-0.278 (0.293) [-0.841, 0.282]	-0.057 (0.193) [-0.45, 0.318]
8. favor Germany	0.245 (0.084) [0.122, 0.49]	0.237 (0.107) [0.096, 0.588]	0.015 (0.283) [-0.539, 0.57]	-0.142 (0.178) [-0.5, 0.216]	0.292 (0.11) [0.12, 0.514]	0.33 (0.164) [0.092, 0.663]	-0.037 (0.314) [-0.603, 0.548]	-0.147 (0.201) [-0.551, 0.236]
9. favor China	0.174 (0.064) [0.083, 0.365]	0.27 (0.118) [0.112, 0.654]	-0.221 (0.286) [-0.781, 0.339]	-0.142 (0.177) [-0.498, 0.214]	0.205 (0.077) [0.083, 0.361]	0.378 (0.196) [0.116, 0.787]	-0.282 (0.293) [-0.869, 0.245]	-0.15 (0.19) [-0.549, 0.19]
Whether missing on international questions (constructed)								

10. missing								
on favor	0.847	0.252	**0.606**	-0.294	0.995	0.343	**0.557**	-0.298
US	(0.264)	(0.109)	**(0.267)**	(0.228)	(0.342)	(0.169)	**(0.279)**	(0.255)
	[0.451,	[0.105,	**[0.082,**	[-0.753,	[0.48,	[0.104,	**[0.014,**	[-0.805,
	1.589]	0.605]	**1.13]**	0.166]	1.71]	0.668]	**1.121]**	0.172]
11. missing								
on favor	0.753	0.357	0.373	-0.273	0.844	0.464	0.324	-0.287
Germany	(0.223)	(0.151)	(0.258)	(0.236)	(0.254)	(0.225)	(0.26)	(0.24)
	[0.414,	[0.153,	[-0.132,	[-0.75,	[0.404,	[0.16,	[-0.169,	[-0.765,
	1.37]	0.836]	0.878]	0.204]	1.324]	0.857]	0.839]	0.149]
12. missing								
on favor	0.832	0.316	0.484	-0.25	0.936	0.452	0.398	-0.244
China	(0.245)	(0.135)	(0.259)	(0.236)	(0.303)	(0.243)	(0.28)	(0.257)
	[0.458,	[0.134,	[-0.024,	[-0.726,	[0.434,	[0.118,	[-0.134,	[-0.73,
	1.508]	0.746]	0.991]	0.226]	1.552]	0.933]	0.949]	0.229]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,int}$ is the interviewer intraclass correlation associated with the FTF mode. $\rho_{t,int}$ is the interviewer intraclass correlation associated with the TEL mode. α refers to the log differences between the FTF and TEL interviewer variances.

Appendix D: Outcome Variables Used in the Health and Retirement Study

Table 3.11: Outcome Variables Used in the Health and Retirement Study

Questions	Original response categories	Response categories used in the study
	CESD questions	
Much of the time during the past week, you felt depressed.	1. Yes 5. No	1. Yes 0. No
Much of the time during the past week, you felt that everything you did was an effort.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, your sleep was restless.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, you were happy.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, you felt lonely.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, you enjoyed life.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, you felt sad.	1.Yes 5. No	1. Yes 0. No
Much of the time during the past week, you could not get going.	1.Yes 5. No	1. Yes 0. No
	Interviewer Observations	
How attentive was the respondent to the questions during the interview?	1. Not at all attentive 2. Somewhat attentive 3. Very attentive	1. Very attentive 0. Not at all or somewhat attentive

How was the respondent's understanding of the questions?	1. Excellent 2. Good 3. Fair 4. Poor	1. Excellent 0. Good, fair, or poor
How was the respondent's cooperation during the interview?	1. Excellent 2. Good 3. Fair 4. Poor	1. Excellent 0. Good, fair, or poor
How much difficulty did the respondent have remembering things that you asked him/her about?	1. No difficulty 2. A little difficulty 3. Some difficulty 4. A lot of difficulty 5. Could not do at all	1. No difficulty 0. A little/some/lot of difficulty or could not do at all
How much difficulty did the respondent have hearing you when you talked to him/her?	1. No difficulty 2. A little difficulty 3. Some difficulty 4. A lot of difficulty 5. Could not do at all	1. No difficulty 0. A little/some/lot of difficulty or could not do at all
Overall, what is your opinion of the quality of this interview? Was it of:	1. High quality 2. Adequate quality 3. Questionable quality	1. High quality 0. Adequate or questionable quality
	Physical activity	
How often do you take part in sports or activities that are vigorous, such as running or jogging, swimming, cycling, aerobics or gym workout, tennis, or digging with a spade or shovel	1. More than once a week 2. Once a week 3. One to three times a month 4. Hardly ever or never 7. (VOL) Every day	1. At least once a week 0. Less than once a week

And how often do you take part in sports or activities that are moderately energetic such as, gardening, cleaning the car, walking at a moderate pace, dancing, floor or stretching exercises:	1. More than once a week 2. Once a week 3. One to three times a month 4. Hardly ever or never 7. (VOL) Every day	1. At least once a week 0. Less than once a week
And how often do you take part in sports or activities that are mildly energetic, such as vacuuming, laundry, home repairs:	1. More than once a week 2. Once a week 3. One to three times a month 4. Hardly ever or never 7. (VOL) Every day	1. At least once a week 0. Less than once a week

Appendix E: Full Results on Interviewer Variances in the Health and Retirement Study

Table 3.12: Interviewer Variances Per Mode for Selected Items in Health and Retirement Study Adjusting for Covariates

Questions	Likelihood					Bayesian				
	σ_f^2	σ_t^2	α	β_1	ρ	σ_f^2	σ_t^2	α	β_1	ρ
CESD questions										
felt depressed	0.012 (0.006) [0.004, 0.032]	0.015 (0.009) [0.005, 0.046]	-0.111 (0.378) [-0.853, 0.63]	0.047 (0.029) [-0.01, 0.104]	0.31 (0.443) [-0.565, 0.857]	0.015 (0.023) [0, 0.036]	0.012 (0.009) [0, 0.029]	-0.038 (1.114) [-1.989, 1.883]	0.043 (0.037) [-0.024, 0.098]	**0.748 (0.123) [0.505, 0.956]**
everything was an effort	0.021 (0.006) [0.012, 0.037]	0.005 (0.006) [0.001, 0.042]	0.7 (0.54) [-0.358, 1.758]	**0.11 (0.025) [0.06, 0.16]**	0.078 (0.545) [-0.76, 0.819]	0.022 (0.008) [0.012, 0.033]	0.006 (0.006) [0.001, 0.015]	**0.866 (0.475) [0.171, 1.706]**	**0.11 (0.025) [0.061, 0.158]**	**0.783 (0.243) [0.169, 0.989]**
restless sleep	0.003 (0.003) [0, 0.026]	0.004 (0.004) [0, 0.034]	-0.231 (0.805) [-1.809, 1.346]	**0.056 (0.022) [0.013, 0.1]**	-0.67 (1.144) [-1, 0.997]	0.013 (0.063) [0, 0.013]	0.002 (0.003) [0, 0.009]	0.667 (0.823) [-0.858, 2.083]	**0.057 (0.019) [0.02, 0.096]**	0.198 (0.106) [-0.062, 0.368]
happy	N/A (N/A) [N/A, N/A]	N/A (N/A) [N/A, N/A]	N/A (N/A) [N/A, N/A]	N/A (N/A) [N/A, N/A]	N/A (N/A) [N/A, N/A]	0.014 (0.06) [0, 0.053]	0.012 (0.024) [0, 0.021]	-1.99 (1.887) [-4.304, 1.102]	0 (0.027) [-0.051, 0.048]	**0.6 (0.049) [0.498, 0.702]**

lonely	0.005 (0.004) [0.001, 0.027]	0 (0.004) [0, N/A]	N/A (N/A) [N/A, N/A]	0.04 (0.025) [-0.009, 0.09]	-0.569 (4.224) [-1, 1]	0.013 (0.052) [0, 0.018]	0.004 (0.004) [0, 0.013]	0.358 (0.663) [-0.914, 1.679]	0.038 (0.027) [-0.018, 0.081]	-0.086 (0.102) [-0.263, 0.108]
enjoyed life	0.009 (0.006) [0.002, 0.038]	0.012 (0.009) [0.003, 0.052]	-0.155 (0.531) [-1.196, 0.887]	**0.064** **(0.032)** **[0, 0.127]**	-0.599 (0.705) [-0.993, 0.898]	0.011 (0.044) [0, 0.021]	0.011 (0.012) [0.001, 0.028]	-0.218 (0.662) [-1.48, 1.042]	**0.055** **(0.031)** **[0.002, 0.122]**	**0.446** **(0.145)** **[0.194, 0.685]**
felt sad	0.031 (0.007) [0.02, 0.049]	0 (0.001) [0, 0.691]	2.426 (2.026) [-1.545, 6.397]	0.042 (0.026) [-0.009, 0.094]	1 (N/A) [N/A, N/A]	0.033 (0.008) [0.02, 0.048]	0.004 (0.004) [0, 0.011]	**1.222** **(0.501)** **[0.332, 2.068]**	0.043 (0.025) [-0.009, 0.09]	-0.009 (0.249) [-0.445, 0.488]
could not get going	0.013 (0.005) [0.006, 0.028]	0.019 (0.008) [0.008, 0.045]	-0.165 (0.29) [-0.734, 0.404]	0.043 (0.027) [-0.01, 0.095]	0.381 (0.337) [-0.354, 0.826]	0.016 (0.027) [0.005, 0.028]	0.022 (0.01) [0.005, 0.037]	-0.184 (0.354) [-0.815, 0.584]	0.037 (0.037) [-0.015, 0.09]	0.252 (0.186) [-0.057, 0.695]
overall indicator	0.012 (0.005) [0.005, 0.029]	0.012 (0.007) [0.004, 0.041]	0.008 (0.377) [-0.732, 0.747]	**0.134** **(0.028)** **[0.079, 0.189]**	-0.194 (0.443) [-0.8, 0.608]	0.014 (0.018) [0.002, 0.024]	0.011 (0.007) [0.001, 0.023]	0.076 (0.455) [-0.799, 1.028]	**0.132** **(0.033)** **[0.078, 0.192]**	**0.29** **(0.185)** **[0.036, 0.699]**
Interviewer Observations										

attentive	0.291 (0.033) [0.234, 0.363]	0.338 (0.043) [0.263, 0.434]	-0.075 (0.064) [-0.199, 0.05]	0.024 (0.035) [-0.045, 0.093]	**0.89** **(0.038)** **[0.788,** **0.944]**	0.296 (0.033) [0.233, 0.355]	0.345 (0.043) [0.263, 0.427]	-0.075 (0.065) [-0.198, 0.05]	0.023 (0.038) [-0.049, 0.097]	**0.871** **(0.051)** **[0.789,** **0.943]**
understanding	0.411 (0.04) [0.339, 0.498]	0.46 (0.05) [0.371, 0.571]	-0.057 (0.049) [-0.153, 0.04]	0 (0.032) [-0.063, 0.063]	**0.922** **(0.024)** **[0.86,** **0.957]**	0.417 (0.046) [0.345, 0.493]	0.46 (0.046) [0.37, 0.55]	-0.048 (0.058) [-0.139, 0.053]	-0.001 (0.032) [-0.063, 0.062]	**0.911** **(0.043)** **[0.861,** **0.961]**
cooperation	0.451 (0.043) [0.374, 0.545]	0.396 (0.044) [0.319, 0.492]	0.065 (0.048) [-0.03, 0.16]	**0.184** **(0.032)** **[0.122,** **0.246]**	**0.935** **(0.022)** **[0.874,** **0.967]**	0.459 (0.044) [0.381, 0.543]	0.394 (0.044) [0.313, 0.482]	0.077 (0.049) [-0.022, 0.168]	**0.186** **(0.033)** **[0.12,** **0.253]**	**0.928** **(0.023)** **[0.88,** **0.968]**
remembering	0.484 (0.047) [0.399, 0.586]	0.59 (0.065) [0.475, 0.731]	**-0.099** **(0.047)** **[-0.191,** **-0.008]**	-0.056 (0.032) [-0.119, 0.006]	**0.95** **(0.018)** **[0.898,** **0.975]**	0.476 (0.052) [0.392, 0.566]	0.572 (0.057) [0.466, 0.69]	-0.092 (0.055) [-0.192, 0.004]	-0.054 (0.04) [-0.124, 0.009]	**0.937** **(0.041)** **[0.902,** **0.979]**
hearing	0.272 (0.031) [0.218, 0.339]	0.376 (0.047) [0.293, 0.482]	**-0.161** **(0.064)** **[-0.288,** **-0.035]**	**0.153** **(0.035)** **[0.084,** **0.221]**	**0.88** **(0.036)** **[0.786,** **0.935]**	0.28 (0.031) [0.216, 0.339]	0.391 (0.045) [0.302, 0.476]	**-0.166** **(0.063)** **[-0.291,** **-0.047]**	**0.159** **(0.037)** **[0.088,** **0.228]**	**0.869** **(0.038)** **[0.79,** **0.933]**
Overall quality	0.885 (0.08) [0.74, 1.058]	0.778 (0.079) [0.637, 0.949]	0.064 (0.04) [-0.014, 0.143]	**0.087** **(0.034)** **[0.02,** **0.154]**	**0.961** **(0.013)** **[0.924,** **0.981]**	0.88 (0.084) [0.722, 1.049]	0.781 (0.085) [0.623, 0.951]	0.06 (0.042) [-0.016, 0.149]	**0.087** **(0.035)** **[0.02,** **0.159]**	**0.95** **(0.051)** **[0.926,** **0.981]**

	Physical activity									
vigorous sports	0.014 (0.004) [0.007, 0.026]	0.01 (0.006) [0.003, 0.033]	0.174 (0.345) [-0.502, 0.85]	-0.035 (0.022) [-0.078, 0.009]	0.665 (0.369) [-0.459, 0.97]	0.016 (0.01) [0.006, 0.024]	0.01 (0.007) [0.001, 0.022]	0.327 (0.383) [-0.428, 1.02]	-0.038 (0.023) [-0.08, 0.006]	**0.586** **(0.228)** **[0.156,** **0.968]**
moderately energetic sports	0.013 (0.004) [0.007, 0.026]	0.017 (0.007) [0.008, 0.039]	-0.143 (0.265) [-0.663, 0.378]	0.022 (0.023) [-0.024, 0.068]	0.429 (0.306) [-0.269, 0.832]	0.016 (0.016) [0.006, 0.026]	0.019 (0.007) [0.007, 0.033]	-0.127 (0.283) [-0.654, 0.42]	0.022 (0.023) [-0.02, 0.072]	**0.31** **(0.139)** **[0.051,** **0.564]**
mildly energetic sports	0.017 (0.006) [0.008, 0.033]	0.027 (0.009) [0.014, 0.053]	-0.239 (0.243) [-0.716, 0.239]	**0.112** **(0.028)** **[0.057,** **0.168]**	0.117 (0.28) [-0.412, 0.587]	0.018 (0.008) [0.006, 0.03]	0.028 (0.011) [0.007, 0.047]	-0.221 (0.283) [-0.79, 0.328]	**0.11** **(0.029)** **[0.052,** **0.162]**	0.15 (0.316) [-0.312, 0.797]

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. $\rho_{f,int}$ is the interviewer intraclass correlation associated with the FTF mode. $\rho_{t,int}$ is the interviewer intraclass correlation associated with the TEL mode. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects. We use "N/A" for two purposes: 1) to indicate nonconvergence, and 2) to mask estimates that are unstable (with a standard error greater than 5) due to numerical difficulties.

