Electronic Journal of Statistics
Vol. 17 (2023) 1492–1546
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2137

Pretest estimation in combining probability and non-probability samples*

Chenyin Gao and Shu Yang

Department of Statistics, North Carolina State University, Raleigh, NC 27695, USA
e-mail: cgao6@ncsu.edu; syang24@ncsu.edu

Abstract: Multiple heterogeneous data sources are becoming increasingly available for statistical analyses in the era of big data. As an important example in finite-population inference, we develop a unified framework of the test-and-pool approach to general parameter estimation by combining gold-standard probability and non-probability samples. We focus on the case when the study variable is observed in both datasets for estimating the target parameters, and each contains other auxiliary variables. Utilizing the probability design, we conduct a pretest procedure to determine the comparability of the non-probability data with the probability data and decide whether or not to leverage the non-probability data in a pooled analysis. When the probability and non-probability data are comparable, our approach combines both data for efficient estimation. Otherwise, we retain only the probability data for estimation. We also characterize the asymptotic distribution of the proposed test-and-pool estimator under a local alternative and provide a data-adaptive procedure to select the critical tuning parameters that target the smallest mean square error of the test-and-pool estimator. Lastly, to deal with the non-regularity of the test-and-pool estimator, we construct a robust confidence interval that has a good finite-sample coverage property.

MSC2020 subject classifications: Primary 62D05; secondary 62E20, 62F03, 62F35.
Keywords and phrases: Data integration, dynamic borrowing, non-regularity, Pretest estimator.

Received October 2022.

Contents

*Yang's research is partially supported by the NIH 1R01AG066883 and 1R01ES031651.

1. Introduction

It has been widely accepted that probability sampling, where each selected sample is treated as a representative sample to the target population, is the best vehicle for finite-population inference. Since the sampling mechanism is known based on survey design, each weight-calibrated sample can be used to obtain consistent estimators for the target population; see [53], [15] and [24] for textbook discussions. However, complex and ambitious surveys are facing more and more hurdles and concerns recently, such as costly intervention strategies and lower participation rates. [2] address some of the current challenges in using probability samples for finite-population inference. On the other hand, higher demands of small area estimation and other more factors have led researchers to seek out alternative data collection with less program budget [69, 28]. In particular, lots of attention has been drawn to the studies of non-probability samples.

Non-probability samples are sets of selected objects where the sampling mechanism is unknown. First of all, non-probability samples are readily available from many data sources, such as satellite information [36], mobile sensor data [40], and web survey panels [62]. In addition, these non-representative samples are far more cost-effective compared to probability samples and have the potential of providing estimates in near real-time, unlike the traditional inferences derived from probability samples [42]. Based on these big and easy-accessible data, a wealth of literature has been proposed which enunciates the bright future while properly utilizing such amount of data (e.g., 18, 14, 61, and 41).

However, the naive use of such data cannot ensure the statistical validity of the resulting estimators because such non-probability samples are often selected without sophisticated supervision. Therefore, the acquisition of large whereas highly unrepresentative data is likely to produce erroneous conclusions. [17] and [23] present more recent examples where non-probability samples can often lead to estimates with significant selection biases. To overcome these challenges, it is essential to establish appropriate statistical tools to draw valid inferences when integrating data from the probability and non-probability samples. Various data integration methods have been proposed in the literature to leverage the unique strengths of the probability and non-probability samples; see [73] for a review, and the existing methods for data integration can be categorized into three types including the inverse propensity score adjustment [50, 22], calibration weighting [19, 31], and mass imputation [46, 30, 74, 11].

But most of the works assume that the non-probability sample is comparable to the probability sample in terms of estimating the finite-population parameters, which may not be satisfied in many applications due to the unknown sampling mechanism of the non-probability samples. Thus, the non-probability samples with unknown sampling mechanisms may bias the estimators for the target parameters. To resolve this issue, [47] propose a pretest to gauge the statistical adequacy of integrating the probability and non-probability samples in an application. The pretesting procedure has been broadly practiced in econometrics and medicine, and its implications are of considerable interests (e.g., [68, 63, 3, 72]). Essentially, the final value of the estimator depends on the outcome of a random testing event and therefore is a stochastic mixture of two different estimators. Despite the long history of the application of the pretest, few literature investigates the theoretical properties of the underlying non-smooth distribution for the pretest estimators.

In this paper, we establish a general statistical framework for the test-and-pool analysis of the probability and non-probability samples by constructing a test to gauge the comparability of the non-probability data and decide whether or not to use non-probability data in a pooled analysis. In addition, we consider the null, fixed, and local alternative hypotheses for the pre-testing, representing different levels of comparability of the non-probability data with the probability data. In particular, the non-probability sample is perfectly comparable under the null hypothesis, whereas it is starkly incomparable under the fixed alternative. Therefore, the fixed alternative cannot adequately capture the finite-sample behavior of the pre-testing estimator, under which the test statistic will diverge to infinity as the sample size increases. Toward this end, we establish the asymptotic distribution of the proposed estimator under local alternatives, which provides a better approximation of the finite-sample behavior of the pretest estimator when the idealistic assumption required for the non-probability data is weakly violated. Also, we provide a data-adaptive procedure to select the optimal values of the tuning parameters achieving the smallest mean square error of the pretest estimator. Lastly, we construct a robust confidence interval accounting for the non-regularity of the estimator, which has a valid coverage property.

The rest of the paper is organized as follows. Section 2 lays out the basic setup and presents an efficient estimator for combing the non-probability sample and the probability sample. Section 3 proposes a test statistic and the test-and-pool estimator. In Section 4, we present the asymptotic properties of the test-and-pool estimator, an adaptive inference procedure, and lastly a data-adaptive selection scheme of the tuning parameters. Section 5 presents a simulation study to evaluate the performance of our test-and-pool estimator. Section 6 provides a real-data illustration. All proofs are given in the Appendix.

2. Basic setup

2.1. Notation: two data sources

Let $\mathcal{F}_N = \{V_i = (X_i^\intercal, Y_i)^\intercal : i \in U\}$ with $U = \{1, \ldots, N\}$ denote a finite population of size N, where X_i is a vector of covariates and Y_i is the study variable. We assume that F_N is a random sample from a superpopulation model ζ and our objective is to estimate the finite-population parameter $\mu_g \in \mathbb{R}^l$, defined as the solution to

$$\frac{1}{N}\sum_{i=1}^{N} S(V_i; \mu) = 0, \tag{2.1}$$

where $S(V_i; \mu)$ is a l-dimensional estimating function. The class of parameters is fairly general. For example, if $S(V; \mu) = Y - \mu$, $\mu_g = \overline{Y}_N = N^{-1}\sum_{i=1}^{N} Y_i$ is the population mean of Y_i. If $S(V; \mu) = \mathbf{1}(Y < c) - \mu$ for some constant c, where $\mathbf{1}(\cdot)$ is an indicator function, $\mu_g = N^{-1}\sum_{i=1}^{N} \mathbf{1}(Y_i < c)$ is the population proportion of Y_i less than c. If $S(V; \mu) = X(Y - X^\intercal\mu)$, $\mu_g = (\sum_{i=1}^{N} X_i X_i^\intercal)^{-1}(\sum_{i=1}^{N} X_i Y_i)$ is the coefficient of the finite-population regression projection of Y_i onto X_i.

Suppose that there are two data sources, one from a probability sample, referred to as Sample A, and the other from a non-probability sample, referred to as Sample B. Assume Sample A to be independent of Sample B, and the observed units can be envisioned as being generated through two phases of sampling [12]. Firstly, a superpopulation model ζ generates the finite population $\mathcal{F}_N$. Then, the probability (or non-probability) sample is selected from it using some known (or unknown) sampling schemes. Hence, the considered total variance of estimators is based on the randomness induced by both the superpopulation model and the sampling mechanisms; see Table 1 for the notations of probability order, expectation and (co-)variance. For example, $\mathbb{E}_{\mathrm{p}}(\cdot \mid \mathcal{F}_N)$ is the average over all possible samples under the probability design for particular finite population $\mathcal{F}_N$, and $\mathbb{E}(\cdot)$ is the average over all possible samples from all possible finite populations.

Table 1
Notation and definitions.

Randomness	order notation	expectation	(co-)variance
probability design	$o_{\mathrm{p}}(1), O_{\mathrm{p}}(1)$	$\mathbb{E}_{\mathrm{p}}\left(\cdot \mid \mathcal{F}_N\right)$	$\operatorname{var}_{\mathrm{p}}\left(\cdot \mid \mathcal{F}_N\right), \operatorname{cov}_{\mathrm{p}}\left(\cdot \mid \mathcal{F}_N\right)$
non-probability design	$o_{\mathrm{np}}(1), O_{\mathrm{np}}(1)$	$\mathbb{E}_{\mathrm{np}}\left(\cdot \mid \mathcal{F}_N\right)$	$\operatorname{var}_{\mathrm{np}}\left(\cdot \mid \mathcal{F}_N\right), \operatorname{cov}_{\mathrm{np}}\left(\cdot \mid \mathcal{F}_N\right)$
ζ model	$o_{\zeta}(1), O_{\zeta}(1)$	$\mathbb{E}_{\zeta}\left(\cdot\right)$	$\operatorname{var}_{\zeta}\left(\cdot\right), \operatorname{cov}_{\zeta}\left(\cdot\right)$
total variance	$o_{\zeta\text{-p-np}}(1), O_{\zeta\text{-p-np}}(1)$	$\mathbb{E}\left(\cdot\right)$	$\operatorname{var}\left(\cdot\right), \operatorname{cov}\left(\cdot\right)$

Thus far, our focus has been on the setting where the covariates X and the study variable Y are available in both the probability and non-probability samples, which has also been considered in [21] and [20]. The sampling indicators are denoted by $\delta_{A,i}$ and $\delta_{B,i}$, respectively; e.g., $\delta_{A,i} = 1$ if unit i is selected into Sample A and zero otherwise. Sample A contains observations $\mathcal{O}_A = \{(d_i = \pi_{A,i}^{-1}, X_i, Y_i) : i \in \mathcal{A}\}$ with sample size n_A, where $\pi_{A,i}$ is the known first-order inclusion probability for Sample A, and Sample B contains observations $\mathcal{O}_B = \{(X_i, Y_i) : i \in \mathcal{B}\}$ with sample size n_B. The unknown propensity score for being selected into Sample B is denoted by $\pi_{B,i}$. Here, $\mathcal{A}$ and $\mathcal{B}$ denote the indexes of units in Samples A and B with total sample size $n = n_A + n_B$ and negligible sampling fractions, i.e., $n/N = o(1)$. Let the limits of the fractions of Sample A and B be $f_A = \lim_{n\to\infty} n_A/n$ and $f_B = \lim_{n\to\infty} n_B/n$ with $0 < f_A, f_B < 1$.

2.2. Assumptions and separate estimators

As observing (X_i, Y_i) for all units i in U is usually not feasible in practice, we can estimate the population estimating equation (2.1) by the design-weighted sample analog under the probability sampling design

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\delta_{A,i}}{\pi_{A,i}}S(V_i;\mu) = 0, \tag{2.2}$$

yielding a design-weighted Z-estimator $\widehat{\mu}_A$ [65]. When $S(V;\mu)$ is a score function, the resulting estimator will be a pseudo maximum likelihood estimator [58]. For example, for estimating $\overline{Y}_N$, we have $S(V;\mu) = Y - \mu$, which leads to $\widehat{\mu}_A = (\sum_{i=1}^N \delta_{A,i}\pi_{A,i}^{-1})^{-1}\sum_{i=1}^N \delta_{A,i}\pi_{A,i}^{-1}Y_i$. We now make the following assumption for the design-weighted Z-estimator.

Assumption 2.1 (Design consistency and central limit theorem)**.** *Let $\widehat{\mu}_A$ be the corresponding design-weighted Z-estimator of μ_g, which satisfies that $\operatorname{var}_{\mathrm{p}}(\widehat{\mu}_A \mid \mathcal{F}_N) = O_\zeta(n_A^{-1})$ and $\{\operatorname{var}_{\mathrm{p}}(\widehat{\mu}_A)\}^{-1/2} \times (\widehat{\mu}_A - \mu_g) \mid \mathcal{F}_N \to \mathcal{N}(0,1)$ in distribution as $n_A \to \infty$.*

Under the typical regularity conditions [24], Assumption 2.1 holds for many common sampling designs such as probability proportional to size and stratified simple random sampling. Under Assumption 2.1, $\widehat{\mu}_A$ is design-consistent and does not rely on any modeling assumptions. This explains why the probability sampling has been the gold standard approach for finite-population inference, and we make this assumption throughout this article.