Appendix F: Results Testing Sensitivity to Rho for the Depression Items in the Health and Retirement Study

Table 3.13: Interviewer Variances Per Mode for Depression Items in Health and Retirement Study Adjusting for Covariates Using Likelihood Estimation

Questions	$\rho = 0$				$\rho = \hat{\rho}_{Bayesian}$			
	σ_f^2	σ_t^2	α	β_1	σ_f^2	σ_t^2	α	β_1
felt depressed	0.012 (0.006) [0.005, 0.032]	0.015 (0.009) [0.005, 0.047]	-0.111 (0.38) [-0.855, 0.633]	0.047 (0.03) [-0.011, 0.106]	0.011 (0.006) [0.004, 0.031]	0.013 (0.008) [0.004, 0.045]	-0.087 (0.429) [-0.928, 0.753]	0.046 (0.028) [-0.01, 0.101]
everything was an effort	0.022 (0.006) [0.013, 0.037]	0.005 (0.006) [0.001, 0.041]	0.7 (0.541) [-0.36, 1.759]	**0.11** **(0.025)** **[0.06, 0.16]**	0.021 (0.006) [0.012, 0.037]	0.001 (0.004) [0, 0.786]	1.431 (1.658) [-1.818, 4.68]	**0.107** **(0.024)** **[0.059, 0.154]**
restless sleep	0.002 (0.003) [0, 0.027]	0.004 (0.004) [0.001, 0.034]	-0.262 (0.813) [-1.856, 1.332]	**0.057** **(0.022)** **[0.014, 0.1]**	0.002 (0.003) [0, 0.033]	0.004 (0.004) [0, 0.037]	-0.285 (0.902) [-2.052, 1.483]	**0.057** **(0.022)** **[0.014, 0.099]**
happy	0.008 (0.006) [0.002, 0.033]	0.007 (0.007) [0.001, 0.052]	0.087 (0.631) [-1.15, 1.324]	0.021 (0.027) [-0.032, 0.074]	0 (0) [0, N/A]	0.007 (0.007) [0.001, 0.053]	N/A (N/A) [N/A, N/A]	0.022 (0.026) [-0.029, 0.073]

lonely	0.005 (0.004) [0.001, 0.027]	0 (0.006) [N/A, N/A]	1.305 (N/A) [N/A, N/A]	0.04 (0.025) [-0.009, 0.089]	0.005 (0.004) [0.001, 0.027]	0.001 (0.005) [0, N/A]	1.122 (N/A) [N/A, N/A]	0.04 (0.025) [-0.009, 0.089]
enjoyed life	0.009 (0.006) [0.002, 0.037]	0.012 (0.009) [0.003, 0.052]	-0.151 (0.527) [-1.184, 0.882]	**0.062 (0.032) [0, 0.125]**	0.007 (0.007) [0.001, 0.046]	0.009 (0.009) [0.001, 0.064]	-0.136 (0.675) [-1.46, 1.188]	0.06 (0.031) [-0.001, 0.12]
felt sad	0.031 (0.007) [0.019, 0.05]	0 (0) [N/A, N/A]	N/A (N/A) [N/A, N/A]	0.044 (0.026) [-0.008, 0.095]	0.031 (0.007) [0.019, 0.05]	0 (0) [0, N/A]	N/A (N/A) [N/A, N/A]	0.044 (0.026) [-0.008, 0.095]
could not get going	0.014 (0.005) [0.007, 0.028]	0.019 (0.008) [0.008, 0.045]	-0.155 (0.29) [-0.724, 0.414]	0.044 (0.027) [-0.01, 0.098]	0.014 (0.005) [0.007, 0.028]	0.019 (0.008) [0.008, 0.045]	-0.165 (0.286) [-0.726, 0.396]	0.043 (0.027) [-0.01, 0.096]
overall indicator	0.012 (0.005) [0.005, 0.029]	0.012 (0.007) [0.004, 0.041]	0.011 (0.38) [-0.734, 0.757]	**0.134 (0.028) [0.079, 0.188]**	0.012 (0.005) [0.005, 0.029]	0.011 (0.007) [0.003, 0.041]	0.047 (0.427) [-0.791, 0.884]	**0.133 (0.027) [0.08, 0.186]**

Notes: β_1 is the mode effects in means, computed as the mean of the FTF estimate minus the mean of the TEL estimate. σ_f^2 is the FTF interviewer variances. σ_t^2 is the TEL interviewer variance. α refers to the log differences between the FTF and TEL interviewer variances. ρ is the correlation between the FTF and TEL random interviewer effects. We use "N/A" for two purposes: 1) to indicate nonconvergence, and 2) to mask estimates that are unstable (with a standard error greater than 5) due to numerical difficulties.

CHAPTER 4

Using Principal Stratification to Detect Mode Effects across Waves in a Longitudinal Study

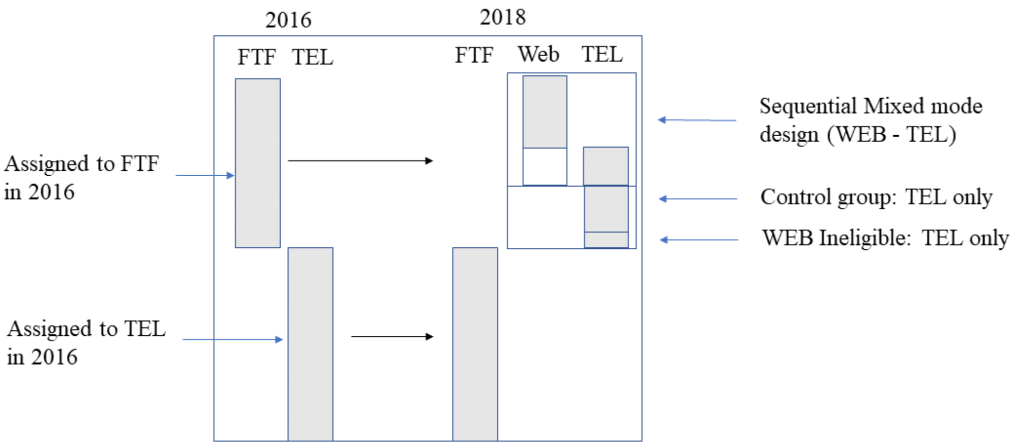
4.1 Introduction

Because mixed-mode designs can save costs and improve representiveness, they have been used in many large-scale longitudinal studies, such as the Panel Study of Income Dynamics (PSID), the UK Household Longitudinal Study, and the German Socio-Economic Panel et al. (see details in [52]). However, embedding mixed-mode designs in a longitudinal setting brings unique challenges for inference. The primary focus of longitudinal studies is to provide estimates comparable over time. To achieve the purpose, researchers typically assume measurement error remains constant across waves [53]. However, the introduction of mixed-mode design may violate the assumption because much literature has noted that different modes are associated with different measurement errors [54]. For example, a larger social desirability bias is reported in interviewer-administered modes compared to self-administered modes [11, 10].

Despite the challenges brought by applying mixed-mode designs in panel studies, the issue has not received adequate scholarly attention, with only five papers that we are aware of in the literature [55, 56, 57, 58, 59]. Three of these studies utilized a randomized design where some respondents were contacted in a single mode, and the remaining participants were randomly assigned to a sequential mixed-mode design. They focused on intent-to-treat analyses, answering the question of how estimates produced from the sequential mixed-mode design differ from the ones computed from the single mode [56, 57, 58]. Cernat et al. estimated mode effects across face-to-face, telephone, and Web while teasing out time effects using a crossover mixed-mode design implemented in the Health and Retirement Study (HRS) 2010, 2011, and 2012 waves [55]. In a more recent paper, Cernat and Sakshaug use selection weights to control for potential selection effects and latent growth modeling to examine the impact of mode design (single mode versus mixed-mode) on estimates of changes [59]. In addition, a few calibration methods have been proposed to account for the mode effects in sequential mixed-mode designs [60, 44].

We aim to separate mode effects and time effects when a single mode in the previous wave is followed by a sequential mixed-mode design in the later wave, using the HRS 2016 and 2018 core surveys. Since 2006, HRS has started to rotate enhanced FTF and TEL across waves for individuals under the age of 80, such that a random half of the sample receives FTF interviews and the other half receives TEL interviews and then switch modes each wave. In 2018, respondents originally scheduled for telephone interviews and who met a series of eligibility requirement were assigned to an experiment where about 2/3 of them were in a sequential mixed-mode design (WEB-TEL), and the remaining 1/3 of panelists were contacted via TEL alone. Meanwhile, participants scheduled for FTF interviews were interviewed as planned. See Figure 4.1 for an illustration of the HRS 2016-2018 mixed-mode designs.

Multiple comparisons can be informative under the HRS design. For example, evaluating whether the sequential mixed-mode design produces comparable estimates to the TEL-only group [58] can demonstrate if the sequential design can be used as an alternative to the TEL-only design. Understanding how the estimates provided by the Web respondents differ from those given by the TEL respondents in the sequential mixed-mode group can imply the magnitude of mode effects, including both the measurement and selection effects. Moreover, comparing the Web participants in the sequential design with the whole sequential



HRS mode experiment design

Figure 4.1: Visualization of the HRS 2016-2018 Mixed-mode Designs

Notes: This figure only includes HRS participants in the randomized mixed-mode design, who would switch between FTF and TEL across waves. It excludes the late baby boomer cohort added in HRS 2016 and the older populations (over 80 years old) who would only be interviewed in FTF. The numbers noted in the boxes are the achieved sample sizes of the groups. Boxes filled in grey reflect the actual modes used for data collection, while boxes in white indicate the mode used to contact panelists but was not the final mode in which they responded (i.e., the Web mode). Nonrespondents were not shown as the figure focuses on illustrating the design.

design group (WEB-TEL) will suggest the added benefits of a TEL nonresponse follow-up. While these comparisons are highly beneficial, they avoid the complexities of across-wave comparisons, which are the biggest challenge faced by applying a new mixed-mode design in longitudinal panels.

Borrowing strength from causal inference literature, this paper uses principal stratification to account for whether respondents used the assigned mode to respond as post-treatment covariates. Principal stratification is commonly used in causal inference to adjust for post-treatment covariates when estimating treatment effects [61]. In our context, the planned mode for administering the survey to an HRS panelist is considered the treatment, while whether the individual responds via that mode serves as the post-treatment covariate. In addition, we employ a potential outcome framework [62, 63] where we impute the response statuses and the outcomes that would have been observed through a mode different from the mode actually used. The potential outcome framework and imputation methods have been considered in mixed-mode inference literature [17, 26, 29]. Suzer-Gurtekin et al [17, 44] first used multiple imputation in mixed-mode inferences where they conceptualized sequential mixed-mode designs as a missing data problem. They considered a selection model that estimate the probability that sampled members use which mode to respond and incorporated the probability in the imputation model. Kolenikov and Kennedy [26] also used the potential outcome framework and applied multiple imputation to propagate the uncertainty. Park, Kim, and Park [29] considered a measurement error model and used fractional imputation to impute the missing values given the observed values from the same participants and their covariates.

This paper differentiates itself from prior studies and contributes to mixed-mode inferences by acknowledging that certain participants might not be reachable or recruitable via a particular mode, which is a more realistic assumption given that each mode and its associated sampling frame have their own coverage and nonresponse attributes. Given this, we employ the principal stratification method to examine mode effects within the subset of participants who potentially would respond under the modes to be compared. This approach is expected to provide a more conceptually grounded estimate of the mode effect, separate from mode selection effects. Additionally, it can provide insights on how to better allocate modes to sampled units when designing mixed-mode studies.

The remainder of the paper is organized as follows: In Section 2, we first introduce the framework, the imputation approach, and the analysis procedures for the cross-sectional data (HRS 2016 and 2018). Following this, we illustrate the methods considered for the longitudinal sample (HRS 2016 - 2018). In Section 3, we present the findings for both the cross-sectional and longitudinal analyses. Finally, in Section 4, we discuss the implications of this study, along with limitations and potential future research.

4.2 Methods

In this section, we begin by describing the framework and notation, based on the mixed-mode design of the HRS, and outline the covariates and outcomes considered in the dataset. Next, we introduce the methods applied to the cross-sectional and longitudinal data, respectively. Given the use of principal stratification throughout the imputation and analytical steps, we also detail the principal strata alongside the introduction of these methods.

4.2.1 Framework

In a population with size N, we are interested in outcome variables denoted by Y on two time points (indexed by $t, t = 1, 2$). We observe some time-varying covariates that are invariate to modes, represented by X. We use a sample of size n and employ mixed-mode designs to collect information from this population. The mode used for participant i $(i = 1, 2, ..., n)$ is denoted by m, $m = 1, 2, 3$, where 1 represents FTF, 2 represents TEL, and 3 represents WEB. Given that each mode has its own distinct nonresponse properties, we treat the response status as mode-specific. We use A_t to indicate mode assignment at time t. We use R_t^m to represent the potential response status of a HRS panelist, when invited to administer the survey using mode m, at time t. The potential outcome variable Y observed when a panelist is a mode m respondent ($R_t^m = 1$) at time t is denoted as Y_t^m.

4.2.2 Data

In this study, we analyze data from the 2016-2018 core surveys of the HRS. For the 2016 analysis, our analytical sample includes respondents who were randomly assigned to either enhanced face-to-face (FTF) interviews or telephone interviews in 2016, were not proxy respondents, had non-zero sample weights, and had non-missing assigned modes. We excluded households with members aged 80 or above in 2016 and new panelists from the mode comparison, as they could only be assigned to the enhanced FTF mode in HRS 2016. This criteria resulted in a sample size of 11,383 for HRS 2016. For the HRS 2018 analysis, we excluded respondents who were in households that had members older than 80 in 2018, were newly added to the panel, had zero weights, were missing assigned mode information, were proxy respondents, lived in nursing homes, or were Spanish speakers, prior reports of having no WEB access, since the last four conditions made them ineligible for Web mode invitations. The sample size for HRS 2018 is 8,466. For the longitudinal analysis covering 2016-2018, we included panelists who responded to both waves, were under randomized assignment between FTF and TEL, were not assigned to the WEB in HRS 2018, were self-respondents, had non-zero weights and non-missing assigned modes for both waves. The sample size for this longitudinal analysis is 4,911.

We consider four outcomes measured in the HRS core survey: 1) cognitive function measured by the total number of words recalled (immediate + delayed, ranging from 0 to 20), 2) a binary variable indicating whether depressed, determined by if respondents provide affirmative responses to at least four out of eight CESD items, 3) BMI index (ranging from 10.2 to 70.7), and 4) a binary variable indicating whether reporting very good or excellent health. These four items are frequently used in social science and public health research and literature have suggested evidence about mode effects for these variables. For example, several studies have suggested that depression scores are higher in self-administered modes, compared to interviewer-administered modes [64, 65, 66]. Obesity (BMI $\geq$ 30) has been found to be more likely reported in in-person interviews than in TEL [67]. Additionally, research has shown that respondents generally perform better on cognitive tests when completing surveys online compared to interviewer-administered modes [68, 69, 70]. Table 4.1 provides the descriptive statistics of these outcome variables.

Table 4.1: Descriptive Statistics of the HRS sample (Unweighted)

Variables	HRS 2016				HRS 2018					
	FTF		TEL		FTF		TEL		WEB	
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
Outcome variables										
Number of words recalled	9.694	0.044	10.277	0.047	11.375	0.051	11.235	0.063	11.499	0.116
Depressed	0.155	0.005	0.141	0.004	0.120	0.006	0.132	0.007	0.095	0.009
BMI	29.520	0.088	29.185	0.080	29.681	0.109	29.818	0.130	29.433	0.190
Very good or above health	0.356	0.007	0.391	0.006	0.463	0.009	0.438	0.010	0.514	0.015
Predictors										
Age	65.550	0.097	65.694	0.090	63.561	0.126	62.158	0.143	66.110	0.189
Schooling years	13.463	0.110	13.503	0.094	14.236	0.040	14.114	0.048	14.420	0.062
Male	0.424	0.007	0.415	0.006	0.426	0.009	0.417	0.010	0.421	0.015
English-speaking Hispanics	0.172	0.005	0.156	0.005	0.069	0.004	0.089	0.006	0.054	0.007
Non-Hispanic Black	0.237	0.006	0.208	0.005	0.183	0.007	0.238	0.008	0.140	0.011
Coupled	0.638	0.007	0.655	0.006	0.666	0.008	0.645	0.009	0.710	0.014
Working	0.398	0.007	0.394	0.006	0.504	0.009	0.547	0.010	0.450	0.015
Vision	0.058	0.003	0.049	0.003	0.038	0.003	0.037	0.004	0.053	0.007
Born in US	0.834	0.005	0.843	0.005	0.911	0.005	0.909	0.006	0.938	0.007

Notes: We use the HRS 2016 cross-sectional sample and the 2018 cross-sectional sample to compute the unweighted descriptive statistics.

In this paper, we examine a range of covariates to predict response status and outcome variables, including age, education, gender, race ethnicity, whether coupled, whether born in US, vision, and whether working. The descriptive statistics of these predictor variables are provided in Table 4.1. To ensure comparability across different waves and outcomes, we employ a consistent set of predictors in all models. Since the missing data rates in the covariates and outcome variables are minor (5%), we use predictive mean matching and classification and regression trees to singly impute the missing cases in these predictors and outcomes prior to our analysis. The imputation for the item missing data is implemented using mice package [71] from the R prgramming language [72].

Figure 4.2 illustrates the observed data structure based on the HRS 2016 and 2018 design using the notation introduced earlier. Because we restrict the analysis to those who participated both in the 2016 and the 2018 waves, covariates (X) are observed for every panelist. In our analytical sample, the response indicator for mode m equals 0 when participants did not respond via the assigned mode m, but instead responded via another mode (m', where $m' \neq m$; in this case, $R^{m'} = 1$). We only observe the potential mode m response indicator (R^m) for those assigned to mode m and those assigned to mode m' but choosing to use mode m to respond. If the potential response indicator is 0 ($R^m = 0$), it means that the panelist must have used another mode to respond, since we only include participants who have responded in both waves in our analytical sample. We consider this sample inclusion criterion because estimating changes across time is only applicable when individuals responded in both waves.

In addition, for the HRS 2018 design, for participants assigned to the sequential mixed-mode design (WEB-TEL), if they responded via the WEB, their response status in the WEB mode is known and their response status in the TEL mode is unknown. If they did not respond in the WEB and were then invited to the TEL mode, their response status for both the WEB and the TEL becomes known. Therefore, the observed areas of R_2^t and R_2^w overlap in Figure 4.2 for those WEB nonrespondents.

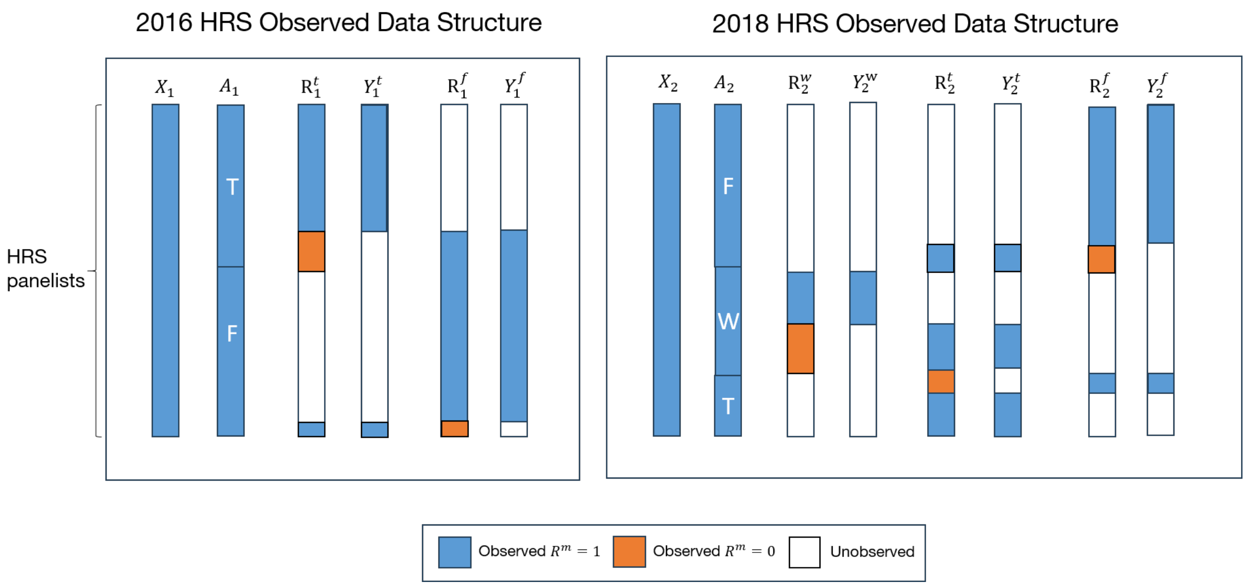


Figure 4.2: Observed Data Structure of HRS 2016 and 2018

Notes: Despite that there are multiple outcomes considered in this paper, we illustrate the imputation plan in the figure only with one outcome variable, as the pattern applies to all outcomes.

4.2.3 Cross-sectional Analysis

4.2.3.1 The HRS 2016

For the HRS 2016 analysis, R_1^f is missing for those assigned to and responded in TEL; similarly, R_1^t is missing for those assigned to and responded in FTF. We impute $R_1^f, Y_1^f, R_1^t, Y_1^t$ based on the their joint distribution conditional on X_1 while considering that $Y_1^m, m = f$ or t would only be imputed if R_1^m is originally missing and imputed to be 1. We consider the following decomposition of the joint conditional probability:

$$f(R_1^f, Y_1^f, R_1^t, Y_1^t|X_1) = f(R_1^f|X_1)f(R_1^t|X_1, R_1^t)f(Y_1^f|X_1, R_1^t, R_1^f)f(Y_1^t|X_1, R_1^t, R_1^f, Y_1^f) \tag{4.1}$$

$$\doteq f(R_1^f|X_1)f(R_1^t|X_1)f(Y_1^f|X_1, R_1^t, R_1^f)f(Y_1^t|X_1, R_1^t, R_1^f) \tag{4.2}$$

Because there is no data to estimate the correlation between R_1^t and R_1^f, Y_1^t and Y_1^f, and research has shown that there is little sensitivity to the correlation when making inference about the treatment effects [73], we simplify equation 4.1 to 4.2.

A Gibbs sampler, which iteratively samples from the full conditional posterior distributions - $P(R_1^f|Y_1^f, R_1^t, Y_1^t, X_1), P(R_1^t|Y_1^f, R_1^f, Y_1^t, X_1), P(Y_1^f|R_1^f, R_1^t, Y_1^t, X_1)$, and $P(Y_1^t|R_1^f, R_1^t, Y_1^f, X_1)$ can be used for the imputation. Take $P(R_1^f|Y_1^f, R_1^t, Y_1^t, X_1)$ for example, we have

$$P(R_1^f|Y_1^f, R_1^t, Y_1^t, X_1) \propto P(R_1^f|X_1)P(Y_1^f|X_1, R_1^t, R_1^f)P(Y_1^t|X_1, R_1^t, R_1^f). \tag{4.3}$$

We assume that $P(R_1^f|X_1)$ follows a probit regression model: $P(R_1^f = 1|X_1) = \Phi(\beta_1^f X_1)$, where Φ is the cumulative distribution function of the standard normal distribution. If the outcome is continuous, we assume that $P(Y_1^f|X_1, R_1^t, R_1^f)$ and $P(Y_1^t|X_1, R_1^t, R_1^f)$ follow normal distributions in a principal stratum defined by R_1^t and R_1^f. If the outcome is binary, we consider them to follow probit regression models.

Specifically, for $f(Y_1^f|X_1, R_1^t, R_1^f)$ and $f(Y_1^t|X_1, R_1^t, R_1^f)$, we categorize respondents into three principal strata (PS, indexed by h) defined by R_1^t and R_1^f: 1) FTF & TEL ($R_1^f = 1$ & $R_1^t = 1$), 2) FTF only ($R_1^f = 1$ & $R_1^t = 0$), and 3) TEL only ($R_1^f = 0$ & $R_1^t = 1$). We assume that cases within the same principal stratum share the same multivariate associations between the potential response indicators and covariates, and between the potential outcomes and covariates. We specify the imputation models for the potential outcomes in Table 4.2.

Table 4.2: Models Per Principal Stratum for the HRS 2016

Principal Strata	Conditions	Models
1) FTF & TEL	$R_1^f = 1 \& R_1^t = 1$	$f(Y_{1,h1}^f|X_{1,h1}, R_1^f, R_1^t) = \beta_{1,h1}^f X_{1,h1}$, $f(Y_{1,h1}^t|X_{1,h1}, R_1^f, R_1^t) = \beta_{1,h1}^t X_{1,h1}$
2) FTF only	$R_1^f = 1 \& R_1^t = 0$	$f(Y_{1,h2}^f|X_{1,h2}, R_1^f, R_1^t) = \beta_{1,h2}^f X_{1,h2}$
2) TEL only	$R_1^f = 0 \& R_1^t = 1$	$f(Y_{1,h3}^t|X_{1,h3}, R_1^f, R_1^t) = \beta_{1,h3}^t X_{1,h3}$

Notes: If outcomes are continuous, we fit the models using normal distributions, assuming $Y_{1,h}^m|X_{1,h} \sim N(\beta_{1,h}^m X_{1,h}, \sigma_m^2)$, where m indicates modes and h indexes PS. If outcomes are binary, we fit probit models.

Then we plug $R_1^f = 1$ and $R_1^f = 0$ to 4.3 to have

$$
\begin{aligned}
&P(R_1^f = 1|Y_1^f, R_1^t, Y_1^t, X_1) \\
&= \frac{P(R_1^f = 1|X_1)P(Y_1^f|X_1, R_1^t, R_1^f = 1)P(Y_1^t|X_1, R_1^t, R_1^f = 1)}{P(R_1^f = 1|X_1)P(Y_1^f|X_1, R_1^t, R_1^f = 1)P(Y_1^t|X_1, R_1^t, R_1^f = 1) + P(R_1^f = 0|X_1)P(Y_1^f|X_1, R_1^t, R_1^f = 0)P(Y_1^t|X_1, R_1^t, R_1^f = 0)}.
\end{aligned}
\tag{4.4}
$$

We compute $P(R_1^t = 1|Y_1^f, R_1^f, Y_1^t, X_1)$ in a similar fashion as in 4.4. Next, the remaining conditional distributions, $P(Y_1^f|R_1^f, R_1^t, Y_1^t, X_1)$ and $P(Y_1^t|R_1^f, R_1^t, Y_1^f, X_1)$, which can be reduced to $P(Y_1^f|R_1^f, R_1^t, X_1)$ and $P(Y_1^t|R_1^f, R_1^t, X_1)$, are modeled in the same way as described in 4.2.