Let $f(Y \mid X)$ be the conditional density function of Y given X in the superpopulation model ζ, and let $f(X)$ and $f(X \mid \delta_B = 1)$ be the density function of X in the finite population and the non-probability sample, respectively. To correct for the selection bias of the non-probability sample, most of the existing literature considers the following assumptions [e.g., 46, 66, 12].

Assumption 2.2 (Common support and ignorability of sampling)**.** *(i) The vector of covariates X has a compact and convex support, with its density bounded and bounded away from zero. Also, there exist positive constants C_l and C_u such that $C_l \leq f(X)/f(X \mid \delta_B = 1) \leq C_u$ almost surely. (ii) Conditional on X, the density of Y in the non-probability sample follows the superpopulation model; i.e., $f(Y \mid X, \delta_B = 1) = f(Y \mid X)$. (iii) The sample inclusion indicator $\delta_{B,i}$ and $\delta_{B,j}$ are independent given X_i and X_j for $i \neq j$.*

Assumption 2.2 (i) and (ii) constitute the strong sampling ignorability condition [50]. Assumption 2.2 (i) implies that the support of X in the non-probability sample is the same as that in the finite population, and it can also be formulated as a positivity assumption that $\mathbb{P}(\delta_B = 1 \mid X) > 0$ for all X. This assumption does not hold if certain units would never be included in the non-probability sample. Assumption 2.2 (ii) is equivalent to the ignorability of the sampling mechanism for the non-probability sample conditional on the covariates X, i.e., $\mathbb{P}(\delta_B = 1 \mid X, Y) = \mathbb{P}(\delta_B = 1 \mid X)$ [34]. This assumption holds if the set of covariates contain all the outcome predictors that affect the possibility of being selected into the non-probability sample. Assumption 2.2 (iii) is a critical condition to employ the weak law of large numbers under the non-probability sampling design [12]. Under Assumption 2.2, the non-probability sample can be used to produce consistent estimators. However, this assumption may be unrealistic if the non-probability data collection suffers from uncontrolled selection biases [6], measurement errors [17], or other error-prone issues. Thus, we consider Assumption 2.2 as an idealistic assumption, which may be violated and require pretesting.

Under Assumptions 2.1 and 2.2, let $\Phi_A(V, \delta_A; \mu)$ and $\Phi_B(V, \delta_A, \delta_B; \mu)$ be two l-dimensional estimating functions for the target parameter μ_g when using the probability sample and the combined samples, respectively. In practice, $\Phi_A(\cdot)$ and $\Phi_B(\cdot)$ may depend on unknown nuisance functions, and solving $\mathbb{E}\{\Phi_A(V, \delta_A; \mu)\} = 0$ and $\mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \mu)\} = 0$ is not feasible. By replacing the nuisance functions with their estimated counterparts, and the expectations with the empirical averages, we obtain $\widehat{\mu}_A$ and $\widehat{\mu}_B$ by solving

$$\frac{1}{N}\sum_{i=1}^{N} \widehat{\Phi}_A(V_i, \delta_{A,i}; \mu) = 0, \quad \frac{1}{N}\sum_{i=1}^{N} \widehat{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu) = 0, \tag{2.3}$$

respectively, where $\{\widehat{\Phi}_A(\cdot), \widehat{\Phi}_B(\cdot)\}$ are the estimated version of $\{\Phi_A(\cdot), \Phi_B(\cdot)\}$.

Remark 2.1. *For estimating the finite population means, that is, $\mu_g = \overline{Y}_N$, $\Phi_A(\cdot)$ and $\Phi_B(\cdot)$ are commonly chosen as*

$$\Phi_A(V, \delta_A; \mu) = \frac{\delta_A}{\pi_A}(Y - \mu), \tag{2.4}$$

$$\Phi_B(V, \delta_A, \delta_B; \mu) = \frac{\delta_B}{\pi_B(X)} \{Y - m(X)\} + \frac{\delta_A}{\pi_A} m(X) - \mu, \tag{2.5}$$

where $\pi_B(X) = \mathbb{P}(\delta_B = 1 \mid X)$ *and* $m(X) = \mathbb{E}(Y \mid X, \delta_B = 1)$. *To obtain the estimators* $\widehat{\mu}_A$ *and* $\widehat{\mu}_B$, *parametric models* $\pi_B(X; \alpha)$ *and* $m(X; \beta)$ *can be posited for the nuisance functions* $\pi_B(X)$ *and* $m(X)$, *respectively.*

In addition, researchers might be interested in estimating the individual-level outcomes rather than the population-level outcomes. In this case, $\Phi_A(\cdot)$ *and* $\Phi_B(\cdot)$ *can be specified for estimating the outcome model* $m(X; \beta)$ *as:*

$$\Phi_A(V, \delta_A; \beta) = \frac{\delta_A}{\pi_A} \frac{\partial m(X; \beta)}{\partial \beta} \{Y - m(X; \beta)\}$$
$$\Phi_B(V, \delta_A, \delta_B; \beta) = \left(\frac{\delta_A}{\pi_A} + \frac{\delta_B}{\pi_B(X)} \right) \frac{\partial m(X; \beta)}{\partial \beta} \{Y - m(X; \beta)\}.$$

Next, we adopt the model-design-based framework for inference, which incorporates the randomness over the two phases of sampling [27, 37, 7, 70]. The asymptotic properties for $\widehat{\mu}_A$ and $\widehat{\mu}_B$ can be derived using the standard M-estimation theory under suitable moment conditions.

Lemma 2.1. *Suppose Assumptions 2.1, 2.2 and additional regularity conditions A.1 hold. Then, we have*

$$n^{1/2} \begin{pmatrix} \widehat{\mu}_A - \mu_g \\ \widehat{\mu}_B - \mu_g \end{pmatrix} \to \mathcal{N} \left\{ \begin{pmatrix} 0_{l\times 1} \\ 0_{l\times 1} \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \right\}, \tag{2.6}$$

where V_A, V_B, *and* Γ *are defined explicitly in the Appendix.*

In Lemma 2.1, we extend the conditional normality to unconditional as in [55], which implies that the asymptotic (co-)variances terms V_A, V_B and Γ refer to all the sources of uncertainty over the two phases.

2.3. Efficient estimator

Under Assumptions 2.1 and 2.2, both $\widehat{\mu}_A$ and $\widehat{\mu}_B$ are consistent, and it is appealing to combine $\widehat{\mu}_A$ with $\widehat{\mu}_B$ to achieve efficient estimation. We consider a class of linear combinations of the functions in (2.3):

$$\sum_{i=1}^{N} \{\widehat{\Phi}_A(V_i, \delta_{A,i}; \mu) + \Lambda \widehat{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu)\} = 0, \tag{2.7}$$

where Λ is the linear coefficient that gauges how much information of the non-probability sample should be integrated with the probability sample. Equation (2.7) leads to a class of composite estimators which is a weighted average of $\widehat{\mu}_A$ and $\widehat{\mu}_B$ with Λ-indexed weight ω_A and ω_B. When $\Lambda = 0$, (2.7) provides the design-consistent estimator $\widehat{\mu}_A$. The optimal choice Λ_{eff} can be empirically tuned to minimize the asymptotic variance of the composite estimator, leading to the efficient estimator $\widehat{\mu}_{\text{eff}}$. However, the major concern for $\widehat{\mu}_{\text{eff}}$ is the possible bias due to the violation of Assumption 2.2 (ii) for the non-probability sample. When it is violated, it is reasonable to choose $\Lambda = 0$ and prevent any bias associated with the non-probability sample.

3. Test-and-pool estimator

Motivated by the above reasoning, we develop a strategy that pretests the comparability of the non-probability sample with the probability sample first and then decides whether or not we should combine them for efficient estimation. We formulate the hypothesis test in Section 3.1, and construct the test-and-pool estimator in Section 3.2.

3.1. Hypothesis and test

We formalize the null hypothesis H_0 when Assumption 2.2 holds, and the fixed and local alternatives H_a and $H_{a,n}$ when Assumption 2.2 is violated. To be specific, we consider

$$H_0 : \mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \mu_{g,0})\} = 0, \tag{3.1}$$

$$H_a : \mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \mu_{g,0})\} = \eta_{\text{fix}}, \tag{3.2}$$

$$H_{a,n} : \mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \mu_{g,0})\} = n_B^{-1/2}\eta, \tag{3.3}$$

where $\mu_{g,0} = \mathbb{E}_\zeta(\mu_g)$, $\mu_g = \mu_{g,0} + O_\zeta(N^{-1/2})$, and η_{fix}, η are two fixed parameters. The fixed alternative H_a is commonly considered in the standard hypothesis testing framework. However, it enforces the bias of the estimating function $\Phi_B(\cdot)$ to be fixed and indicates a strong violation of Assumption 2.2, under which the test statistic T will diverge to infinity with the sample size. Moreover, the inference under the fixed alternative can not capture the finite-sample behavior of the test well and lacks uniform validity. On the contrary, the local alternative provides a useful tool to study the finite-sample distribution of non-regular estimators when the signal of violation is weak, i.e., in the $n_B^{-1/2}$ neighborhood of zero. In such cases, we allow the existence of a set of unmeasured covariates whose association with either the possibility of being selected into Sample B or the outcome is small. Also, the local alternative $H_{a,n}$ is more general in the sense that it reduces to H_a with $\eta = \pm\infty$, and has been widely employed to illustrate the non-regularity settings, such as weak instrumental variables regression [59], regression estimators of weakly identified parameters [13] and test errors in classification [33]. We will mainly exploit the local alternative to show the inherent non-regularity of the pretest estimator.

Under the null hypothesis (3.1), $\widehat{\mu}_B$ is consistent, and hence, it is reasonable to combine $\widehat{\mu}_A$ and $\widehat{\mu}_B$ for efficient estimation. However, when the null hypothesis is violated as in (3.3), the efficient estimator is biased. Lemma 3.1 presents the asymptotic properties of the separate and efficient estimators under $H_{a,n}$.

Lemma 3.1. *Suppose Assumptions 2.1, 2.2 (i) and (iii), and all the regularity conditions in Lemma 2.1 hold. Then, under the local alternative $H_{a,n}$, the asymptotic distributions for $\widehat{\mu}_A$ and $\widehat{\mu}_B$ are*

$$\begin{aligned} & n^{1/2}\begin{pmatrix} \widehat{\mu}_A - \mu_g \\ \widehat{\mu}_B - \mu_g \end{pmatrix} \\ & \to N\left\{ \begin{pmatrix} 0_{l\times 1} \\ -f_B^{-1/2}\mathbb{E}\left\{\partial\Phi_B(V,\delta_A,\delta_B;\mu_{g,0})/\partial\mu\right\}^{-1}\eta \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \right\}. \end{aligned} \tag{3.4}$$

The asymptotic distribution of the efficient estimator $\widehat{\mu}_{\rm eff}$ is

$$n^{1/2}(\widehat{\mu}_{\rm eff} - \mu_g) \to \mathcal{N}\left\{b_{\rm eff}(\eta), V_{\rm eff}\right\},$$

where $b_{\rm eff}(\eta) = -f_B^{-1/2}\omega_B(\Lambda_{\rm eff})\mathbb{E}\left\{\partial\Phi_B(V,\delta_A,\delta_B;\mu_{g,0})/\partial\mu\right\}^{-1}\eta$. The exact form of $\omega_B(\Lambda_{\rm eff})$ and $V_{\rm eff}$ are presented in Lemma A.3.

By Lemma 3.1, among the three estimators $\widehat{\mu}_A$, $\widehat{\mu}_B$ and $\widehat{\mu}_{\rm eff}$, when H_0 holds, $\widehat{\mu}_{\rm eff}$ is optimal because it is consistent and the most efficient; while when H_0 is violated, $\widehat{\mu}_A$ is optimal because it is consistent but the other two estimators are not.

We now use pretesting to guide choosing the estimators. To test H_0, the key insight is that $\widehat{\mu}_A$ is always consistent for μ_g by Assumption 2.1, and if H_0 holds, $\widehat{\Phi}_{B,n}(\widehat{\mu}_A) = n_B^{1/2}N^{-1}\sum_{i=1}^N \widehat{\Phi}_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}_A)$ should behave as a mean-zero random vector asymptotically. Thus, we construct the test statistic T as

$$T = \left\{\widehat{\Phi}_{B,n}(\widehat{\mu}_A)\right\}^\intercal \widehat{\Sigma}_T^{-1}\left\{\widehat{\Phi}_{B,n}(\widehat{\mu}_A)\right\}, \tag{3.5}$$

where Σ_T is the asymptotic variance of $\Phi_{B,n}(\widehat{\mu}_A,\widehat{\tau})$, and $\widehat{\Sigma}_T$ is a consistent estimator of Σ_T. The exact form of Σ_T in (A.15) involves V_A, V_B, and Γ. Thus, $\widehat{\Sigma}_T$ can be obtained by replacing the unknown components in the expression of Σ_T with their estimated counterparts, and the expectations with the empirical averages. In addition, we can consider the replication-based method for variance estimation in Algorithm B.1 adapted from [35].