After initialization using fixed or random draws of the unobserved R_1^f, R_1^t, Y_1^f, and Y_1^t, we can apply the Gibbs sampling algorithm described above. Drawing from posterior distributions of the coefficients in the probit models for $f(R_1^f|X_1)$ and $f(R_1^t|X_1)$, as well as the normal or probit models for $f(Y_1^f|X_1, R_1^f, R_1^t)$ and $f(Y_1^t|X_1, R_1^f, R_1^t)$, we can derive the posterior distributions of the full conditionals and then impute the unobserved R_1^f, R_1^t, Y_1^f, and Y_1^t.

We implement the models using Proc MCMC in the SAS programming language. We consider weakly informative priors for the coefficients by specifying a normal distribution with mean 0 and variance 100 ($\beta_{1,h}^m \sim N(0, 100)$). For continuous outcomes, we additionally use a half-T prior with mean 0, standard deviation(SD) 25, and degrees of freedom 3 for the SD of the normal distribution. We consider 100000 posterior draws with thinning rate as 50 after 40000 burn-in samples, resulting in 2000 sets of draws. We refer to each set of posterior draws of R_1^t, R_1^f, Y_1^t, and Y_1^f as an imputed dataset (indexed using l).

After the imputation, we construct the PS as outlined in Table 4.2 and compute stratum-specific estimates for each imputed dataset l: stratum mean ($\bar{y}_{1,lh}^m$) and stratum sampling variance ($V(\bar{y}_{1,lh}^m)$), accounting for the complex survey design features. For Stratum 1, where respondents would have responded to both modes, we can compute the individual mode effects as the difference between the potential outcome obtained when interviewed via FTF and the potential outcome obtained via TEL $\delta^{ft}_{1,lhi} = Y^f_{1,lhi} - Y^t_{1,lhi}$, and then sum over all cases in Stratum 1 in dataset l to get a point estimate of mode effects $(\delta^{ft}_{1,l})$ and sampling variance of the estimate $(V(\delta^{ft}_{1,l}))$ for this dataset . To compute the estimates averaged across all imputed datasets, we apply Rubin's combining rules [74] and get the point estimate of mode effects $(\hat{\delta}^{ft}_1)$ and its standard error $(SE(\hat{\delta}^{ft}_1))$ using 4.5, where L is the total number of imputed datasets.

$$\begin{aligned} \hat{\delta}^{ft}_1 &= \frac{\sum_l \hat{\delta}^{ft}_{1,l}}{L}, \\ SE(\hat{\delta}^{ft}_1) &= \sqrt{\frac{1}{L}\sum_l V(\hat{\delta}^{ft}_{1,l}) + \frac{L+1}{L(L-1)}\sum_l (\hat{\delta}^{ft}_{1,l} - \hat{\delta}^{ft}_1)^2}, \end{aligned} \tag{4.5}$$

Finally, we determine if there are mode effects in the HRS 2016 by computing the paired t-test statistic as $T = \frac{\hat{\delta}^{ft}_1}{SE(\hat{\delta}^{ft}_1)}$. We refer the statistic to a t-distribution with degrees of freedom (df) given by:

$$df = (L-1)(1+\frac{1}{r})^2, r = \frac{\frac{L+1}{L(L-1)}\sum_l (\hat{\delta}^{ft}_{1,l} - \hat{\delta}^{ft}_1)^2}{\frac{1}{L}\sum_l V(\hat{\delta}^{ft}_{1,l})}$$

We reject the null hypothesis of no mode effects if the probability of observing the T statistic under the null hypothesis is less than 0.05.

4.2.3.2 The HRS 2018

Compared to the HRS 2016, one major difference for HRS 2018 is the introduction of the Web-TEL sequential mixed-mode design. We assume that respondents' propensity to respond to one mode is not changed after having been invited to and refused the other modes. Take the TEL respondents in the sequential mixed-mode group for example, we assume $f(R^T|R^W = 0, A = (W,T)) = f(R^T|A = T)$, where $A = (W,T)$ indicates the assignment to WEB-TEL sequential mixed-mode group. This relates to the monotonicity assumption in the causal inference literature [75, 76], which says that there is no one who does the opposite of the assignment. In this context, when a respondent chooses a mode after being initially assigned to a different one, it is considered that they would also respond to the mode if it had been their initial assignment. In sequential mixed-mode designs, respondents are informed in advance that if they do not complete the web survey, they would be contacted again using the TEL mode. In this paper, we handle potential response indicators similarly for this scenario, as for others, when respondents are not initially provided the option to do the survey in other mode but still end up finishing the survey in another mode. This assumption is used in all three analysis in ths paper for the imputation.

To account for the WEB mode used in the HRS 2018, we consider 7 PS during the imputation and the analysis phases. For the imputation process, we impute $R_2^f, R_2^t, R_2^w, Y_2^f, Y_2^t, Y_2^w$ based on the joint distribution of $f(R_2^f, R_2^t, R_2^w, Y_2^f, Y_2^t, Y_2^w|X_2)$. We decompose the joint distribution in equation 4.6.

$$\begin{aligned} f(R_2^f, R_2^t, R_2^w, Y_2^f, Y_2^t, Y_2^w|X_2) &= f(R_2^f|X_2)f(R_2^t|X_2, R_2^f)f(R_2^w|X_2, R_2^f, R_2^t)\times \\ &f(Y_2^f|X_2, R_2^f, R_2^t, R_2^w)f(Y_2^t|X_2, R_2^f, R_2^t, R_2^w, Y_2^f)\times \\ &f(Y_2^w|X_2, R_2^f, R_2^t, R_2^w, Y_2^f, Y_2^t) \\ &\doteq f(R_2^f|X_2)f(R_2^t|X_2)f(R_2^w|X_2)\times \\ &f(Y_2^f|X_2, R_2^f, R_2^t, R_2^w)f(Y_2^t|X_2, R_2^f, R_2^t, R_2^w)\times \\ &f(Y_2^w|X_2, R_2^f, R_2^t, R_2^w) \end{aligned} \tag{4.6}$$

For the imputation of potential outcomes, we create 7 PS, illustrated in Table 4.3, based on the different response status when interviewed via FTF, TEL, and WEB. The imputation of the unobserved variables $R_2^f, R_2^t, R_2^w, Y_2^f, Y_2^t$, and Y_2^w proceeds as described in Section 4.2.3.1, with the addition of a probit regression model for R_2^w and linear or probit regression models for Y_2^w.

We consider three mode comparisons for the HRS 2018: FTF vs TEL, FTF VS WEB, and TEL VS WEB. In order to make these comparisons, we will need to combine some stratum-specific estimates to obtain a net comparison of the modes. For example, to compute mode effects between FTF and TEL, we combine strata 1 and 2 to compare FTF and TEL responses, regardless of whether these participants would participate via WEB or not. Similarly, we combine strata 1 and 3 to compare FTF and WEB and combine strata 1 and 4 to compare TEL and WEB. The resulting mode effect estimates ($\hat{\delta}^{ft}_{2,l}$, $\hat{\delta}^{fw}_{2,l}$, $\hat{\delta}^{tw}_{2,l}$) and their variances ($V(\hat{\delta}^{ft}_{2,l})$,$V(\hat{\delta}^{fw}_{2,l})$,$V(\hat{\delta}^{tw}_{2,l})$) for an imputed dataset l are computed using 4.7, where $\hat{\delta}^{mm^*}_{2,h,l}$ is the stratum h mode effects (m indexes one mode, m^* indexes another mode ($m \neq m^*$)). Note that $\hat{N}_{2,hl}$ in 4.7 are the estimated population size for principal strata h in imputed dataset l and it is design-based estimates based on the achieved sample size per principal stratum ($\hat{N}_{2,hl} = \sum_i^{n_{2,hl}} w_{2,hli}$, where $w_{2,hli}$ is the HRS 2018 individual weight). Because the PS are defined using predicted response indicators, the achieved sample size per stratum differs across imputed datasets and outcomes.

$$\begin{aligned}
\hat{\delta}^{ft}_{2,l} &= \frac{\hat{N}_{2,h1,l}\hat{\delta}^{ft}_{2,h1,l} + \hat{N}_{2,h2,l}\hat{\delta}^{ft}_{2,h2,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h2,l}}, V(\hat{\delta}^{ft}_{2,l}) = (\frac{\hat{N}_{2,h1,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h2,l}})^2 V(\hat{\delta}^{ft}_{2,h1,l}) + (\frac{\hat{N}_{2,h2,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h2,l}})^2 V(\hat{\delta}^{ft}_{2,h2,l}) \\
\hat{\delta}^{fw}_{2,l} &= \frac{\hat{N}_{2,h1,l}\hat{\delta}^{fw}_{2,h1,l} + \hat{N}_{2,h3,l}\hat{\delta}^{fw}_{2,h3,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h3,l}}, V(\hat{\delta}^{fw}_{2,l}) = (\frac{\hat{N}_{2,h1,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h3,l}})^2 V(\hat{\delta}^{fw}_{2,h1,l}) + (\frac{\hat{N}_{2,h3,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h3,l}})^2 V(\hat{\delta}^{fw}_{2,h3,l}) \\
\hat{\delta}^{tw}_{2,l} &= \frac{\hat{N}_{2,h1,l}\hat{\delta}^{tw}_{2,h1,l} + \hat{N}_{2,h4,l}\hat{\delta}^{tw}_{2,h4,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h4,l}}, V(\hat{\delta}^{tw}_{2,l}) = (\frac{\hat{N}_{2,h1,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h4,l}})^2 V(\hat{\delta}^{tw}_{2,h1,l}) + (\frac{\hat{N}_{2,h4,l}}{\hat{N}_{2,h1,l} + \hat{N}_{2,h4,l}})^2 V(\hat{\delta}^{tw}_{2,h4,l})
\end{aligned} \tag{4.7}$$

Next, we apply paired-T tests to the mode effect estimates, in a similar approach as illustrated in the HRS 2016 analysis.

Table 4.3: Principal Strata considered for HRS 2018

Principal Strata	Conditions	Imputation Models	F vs T	F vs W	T vs W
1) F & T & W	$\boldsymbol{R}_2^f = 1, \boldsymbol{R}_2^t = 1, \boldsymbol{R}_2^w = 1$	$f(\boldsymbol{Y}_{2,h1}^f \vert \boldsymbol{X}_{2,h1}), f(\boldsymbol{Y}_{2,h1}^t \vert \boldsymbol{X}_{2,h1}), f(\boldsymbol{Y}_{2,h1}^w \vert \boldsymbol{X}_{2,h1})$	✓	✓	✓
2) F & T only	$\boldsymbol{R}_2^f = 1, \boldsymbol{R}_2^t = 1, \boldsymbol{R}_2^w = 0$	$f(\boldsymbol{Y}_{2,h2}^f \vert \boldsymbol{X}_{2,h2}), f(\boldsymbol{Y}_{2,h2}^t \vert \boldsymbol{X}_{2,h2})$	✓		
3) F & W only	$\boldsymbol{R}_2^f = 1, \boldsymbol{R}_2^t = 0, \boldsymbol{R}_2^w = 1$	$f(\boldsymbol{Y}_{2,h3}^f \vert \boldsymbol{X}_{2,h3}), f(\boldsymbol{Y}_{2,h3}^w \vert \boldsymbol{X}_{2,h3})$		✓	
4) T & W only	$\boldsymbol{R}_2^f = 0, \boldsymbol{R}_2^t = 1, \boldsymbol{R}_2^w = 1$	$f(\boldsymbol{Y}_{2,h4}^t \vert \boldsymbol{X}_{2,h4}), f(\boldsymbol{Y}_{2,h4}^w \vert \boldsymbol{X}_{2,h4})$			✓
5) F only	$\boldsymbol{R}_2^f = 1, \boldsymbol{R}_2^t = 0, \boldsymbol{R}_2^w = 0$	$f(\boldsymbol{Y}_{2,h5}^f \vert \boldsymbol{X}_{2,h5})$			
6) T only	$\boldsymbol{R}_2^f = 0, \boldsymbol{R}_2^t = 1, \boldsymbol{R}_2^w = 0$	$f(\boldsymbol{Y}_{2,h6}^t \vert \boldsymbol{X}_{2,h6})$			
7) W only	$\boldsymbol{R}_2^f = 0, \boldsymbol{R}_2^t = 0, \boldsymbol{R}_2^w = 1$	$f(\boldsymbol{Y}_{2,h7}^w \vert \boldsymbol{X}_{2,h7})$			

Notes: F stands for FTF, T means TEL, and W indicates WEB.

4.2.4 Longitudinal Analysis

For the 16-18 longitudinal analysis, we focus on analyzing the potential response indicators and the potential outcomes for the FTF and the TEL in both time points. Similarly, for the imputation, we consider the joint distribution of these parameters conditional on covariates measured at both waves (4.8).

$$\begin{aligned}
f(R_1^f, R_1^t, R_2^f, R_2^t, Y_1^f, Y_1^t, Y_2^f, Y_2^t | X_1, X_2) \doteq\ & f(R_1^f|X_1, X_2) f(R_1^t|X_1, X_2) f(R_2^f|X_1, X_2, R_1^f, R_1^t) \times \\
& f(R_2^t|X_1, X_2, R_1^f, R_1^t) \times \\
& f(Y_1^f|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t) \times \\
& f(Y_1^t|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t) \times \\
& f(Y_2^f|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t, Y_1^f, Y_1^t) \times \\
& f(Y_2^t|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t, Y_1^f, Y_1^t)
\end{aligned} \tag{4.8}$$

$$\begin{aligned}
\doteq\ & f(R_1^f|X_1, X_2) f(R_1^t|X_1, X_2) f(R_2^f|X_1, X_2) \times \\
& f(R_2^t|X_1, X_2) f(Y_1^f|X_1, X_2, R_1^f, R_1^t) \times \\
& f(Y_1^t|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t) \times \\
& f(Y_2^f|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t) \times \\
& f(Y_2^t|X_1, X_2, R_1^f, R_1^t, R_2^f, R_2^t)
\end{aligned} \tag{4.9}$$

Due to the HRS crossover design, we cannot estimate the correlation between Y_2^m and Y_1^m, where m=FTF or TEL. To disentangle mode effects from time effects, we remove $Y_1^{m'}$ when writing the conditional distribution for Y_1^m, where $m \neq m'$. Therefore, we write 4.8 as 4.9. Depending on whether respondents would participate via FTF and (or) TEL in the two waves, we categorize the respondents into 9 PS and specify corresponding imputation models within each stratum. Table 4.4 summarizes the PS and illustrates which strata are necessary for specific comparisons. Again, the imputation of the unobserved $R_1^f, R_1^t, R_2^f, R_2^t, Y_1^f, Y_2^f, Y_1^t$, and Y_2^2 proceed as in 2.3.1.