Lemma 3.2 serves as the foundation for our data-driven pooling step in Section 3.2.

Lemma 3.2. *Suppose Assumptions 2.1, 2.2 (i) and (iii), and all the regularity conditions in Lemma 2.1 hold. Under H_0, the test statistic $T \to \chi_l^2$, i.e., a chi-square distribution with degree of freedom l. Under $H_{a,n}$, $T \to \chi_l^2(\eta^\intercal\Sigma_T^{-1}\eta/2)$ with non-central parameter $\eta^\intercal\Sigma_T^{-1}\eta/2$ as $n \to \infty$.*

3.2. Data-driven pooling

If T is large, it indicates that H_0 may be violated and thus it is desirable to retain only the probability sample for estimation. If T is small, it indicates that H_0 may be accepted and suggests combining the probability and non-probability samples for efficient estimation. This strategy leads to the test-and-pool estimator $\widehat{\mu}_{\text{tap}}$ as the solution to

$$\sum_{i=1}^{N}\{\widehat{\Phi}_A(V_i,\delta_{A,i};\mu)+\mathbf{1}(T<c_\gamma)\Lambda\widehat{\Phi}_B(V_i,\delta_{A,i},\delta_{B,i};\mu)\}=0, \tag{3.6}$$

where c_γ is the $(1-\gamma)$ critical value of χ^2_l. In (3.6), we can fix Λ to be the optimal form Λ_{eff} leading to an efficient estimator under H_0 in Section 2.3. Alternatively, we view c_γ and Λ jointly as tuning parameters that determine how much information from the non-probability sample can be borrowed in pooling. Larger c_γ and Λ borrow more information from the non-probability sample, leading to more efficient but more error-prone estimators, and vice versa. We will use a data-adaptive rule to select (Λ, c_γ) that minimizes the mean squared error of $\widehat{\mu}_{\text{tap}}$.

Remark 3.1. *Compare to the t-test-based pooling estimator in [38], our proposed method is more general in the sense that (a) the auxiliary covariates are used to provide a more informative model of* μ_g*; (b) our test statistic* T *is motivated by the estimating function, which can be more robust to model misspecification, and (c) a data-adaptive selection of* (Λ, c_γ) *is adopted for minimizing the post-integration mean squared error.*

4. Asymptotic properties of the test-and-pool estimator

In this section, we characterize the asymptotic properties of $\widehat{\mu}_{\text{tap}}$. Before proceeding further, we introduce more notations. Let $I_{l\times l}$ be a $l\times l$ identify matrix, $F_l(\cdot;\eta)$ be the cumulative distribution function for χ^2_l with non-central parameter η, and $F_l(\cdot)=F_l(\cdot;0)$. Denote $V_{\text{A-eff}}=V_A-V_{\text{eff}}$ and $V_{\text{B-eff}}=V_B-V_{\text{eff}}$, which are both positive-definite.

4.1. Asymptotic distribution

By construction, the estimator $\widehat{\mu}_{\text{tap}}$ is a pretest estimator that first constructs T for pretesting H_0 and then forms the test-based weights for combining $\widehat{\mu}_A$ and $\widehat{\mu}_B$. It is challenging to derive the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$ because it is involved with the test statistic T and two asymptotically dependent components $\widehat{\mu}_A$ and $\widehat{\mu}_B$. In order to formally characterize the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$, we decompose the asymptotic representation of $\widehat{\mu}_{\text{tap}}$ by two orthogonal components, one is affected by the testing and the other is not.

First, by Lemma 3.1, let $n^{1/2}(\widehat{\mu}_A-\mu_g)\to Z_1$ and $n^{1/2}(\widehat{\mu}_B-\mu_g)\to Z_2$, where Z_1 and Z_2 are multivariate normal random vectors as in (3.4).

Second, by Lemma 3.2, asymptotically, we write T as a quadratic form $W_2^T W_2$ with $W_2=-f_B^{1/2}\Sigma_T^{-1/2}\mathbb{E}\left\{\partial\Phi_B(\mu_{g,0},\tau_0)/\partial\mu\right\}^{-1}(Z_1-Z_2)$. We then find another standardized l-variate normal vector $W_1=f_B^{1/2}\Sigma_S^{-1/2}\{(\Gamma^{\intercal}-V_B)(\Gamma-V_A)^{-1}Z_1+Z_2\}$ that is orthogonal to W_2, where $\text{cov}(W_1,W_2)=0_{l\times l}$, $\mathbb{E}(W_1)=$

$\mu_1, \text{var}(W_1) = I_{l\times l}$ and $\mathbb{E}(W_2) = \mu_2, \text{var}(W_2) = I_{l\times l}$, Σ_S is introduced for the purpose of standardization.

Third, $\widehat{\mu}_{\text{tap}}$ can be asymptotically represented by two components involving W_1 and W_2, respectively, one component is affected by the test constraint and the other component is not. Following the above steps, Theorem 4.1 characterizes the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$.

Theorem 4.1. *Suppose the assumptions in Lemma 3.1 hold except that Assumption 2.2 (ii) may be violated as dictated by $H_{a,n}$ in (3.3). Let W_1 and W_2 to be independent normal random vectors with mean μ_1 and μ_2 (given below, which vary by hypothesis) and variance matrices $I_{l\times l}$. The test-and-pool estimator $\widehat{\mu}_{\text{tap}}$ follows the following asymptotic distribution*

$$n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \to \begin{cases} -V_{\text{eff}}^{1/2}W_1 + (\omega_A V_{A-\text{eff}}^{1/2} - \omega_B V_{B-\text{eff}}^{1/2})W_{[0,c_\gamma]}^t & w.p.\ \xi, \\ -V_{\text{eff}}^{1/2}W_1 + V_{A-\text{eff}}^{1/2}W_{[c_\gamma,\infty]}^t & w.p.\ 1-\xi, \end{cases}$$

where $W_{[a,b]}^t$ is the truncated normal distribution $W_2 \mid (a \le W_2^\intercal W_2 \le b)$ and $\xi = F_l(c_\gamma; \mu_2^\intercal \mu_2/2)$.

(a) Under H_0, $\mu_1 = \mu_2 = 0, \xi = F_l(c_\gamma; 0) = \gamma$.

(b) Under $H_{a,n}$, $\mu_1 = -\Sigma_S^{-1/2}\mathbb{E}\left\{\partial\Phi_B(\mu_{g,0}, \tau_0)/\partial\mu\right\}^{-1}\eta$, $\mu_2 = -\Sigma_T^{-1/2}\eta$ and $\xi = F_1(c_\gamma; \mu_2^T\mu_2/2)$.

Theorem 4.1 reveals that the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$ depends on the local parameter η and thus characterizes the non-regularity of the pretest estimator. When H_0 is violated weakly (a small perturbation in the true data generating model), the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$ can change abruptly depending on η. The non-regularity of $\widehat{\mu}_{\text{tap}}$ also poses challenges for inference as shown in Section 4.3. Based on Theorem 4.1, we derive the asymptotic biases and mean squared errors of $\widehat{\mu}_{\text{tap}}$ under H_0 and $H_{a,n}$, which serve as the stepping stone to a data-driven procedure to select the tuning parameters Λ and c_γ.

4.2. Asymptotic bias and mean squared error

Based on the Theorem 4.1, the asymptotic distribution of $\widehat{\mu}_{\text{tap}}$ involves elliptical truncated normal distributions [60, 4]. To understand the asymptotic behavior of our proposed estimator, it is crucial to comprehend the essential properties of elliptical truncated multivariate normal distributions. We derive the moment generating function and subsequently the mean square error of the estimator $\widehat{\mu}_{\text{tap}}$. The exact form of mean squared error given by $\text{mse}(\Lambda, c_\gamma; \eta)$ in (B.13), albeit complicated, reveals that the amount of information borrowed from the non-probability sample (controlled by Λ and c_γ) should tailor to the strength of violation of H_0 (dictated by local parameter η). For illustration, we consider a toy example in the supplemental material.

We search for the optimal values (Λ^*, c_γ^*) that minimize $\text{mse}(\Lambda, c_\gamma; \widehat{\eta})$ using standard numerical optimization algorithm [39], where $\widehat{\eta} = \Phi_{B,n}(\widehat{\mu}_A, \widehat{\tau})$. Note that the decision of rejecting H_0 or not is subject to the hypothesis testing errors, namely the Type I error and Type II error. That is, the test statistic T can be larger than c_γ even when H_0 holds; similarly, it can be small when $H_{a,n}$ holds. However, the data-adaptive tuning procedure aims at minimizing the mean squared error of the estimator $\widehat{\mu}_{\text{tap}}$, which implicitly restricts these two testing errors to be small.

4.3. Adaptive inference

Standard approaches to inference, e.g., the nonparametric bootstrap, require the estimators to be regular [56]. In non-regular settings, researchers have proposed alternative approaches such as the m-out-n bootstrap or subsampling. However, these approaches critically rely on a proper choice of m or the subsample size; otherwise, the small sample performances can be poor. The non-regularity is induced because the asymptotic distribution of the estimator $\widehat{\mu}_{\text{tap}}$ depends on the local parameter, thus, it does not converge uniformly over the parameter space. [33] propose adaptive confidence intervals for test errors in the classification problems. Following this idea, we construct the bound-based adaptive confidence interval (BACI) for the estimator $\widehat{\mu}_{\text{tap}}$ that guarantees good coverage properties. To avoid the non-regularity, our general strategy is to derive two smooth functionals that bound the estimator $\widehat{\mu}_{\text{tap}}$. Because these two functionals are regular, standard approaches to inference can be adopted and valid confidence intervals follow.

To be concrete, we construct a bound-based adaptive confidence interval for $a^\intercal\mu_g$, where $a \in \mathbb{R}^l$ is fixed. By Theorem 4.1, we can reparametrize the asymptotic distribution of $a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ as

$$a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \to R_n + a^\intercal\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})U_n, \tag{4.1}$$

where

$$R_n = -a^\intercal V_{\text{eff}}^{1/2}W_1 + a^\intercal(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2})W_2 + a^\intercal\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})\mu^t_{[c_\gamma,\infty)},$$
$$U_n = W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)},$$

and $\mu^t_{[c_\gamma,\infty)} = \mu_2 \mathbf{1}_{\mu_2^\intercal\mu_2 > c_\gamma}$. By construction, R_n is regular and asymptotically normal, but U_n is nonsmooth. Nonsmoothness and nonregularity are interrelated. To illustrate, if $\mu_2 = 0$, U_n follows a standard truncated normal distribution with truncated probability $\mathbb{P}(W_2^\intercal W_2 \leq c_\gamma \mid \mu_2 = 0)$; whereas, if $|\mu_2| \to \infty$, $\mathbb{P}(W_2^\intercal W_2 \leq c_\gamma \mid \mu_2)$ diminishes to zero, implying that U_n follows a standard normal distribution. Thus, the limiting distribution of $a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ is not uniform over local parameter μ_2 (or equivalently η).

Our goal is to form the least conservative smooth upper and lower bounds. An important observation is that if $|\mu_2|$ is sufficiently large, we may treat U_n as regular. Thus, we define $\mathbb{B}$ as the nonregular zone for $\mu_2^\intercal\mu_2$ such that $\max_{\mu_2^\intercal\mu_2 \in \mathbb{B}} \mathbb{P}(W_2^\intercal W_2 \geq c_\gamma \mid \mu_2) \leq 1 - \varepsilon$ for small $\epsilon > 0$ and $\mathbb{B}^\complement$ the regular zone. When $\mu_2^\intercal\mu_2 \in \mathbb{B}^\complement$, standard inference can apply, and bounds are only

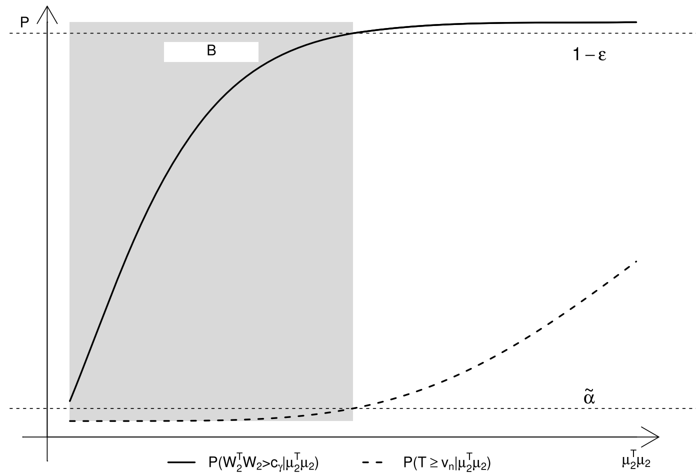

Fig 1. *Illustration of the nonregular zone $\mathbb{B}$ (shaded) and two power functions: the solid and dash lines are $\mathbb{P}(W_2^\intercal W_2 > c_\gamma \mid \mu_2^\intercal\mu_2)$ and $\mathbb{P}(T \geq v_n \mid \mu_2^\intercal\mu_2)$ as functions of $\mu_2^\intercal\mu_2$, respectively.*

needed when $\mu_2^\intercal\mu_2 \in \mathbb{B}$ to avoid the inference procedure to be overly conservative. We then require another test procedure to test $\mu_2^\intercal\mu_2 \in \mathbb{B}$ against $\mu_2^\intercal\mu_2 \in \mathbb{B}^\complement$. Toward this end, we use $T \geq v_n$, where v_n is chosen such that $\max_{\mu_2^\intercal\mu_2\in\mathbb{B}} \mathbb{P}(T \geq v_n \mid \mu_2) = \tilde{\alpha}$ for a pre-specified $\tilde{\alpha}$. Figure 1 illustrates the regular and nonregular zones and the test. If $T \geq \nu_n$, we conclude the regularity of the estimator $\widehat{\mu}_{\text{tap}}$ and construct a normal confidence interval, but if $T < \nu_n$, we construct the least favorable confidence interval by taking the union for all $\mu_2 \in \mathbb{R}^l$. In practice, v_n can be determined by the double bootstrapping satisfying the regularity condition that $\lim_{n\to\infty} v_n/n = 0$; see Section B.4 of the supplemental material for more details.

Accordingly, U_n can be decomposed into two components $U_n = (W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)})\mathbf{1}_{T\geq v_n} + (W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)})\mathbf{1}_{T<v_n}$ and only regularize (i.e., deriving bounds for) the latter component. Continuing with (4.1), we can take the supremum over all μ_2 in the nonregular zone to construct the upper bound $U(a)$,

$$
\begin{aligned}
U(a) =& R_n + a^\intercal\omega_B(V^{1/2}_{\text{B-eff}} + V^{1/2}_{\text{A-eff}})(W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)})\mathbf{1}_{T\geq v_n} \\
&+ \sup_{\mu_2\in\mathbb{R}^l}\left\{a^\intercal\omega_B(V^{1/2}_{\text{B-eff}} + V^{1/2}_{\text{A-eff}})(W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)})\right\}\mathbf{1}_{T<v_n}
\end{aligned}
\tag{4.2}
$$

The lower bound $L(a)$ for $a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ can be computed in an analogous way by replacing sup with inf in (4.2). Taking the supremum and the infimum of μ_2 over $\mathbb{R}^l$ renders the two bounds $U(a)$ and $L(a)$ smooth and regular. The limiting distribution of $U(a)$ is

$$
U(a) \to R + a^\intercal\omega_B(V^{1/2}_{\text{B-eff}} + V^{1/2}_{\text{A-eff}})(W^t_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)})\mathbf{1}_{\mu_2^\intercal\mu_2\in\mathbb{B}^\complement}
$$

$$+ \sup_{\mu_2 \in \mathbb{R}^l} \left\{ a^\intercal \omega_B (V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})(W_{[c_\gamma,\infty)}^t - \mu_{[c_\gamma,\infty)}^t) \right\} \mathbf{1}_{\mu_2^\intercal \mu_2 \in \mathbb{B}}. \tag{4.3}$$

Similarly, the limiting distribution of $L(a)$ is (4.3) by replacing sup with inf. Based on the limiting distribution of $U(a)$ and $L(a)$, if $\mathbb{P}(\mu_2^\intercal \mu_2 \in \mathbb{B}) = 0$, $U(a)$ and $L(a)$ have approximately the same limiting distributions as $a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$. However, if $\mathbb{P}(\mu_2^\intercal \mu_2 \in \mathbb{B}) \neq 0$, $U(a)$ is stochastically larger and $L(a)$ is stochastically smaller than $a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$.

Based on the regular bounds $U(a)$ and $L(a)$, we construct the $(1-\alpha) \times 100\%$ bound-based adaptive confidence interval of $a^\intercal \mu_g$ as

$$\mathbb{C}_{\mu_g,1-\alpha}^{\text{BACI}}(a) = \left[a^\intercal \widehat{\mu}_{\text{tap}} - \widehat{U}_{1-\alpha/2}(a)/\sqrt{n}, a^\intercal \widehat{\mu}_{\text{tap}} - \widehat{L}_{\alpha/2}(a)/\sqrt{n} \right], \tag{4.4}$$

where $\widehat{U}_d(a)$ and $\widehat{L}_d(a)$ approximate the d-th quantiles of the distribution of $U(a)$ and $L(a)$, respectively, which can be obtained by the nonparametric bootstrap method.

Theorem 4.2. *Assume the conditions in Theorem 4.1 hold true. Furthermore, assume matrices Σ_T, Σ_S in Lemma 3.1 and their consistent estimates $\widehat{\Sigma}_T, \widehat{\Sigma}_S$ are strictly positive-definite, and the sequence v_n satisfies $v_n \to \infty$ and $v_n/n \to 0$ with probability one. The asymptotic coverage rate of (4.4) satisfies*

$$\mathbb{P}\left\{ a^\intercal \mu_g \in \mathbb{C}_{\mu_g,1-\alpha}^{\text{BACI}}(a) \right\} \geq 1 - \alpha. \tag{4.5}$$

In particular, if Assumption 2.2 is strongly violated with $\mathbb{P}(\mu_2^\intercal \mu_2 \in \mathbb{B}^\complement) = 1$, the inequality in (4.5) becomes equality.

Remark 4.1. *We discuss an alternative approach to construct valid confidence intervals for the non-regular estimators using projection sets [48] (referred to as projection-based adaptive confidence intervals (PACI), $\mathbb{C}_{\mu_g,1-\alpha}^{\text{PACI}}(a)$). The basic idea is as follows. For a given μ_2, the limiting distribution of $\widehat{\mu}_{\text{tap}}$ is known and a regular $(1-\tilde{\alpha}_1) \times 100\%$ confidence interval $\mathbb{C}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2)$ of $a^\intercal \mu_g$ can be formed through the standard procedure. Since μ_2 is unknown, a $(1-\alpha) \times 100\%$ projection confidence interval of μ_g can be conservatively constructed as the union of all $\mathbb{C}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2)$ over μ_2 in its $(1-\tilde{\alpha}_2) \times 100\%$ confidence region, where $\alpha = \tilde{\alpha}_1 + \tilde{\alpha}_2$. Such strategy may be overly conservative, and in that way, the projection-based adaptive confidence interval then introduces a pretest in order to mitigate the conservatism. If the pretest rejects $H_0 : \mu_2^\intercal \mu_2 \in \mathbb{B}$, $\mathbb{C}_{\mu_g,1-\tilde{\alpha}_1}(a;\widehat{\mu}_2)$ is used; otherwise, the union of $\mathbb{C}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2)$ is used. The technical details for the $\mathbb{C}_{\mu_g,1-\alpha}^{\text{PACI}}(a)$ are presented in the supplemental material. Our simulation study later shows that the $\mathbb{C}_{\mu_g,1-\alpha}^{\text{PACI}}(a)$ is more conservative than the proposed $\mathbb{C}_{\mu_g,1-\alpha}^{\text{BACI}}(a)$.*

5. Simulation study

In this section, we evaluate the finite-sample performances of the proposed estimator $\widehat{\mu}_{\text{tap}}$ and $\mathbb{C}_{\mu_g,1-\alpha}^{\text{BACI}}(a)$. First, we generate the finite population $\mathcal{F}_N$ with size

$N = 10^5$. For each subject i, generate $X_i = (1, X_{1,i}, X_{2,i})^\intercal$, where $X_{1,i} \sim \mathcal{N}(0,1)$ and $X_{2,i} \sim \mathcal{N}(1,1)$, and generate Y_i by $Y_i = 1 + X_{1,i} + X_{2,i} + u_i + u_i^2 + \varepsilon_i$, where $u_i \sim \mathcal{N}(0,1)$ and $\epsilon_i \sim \mathcal{N}(0,1)$. Generate samples from the finite population $\mathcal{F}_N$ by Bernoulli sampling with specified inclusion probabilities

$$\log\left(\frac{\pi_{A,i}}{1-\pi_{A,i}}\right) \mid X_i = \nu_A + .2X_{1,i} + .1X_{2,i},$$
$$\log\left(\frac{\pi_{B,i}}{1-\pi_{B,i}}\right) \mid X_i = \nu_B + .1X_{1,i} + .2X_{2,i} + .5n_B^{-1/2} b u_i,$$

where ν_A and ν_B are adaptively chosen to ensure the target sample sizes $n_A \approx 600$ and $n_B \approx 5000$. We assume that (X_i, Y_i) are observed but u_i is unobserved, and we vary b in $\{0, 10, 100\}$ to represent the scenarios where H_0 holds, is slightly violated or strongly violated, respectively.

We compare the estimator $\widehat{\mu}_{\text{tap}}$ with other estimators: (a) $\widehat{\mu}_A$: the solution to $\sum_{i=1}^N \Phi_A(V_i, \delta_{A,i}; \mu) = 0$ with $\Phi_A(V_i, \delta_{A,i}; \mu)$ defined in (2.4). (b) $\overline{\mu}_B$: the naive sample mean $\overline{\mu}_B = (\sum_{i=1}^N \delta_{B,i})^{-1} \sum_{i=1}^N \delta_{B,i} Y_i$. (c) $\widehat{\mu}_{\text{dr}}$: the solution to $\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu, \alpha, \beta) = 0$ with $\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu, \alpha, \beta)$ defined in (2.5), where (α, β) are estimated by using the maximum pseudo-likelihood estimator $\widehat{\alpha}$ and the ordinary least square estimator $\widehat{\beta}$ [26]; see Equations (B.2) and (B.4). (d) $\widehat{\mu}_{\text{eff}}$: the solution to (2.7) with the optimal choice Λ_{eff} specified in (A.11) and the consistent estimators $(\widehat{\alpha}, \widehat{\beta})$ obtained from (c). (e) $\widehat{\mu}_{\text{eff}:B}$: $\widehat{\mu}_{\text{eff}}$, where α is estimated in the same manner as (c) but β is estimated solely based on the non-probability sample; see Equation (B.3). (f) $\widehat{\mu}_{\text{eff:KH}}$: $\widehat{\mu}_{\text{eff}}$, where (α, β) are estimated simultaneously by adopting the methods proposed by [29]; see Equations (B.5) and (B.6). (g) $\widehat{\mu}_{\text{tap}}$, $\widehat{\mu}_{\text{tap}:B}$, $\widehat{\mu}_{\text{tap:KH}}$: the solution to (3.6), where (Λ, c_γ) are chosen by our data-adaptive procedure with $(\widehat{\alpha}, \widehat{\beta})$ obtained from (d), (e), (f), respectively. (h) $\widehat{\mu}_{\text{Bayes:1}}$, $\widehat{\mu}_{\text{Bayes:2}}$, $\widehat{\mu}_{\text{Bayes:3}}$: the Bayesian approaches for combining the non-probability sample with the probability sample assuming different informative priors [52].

For all estimators, we specify the model $\pi_B(X; \alpha)$ to be a logistic regression model with X_i and the outcome mean model $m(X; \beta)$ to be a linear regression model with X_i. For non-regular estimators $\widehat{\mu}_{\text{tap}}$, $\widehat{\mu}_{\text{tap}:B}$ and $\widehat{\mu}_{\text{eff:KH}}$, we construct the $\mathbb{C}^{\text{BACI}}_{\mu_g, 1-\alpha}(a)$ in (4.4) with a data-adaptiv choice of ν_n, the $\mathbb{C}^{\text{BACI}}_{\mu_g, 1-\alpha}(a)$ with a fixed $v_n = \log\log n\{\mathbb{C}^{\text{BACI:F}}_{\mu_g, 1-\alpha}(a)\}$ (BACI$_F$), and the $\mathbb{C}^{\text{PACI}}_{\mu_g, 1-\alpha}(a)$. For any confidence intervals requiring the nonparametric bootstrap, the bootstrap size is 2000. For the Bayesian estimators, the point estimates are obtained by the Markov chain Monte Carlo sampling with size 2000 after additional 500 burn-in samples.