With the longitudinal design, we can make five comparisons: 1) FTF and TEL mode effects for the 2016 (F VS T 16), 2) FTF and TEL mode effects for the 2018 (F VS T 18), 3) FTF and TEL mode effects 16-18 time difference (FT 16 VS 18), 4) FTF 16-18 time difference (F 16 VS 18), and 5) TEL 16-18 time difference (T 16 VS 18). These estimates can be obtained by combining the relevant strata (see 4.4), except for the mode effects across time (comparison 3), which can only be estimated using stratum 1.

We combine the strata following a similar approach as described in 4.7, using the estimated population size of the stratum as weights. In this case, the estimated population size should incorporate the probability that a HRS panelist is in both the 2016 and the 2018 wave. To address this, we initially compute nonresponse weights ($w_{nonresp}$) as the inverse of the 2018 response propensity conditioned on the response in the HRS 2016. The nonresponse propensity model incorporates the 2016 survey weight and a set of demographic, socio-economic status, and health-related variables measured in the HRS 2016. Specifically, the covariates include age, income, wealth, number of health conditions, functional limitation measures (mobile, muscles, activity of daily life, instrumental activity of daily life, number of words recalled, and BMI). We then compute the weight for the longitudinal analysis as $w_{1618} = w_{nonresp} \times w_{16}$ to account for the probability that a HRS panelist participate to both the 2016 and 2018 waves.

Finally, we apply the combining rules as described earlier to get the longitudinal estimates across the imputed datasets and use the paired t-test to examine whether the differences in the changes between FTF and TEL are significant.

Table 4.4: Principal Strata considered for HRS 2016-2018

Principal Strata	R_1^f	R_1^t	R_2^f	R_2^t	Imputation Models	F VS T 16	F VS T 18	FT 16 VS 18	F 16 VS 18	T 16 VS 18
1) FT1618	1	1	1	1	$f(Y_{1,h1}^f \mid X_{1,h1}), f(Y_{1,h1}^t \mid X_{1,h1}),$ $f(Y_{2,h1}^f \mid X_{2,h1}), f(Y_{2,h1}^t \mid X_{2,h1})$	✓	✓	✓	✓	✓
2) FT16,T18	1	1	0	1	$f(Y_{1,h2}^f \mid X_{1,h2}), f(Y_{1,h2}^t \mid X_{1,h2}),$ $f(Y_{2,h2}^t \mid X_{2,h2})$	✓				✓
3) FT16,T18	1	1	1	0	$f(Y_{1,h3}^f \mid X_{1,h3}), f(Y_{1,h3}^t \mid X_{1,h3}),$ $f(Y_{2,h3}^f \mid X_{2,h3})$	✓			✓	
4) F16,FT18	1	0	1	1	$f(Y_{1,h4}^f \mid X_{1,h4}),$ $f(Y_{2,h4}^f \mid X_{2,h4}), f(Y_{2,h4}^t \mid X_{2,h4})$		✓		✓	
5) T16,FT18	0	1	1	1	$f(Y_{1,h5}^t \mid X_{1,h5}),$ $f(Y_{2,h5}^f \mid X_{2,h5}), f(Y_{2,h5}^t \mid X_{2,h5})$		✓			✓
6) T16,T18	0	1	0	1	$f(Y_{1,h6}^t \mid X_{1,h6}),$ $f(Y_{2,h6}^t \mid X_{2,h6})$					✓
7) F16,F18	1	0	1	0	$f(Y_{1,h7}^f \mid X_{1,h7}), f(Y_{2,h7}^f \mid X_{2,h7})$				✓	
8) F16,T18	1	0	0	1	$f(Y_{1,h8}^f \mid X_{1,h8}), f(Y_{2,h8}^t \mid X_{2,h8})$					
9) T16,F18	0	1	1	0	$f(Y_{1,h9}^t \mid X_{1,h9}), f(Y_{2,h9}^f \mid X_{2,h9})$					

Notes: "F" stands for FTF, "T" means TEL, and "W" indicates WEB. "F VS T 16" denotes the mode comparison between FTF and TEL using the HRS 2016, represented by $Y_1^f - Y_1^t$. Similarly, "F VS T 18" refers to the mode effects for the HRS 2018. "FT 16 VS 18" indicates the difference between the 2016 mode effects and the 2018 effects. "F 16 VS 18" describes the time effect between the HRS 2016 and the HRS 2018 using the FTF mode. "T 16 VS 18" outlines the time effect between the HRS 2016 and the HRS 2018 using the TEL mode. For the "Principal Strata" column: If a respondent can participate via both FTF and TEL in the 2016 and 2018 waves, they belong to the "FT1618"stratum. If they are predicted to be a non-respondent in some mode for some wave, the naming of the stratum omits the mode and wave information. For example, a respondent predicted as a FTF non-respondent in the 2018 wave and anticipated to participate via FTF in the 2016 wave and TEL in both waves would be categorized under the "FT16, T18" stratum.

4.3 Results

In this section, we present the results for the 2016, 2018, and the 16-18 analyses, respectively. For each analysis, we begin by presenting the stratum-level estimates of the outcome variables and the comparisons between modes. Next, we present stratum-level information, including the sample size and the fraction of missing information for each outcome. Last, we show the mode effect estimates combining the strata and imputed datasets.

4.3.1 The HRS 2016

Table 4.5 summarizes the mode effect estimates for each stratum, mode, and outcome. Significant mode effects were observed for two items: the number of words recalled and self-reported health. The TEL mode was associated with better word recall ability and improved self-reported health, after controlling for individual-level covariates and principal strata. The enhanced cognitive function observed in interviews conducted via TEL may be attributed to the absence of interviewers in front of the respondents, allowing them to take notes or use various forms of cues to facilitate recall. From this perspective, estimates for cognitive function produced from FTF interviews can be considered closer to the truth. Regarding self-reported health, this finding aligns with previous research indicating that respondents tend to provide more socially desirable responses when interviewed via TEL [22]. However, a similar pattern was not found for depression, another potentially sensitive item.

Table 4.5 also suggests substantial mode selection effects, as the estimates of the population mean in different principal strata vary greatly, often exceeding the differences within each stratum. For instance, respondents who can only respond in the FTF mode were 10% more likely to be depressed and 15% less likely to report very good or excellent health compared to respondents who could respond in both the FTF and TEL modes. By comparing the results row-wise and column-wise in the table, we can disentangle the mode selection effects from the measurement effects. While participants who can only join via FTF differ substantially in all examined outcomes from those who can respond in both modes, participants who can only respond via TEL provide estimates closer to the group that responds in both modes.

Table 4.5: Mode Effect Estimates for the HRS 2016

Stratum	FTF	TEL	FT mode effects
		Number of Words Recalled	
FTF&TEL	10.413 [0.071]	10.619 [0.072]	-0.206 [0.079] **
FTF only	7.798 [0.276]		
TEL only		10.651 [0.207]	
		Depression	
FTF&TEL	0.125 [0.006]	0.123 [0.007]	0.002 [0.009]
FTF only	0.229 [0.029]		
TEL only		0.131 [0.02]	
		BMI	
FTF&TEL	28.591 [0.146]	28.356 [0.119]	0.235 [0.18]
FTF only	31.049 [0.721]		
TEL only		31.06 [0.491]	
		Self-reported Health	
FTF&TEL	0.42 [0.01]	0.459 [0.01]	-0.039 [0.013] **
FTF only	0.264 [0.032]		
TEL only		0.415 [0.029]	

Notes: $p < 0.001$, ***. $p < 0.01$, **. $p < 0.05$, *.

Table 4.6 presents the fraction of missing information (FMI) for the estimates and the achieved sample sizes. Since the potential response indicators are imputed jointly with the outcomes, these indicators may vary depending on the outcomes. For the four outcomes considered, approximately 85% of the sample belongs to Stratum 1, where participants can respond via both modes; the second-largest stratum is the TEL only stratum, followed by the FTF only stratum. The FMI approximately ranges from 0.3 to 0.5 for the first and fourth outcomes, and from 0.4 to 0.6 for the second and third outcomes. To reflect the uncertainty in the stratum membership, we note the range of sample sizes within each stratum in Table 4.6. The stratum-specific sample sizes vary by outcome, as we re-impute the response indicators for each outcome, taking into account the correlation between potential outcomes and response indicators. Additionally, we provide the actual number of respondents participating via the FTF and TEL modes, indicating the respective contributions of data collected through each mode to the estimation.

4.3.2 The HRS 2018

We present the stratum estimates for the HRS 2018 in Table 4.7. First, we observe that among respondents who are able to participate through all three modes (h1), there is a significant decrease (6.5%, 8.9%, $p < 0.001$) in depression when using WEB reporting, compared to the FTF and TEL modes. Additionally, respondents recall fewer words in the TEL mode compared to the WEB mode (-0.453, $p < 0.05$). In strata where participants can only respond via limited modes, estimates vary substantially between the WEB mode and other modes (see Strata 3 and 4). For example, in Stratum 4, where panelists can only respond via TEL and WEB, respondents recall fewer words, report lower BMI, and are more likely to report depression in the WEB mode than in TEL. According to Table 4.8, on average, fewer than 25 actual WEB respondents are in Stratum 3 and fewer than 100 are in Stratum 4, making the WEB estimates in these strata less reliable.

Moreover, the FTF and TEL modes show differences in the number of words recalled (Stratum 1) and the proportion of depression (Stratum 2). In Stratum 1, where respondents can use all modes, the FTF interviews yield slightly better recall. In Stratum 2, where only the FTF and TEL are possible (excluding the WEB), the FTF is associated with lower proportions of depression. Although not significant, we note that the direction of the mode effects between FTF and TEL is reversed in three outcomes. Since Strata 1 and 2 differ only in whether respondents can participate via the WEB, we speculate that individuals in Stratum 2 may systematically differ from those in Stratum 1 in demographic or socioeconomic characteristics, which might explain their reluctance to use the WEB mode. The mixed results underscore the heterogeneity of mode effects among various subgroups and emphasize the importance of considering stratification.

Table 4.8 reveals that the FMI is quite large in the imputation of the HRS 2018. This is largely attributed to the high number of cases requiring imputation; for instance, to derive estimates in WEB mode, we must impute data for nearly 90% of the sample, given the small fraction of respondents assigned to and actually responding via WEB. Furthermore, the FMI varies across outcomes and modes, possibly due to the unequal sample sizes of different modes and the varying predictive power of our imputation models across outcomes and modes. Despite using the same set of covariates, their associations with mode-specific outcomes can differ. In addition, we find that the sample sizes of the principal strata vary substantially between outcomes. For example, more samples are categorized in Stratum 1 for the number of words recalled and BMI, compared to depression and self-reported health, and fewer samples are categorized in Stratum 2 for BMI.

After combining the strata, we observe in Table 4.9 that, compared to the WEB mode, the FTF mode is associated with more cases of depression (7%, $p < 0.001$), and the TEL mode is linked to a higher BMI (1.358, $p < 0.001$) and to recalling more words (0.47, $p < 0.01$). These findings contradict the prior belief that self-administered modes (e.g., WEB) reduce socially desirable reporting compared to interviewer-administered modes [10, 11], and challenge literature that finds WEB respondents perform better on cognitive tests than those using FTF and TEL modes [68, 58]. Some differences reported in the stratum-specific analysis (e.g., differences in cognitive performance) diminished after combining the relevant strata. This once again highlights the importance of considering heterogeneity in mode effects across different strata. Overall, the results of the HRS 2018 analysis suggest minimal mode effects between FTF and TEL and a few differences between FTF and WEB, and between TEL and WEB.

Table 4.6: Fraction of Missing Information and Sample Size for the Principal Strata in the HRS 2016

Stratum	FTF FMI	TEL FMI	FT FMI	Sample Size	FTF Rs	TEL Rs
			Number of Words Recalled			
FTF&TEL	0.432	0.372	0.492	9677 [9571,9787]	4735 [4666,4806]	4942 [4847,5033]
FTF only	0.28			670 [599,739]	670 [599,739]	
TEL only		0.368		1036 [945,1131]		1036 [945,1131]
			Depression			
FTF&TEL	0.493	0.443	0.526	9614 [9478,9727]	4715 [4634,4781]	4900 [4786,5009]
FTF only	0.376			690 [624,771]	690 [624,771]	
TEL only		0.44		1078 [969,1192]		1078 [969,1192]
			BMI			
FTF&TEL	0.441	0.547	0.57	9751 [9518,9924]	4771 [4669,4839]	4980 [4829,5094]
FTF only	0.451			634 [566,736]	634 [566,736]	
TEL only		0.429		998 [884,1149]		998 [884,1149]
			Self-reported Health			
FTF&TEL	0.423	0.387	0.517	9615 [9499,9722]	4711 [4618,4786]	4905 [4802,5002]
FTF only	0.423			694 [619,787]	694 [619,787]	
TEL only		0.352		1073 [976,1176]		1073 [976,1176]

Notes: “FTF Rs” means the number of respondents actually responding in FTF, whereas “TEL Rs” means the number of respondents actually responding in TEL. To reflect the uncertainty in the sample size and the number of respondents who actually responded in the FTF and TEL surveys across the imputed datasets, we provided the range of sample sizes in brackets.