Table 2 reports the bias, variance and mean squared error of each estimator over 2000 simulated datasets. The benchmark estimators $\widehat{\mu}_A$ have small biases across all scenarios, guaranteed by the probability sampling design. On the other hand, the non-probability-only estimators $\overline{\mu}_B$ exhibit high biases in all cases, mainly due to the effect of selection bias. When the impact of the unmeasured confounder b increases, the pooled estimators $\widehat{\mu}_{\text{eff}}, \widehat{\mu}_{\text{eff}:B}$ and $\widehat{\mu}_{\text{eff:KH}}$ are be-

TABLE 2

Simulation results for bias $(\times 10^{-3})$*, variance (var)* $(\times 10^{-3})$ *and mean squared error (MSE)* $(\times 10^{-3})$ *of* $\widehat{\mu}_A, \overline{\mu}_B, \widehat{\mu}_{\mathrm{dr}}, \widehat{\mu}_{eff}, \widehat{\mu}_{Bayes}$ *and* $\widehat{\mu}_{\mathrm{tap}}$ *when* H_0 *holds, is slightly violated or strongly violated.*

H_0		holds			slightly violated			strongly violated		
		bias	var	MSE	bias	var	MSE	bias	var	MSE
Regular	$\widehat{\mu}_A$	−4.1	10.4	10.4	−4.1	10.4	10.4	−4.1	10.4	10.4
	$\overline{\mu}_B$	284.1	1.2	81.9	355.3	1.2	127.4	1318.8	2.0	1741.4
	$\widehat{\mu}_{\mathrm{dr}}$	−0.4	4.2	4.2	71.0	4.3	9.3	1048.0	5.0	1103.2
	$\widehat{\mu}_{\mathrm{eff}}$	−0.9	4.1	4.1	62.3	4.2	8.1	851.5	6.6	731.7
	$\widehat{\mu}_{\mathrm{eff}:B}$	−0.9	4.1	4.1	62.3	4.2	8.1	851.7	6.6	732.1
	$\widehat{\mu}_{\mathrm{eff:KH}}$	−0.9	4.1	4.1	62.3	4.2	8.1	851.5	6.7	731.7
Bayes	$\widehat{\mu}_{\mathrm{Bayes:1}}$	−3.7	14.1	14.1	1.0	14.0	14.0	−4.3	14.1	14.1
	$\widehat{\mu}_{\mathrm{Bayes:2}}$	−4.1	10.8	10.8	17.1	11.1	11.4	7.0	13.8	13.8
	$\widehat{\mu}_{\mathrm{Bayes:3}}$	−2.4	8.9	8.9	51.2	9.0	11.6	614.0	10.8	387.9
TAP	$\widehat{\mu}_{\mathrm{tap}}$	−4.8	7.6	7.6	10.1	9.3	9.4	−4.1	10.4	10.4
	$\widehat{\mu}_{\mathrm{tap}:B}$	−4.8	7.6	7.6	10.1	9.3	9.4	−4.1	10.4	10.4
	$\widehat{\mu}_{\mathrm{tap:KH}}$	−4.8	7.6	7.6	10.1	9.3	9.4	−4.1	10.4	10.4

TABLE 3

Simulation results for coverage rates (CR) $(\times 10^{-2})$ *and widths* $(\times 10^{-3})$ *for 95% confidence intervals when* H_0 *holds, is slightly violated or strongly violated.*

H_0		holds		slightly violated		strongly violated	
	CIs	CR	width	CR	width	CR	width
$\widehat{\mu}_A$	Wald	95.2	404.1	95.3	404.1	95.2	404.0
$\overline{\mu}_B$		0.0	135.5	0.0	138.8	0.0	173.7
$\widehat{\mu}_{\mathrm{dr}}$		95.9	262.8	81.8	264.4	0.0	282.4
$\widehat{\mu}_{\mathrm{eff}}$		95.9	259.5	85.1	260.9	0.0	273.6
$\widehat{\mu}_{\mathrm{Bayes:1}}$	HPDI	98.3	463.0	97.5	461.5	97.3	462.8
$\widehat{\mu}_{\mathrm{Bayes:2}}$		97.8	404.2	97.4	409.8	97.5	458.3
$\widehat{\mu}_{\mathrm{Bayes:3}}$		99.3	368.2	97.4	370.6	0.0	407.0
$\widehat{\mu}_{\mathrm{tap}}$	PACI	98.4	558.7	98.4	535.7	99.2	541.0
	BACI$_F$	94.7	399.1	95.9	402.3	94.7	402.6
	BACI	92.1	363.1	93.3	367.2	94.8	402.8

coming more biased. Additionally, the Bayesian methods, particularly $\widehat{\mu}_{\mathrm{Bayes:2}}$, perform reasonably well when H_0 holds or is slightly violated, but it tends to have large biases when H_0 is strongly violated. Whereas the proposed estimators $\widehat{\mu}_{\mathrm{tap}}, \widehat{\mu}_{\mathrm{tap}:B}$ and $\widehat{\mu}_{\mathrm{tap:KH}}$ have small biases regardless of the strength of the unmeasured confounder. When H_0 is slightly violated, our proposed estimators have slightly larger biases but smaller mean squared errors than $\widehat{\mu}_A$ by integrating the non-probability sample. When H_0 is strongly violated, the proposed estimators perform similarly to $\widehat{\mu}_A$ with the protection of pretesting.

Table 3 reports the properties of 95% Wald confidence intervals for the regular estimators, the highest posterior density intervals (HPDIs) for the Bayesian estimators, and various adaptive confidence intervals for the non-regular estimators $\widehat{\mu}_{\mathrm{tap}}$, where the Wald confidence intervals are constructed, and the Bayesian credible intervals are constructed based on the posterior samples after burn-in. Because the confidence intervals (and the point estimates; see Table 2) are not sensitive to the methods of estimating the nuisance parameters (α, β), we only present the confidence intervals for $\widehat{\mu}_{\text{eff:KH}}$ and $\widehat{\mu}_{\text{tap:KH}}$ for simplicity. Based on Table 3, $\mathbb{C}^{\text{PACI}}_{\mu_g,1-\alpha}$ tend to overestimate the uncertainty, leading to over-conservative confidence intervals. $\mathbb{C}^{\text{BACI}}_{\mu_g,1-\alpha}$ and $\mathbb{C}^{\text{BACI:F}}_{\mu_g,1-\alpha}$ are less conservative and alleviate the over-coverage issues; thus, the empirical coverage rates are close to the nominal level in all cases. Moreover, $\mathbb{C}^{\text{BACI}}_{\mu_g,1-\alpha}$ have narrower intervals than $\mathbb{C}^{\text{BACI:F}}_{\mu_g,1-\alpha}$ by using the double bootstrap procedure to select v_n at the expense of computational burden. When H_0 holds, the $\mathbb{C}^{\text{BACI}}_{\mu_g,1-\alpha}$ are narrower than the Wald for the probability-only estimator $\widehat{\mu}_A$, indicating the advantages of implementing the test-and-pool strategy in these cases. When H_0 is slightly violated, the benefit in coverage rate is not significantly observed under similar coverage rates. When H_0 is strongly violated, the adaptive confidence interval $\mathbb{C}^{\text{BACI}}_{\mu_g,1-\alpha}$ reduces to the Wald confidence intervals for $\widehat{\mu}_A$. Lastly, the credible intervals for the Bayesian estimators do not have satisfactory coverage properties as the model misspecification persists across scenarios, which is aligned with the Bernstein-von Mises Theorem [65, Chapter 10.2].

6. A real-data illustration

To demonstrate the practical use, we apply the proposed method to a probability sample from the 2015 Current Population Survey (CPS) and a non-probability sample from the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. Note that the Behavioral Risk Factor Surveillance System survey itself is a probability sample and we manually discard its sampling weights to recast it as a non-probability sample for illustrating our proposed method.

To apply the proposed method, we use a two-phase sampling survey data with sizes $n_A = 1000$ and $n_B = 8459$. We focus on two outcome variables of interest: employment (percentages of working and retired) and educational attainment (high school or less as h.s.o.l, and college or above as c.o.a.). Both datasets provide measurements on the outcomes of interest and some common covariates including age, sex (female or not), race (white and black), origin (Hispanic or not), region (northeast, south, or west), and marital status (married or not). To illustrate the heterogeneity in the study populations, Table 4 contrasts the means of variables from the CPS sample (design-weighted averages) and the BRFSS sample (simple averages). Based on Table 4, the BRFSS sample may not be representative of the target population, and the pretesting procedures before pooling should be expected.

Table 5 presents the results. For all estimators, we specify the propensity score model to be a logistic regression model with the covariates (all variables excluding the outcome variable) and the outcome mean model to be a logistic regression model with the covariates. The efficient estimator $\widehat{\mu}_{\text{eff}}$ gains efficiency in all estimators compared to both $\widehat{\mu}_A$ and $\widehat{\mu}_{\text{dr}}$; however, it may be subject to biases if the non-probability sample does not satisfy the required assumptions. In the test-and-pool analysis, the pretesting rejects the use of the non-probability

TABLE 4
The covariate means by two samples: CPS sample (a probability sample) and BRFSS sample (a hypothetical non-probability sample.)

Data source	age	%sex	%white	%black	%hispanic	%northeast	%south
CPS	47.5	56.5	81.9	11.0	13.3	18.1	37.7
BRFSS	48.3	54.2	83.2	8.4	8.3	20.0	27.6
	%west	%married	%working	%retired	%h.s.o.l.	%c.o.a.	
CPS	24.1	52.5	58.7	13.6	39.4	30.3	
BRFSS	29.5	50.8	52.2	24.5	21.2	41.9	

TABLE 5
Estimated population mean (EST), standard errors (SE) and confidence intervals of μ_g for selected covariates when combining two datasets.

Outcome Y		%working	%retire	%h.s.o.l.	%c.o.a.
$\widehat{\mu}_A$	EST	58.7	13.6	39.4	30.3
	SE	1.51	1.17	1.60	1.59
	Wald	(54.8,62.3)	(11.6,16.2)	(35.7,43.0)	(27.2,33.7)
$\widehat{\mu}_{\text{dr}}$	EST	56.5	20.0	25.8	32.3
	SE	1.03	1.24	0.93	1.20
	Wald	(54.2,58.8)	(17.9,22.4)	(234.0,27.5)	(30.3,34.5)
$\widehat{\mu}_{\text{eff:KH}}$	EST	56.6	17.3	26.4	32.1
	SE	0.80	0.19	0.87	0.62
	Wald	(54.3,58.9)	(15.4,19.6)	(24.6,28.1)	(30.1,34.3)
$\widehat{\mu}_{\text{Bayes:1}}$	EST	59.8	14.1	40.5	30.7
	SE	1.97	1.37	2.00	1.84
	HPDI	(56.0, 63.6)	(11.4,16.8)	(36.6,44.4)	(27.2,34.4)
$\widehat{\mu}_{\text{Bayes:2}}$	EST	59.8	14.0	40.3	30.9
	SE	2.01	1.33	1.92	1.84
	HPDI	(56.1,63.9)	(11.4,16.4)	(36.4,44.0)	(27.2,34.5)
$\widehat{\mu}_{\text{Bayes:3}}$	EST	58.6	14.1	37.6	31.1
	SE	1.94	1.30	1.92	1.76
	HPDI	(54.7, 62.4)	(11.6,16.7)	(33.7,41.4)	(27.7,34.7)
$\widehat{\mu}_{\text{tap:KH}}$	EST	58.7	13.6	39.0	31.7
	SE	1.51	1.17	1.55	0.64
	BACI	(54.9,62.6)	(11.6,15.8)	(35.8,42.6)	(31.0,33.6)

sample for the employment variables "working" and "retired" but accepts the use of the non-probability sample for the education variables "high school or less" and "college or above". Thus, for the employment variables, $\widehat{\mu}_{\text{tap}} = \widehat{\mu}_A$, and for the educational attainment variables, $\widehat{\mu}_{\text{tap}}$ gains efficiency over $\widehat{\mu}_A$. The Bayesian estimators with the informative priors 2 and 3 are more efficient than the prior 1. However, they still yield larger standard errors compared to the probability-only estimator $\widehat{\mu}_A$ perhaps because the non-probability-based informative priors are biased for the model parameters for the probability sample. From the test-and-pool analysis, the employment rate and the retirement rate are 58.7% and 13.6%, respectively, the percentage of the U.S. population with a high school education or less is 39.0% and the percentage of the population with a college education or above is 31.7% in 2015.

7. Concluding remarks

When utilizing the non-probability samples, researchers often assume that the observed covariates contain all the information needed for recovering the sampling mechanism. However, this assumption may be violated, and hence the integration of the probability and non-probability samples is subject to biases. In this paper, we propose the test-and-pool estimator that firstly scrutinizes the assumption required for combining by hypothesis testing and carefully combines the probability and non-probability samples by a data-driven procedure to achieve the minimum mean squared error. In theoretical development, we treat (Λ, c_γ) jointly as two tuning parameters and establish the asymptotic distribution of the pretesting estimator without taking their uncertainties into account. The non-regularity of the pretest estimator invalidates the conventional method for generating reliable inferences. To address this issue, the proposed adaptive confidence interval has been designed to effectively handle the non-smoothness of the pretest estimator and ensure uniform validity of inferences. It is important to note, however, that this approach may result in a little gain in the precision of the confidence interval, although the point estimator might have a significant gain in the MSE compared to the estimator based only on the probability sample. Further research is required to develop a valid post-testing confidence interval that offers reduced conservatism.

Pretest estimation is the norm rather than the exception in applied research, so the theories that we have established are highly relevant to researchers who engage in applied work. The proposed framework can be extended in the following directions. First, in this work, we study the implications of pretesting on estimation and inference under one single pretest. In practice, researchers may engage in multiple presetting. For example, in the data integration context, one can encounter multiple data sources [51, 71, 16], requiring pretesting of the comparability of each data source and the benchmark. Multiple presetting alters the current asymptotic results and is an important future research topic. Second, our framework considers a fixed number of covariates; however, in reality, practitioners often collect a rich set of auxiliary variables, rendering variable selection imperative [75]. Developing a valid statistical framework to deal with issues arising from selective inference is a challenging but important topic for further investigation. Third, small area estimation has received a lot of attention in the data integration context [44, 28, 25]. The typical estimator in small area estimation is a weighted average of the design-based estimator and a model-based synthetic estimator. [5] discussed the trade-off of the efficiency gain from invoking model assumptions and the risk that these assumptions do not hold. Thus, pretesting can be potentially useful for small-area estimation, which we will investigate in the future.

Appendix A: Proofs

A.1. Regularity conditions

Let $\mathcal{F}_N = \{V_i = (X_i^\intercal, Y_i)^\intercal : i \in U\}$, $\Phi_A(V, \delta_A; \mu)$ and $\Phi_B(V, \delta_A, \delta_B; \mu, \tau)$ be l-dimensional estimating functions for the parameter $\mu_g \in \mathbb{R}^l$ when using the probability sample and the combined samples, respectively. Let $\Phi_\tau(V, \delta_A, \delta_B; \tau)$ be the k-dimensional estimating equations for the nuisance parameter $\tau_0 \in \mathbb{R}^k$. Then, we construct one stacked estimating equation system $\Phi(V, \delta_A, \delta_B; \theta)$ with $\theta = (\mu_A^\intercal, \mu_B^\intercal, \tau^\intercal)^\intercal$ and $\dim(\theta) = 2l + k$. For establishing our stochastic statements, we require the following regularity conditions.

Assumption A.1. *The following regularity conditions hold.*

a) The parameter $\theta = (\mu_A^\intercal, \mu_B^\intercal, \tau^\intercal)^\intercal$ belongs to a compact parameter spaces Θ in $\mathbb{R}^{2l+k}$.
b) There exist a unique solution $\theta_0 = (\mu_{A,0}^\intercal, \mu_{B,0}^\intercal, \tau_0^\intercal)^\intercal$ lying in the interior of the compact space Θ such that

$$\mathbb{E}\{\Phi_A(V, \delta_A; \theta_0)\} = \mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \theta_0)\} = \mathbb{E}\{\Phi_\tau(V, \delta_A, \delta_B; \theta_0)\} = 0.$$

c) $\Phi(V, \delta_A, \delta_B; \theta)$ is integrable with respect to the joint distribution of (V, δ_A, δ_B) for all θ in a neighborhood of θ_0.
d) The first two partial derivatives of $\mathbb{E}\{\Phi(V, \delta_A, \delta_B; \theta)\}$ and their empirical estimators are invertible for all θ in a neighborhood of θ_0.
e) For all $j, k, l \in \{1, \cdots, 2l+k\}$, there is an integrable function $B(V, \delta_A, \delta_B)$ such that

$$|\partial \Phi_j(V, \delta_A, \delta_B; \theta)/\partial\theta_k\partial\theta_l| \le B(V, \delta_A, \delta_B), \quad \mathbb{E}\{B(V, \delta_A, \delta_B)\} < \infty,$$

for all θ in a neighborhood of θ_0 almost surely.
f) $\{V_i : i \in \mathcal{U}\}$ are a set of i.i.d. random variables s.t. $\mathbb{E}\{|\Phi(V, \delta_A, \delta_B; \theta)|^{2+\delta}\}$ is uniformly bounded for θ in a neighborhood of θ_0.
g) The sample sizes n_A and n_B are in the same order of magnitude, i.e., $n_A = O(n_B)$. The sampling fractions for both Sample A and B are negligible, i.e., $n/N = o(1)$, where $n = n_A + n_B$.
h) There exist C_1 and C_2 such that $0 < C_1 \le N\pi_{A,i}/n_A \le C_2$ and $0 < C_1 \le N\pi_{B,i}/n_B \le C_2$ for all $i \in \mathcal{U}$.

Assumption A.1 a)-e) are typical finite moment conditions to ensure the consistency of the solution to the estimating functions [49, Appendix B], [64, Section 3.2], [9, page 293] and [67, Appendix C]. Assumption A.1 f) is required for obtaining the asymptotic normality of μ_g under superpopulation. Assumption A.1 g) states that the sampling fraction is negligible, which is helpful for subsequent variance estimation, and we can use $O(n_A^{-1/2})$, $O(n_B^{-1/2})$ and $O(n^{-1/2})$ interchangeably. Assumption A.1 h) implies that the inclusion probabilities for Samples A and B are in the order of n/N, which is necessary to establish their root-n consistency.

It is noteworthy that in Assumption 2.1, the asymptotic normality is ascertained for the design-weighted estimators given the finite population $\mathcal{F}_N$. Hereby, we extend the conditional normality to the unconditional one, which averages over all possible finite populations satisfying the Assumption A.1 (f). The following lemma plays a key role to establish the stochastic statements [24, Theorem 1.3.6.].

Lemma A.1. *Under Assumption 2.1 and Assumption A.1 (f), let $\{\mathcal{F}_N\}$ be a sequence of finite populations and $\mathcal{A}_N$ be a sample selected from the Nth population by PR design with size n_N. Assume that*

$$\lim_{N\to\infty} n_N = \infty, \quad \lim_{N\to\infty} N - n_N = \infty.$$

We know that the distribution of the design-weighted estimator $\widehat{\mu}_g$ and finite-population estimator μ_g are both asymptotically normal distributed such that

$$\widehat{\mu}_g \mid \mathcal{F}_N \dot{\sim} \mathcal{N}(\mu_g, V_1), \quad \mu_g \dot{\sim} \mathcal{N}(\mu_{g,0}, V_2),$$

where $\dot{\sim}$ denotes the asymptotic distribution. Then, $\widehat{\mu}_g - \mu_g$ is also asymptotically normal.

By lemma A.1, the sampling fraction is negligible, and therefore the limiting variance of $\lim_{N\to\infty} n_N^{1/2}(\mu_g - \mu_{g,0})$ is 0, indicating that the intermediate step of producing the finite population is of little significance.

A.2. Proof of Lemmas 2.1 and 3.1

In the general case, we begin to investigate the statistical properties of

$$\Phi_{A,n}(\widehat{\mu}_A, \widehat{\tau}) = n^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_A(V_i, \delta_{A,i}; \widehat{\mu}_A, \widehat{\tau})$$

and

$$\Phi_{B,n}(\widehat{\mu}_B, \widehat{\tau}) = n^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_B, \widehat{\tau}).$$

First, to simplify our notations, let

$$\begin{aligned}
\dot{\Phi}_A(V, \delta_A; \mu, \tau) &= \partial \Phi_A(V, \delta_A; \mu, \tau)/\partial \mu, \\
\dot{\Phi}_B(V, \delta_A, \delta_B; \mu, \tau) &= \partial \Phi_B(V, \delta_A, \delta_B; \mu, \tau)/\partial \mu, \\
\phi_{B,\tau}(V, \delta_A, \delta_B; \mu, \tau) &= \partial \Phi_B(V, \delta_A, \delta_B; \mu, \tau)/\partial \tau, \\
\phi_\tau(V, \delta_A, \delta_B; \tau) &= \partial \Phi_\tau(V, \delta_A, \delta_B; \tau)/\partial \tau.
\end{aligned}$$

By the Taylor expansion of $\Phi_{B,n}(\widehat{\mu}_B, \widehat{\tau})$ at (μ_g, τ_0), we have

$$\begin{aligned}
0 &= \Phi_{B,n}(\widehat{\mu}_B, \widehat{\tau}) \\
&= n^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_g, \tau_0) \\
&\quad + n^{1/2} N^{-1} \sum_{i=1}^{N} \phi_{B,\tau}(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_B^*, \widehat{\tau}^*)(\widehat{\tau} - \tau_0)
\end{aligned}$$

$$+n^{1/2}N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}_B^*,\widehat{\tau}^*)(\widehat{\mu}_B-\mu_g), \tag{A.1}$$

for some $(\widehat{\mu}_B^*,\widehat{\tau}^*)$ lying between $(\widehat{\mu}_B,\widehat{\tau})$ and (μ_g,τ_0), which leads to

$$\begin{aligned}
&-n^{1/2}N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i;\widehat{\mu}_B^*,\widehat{\tau}^*)(\widehat{\mu}_B-\mu_g) \\
&=n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu_g,\tau_0) \\
&+n^{1/2}N^{-1}\sum_{i=1}^{N}\phi_{B,\tau}(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}_B^*,\widehat{\tau}^*)(\widehat{\tau}-\tau_0).
\end{aligned} \tag{A.2}$$

Also, under Assumption A.1 a), b) and c), by the Taylor expansion, we have

$$\begin{aligned}
n^{1/2}(\widehat{\tau}-\tau_0)=&-\left\{\frac{1}{N}\sum_{i=1}^{N}\phi_\tau(V_i;\tau_0)\right\}^{-1} \\
&\times\left\{n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_\tau(V_i,\delta_{A,i},\delta_{B,i};\tau_0)\right\}+o_{\zeta\text{-p-np}}(1),
\end{aligned} \tag{A.3}$$

as $\widehat{\tau}\to\tau_0$. Also, under Assumption A.1 (e), we know that

$$\begin{aligned}
&N^{-1}\sum_{i=1}^{N}\dot{\Phi}_A(V_i;\widehat{\mu}_A^*,\widehat{\tau}^*)\to\mathbb{E}\{\dot{\Phi}_A(V;\mu_{g,0},\tau_0)\}, \\
&N^{-1}\sum_{i=1}^{N}\phi_r(V_i;\tau_0)\to\mathbb{E}\left\{\phi_\tau(V;\tau_0)\right\}, \\
&N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i;\widehat{\mu}_B^*,\widehat{\tau}^*)\to\mathbb{E}\{\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\}, \\
&N^{-1}\sum_{i=1}^{N}\phi_{B,\tau}(V_i;\widehat{\mu}_B^*,\widehat{\tau}^*)\to\mathbb{E}\left\{\phi_{B,\tau}(V;\mu_{g,0},\tau_0)\right\},
\end{aligned} \tag{A.4}$$

where the first two probability convergence can be straightforward to obtain by Weak Law of Large Numbers under Assumption A.1 f) and continuous mapping theorem as $\mu_g\to\mu_{g,0}$, $(\widehat{\mu}_A,\widehat{\tau})\to(\mu_{g,0},\tau_0)$ by design and $(\widehat{\mu}_A^*,\widehat{\tau}^*)$ is lying between $(\widehat{\mu}_A,\widehat{\tau})$ and $(\mu_{g,0},\tau_0)$. As for the third and fourth probability convergence, we first prove that $\mu_{B,0}-\mu_{g,0}=o_{\text{np-p-}\zeta}(1)$ under the local alternative $\mathbb{E}\{\Phi_B(V,\delta_A,\delta_B;\mu_{g,0},\tau_0)\}=n_B^{-1/2}\eta$ in Lemma A.2.