Table 4.7: Stratum-specific Mode Effect Estimates for the HRS 2018

Stratum	FTF	TEL	WEB	F VS T	F VS W	T VS W
			Number of Words Recalled			
h1:FTW	11.773 [0.072]	11.465 [0.095]	11.701 [0.12]	0.308 [0.12] *	0.072 [0.142]	-0.236 [0.144]
h2:FT	10.525 [0.357]	11.338 [0.325]		-0.812 [0.434]		
h3:FW	10.618 [0.685]		2.248 [0.516]		8.37 [0.904] ***	
h4:TW		11.582 [0.317]	0.833 [0.109]			10.749 [0.327] ***
h5:F	9.467 [1.802]					
h6:T		8.698 [0.794]				
h7:W			12.064 [2.218]			
			Depression			
h1:FTW	0.113 [0.009]	0.089 [0.013]	0.024 [0.007]	0.024 [0.016]	0.089 [0.012] ***	0.065 [0.013] ***
h2:FT	0.059 [0.024]	0.133 [0.028]		-0.074 [0.035] *		
h3:FW	0.092 [0.052]		0.877 [0.12]		-0.785 [0.142] ***	
h4:TW		0.084 [0.028]	0.735 [0.093]			-0.651 [0.102] ***
h5:F	0.714 [0.190]					
h6:T		0.487 [0.104]				
h7:W			0.441 [0.435]			

			BMI			
h1:FTW	28.885 [0.172]	28.535 [0.238]	28.479 [0.268]	0.351 [0.226]	0.406 [0.356]	0.056 [0.404]
h2:FT	25.993 [3.075]	30.384 [1.095]		-4.391 [2.619]		
h3:FW	30.635 [1.754]		13.096 [2.754]		17.54 [3.532] ***	
h4:TW		30.107 [0.606]	9.008 [0.727]			21.099 [0.953] ***
h5:F	3.881 [7.042]					
h6:T		26.534 [1.533]				
h7:W			0.076 [1.086]			
			Self-reported Health			
h1:FTW	0.512 [0.022]	0.498 [0.022]	0.516 [0.024]	0.013 [0.032]	-0.004 [0.027]	-0.017 [0.033]
h2:FT	0.466 [0.069]	0.401 [0.044]		0.064 [0.082]		
h3:FW	0.551 [0.099]		0.962 [0.052]		-0.411 [0.117] ***	
h4:TW		0.583 [0.046]	0.837 [0.126]			-0.254 [0.138]
h5:F	0.396 [0.174]					
h6:T		0.227 [0.081]				
h7:W			0.075 [0.248]			

Table 4.8: Fraction of Missing Information and Sample Size for the Principal Strata in the HRS 2018

Stratum	FTF FMI	TEL FMI	WEB FMI	FT FMI	FW FMI	TW FMI	Sample size	FTF Rs	TEL Rs	WEB Rs
Number of Words Recalled										
h1:FTW	0.492	0.734	0.663	0.712	0.655	0.672	5504 [5291,5661]	2809 [2657,2910]	1636 [1541,1717]	1060 [1047,1070]
h2:FT	0.783	0.629		0.639			940 [798,1150]	442 [337,589]	498 [433,576]	
h3:FW	0.393		0.729		0.522		101 [68,137]	95 [67,128]		6 [0,18]
h4:TW		0.292	0.762			0.304	413 [376,468]		395 [359,448]	18 [9,26]
h5:F	0.568						17 [4,36]	17 [4,36]		
h6:T		0.451					67 [39,103]		67 [39,103]	
h7:W			0.43				2 [1,6]			2 [1,6]
Depression										
h1:FTW	0.604	0.826	0.847	0.781	0.722	0.817	4923 [4698,5158]	2528 [2377,2659]	1413 [1303,1528]	982 [951,1008]
h2:FT	0.895	0.832		0.84			1354 [1148,1590]	697 [556,851]	657 [563,767]	
h3:FW	0.641		0.903		0.881		123 [82,173]	101 [68,146]		21 [7,38]
h4:TW		0.684	0.904			0.896	465 [394,530]		384 [320,429]	81 [56,114]

h5:F	0.728						37 [20,62]	37 [20,62]		
h6:T		0.695					140 [105,184]		140 [105,184]	
h7:W			0.8				2 [0,9]			2 [0,9]
BMI										
h1:FTW	0.675	0.825	0.771	0.675	0.82	0.866	5811 [5554,6061]	3028 [2895,3164]	1710 [1569,1839]	1073 [1051,1083]
h2:FT	0.932	0.699		0.84			661 [444,889]	237 [96,367]	424 [338,542]	
h3:FW	0.48		0.693		0.676		102 [79,137]	98 [78,135]		4 [0,15]
h4:TW		0.444	0.547			0.5	416 [372,469]		407 [360,455]	9 [1,23]
h5:F	0.999						0 [0,3]	0 [0,3]		
h6:T		0.69					54 [14,93]		54 [14,93]	
h7:W			1				0 [0,1]			0 [0,1]
Self-reported Health										
h1:FTW	0.813	0.788	0.815	0.819	0.778	0.839	4817 [4561,5033]	2462 [2284,2596]	1376 [1271,1481]	979 [934,1017]
h2:FT	0.93	0.82		0.909			1459 [1250,1725]	761 [625,939]	697 [605,819]	
h3:FW	0.625		0.835		0.712		120 [80,168]	101 [71,137]		20 [6,40]

h4:TW		0.596	0.971			0.93	457 [387,538]		371 [318,431]	85 [48,127]
h5:F	0.456						39 [18,60]	39 [18,60]		
h6:T		0.672					150 [104,192]		150 [104,192]	
h7:W			0.779				2 [0,7]			2 [0,7]

Table 4.9: Combined Mode Effect Estimates for the HRS 2018

Comparison	FTF	TEL	WEB	Differences
		Number of Words Recalled		
FT	11.614 [0.066]	11.45 [0.086]		0.164 [0.104]
FW	11.755 [0.071]		11.551 [0.12]	0.204 [0.141]
TW		11.473 [0.09]	11.003 [0.114]	0.47 [0.137] **
		Depression		
FT	0.103 [0.007]	0.098 [0.009]		0.005 [0.011]
FW	0.113 [0.009]		0.043 [0.007]	0.07 [0.011] ***
TW		0.089 [0.012]	0.081 [0.01]	0.008 [0.013]
		BMI		
FT	28.658 [0.184]	28.716 [0.186]		-0.058 [0.283]
FW	28.913 [0.175]		28.241 [0.279]	0.672 [0.375]
TW		28.632 [0.223]	27.274 [0.252]	1.358 [0.371] ***
		Self-reported Health		
FT	0.502 [0.011]	0.478 [0.015]		0.024 [0.019]
FW	0.513 [0.022]		0.526 [0.023]	-0.014 [0.027]
TW		0.505 [0.02]	0.542 [0.019]	-0.037 [0.027]

4.3.3 The HRS 2016-2018

We first illustrate the stratum-specific mode effects, time effects, and mode- and time-specific population means for four outcomes in Table 4.10.

Based on Table 4.10, we first note increasing mode effects on number of words recalled (1.091, $p < 0.001$) and self-reported health (9.3%, $p < 0.01$), suggesting that mode effects have become larger from HRS 2016 to HRS 2018 for these two items. The increase in the number of words recalled is largely driven by the estimates from the FTF mode (0.937, $p < 0.001$). For the self-reported health, the observed trend is mostly due to poorer health outcomes over time as estimated from the TEL mode (-6.6%, $p < 0.001$). The time differences in mode effects are considerably large, highlighting the importance of such analysis, especially given the minimal FTF-TEL differences found in HRS 2018. Second, for respondents who can only respond to FTF or TEL in a particular wave, their estimates can be very different from those of respondents predicted to be able to respond to both FTF and TEL modes. Specifically, the respondents in strata (“FT16,T18”, “FT16,F18”, “T16,FT18”, and “F16,FT18”) recall significantly fewer words and have much higher proportions of depression than those in “FT1618”; again, this points to potential selection effects.

Table 4.11 summarizes the combined estimates for three comparisons: changes in FTF and TEL mode effects from HRS 2016 to HRS 2018 (“FT1618”), changes in the FTF estimates from 2016 to 2018 (“F1618”), and changes in the TEL estimates from 2016 to 2018 (“T1618”). The first comparison is only supported by the first principal stratum illustrated in Table 4.10; therefore, we again observe increasing mode effects on the number of words recalled and self-reported health. For the BMI measure, we see an increasing BMI across waves in the FTF mode while the trend is not significant in the TEL mode. This might suggest that respondents report BMI more conscientiously in FTF interviews than in TEL interviews, since the interviewer can observe their physical shape, making FTF a better mode for collecting BMI information. For self-reported health, the proportion of respondents reporting very good or excellent health significantly decreases in the TEL mode, whereas the pattern is reversed in the FTF interviews (though not significantly). This leads to increasing mode differences between the FTF and TEL from HRS 2016 to HRS 2018. We provide the stratum-specific FMI and sample size table (Table 4.12) in the appendix.

In summary, we note that time effects can have different directions and magnitudes in FTF and TEL interviews, highlighting the importance of examining mode-specific time effects. However, due to the crossover design in the HRS, it may not be possible to distinguish between time effects and mode effects.

Table 4.10: Stratum-specific Mode Effect Estimates for the HRS 16-18

Stratum	F16	T16	F18	T18	FT1618	F1618	T1618
Number of Words Recalled							
FT1618	10.065 [0.105]	10.811 [0.104]	11.002 [0.087]	10.657 [0.148]	1.091 [0.207] ***	0.937 [0.141] ***	-0.154 [0.162]
FT16,T18	7.321 [0.364]	10.167 [0.242]		10.294 [0.235]			0.128 [0.27]
FT16,T18	7.362 [0.396]	7.68 [0.328]	8.786 [0.289]			1.425 [0.425] **	
F16,FT18	7.637 [0.267]		8.354 [0.293]	4.57 [0.374]		0.717 [0.308] *	
T16,FT18		9.409 [0.266]	5.119 [0.483]	9.765 [0.346]			0.356 [0.357]
T16,T18		6.654 [1.05]		7.532 [1.103]			0.878 [0.875]
F16,F18	3.472 [1.205]		2.327 [1.283]			-1.144 [1.934]	
F16,T18	6.796 [0.908]			7.314 [0.693]			
T16,F18		4.479 [1.388]	5.452 [1.512]				
Depression							
FT1618	0.08 [0.009]	0.069 [0.007]	0.068 [0.007]	0.068 [0.007]	-0.011 [0.016]	-0.012 [0.011]	-0.001 [0.01]
FT16,T18	0.624 [0.093]	0.259 [0.058]		0.326 [0.062]			0.066 [0.055]

FT16,T18	0.452 [0.073]	0.976 [0.028]	0.785 [0.057]			0.333 [0.074] ***	
F16,FT18	0.57 [0.056]		0.363 [0.059]	0.946 [0.06]		-0.208 [0.072] **	
T16,FT18		0.284 [0.074]	0.415 [0.116]	0.223 [0.07]			-0.061 [0.066]
T16,T18		0.054 [0.02]		0.068 [0.021]			0.014 [0.024]
F16,F18	0.034 [0.025]		0.036 [0.024]			0.003 [0.034]	
F16,T18	0.158 [0.088]			0.142 [0.096]			
T16,F18		0.493 [0.441]	0.062 [0.191]				
			BMI				
FT1618	28.363 [0.264]	28.114 [0.217]	29.032 [0.124]	28.716 [0.241]	0.067 [0.359]	0.669 [0.282] *	0.602 [0.339]
FT16,T18	9.865 [0.721]	27.796 [0.47]		27.995 [0.422]			0.198 [0.258]
FT16,T18	28.016 [0.864]	8.12 [1.035]	27.89 [0.802]			-0.126 [0.335]	
F16,FT18	28.179 [0.657]		28.231 [0.686]	9.895 [0.91]		0.052 [0.267]	
T16,FT18		28.662 [0.474]	4.538 [0.565]	28.325 [0.44]			-0.337 [0.172]

T16,T18		28.704 [2.809]		29.206 [3.723]			0.501 [1.39]
F16,F18	24.488 [1.995]		25.919 [2.724]			1.431 [1.755]	
F16,T18	26.645 [1.444]			27.035 [1.115]			
T16,F18		25.382 [2.286]	26.222 [2.393]				
			Self-reported Health				
FT1618	0.402 [0.014]	0.453 [0.012]	0.43 [0.012]	0.387 [0.015]	0.093 [0.029] **	0.028 [0.016]	-0.066 [0.018] ***
FT16,T18	0.313 [0.089]	0.428 [0.047]		0.397 [0.047]			-0.031 [0.046]
FT16,T18	0.55 [0.083]	0.713 [0.134]	0.575 [0.081]			0.025 [0.055]	
F16,FT18	0.419 [0.071]		0.41 [0.068]	0.801 [0.168]		-0.009 [0.061]	
T16,FT18		0.395 [0.059]	0.418 [0.079]	0.386 [0.058]			-0.009 [0.058]
T16,T18		0.383 [0.064]		0.423 [0.061]			0.039 [0.054]
F16,F18	0.053 [0.027]		0.008 [0.005]			-0.046 [0.027]	
F16,T18	0.173 [0.094]			0.253 [0.155]			
T16,F18		0.717 [0.328]	0.308 [0.365]				

Table 4.11: Combined Mode Effect Estimates for the HRS 16-18

Comparison	F16	T16	F18	T18	Differences
			Number of Words Recalled		
FT1618	10.065 [0.105]	10.811 [0.104]	11.002 [0.087]	10.657 [0.148]	1.091 [0.207] ***
F1618	9.851 [0.098]		10.801 [0.079]		0.95 [0.129] ***
T1618		10.691 [0.091]		10.582 [0.128]	-0.109 [0.145]
			Depression		
FT1618	0.08 [0.009]	0.069 [0.007]	0.068 [0.007]	0.068 [0.007]	-0.011 [0.016]
F1618	0.109 [0.008]		0.107 [0.006]		-0.003 [0.01]
T1618		0.083 [0.006]		0.083 [0.007]	0 [0.009]
			BMI		
FT1618	28.363 [0.264]	28.114 [0.217]	29.032 [0.124]	28.716 [0.241]	0.067 [0.359]
F1618	28.341 [0.252]		28.98 [0.12]		0.639 [0.268] *
T1618		28.115 [0.201]		28.669 [0.223]	0.554 [0.312]
			Self-reported Health		
FT1618	0.402 [0.014]	0.453 [0.012]	0.43 [0.012]	0.387 [0.015]	0.093 [0.029] **
F1618	0.398 [0.013]		0.424 [0.011]		0.025 [0.015]
T1618		0.447 [0.011]		0.389 [0.013]	-0.058 [0.015] ***

4.4 Discussion

This paper explores the application of principal stratification in detecting mode effects within a longitudinal study, where there are different mixed-mode designs across waves. The HRS serves as the context for this study, yet the methods proposed are applicable to other longitudinal surveys under certain conditions. Specifically, some randomized designs are needed to allow for the imputation of potential response indicators. The key to this approach is to consider the counterfactual scenario of what would happen if an individual were invited to participate in the alternate modes. Depending on whether they would respond in these hypothetical scenarios, individuals are categorized into different subgroups to estimate the mode effects within each group. To predict these hypothetical scenarios, it is essential to gather data from a similar population. This data helps estimate the relationships between potential outcomes and covariates, allowing us to then extrapolate these relationships to individuals whose potential outcomes need to be imputed. In this paper, the presence of the TEL-only group, as a reference group to the sequential WEB-TEL mixed-mode group, along with the random assignment between FTF and TEL modes, provide the necessary data for such imputation in the cross-sectional analyses. Survey researchers should consider this design aspect during the initial mixed-mode plan phase if they intend to evaluate the mode effects in their designs afterward.

Evidence from the 2016 HRS data suggests that the TEL mode is associated with more words and better self-reported health than the FTF mode. For the 2018 HRS analysis, we observe significant mode effects between WEB and FTF or TEL in three out of four items examined. Specifically, respondents recalled fewer words and had lower BMIs in the WEB mode than in TEL interviews, and reported less depression in the WEB mode than in FTF interviews. Note that our finding contradicts prior literature, which finds that WEB respondents perform better in cognitive tests than those using FTF and TEL modes [68, 58]. This discrepancy may be attributable to the multiple strategies we apply to account for potential selection effects. In this study, we 1) apply sample inclusion criteria to only include samples eligible for random assignment, 2) use principal stratification to group respondents who would have responded to the same modes, and 3) add individual-level covariates to partly explain differential nonresponse. Moreover, our findings suggest that membership in principal stratification can be strongly related to outcomes; therefore, the stratification should be considered an important approach to controlling for selection effects.

In addition, our findings from the HRS 2018 challenge the previous finding that self-administered modes facilitate more honest reporting [55, 11, 10]. One reason may be we considered only experienced panelists who are not newly admitted to the panel and might have already developed trust towards interviewers, thereby being more willing to provide candid responses in the FTF or TEL mode.

Concerning the longitudinal analysis, the HRS crossover design is not optimal for examining FTF and TEL mode effects, since time effects are confounded with mode effects by design. If data permits, we can account for the correlation between different waves by including Y_1^f as a predictor when predicting Y_2^f and the same applies for Y_1^t and Y_2^t. With the current design, the correlation between Y_1^f and Y_1^f cannot be estimated except for the defiers who are nonrespondents of the originally assigned mode. Without adjusting for individual correlations, we may risk overestimating the time effects.

Nevertheless, the categorization of respondents into different principal strata also have applications in the design and implementation of longitudinal mixed-mode surveys. By determining whether a panelist would respond to each mode, survey agencies can optimize mode assignments for panelists, starting with the most cost-effective option. For example, for respondents who are willing to respond via FTF, TEL, and WEB, the WEB can be designated as the initial mode in the sequential mixed-mode design. For those who can only respond via TEL and FTF, TEL can be the first mode attempted. By tailoring sequential mixed-mode designs to each panelist's response propensity, it is possible to reduce field time, save on survey costs, and decrease sample attrition rates.

For this approach to be successful, it is crucial to have highly predictive models that can accurately impute potential response indicators and outcomes. In this study, we apply model diagnostic checks to aid in model selection and also use a large number of imputation iterations to propagate the uncertainty within the approach. For future work, it may be beneficial to use mixed-effects models to leverage strength across principal strata when estimating stratum-specific parameters. This could be particularly advantageous for estimating parameters within strata of small sample sizes. Additionally, exploring modern modeling techniques, such as machine learning models, in the imputation process could also be worthwhile.

This study has three additional limitations. First, although predictive modeling is generally more efficient for estimating treatment effects compared to the weighting approach

[77], it requires making distributional assumptions, which may limit the applicability of this approach. Second, in this study, we only consider mode effects as biases to the means, neglecting other mode effects, such as primacy or recency effects. These would require considering other modeling approaches and utilizing questions with multiple categories. Third, we do not adopt a fully Bayesian approach in the study. During the imputation stage, we consider the data as a simple random sample and only account for the complex survey designs during the analysis stage. For future work, synthetic populations using Bayesian finite population bootstrap methods can be used to support more congenial inference.

4.5 Appendix

4.5.1 Longitudinal Analysis Supplementary Tables

Table 4.12: Information about the Principal Strata for the HRS 16-18

	FMI							Sample Size				
Stratum	F16	T16	F18	T18	FT1618	F1618	T1618	n	F16	T16	F18	T18
Number of Words Recalled												
FT1618	0.644	0.474	0.489	0.737	0.544	0.721	0.604	6289[35]	2537[19]	3752[33]	3945[34]	2344[19]
FT16,T18	0.547	0.28		0.157			0.331	516[15]	172[15]	344[3]		516[15]
FT16,T18	0.447	0.368	0.356			0.478		391[27]	181[4]	210[26]	391[27]	
F16,FT18	0.151		0.284	0.593		0.335		353[12]	353[12]		261[3]	92[11]
T16,FT18		0.355	0.78	0.448			0.374	403[24]		403[24]	168[24]	234[2]
T16,T18		0.164		0.278			0.277	20[4]		20[4]		20[4]
F16,F18	0.524		0.574			0.524		5[2]	5[2]		5[2]	
F16,T18	0.165			0.382				36[4]	36[4]			36[4]
T16,F18		0.629	0.572					9[3]		9[3]	9[3]	
Depression												
FT1618	0.72	0.532	0.512	0.678	0.663	0.66	0.634	6268[50]	2533[24]	3736[33]	3930[34]	2339[23]
FT16,T18	0.842	0.646		0.657			0.384	299[35]	137[17]	163[21]		299[35]
FT16,T18	0.698	0.833	0.666			0.615		324[24]	84[7]	240[19]	324[24]	
F16,FT18	0.424		0.553	0.892		0.391		270[17]	270[17]		133[9]	137[13]
T16,FT18		0.628	0.859	0.647			0.662	261[34]		261[34]	162[22]	99[15]
T16,T18		0.372		0.428			0.396	337[33]		337[33]		337[33]
F16,F18	0.516		0.323			0.44		229[11]	229[11]		229[11]	
F16,T18	0.373			0.367				30[3]	30[3]			30[3]
T16,F18		0.216	0.812					3[1]		3[1]	3[1]	
BMI												
FT1618	0.65	0.691	0.264	0.7	0.83	0.903	0.842	6898[7]	2785[6]	4113[3]	4323[4]	2575[6]
FT16,T18	0.536	0.168		0.058			0.336	380[5]	32[5]	348[1]		380[5]
FT16,T18	0.223	0.512	0.163			0.321		163[8]	158[7]	5[2]	163[8]	
F16,FT18	0.066		0.105	0.514		0.253		264[6]	264[6]		251[4]	13[4]
T16,FT18		0.031	0.65	0.037			0.064	239[2]		239[2]	2[2]	237[1]
T16,T18		0.173		0.122			0.095	13[2]		13[2]		13[2]

F16,F18	0.572		0.541			0.514		22[10]	22[10]		22[10]	
F16,T18	0.05			0.113				23[1]	23[1]			23[1]
T16,F18		0.523	0.534					20[3]		20[3]	20[3]	
Self-reported Health												
FT1618	0.594	0.309	0.304	0.622	0.599	0.646	0.6	5826[43]	2445[24]	3381[36]	3574[37]	2252[24]
FT16,T18	0.88	0.457		0.448			0.432	393[20]	185[18]	208[7]		393[20]
FT16,T18	0.557	0.892	0.608			0.494		286[49]	80[5]	206[48]	286[49]	
F16,FT18	0.516		0.418	0.961		0.409		294[21]	294[21]		122[9]	172[16]
T16,FT18		0.613	0.794	0.622			0.609	690[50]		690[50]	539[47]	151[5]
T16,T18		0.549		0.434			0.514	240[10]		240[10]		240[10]
F16,F18	0.386		0.279			0.386		245[11]	245[11]		245[11]	
F16,T18	0.405			0.158				33[4]	33[4]			33[4]
T16,F18		0.759	0.301					14[5]		14[5]	14[5]	

Table 4.13: Fraction of Missing Information the Combined Principal Strata for the HRS 16-18

comparison	F16 FMI	T16 FMI	F18 FMI	T18 FMI	Difference FMI
			Number of words recalled		
FT16	0.635	0.286			0.513
FT18			0.218	0.453	0.637
FT1618	0.61	0.288	0.267	0.5	0.704
F1618	0.586		0.269		0.721
T1618		0.262		0.477	0.665
			Depression		
FT16	0.483	0.268			0.434
FT18			0.282	0.588	0.528
FT1618	0.524	0.501	0.308	0.601	0.489
F1618	0.488		0.242		0.507
T1618		0.464		0.566	0.547
			BMI		
FT16	0.724	0.436			0.602
FT18			0.483	0.738	0.687
FT1618	0.717	0.434	0.539	0.748	0.661
F1618	0.723		0.543		0.706
T1618		0.426		0.742	0.649
			Self-reported Health		
FT16	0.538	0.269			0.521
FT18			0.249	0.567	0.567
FT1618	0.594	0.309	0.304	0.622	0.599
F1618	0.578		0.279		0.637
T1618		0.257		0.587	0.575

CHAPTER 5

Conclusion and Future Work

The three studies in the dissertation are closely related. To make inferences from mixed-mode studies, researchers first need to understand the bias and variance properties of the multiple modes used in a survey. Chapter 2 develops approaches to account for potential mode effects when making inferences from data collected in mixed-mode designs. Chapter 3 focuses on a less understood area—interviewer variances in mixed-mode designs—and explores whether they are consistent across different modes. Chapter 4 applies principal stratification to estimate the bias of multiple modes in a longitudinal mixed-mode study where participants were assigned to a given mode but might choose to respond under a different mode.

Since each mode has its distinct measurement properties, mixing modes can introduce mode-specific biases. To account for the potential mode effects when analyzing data collected with mixed-mode designs, we propose three approaches in Chapter 2: the Testimator, the Bayesian, and the model averaging approaches. Through a simulation study and an application using real survey data, Chapter 2 demonstrates that with a modest sample size and prior information about preferred directions, the three proposed approaches can account for bias due to mode when the direction of the mode bias is assumed known, and appropriately inflate confidence or credible intervals when this direction is not assumed known. We achieve this by embedding a testing procedure within the approaches or by using a model averaging approach to average across all possible models. We illustrate the approaches with randomized mixed-mode data; however, combining with methods to account for mode selection effects, these approaches can be applied to any mixed-mode designs.

Interviewer variance, which can arise when responses from respondents interviewed by the same interviewers are more similar than those collected from other respondents, is a well-documented source of uncertainty in survey estimates [47, 39]. Despite its importance, no research to date has investigated the source of error in mixed-mode surveys. Chapter 3 employs hierarchical models to estimate mode-specific interviewer variances while incorporating the design of interviewer-mode assignment, specifically whether interviewers are nested within modes. Although the power of this analysis is usually limited, we are still able to detect differential interviewer variances for a few items. The variation in interviewer variance may indicate items that require further refinement when administered on instruments, or suggest that interviewers need additional training on a specific mode. This variation can also have implications for sample size allocation as it affects the mode-specific precision. Thus, our findings underscore the importance of incorporating such analysis into survey data quality check routines.

In studies where participants are randomized to a mode, there can be non-compliance – participants may respond using a mode different from the one to which they were randomized. Consequently, evaluating mode effects based on the observed conditions, specifically the actual mode used, does not yield reliable estimates of mode effects. Chapter 4 employs principal stratification to explicitly allow mode comparisons only within individuals who would respond to certain modes. Principal stratification creates strata based on the joint distribution of indicators that describe whether or not a subject would respond if invited to a given mode, thus generate a (latent) pre-treatment variable that can be conditioned on to give causal mode effects under certain assumptions. While researchers are aware of mode selection effects and utilize the selection properties in mixed-mode designs, prior literature typically estimates mode measurement effects either by using all sample members—assuming everyone can participate via all modes, albeit with varying propensities [26, 17] —or only within experimental and control groups (i.e., an "as treated" analysis [55, 58]. There is a lack of a framework to incorporate all sample units while using a model-based approach to isolate ineligible samples when estimating mode measurement effects. This study addresses this gap by presenting an application of principal stratification in two waves of the HRS.

Accounting for the design features when making inferences from the mixed-mode sample is a critical challenge throughout this dissertation. In Chapter 3, how interviewers are assigned to modes affects our modeling approaches. Additionally, the number of interviewers assigned to each mode strongly determines the power of the analysis, as shown by the simulation study. In Chapter 4, the success of the imputation strategy relies on the existence of randomized designs to provide data for imputing unobserved potential outcomes. Moreover, our findings emphasize that having a random sample allocated to the same modes across waves is critical to disentangling mode effects and time effects in longitudinal studies. From the perspective of survey data collection agencies, it is important to design surveys appropriately beforehand to facilitate future methodological investigations. For example, determining the sample sizes across modes in both cross-sectional and longitudinal studies is a critical design consideration that merits further examination.

In addition, controlling for mode selection effects when making inferences is a longstanding question. Although randomized designs can mitigate most selection effects, it is most feasible to apply such designs to a pre-planned small-scale sample. When estimating mode effects with large samples, including individual-level covariates in models [28] or using propensity-based approaches (such as matching, weighting, and prediction) [68, 53] remain common methods. However, determining which covariates should be used and how to ensure that mode selection effects are adequately isolated from the mode measurement effects remain open questions, since the covariates themselves must not be subject to mode measurement effects.

We propose four extensions to this dissertation. First, a natural extension would be to combine the inference strategies in Chapter 2 with the principal stratification and multiple imputation framework in Chapter 4, to provide inference tools in a broader context. Specifically, we can determine if there are mode effects within each stratum, decide which mode to use if substantial differences are present, and then combine estimates across the principal strata. This approach extends the three approaches to account for samples not in the randomized designs and offer the flexibility to work with longitudinal samples.

Second, we plan to investigate whether certain interviewer characteristics are associated with differential interviewer effects between the FTF and TEL modes. In the HRS analysis presented in Chapter 2, we found evidence of varying interviewer variances on a few sensitive questions and one interviewer observation item. Examining which interviewer characteristics can explain the differences in interviewer effects across modes may help tailor a more effective interviewer training strategy.