Lemma A.2. *Under Assumptions 2.1, 2.2 (iii) and suitable moments conditions in Assumption A.1, we have* $\mu_{B,0}-\mu_{g,0}=O_{\text{np-p-}\zeta}(n^{-1/2})$.

Next, we have under Assumption A.1 e),

$$
\begin{aligned}
& N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i;\widehat{\mu}_B^*,\widehat{\tau}^*) \\
& \cong N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i;\mu_{g,0},\tau_0)+N^{-1}\sum_{i=1}^{N}\frac{\partial^2\Phi_B(V_i;\mu_B^{\#},\tau_0)}{\partial\mu\partial\mu^{\intercal}}(\widehat{\mu}_B^*-\mu_{g,0}) \\
& \cong N^{-1}\sum_{i=1}^{N}\dot{\Phi}_B(V_i;\mu_{g,0},\tau_0)+O_{\zeta\text{-p-np}}\{(\widehat{\mu}_B^*-\mu_{B,0})+(\mu_{B,0}-\mu_{g,0})\} \\
& =\mathbb{E}\{\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\}+o_{\zeta\text{-p-np}}(1),
\end{aligned} \tag{A.5}
$$

where $A_n \cong B_n$ means that $A_n = B_n + o_{\zeta\text{-p-np}}(1)$ and $\mu_B^{\#}$ lies between $\widehat{\mu}_B^*$ and $\mu_{g,0}$. Since $\widehat{\mu}_B \to \mu_{B,0}, \mu_g \to \mu_{g,0}$ and $\widehat{\mu}_B^*$ lies between $\widehat{\mu}_B$ and μ_g, we establish the second approximation in (A.5) as

$$
(\widehat{\mu}_B^*-\mu_{B,0})+(\mu_{B,0}-\mu_{g,0})=O_{\mathrm{np}}(n_B^{-1/2})+O_\zeta(N^{-1/2})=o_{\zeta\text{-p-np}}(1),
$$

since $n_B/N = o(1)$. The probability convergence of $N^{-1}\sum_{i=1}^{N}\phi_{B,\tau}(V_i;\widehat{\mu}_B^*,\widehat{\tau}^*)$ can be established similarly and hence we obtain the last two parts of (A.4). By plugging (A.3) and (A.4) into (A.2), we obtain the influence function for $\widehat{\mu}_B$ as

$$
\begin{aligned}
& n^{1/2}(\widehat{\mu}_B-\mu_g) \\
& \cong -\mathbb{E}\left\{\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\right\}^{-1}\times\left[n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu_g,\tau_0)\right. \\
& \left.-\mathbb{E}\left\{\phi_{B,r}(V;\mu_{g,0},\tau_0)\right\}\cdot\mathbb{E}\left\{\phi_\tau(V;\tau_0)\right\}^{-1}\left\{n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_\tau(V_i,\delta_{A,i},\delta_{B,i};\tau_0)\right\}\right] \\
& \cong n^{1/2}N^{-1}\sum_{i=1}^{N}\psi_B(V_i;\mu_g,\tau_0),
\end{aligned} \tag{A.6}
$$

where $\psi_B(V_i;\mu,\tau)$ is the influence function for estimation of $\widehat{\mu}_B$ under H_0. For completeness, we define the influence function $\psi_A(V_i;\mu,\tau)$ for estimator $\widehat{\mu}_A$ in an analogous way as

$$
\begin{aligned}
& n^{1/2}(\widehat{\mu}_A-\mu_g)\cong -n^{1/2}N^{-1}\sum_{i=1}^{N}\left\{N^{-1}\sum_{i=1}^{N}\dot{\Phi}_A(V_i,\delta_{A,i};\widehat{\mu}_A^*,\widehat{\tau}^*)\right\}^{-1} \\
& \times\left\{\Phi_A(V_i,\delta_{A,i};\mu_g,\tau_0)+\phi_{A,\tau}(V_i;\widehat{\mu}_A^*,\widehat{\tau}^*)\cdot(\widehat{\tau}-\tau_0)\right\} \\
& \cong -\mathbb{E}\left\{\dot{\Phi}_A(V;\mu_{g,0},\tau_0)\right\}^{-1}\times\left[n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_A(V_i,\delta_{A,i};\mu_g,\tau_0)\right. \\
& \left.-\mathbb{E}\left\{\phi_{A,\tau}(V;\mu_{g,0},\tau_0)\right\}\cdot\mathbb{E}\left\{\phi_\tau(V;\tau_0)\right\}^{-1}\left\{n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi_\tau(V_i,\delta_{A,i},\delta_{B,i};\tau_0)\right\}\right]
\end{aligned} \tag{A.7}
$$

$$\cong n^{1/2} N^{-1} \sum_{i=1}^{N} \psi_A(V_i; \mu_g, \tau_0), \tag{A.8}$$

where $\phi_{A,r}(V; \mu, \tau) = \partial \Phi_A(V, \delta_A; \mu, \tau)/\partial \tau$. By Lemma A.1, the joint asymptotic distribution for $n^{1/2}(\widehat{\mu}_A - \mu_g)$ and $n^{1/2}(\widehat{\mu}_B - \mu_g)$ would be

$$n^{1/2} \begin{pmatrix} \widehat{\mu}_A - \mu_g \\ \widehat{\mu}_B - \mu_g \end{pmatrix} \to$$
$$\mathcal{N} \left\{ \begin{pmatrix} 0_{l\times 1} \\ -f_B^{-1/2} \left[\mathbb{E} \left\{ \partial \Phi_B(V_i; \mu_{g,0}, \tau_0)/\partial \mu \right\} \right]^{-1} \eta \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^{\intercal} & V_B \end{pmatrix} \right\},$$

where V_A, Γ and V_B are the total (co-)variance of two-phase design averaging over the finite populations:

$$\begin{aligned}
V_A &= nN^{-2} \mathbb{E}_\zeta \left[\mathrm{var}_{\mathrm{p}} \left\{ \sum_{i=1}^{N} \psi_A(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right] \\
&\quad + nN^{-2} \mathrm{var}_\zeta \left[\mathbb{E}_{\mathrm{p}} \left\{ \sum_{i=1}^{N} \psi_A(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right], \\
V_B &= nN^{-2} \mathbb{E}_\zeta \left[\mathrm{var}_{\text{p-np}} \left\{ \sum_{i=1}^{N} \psi_B(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right] \\
&\quad + nN^{-2} \mathrm{var}_\zeta \left[\mathbb{E}_{\text{p-np}} \left\{ \sum_{i=1}^{N} \psi_B(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right], \\
\Gamma &= nN^{-2} \mathbb{E}_\zeta \left[\mathrm{cov}_{\text{p-np}} \left\{ \sum_{i=1}^{N} \psi_A(V_i; \mu_g, \tau_0), \sum_{i=1}^{N} \psi_B(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right] \\
&\quad + nN^{-2} \\
&\quad \times \mathrm{var}_\zeta \left[\mathbb{E}_{\mathrm{p}} \left\{ \sum_{i=1}^{N} \psi_A(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\}, \mathbb{E}_{\text{p-np}} \left\{ \sum_{i=1}^{N} \psi_B(V_i; \mu_g, \tau_0) \mid \mathcal{F}_N \right\} \right],
\end{aligned}$$

where the first term is attributed to the randomness of probability (and non-probability) sample designs, and the second term is attributed to the randomness of the superpopulation model. The rest of the proof is summarized in Lemma A.3.

Lemma A.3. *Under the Assumption A.1 and the asymptotic joint distribution for $\widehat{\mu}_A$ and $\widehat{\mu}_B$ in Lemma 3.1, the form of $\widehat{\mu}_{\mathrm{eff}}$ which maximizes the variance reduction under H_0 would be*

$$n^{1/2}(\widehat{\mu}_{\mathrm{eff}} - \mu_0) \cong n^{1/2} \{ \omega_A(\Lambda_{\mathrm{eff}})(\widehat{\mu}_A - \mu_g) + \omega_B(\Lambda_{\mathrm{eff}})(\widehat{\mu}_B - \mu_g) \},$$

where the weight functions are

$$\omega_A(\Lambda) = \mathbb{E} \left\{ \dot{\Phi}_{A,B,n}(\Lambda, \mu_{g,0}, \tau_0) \right\}^{-1} \mathbb{E} \left\{ \dot{\Phi}_A(V_i, \delta_{A,i}; \mu_{g,0}, \tau_0) \right\}, \tag{A.9}$$

$$\omega_B(\Lambda) = \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda, \mu_{g,0}, \tau_0)\right\}^{-1} \Lambda\mathbb{E}\left\{\dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)\right\}, \quad \text{(A.10)}$$

where $\dot{\Phi}_{A,B,n}(\Lambda, \mu_{g,0}, \tau_0) = \dot{\Phi}_A(V_i, \delta_{A,i}; \mu_{g,0}, \tau_0) + \Lambda\dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)$. *The most efficient estimator* $\widehat{\mu}_{\mathrm{eff}}$ *with*

$$\Lambda_{\mathrm{eff}} = \mathbb{E}\left\{\dot{\Phi}_A(V_i; \mu_{g,0}, \tau_0)\right\}(V_A - \Gamma)(V_B - \Gamma^{\intercal})^{-1}\mathbb{E}\left\{\dot{\Phi}_B(V_i; \mu_{g,0}, \tau_0)\right\}^{-1} \quad \text{(A.11)}$$

has the asymptotic distribution under $H_{a,n}$ *as*

$$n^{1/2}(\widehat{\mu}_{\mathrm{eff}} - \mu_g) \to \mathcal{N}\{b_{\mathrm{eff}}(\eta), V_{\mathrm{eff}}\},$$

where $b_{\mathrm{eff}}(\eta) = -f_B^{-1/2}\omega_B(\Lambda_{\mathrm{eff}})\left\{\mathbb{E}\partial\Phi_B(\mu_{g,0}, \tau_0)/\partial\mu\right\}^{-1}\eta$ *and*

$$V_{\mathrm{eff}} = \begin{pmatrix} \omega_A^{\intercal}(\Lambda_{\mathrm{eff}}) \\ \omega_B^{\intercal}(\Lambda_{\mathrm{eff}}) \end{pmatrix}^{\intercal} \begin{pmatrix} V_A & \Gamma \\ \Gamma^{\intercal} & V_B \end{pmatrix} \begin{pmatrix} \omega_A(\Lambda_{\mathrm{eff}}) \\ \omega_B(\Lambda_{\mathrm{eff}}) \end{pmatrix}.$$

When μ_A *and* μ_B *are both scalar,* V_{eff} *would reduce to*

$$V_{\mathrm{eff}} = (V_A V_B - \Gamma^2)(V_A + V_B - 2\Gamma)^{-1} = V_A - V_\Delta,$$

where $V_\Delta = (V_A - \Gamma)^2(V_A + V_B - 2\Gamma)^{-1}$.