Third, the principal stratification and the multiple imputation framework used in Chapter 4 have the potential to improve future mixed-mode designs. For example, in adaptive designs, we can optimize mode allocation based on their principal stratum membership and whether they are susceptible to mode measurement effects (i.e., whether their mode-specific potential outcomes differ a lot). This strategy can potentially reduce survey field time and costs while improving data quality by minimizing mode measurement effects.

Lastly, this dissertation considers design-based inferences to account for complex survey designs, as discussed in Chapters 2 and 4. However, in multiple imputation, it is assumed that the sample is drawn under simple random sampling. To improve the congeniality between the analysis procedure and the imputation models, we can use the weighted finite population Bayesian bootstrap and incorporate the imputation strategies developed in Chapter 4 to create synthetic populations. Next, we can apply the three methods in Chapter 2 to provide population inference that accounts for mode effects and survey features.

In conclusion, this dissertation has made significant efforts in improving mixed-mode inferences, by providing principled approaches to detect mode effects and analyze mixed-mode data. As mixed-mode designs become increasingly prevalent and new data collection modes (e.g., biomarkers, social media data, administrative data) emerge, the development of methods to integrate data collected with multiple modes and designs—while accounting for their unique measurement and selection properties—will be of great value. This dissertation has the potential to significantly enhance survey practices and maximize the utility of data collected through mixed-mode designs, ultimately benefiting the quantitative research community and the broader population.

BIBLIOGRAPHY

[1] Douglas S Massey and Roger Tourangeau. Where do we go from here? nonresponse and social measurement. *The ANNALS of the American Academy of Political and Social Science*, 645(1):222–236, 2013.

[2] Scott Keeter, Carolyn Miller, Andrew Kohut, Robert M Groves, and Stanley Presser. Consequences of reducing nonresponse in a national telephone survey. *Public Opinion Quarterly*, 64(2):125–148, 2000.

[3] Scott Keeter, Courtney Kennedy, Michael Dimock, Jonathan Best, and Peyton Craighill. Gauging the impact of growing nonresponse on estimates from a national rdd telephone survey. *International Journal of Public Opinion Quarterly*, 70(5):759–779, 2006.

[4] Andy Peytchev, Rodney K Baxter, and Lisa R Carley-Baxter. Not all survey effort is equal: Reduction of nonresponse bias and nonresponse error. *Public Opinion Quarterly*, 73(4):785–806, 2009.

[5] Robert JJ Voogt and Willem E Saris. Mixed mode designs: Finding the balance between nonresponse bias and mode effects. *Journal of Official Statistics*, 21(3):367, 2005.

[6] Paul J Lavrakas. *Encyclopedia of Survey Research Methods*. Sage publications, 2008.

[7] Edith Desiree De Leeuw. *Data Quality in Mail, Telephone and Face to Face Surveys.* ERIC, 1992.

[8] Norbert Schwarz, Hans-J Hippler, Brigitte Deutsch, and Fritz Strack. Response scales: Effects of category range on reported behavior and comparative judgments. *Public Opinion Quarterly*, 49(3):388–395, 1985.

[9] Don A Dillman and Leah Melani Christian. Survey mode as a source of instability in responses across surveys. *Field Methods*, 17(1):30–52, 2005.

[10] Roger Tourangeau and Tom W Smith. Asking sensitive questions: The impact of data collection mode, question format, and question context. *Public Opinion Quarterly*, 60(2):275–304, 1996.

[11] Frauke Kreuter, Stanley Presser, and Roger Tourangeau. Social desirability bias in cati, ivr, and web surveys the effects of mode and question sensitivity. *Public Opinion Quarterly*, 72(5):847–865, 2008.

[12] Stanley Presser and Linda Stinson. Data collection mode and social desirability bias in self-reported religious attendance. *American Sociological Review*, pages 137–145, 1998.

[13] Allyson L Holbrook and Jon A Krosnick. Social desirability bias in voter turnout reports: Tests using the item count technique. *Public Opinion Quarterly*, 74(1):37–67, 2010.

[14] Edith D de Leeuw, Z Tuba Suzer-Gurtekin, and Joop J Hox. The design and implementation of mixed-mode surveys. *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, pages 387–409, 2018.

[15] Jorre Vannieuwenhuyze, Geert Loosveldt, and Geert Molenberghs. A method for evaluating mode effects in mixed-mode surveys. *Public Opinion Quarterly*, 74(5):1027–1045, 2010.

[16] Jorre TA Vannieuwenhuyze and Melanie Revilla. Relative mode effects on data quality in mixed-mode surveys by an instrumental variable. In *Survey Research Methods*, volume 7, pages 157–168, 2013.

[17] Z Tuba Suzer-Gurtekin, S Heeringa, and R Vaillant. Investigating the bias of alternative statistical inference methods in sequential mixed-mode surveys. *Proceedings of the JSM, Section on Survey Research Methods*, pages 4711–2, 2012.

[18] Edith D DeLeeuw. Mixed-mode: Past, present, and future. In *Survey Research Methods*, volume 12, pages 75–89, 2018.

[19] Institute for Social Research. Panel study of income dynamics, child development supplement 2020: User guide. Technical report, Institute for Social Research, University of Michigan, 2022.

[20] University of Michigan. Health and retirement study, 2020 hrs final core public use dataset. Produced and distributed by the University of Michigan with funding from the National Institute on Aging (grant number NIA U01AG009740), 2024.

[21] American National Election Studies. Anes 2020 time series study full release [dataset and documentation], 2021. February 10, 2022 version. Available at: www.electionstudies.org.

[22] Allyson L Holbrook, Melanie C Green, and Jon A Krosnick. 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(1):79–125, 2003.

[23] Thomas Klausch, Joop J Hox, and Barry Schouten. Measurement effects of survey mode on the equivalence of attitudinal rating scale questions. *Sociological Methods & Research*, 42(3):227–263, 2013.

[24] Bart Buelens and Jan A van den Brakel. Measurement error calibration in mixed-mode sample surveys. *Sociological Methods & Research*, 44(3):391–426, 2015.

[25] J Michael Brick, Courtney Kennedy, Ismael Cervantes-Flores, and Andrew W Mercer. An adaptive mode adjustment for multimode household surveys. *Journal of Survey Statistics and Methodology*, 2021.

[26] Stanislav Kolenikov and Courtney Kennedy. Evaluating three approaches to statistically adjust for mode effects. *Journal of Survey Statistics and Methodology*, 2(2):126–158, 2014.

[27] Jennifer Robyn Powers, G Mishra, and Anne F Young. 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(2):149–154, 2005.

[28] Marc N Elliott, Alan M Zaslavsky, Elizabeth Goldstein, William Lehrman, Katrin Hambarsoomians, Megan K Beckett, and Laura Giordano. Effects of survey mode, patient mix, and nonresponse on cahps® hospital survey scores. *Health Services Research*, 44(2p1):501–518, 2009.

[29] Seunghwan Park, Jae Kwang Kim, and Sangun Park. An imputation approach for handling mixed-mode surveys. *The Annals of Applied Statistics*, 10(2):1063–1085, 2016.

[30] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical Science*, 14(4):382–417, 1999.

[31] Saralees Nadarajah and Samuel Kotz. Exact distribution of the max/min of two gaussian random variables. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 16(2):210–212, 2008.

[32] Darrel Robinson and Marcus Tannenberg. Self-censorship of regime support in authoritarian states: Evidence from list experiments in china. *Research & Politics*, 6(3):2053168019856449, 2019.

[33] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, 1(3):515–534, 2006.

[34] Brady T West and Annelies G Blom. Explaining interviewer effects: A research synthesis. *Journal of Survey Statistics and Methodology*, 5(2):175–211, 2017.

[35] Howard Schuman and Jean M Converse. The effects of black and white interviewers on black responses in 1968. *Public Opinion Quarterly*, 35(1):44–68, 1971.

[36] Robert H Hanson and Eli S Marks. Influence of the interviewer on the accuracy of survey results. *Journal of the American Statistical Association*, 53(283):635–655, 1958.

[37] June Sachar Ehrlich and David Riesman. Age and authority in the interview. *Public Opinion Quarterly*, pages 39–56, 1961.

[38] Rainer Schnell and Frauke Kreuter. Separating interviewer and sampling-point effects. 2003.

[39] Leslie Kish. Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, 57(297):92–115, 1962.

[40] Robert M Groves and Lou J Magilavy. Measuring and explaining interviewer effects in centralized telephone surveys. *Public Opinion Quarterly*, 50(2):251–266, 1986.

[41] Clyde Tucker. Interviewer effects in telephone surveys. *Public Opinion Quarterly*, 47(1):84–95, 1983.

[42] Robert M Groves and Robert L Kahn. Surveys by telephone; a national comparison with personal interviews. 1979.

[43] Nora Cate Schaeffer, Jennifer Dykema, and Douglas W Maynard. Interviewers and interviewing. *Handbook of Survey Research*, 2:437–471, 2010.

[44] Z Tuba Suzer-Gurtekin, Steven G Heeringa, and Richard Valliant. Investigating the bias of alternative statistical inference methods in mixed-mode surveys. *Proceedings of the JSM, Section on Survey Research Methods*, 2013.

[45] Jorre TA Vannieuwenhuyze. Mode effects on variances, covariances, standard deviations, and correlations. *Journal of Survey Statistics and Methodology*, 3(3):296–316, 2015.

[46] Brady T West, Frauke Kreuter, and Ursula Jaenichen. "interviewer" effects in face-to-face surveys: A function of sampling, measurement error, or nonresponse? *Journal of Official Statistics*, 29(2):277–297, 2013.

[47] Brady T West and Kristen Olson. How much of interviewer variance is really nonresponse error variance? *Public Opinion Quarterly*, 74(5):1004–1026, 2010.

[48] Gwenith G Fisher and Lindsay H Ryan. Overview of the health and retirement study and introduction to the special issue. *Work, Aging and Retirement*, 4(1):1–9, 2018.

[49] Steven G Heeringa, Judith H Connor, et al. Technical description of the health and retirement survey sample design. *Ann Arbor: University of Michigan*, 1995.

[50] Staff HRS. Hrs core interview sample sizes and response rates. Technical report, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, 2023. Available online.

[51] Brady T West, Ai Rene Ong, Frederick G Conrad, Michael F Schober, Kallan M Larsen, and Andrew L Hupp. Interviewer effects in live video and prerecorded video interviewing. *Journal of Survey Statistics and Methodology*, 10(2):317–336, 2022.

[52] Marieke Voorpostel, Oliver Lipps, and Caroline Roberts. Mixing modes in household panel surveys: Recent developments and new findings. *Advances in Longitudinal Survey Methodology*, pages 204–226, 2021.

[53] Alexandru Cernat and Joseph W Sakshaug. Estimating the measurement effects of mixed modes in longitudinal studies: Current practice and issues. *Advances in Longitudinal Survey Methodology*, pages 227–249, 2021.

[54] Joop Hox, Edith De Leeuw, and Thomas Klausch. Mixed mode research: Issues in design and analysis. *Total Survey Error in Practice*, pages 511–530, 2017.

[55] Alexandru Cernat, Mick P Couper, and Mary Beth Ofstedal. Estimation of mode effects in the health and retirement study using measurement models. *Journal of Survey Statistics and Methodology*, 4(4):501–524, 2016.

[56] Alexandru Cernat. The impact of mixing modes on reliability in longitudinal studies. *Sociological Methods & Research*, 44(3):427–457, 2015.

[57] Alexandru Cernat. Impact of mixed modes on measurement errors and estimates of change in panel data. In *Survey Research Methods*, volume 9, pages 83–99, 2015.

[58] Mary Beth Ofstedal, Gábor Kézdi, and Mick P Couper. Data quality and response distributions in a mixed-mode survey. *Longitudinal and Life Course Studies*, pages 1–26, 2022.

[59] Alexandru Cernat and Joseph W Sakshaug. The impact of mixing survey modes on estimates of change: A quasi-experimental study. *Journal of Survey Statistics and Methodology*, 11(5):1110–1132, 2023.

[60] Bart Buelens and Jan A Van den Brakel. Comparing two inferential approaches to handling measurement error in mixed-mode surveys. *Journal of Official Statistics*, 33(2):513–531, 2017.

[61] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[62] Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

[63] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[64] William S Aquilino. Effects of interview mode on measuring depression in younger adults. *Journal of Official Statistics*, 14(1):15, 1998.

[65] Sandra W Geerlings, Aartjan TF Beekman, Dorly JH Deeg, Willem Van Tilburg, and Jan H Smit. The center for epidemiologic studies depression scale (ces-d) in a mixed-mode repeated measurements design: sex and age effects in older adults. *International Journal of Methods in Psychiatric Research*, 8(2):102–109, 1999.

[66] Holly F Levin-Aspenson and David Watson. Mode of administration effects in psychopathology assessment: Analyses of gender, age, and education differences in self-rated versus interview-based depression. *Psychological Assessment*, 30(3):287, 2018.

[67] Yves Beland and Martin St-Pierre. Mode effects in the canadian community health survey: a comparison of cati and capi. *Advances in telephone survey methodology*, pages 297–314, 2008.

[68] Benjamin W Domingue, Ryan J McCammon, Brady T West, Kenneth M Langa, David R Weir, and Jessica Faul. The mode effect of web-based surveying on the 2018 us health and retirement study measure of cognitive functioning. *The Journals of Gerontology: Series B*, 78(9):1466–1473, 2023.

[69] Mary Beth Ofstedal, Colleen A McClain, and Mick P Couper. Measuring cognition in a multi-mode context. *Advances in Longitudinal Survey Methodology*, pages 250–271, 2021.

[70] Tarek Al Baghal. The effect of online and mixed-mode measurement of cognitive ability. *Social Science Computer Review*, 37(1):89–103, 2019.

[71] Stef van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[72] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[73] Wenshan Yu, Michael Elliott, and Trivellore Raghunathan. mice: Multivariate imputation by chained equations in r. *Journal of Survey Statistics and Methodology*, 2024.

[74] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[75] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[76] Joshua Angrist and Guido Imbens. Identification and estimation of local average treatment effects, 1995.

[77] Tingting Zhou, Michael R Elliott, and Roderick JA Little. Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association*, 114(525):1–19, 2019.