A.3. Proof of Lemma 3.2

By applying the Taylor expansion with Lagrange forms of remainder to the asymptotic distribution for $n_B^{1/2}N^{-1}\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau})$ in (3.5) could be shown as

$$n_B^{1/2}N^{-1}\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) = n_B^{1/2}N^{-1}\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_g, \tau_0)$$
$$+ n_B^{1/2}N^{-1}\sum_{i=1}^N \begin{pmatrix} \frac{\partial\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A^*, \tau^*)}{\partial\mu} & \frac{\partial\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A^*, \tau^*)}{\partial\tau} \end{pmatrix} \begin{pmatrix} \widehat{\mu}_A - \mu_g \\ \widehat{\tau} - \tau_0 \end{pmatrix}$$

where $(\widehat{\mu}_A^* \quad \tau^*)^{\intercal}$ is the neighborhood of $(\mu_{g,0}, \tau_0)^{\intercal}$ as $\mathrm{plim}\widehat{\mu}_A = \mu_{g,0}$ and $\mathrm{plim}\widehat{\tau} = \tau_0$. Under the Assumption A.1 e), we have

$$\begin{aligned} & n_B^{1/2}N^{-1}\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) \\ &= n_B^{1/2}N^{-1}\sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_g, \tau_0) \qquad \text{(A.12)} \\ &+ n_B^{1/2}N^{-1}\sum_{i=1}^N \frac{\partial\Phi_{B,j}(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A^*, \tau^*)}{\partial\mu}(\widehat{\mu}_A - \mu_g) \\ &+ n_B^{1/2}N^{-1}\sum_{i=1}^N \frac{\partial\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A^*, \tau^*)}{\partial\tau}(\widehat{\tau} - \tau_0) \end{aligned}$$

$$
\begin{aligned}
&= n_B^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_g, \tau_0) \\
&+ \mathbb{E}\left\{ \frac{\partial \Phi_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0)}{\partial \tau} \right\} n_B^{1/2}(\widehat{\tau} - \tau_0) \\
&+ \mathbb{E}\left\{ \frac{\partial \Phi_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0)}{\partial \mu} \right\} n_B^{1/2}(\widehat{\mu}_A - \mu_g) + o_{\zeta\text{-p-np}}(1).
\end{aligned} \tag{A.13}
$$

Next, by replacing the first two term in Equation (A.13) with Equation (A.2), we have

$$
\begin{aligned}
& n_B^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) \\
&= -\mathbb{E}\left\{ \frac{\partial \Phi_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0)}{\partial \mu} \right\} n_B^{1/2}(\widehat{\mu}_B - \mu_g) \\
&+ \mathbb{E}\left\{ \frac{\partial \Phi_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0)}{\partial \mu} \right\} n_B^{1/2}(\widehat{\mu}_A - \mu_g) + o_{\zeta\text{-p-np}}(1) \\
&= -(n_B/n)^{1/2} \cdot \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau_0) \cdot n^{1/2}(\widehat{\mu}_B - \mu_g) \\
&+ (n_B/n)^{1/2} \cdot \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau_0) \cdot n^{1/2}(\widehat{\mu}_A - \mu_g) + o_{\zeta\text{-p-np}}(1),
\end{aligned}
$$

provided by WLLN under Assumptions 2.1, 2.2 (iii) and Assumption A.1. By the joint distribution of $\widehat{\mu}_A$ and $\widehat{\mu}_B$ in Lemma 3.1, the variance of $\Phi_{B,n}(\widehat{\mu}_A, \widehat{\tau})$ would be

$$
\Sigma_T = f_B \left\{ \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau) \right\} (V_A + V_B - \Gamma^\intercal - \Gamma) \left\{ \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau) \right\}^\intercal .
$$

Thus, the asymptotic distribution for $\Phi_{B,n}(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau})$ would be

$$
\begin{aligned}
& n_B^{1/2} N^{-1} \sum_{i=1}^{N} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) \\
& \to \mathcal{N}\left\{ \eta, f_B \left\{ \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau) \right\} (V_A + V_B - \Gamma^\intercal - \Gamma) \left\{ \mathbb{E}\dot{\Phi}_B(V_i; \mu_{g,0}, \tau) \right\}^\intercal \right\}.
\end{aligned}
$$

A.4. Proof of Theorem 4.1

From Lemma 2.1 and 3.1, we know that the asymptotic joint distribution for $\widehat{\mu}_A$ and $\widehat{\mu}_B$ would be

$$
\begin{aligned}
& n^{1/2} \begin{pmatrix} \widehat{\mu}_A - \mu_g \\ \widehat{\mu}_B - \mu_g \end{pmatrix} \\
& \to \mathcal{N}\left\{ \begin{pmatrix} \mathbf{0}_{l\times 1} \\ -f_B^{-1/2} \left[\mathbb{E}\left\{ \partial \Phi_B(\mu_{g,0}, \tau_0)/\partial \mu \right\} \right]^{-1} \eta \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \right\}.
\end{aligned}
$$

For simplicity, we let $n^{1/2}(\widehat{\mu}_A - \mu_g)$ and $n^{1/2}(\widehat{\mu}_B - \mu_g)$ be asymptotically distributed as Z_1 and Z_2, respectively. Then, $\Phi_{B,n}(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau})$ could be expressed as

$$
\begin{aligned}
& n_B^{1/2} N^{-1} \sum_{i=1}^N \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) \\
\cong & - n_B^{1/2} N^{-1} \sum_{i=1}^N \dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)(\widehat{\mu}_B - \mu_g) \\
& + n_B^{1/2} N^{-1} \sum_{i=1}^N \dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)(\widehat{\mu}_A - \mu_g) \\
\to & f_B^{1/2} \left\{ \mathbb{E}\dot{\Phi}_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0) \right\} (Z_1 - Z_2).
\end{aligned}
$$

Let $U_2 = f_B^{1/2} \left\{ \mathbb{E}\dot{\Phi}_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0) \right\} (Z_1 - Z_2)$. Next step, we attempt to find another linear combination of Z_1 and Z_2 which is orthogonal to U_2. Observed that when $U_1 = f_B^{1/2}\{(\Gamma^\intercal - V_B)(\Gamma - V_A)^{-1} Z_1 + Z_2\}$, it is easy to verify that the covariance of U_1 and U_2 is zero under H_0.

$$
\begin{aligned}
\operatorname{cov}(U_2, U_1) &= f_B \left\{ \mathbb{E}\dot{\Phi}_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0) \right\} \\
&\times \begin{pmatrix} I_{l\times l} \\ -I_{l\times l} \end{pmatrix}^\intercal \times \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \times \begin{pmatrix} (\Gamma^\intercal - V_A)^{-1}(\Gamma - V_B) \\ I_{l\times l} \end{pmatrix} \\
&= f_B \left\{ \mathbb{E}\dot{\Phi}_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0) \right\} \\
&\times \begin{pmatrix} V_A - \Gamma^\intercal & \Gamma - V_B \end{pmatrix} \times \begin{pmatrix} (\Gamma^\intercal - V_A)^{-1}(\Gamma - V_B) \\ I_{l\times l} \end{pmatrix} \\
&= 0_{l\times l}.
\end{aligned}
$$

Also, since U_1 and U_2 are both asymptotically normal distributions, which implies that zero covariance leads to independency. After a few standardization procedures, we have W_1 and W_2 as $W_1 = \Sigma_S^{-1/2} U_1$, $W_2 = \Sigma_T^{-1/2} U_2$ with Σ_S and Σ_T defined as

$$
\Sigma_S = \operatorname{var}(U_1) = f_B \operatorname{var}\{(\Gamma^\intercal - V_B)(\Gamma - V_A)^{-1} Z_1 + Z_2\}, \tag{A.14}
$$

$$
\Sigma_T = f_B \left\{ \mathbb{E}\dot{\Phi}_B(\mu_{g,0}, \tau_0) \right\} (V_A + V_B - \Gamma^\intercal - \Gamma) \left\{ \mathbb{E}\dot{\Phi}_B(\mu_{g,0}, \tau_0) \right\}^\intercal. \tag{A.15}
$$

Therefore, we have the form for the standardized random variables W_1 and W_2 as

$$
\begin{aligned}
W_1 &= \Sigma_S^{-1/2} U_1 = f_B^{1/2} \Sigma_S^{-1/2} \{(\Gamma^\intercal - V_B)(\Gamma - V_A)^{-1} Z_1 + Z_2\}, \\
W_2 &= -\Sigma_T^{-1/2} U_2 = -(V_A + V_B - \Gamma^\intercal - \Gamma)^{-1/2}(Z_1 - Z_2).
\end{aligned}
$$

Here we use $-\Sigma_T^{-1/2}$ to standardize U_2 for the sake of convenience later. Therefore, under the local alternative $H_{a,n} : \mathbb{E}\{\Phi_B(V, \delta_A, \delta_B; \mu_{g,0}, \tau_0)\} = n_B^{-1/2}\eta$, we have that $\mathbb{E}(Z_1) = 0, \mathbb{E}(Z_2) = -f_B^{-1/2} \left\{ \mathbb{E}\dot{\Phi}_B(\mu_{g,0}, \tau_0) \right\}^{-1} \eta$. Combining the above leads to

$$
W_1 \sim N(\mu_1, I_{l\times l}),\ W_2 \sim N(\mu_2, I_{l\times l}),
$$

where

$$\mu_1 = -\Sigma_S^{-1/2}\left\{\mathbb{E}\dot{\Phi}_B(\mu_{g,0},\tau_0)\right\}^{-1}\eta,$$
$$\mu_2 = -f_B^{-1/2}(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1/2}\left\{\mathbb{E}\dot{\Phi}_B(\mu_{g,0},\tau_0)\right\}^{-1}\eta = -\Sigma_T^{-1/2}\eta,$$

and since $W_1 \perp W_2$, we could project out TAP estimator $\widehat{\mu}_{\text{tap}}$ with the optimal tuning parameter $(\Lambda^*, c_{\gamma^*})$ onto these two basis respectively. First, on the condition that

$$T > c_{\gamma^*} = \{\Phi_{B,n}(\widehat{\mu}_A,\widehat{\tau})\}^\intercal\,\widehat{\Sigma}_T^{-1}\{\Phi_{B,n}(\widehat{\mu}_A,\widehat{\tau})\} > c_{\gamma^*} \rightarrow W_2^\intercal W_2 > c_{\gamma^*},$$

we have

$$\begin{aligned}
&n^{1/2}(\widehat{\mu}_{\text{tap}}-\mu_g)\mid T > c_{\gamma^*} = n^{1/2}(\widehat{\mu}_A-\mu_g)\mid T > c_{\gamma^*}\\
&\rightarrow Z_1 | W_2^\intercal W_2 > c_{\gamma^*}\\
&\rightarrow -f_B^{-1/2}(\Gamma-V_A)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}U_1\\
&+f_B^{-1/2}(\Gamma-V_A)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}\left\{\mathbb{E}\dot{\Phi}_B(\mu_{g,0},\tau_0)\right\}^{-1}U_2|W_2^\intercal W_2 > c_{\gamma^*}\\
&\rightarrow -f_B^{-1/2}(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}\Sigma_S^{1/2}W_1\\
&+(\Gamma-V_A)(V_A+V_B-\Gamma-\Gamma^\intercal)^{-1/2}W_2|W_2^\intercal W_2 > c_{\gamma^*}\\
&\rightarrow -V_{\text{eff}}^{1/2}W_1 + V_{\text{A-eff}}^{1/2}W_2|W_2^\intercal W_2 > c_{\gamma^*}.
\end{aligned}$$

Next, on the condition $T = W_2^\intercal W_2 \leq c_{\gamma^*}$, we have

$$\begin{aligned}
&n^{1/2}(\widehat{\mu}_{\text{tap}}-\mu_g)\rightarrow \omega_A^* Z_1 + \omega_B^* Z_2 | W_2^\intercal W_2 \leq c_{\gamma^*}\\
&\rightarrow -f_B^{-1/2}(\Gamma-V_A)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}U_1\\
&+f_B^{-1/2}\omega_A^*(\Gamma-V_A)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}\\
&\times\left\{\mathbb{E}\dot{\Phi}_B(\mu_{g,0},\tau_0)\right\}^{-1}U_2\mid W_2^\intercal W_2 \leq c_{\gamma^*}\\
&-f_B^{-1/2}\omega_B^*(\Gamma^\intercal-V_B)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}\\
&\times\left\{\mathbb{E}\dot{\Phi}_B(\mu_{g,0},\tau_0)\right\}^{-1}U_2\mid W_2^\intercal W_2 \leq c_{\gamma^*}\\
&\rightarrow -f_B^{-1/2}(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1}\Sigma_S^{1/2}W_1\\
&+f_B^{-1/2}\omega_A^*(\Gamma-V_A)(V_A+V_B-\Gamma-\Gamma^\intercal)^{-1/2}W_2\mid W_2^\intercal W_2 \leq c_{\gamma^*}\\
&-f_B^{-1/2}\omega_B^*(\Gamma^\intercal-V_B)(V_A+V_B-\Gamma^\intercal-\Gamma)^{-1/2}W_2\mid W_2^\intercal W_2 \leq c_{\gamma^*}\\
&\rightarrow -V_{\text{eff}}^{1/2}W_1 + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2})W_2|W_2^\intercal W_2 \leq c_{\gamma^*},
\end{aligned}$$

where $W_2^t = W_2|W_2^\intercal W_2 \leq c_\gamma$, and ω_A^*, ω_B^* are the new tuned weighted functions defined in (A.9) and (A.10) with $\Lambda = \Lambda^*$. In this way, we could fully characterize the asymptotic distribution for the TAP estimator $\widehat{\mu}_{\text{tap}}$ under the optimal tuning parameter as,

$$n^{1/2}(\widehat{\mu}_{\text{tap}}-\mu_g)\rightarrow\begin{cases}-V_{\text{eff}}^{1/2}W_1 + (\omega_A V_{\text{A-eff}}^{1/2}-\omega_B V_{\text{B-eff}}^{1/2})W^t_{[0,c_\gamma]} & w.p.\ \xi,\\ -V_{\text{eff}}^{1/2}W_1 + V_{\text{A-eff}}^{1/2}W^t_{[c_{\gamma^*},\infty]} & w.p.\ 1-\xi,\end{cases}$$

where $\xi = \mathbb{P}(W_2^\intercal W_2 < c_{\gamma^*})$.

A.5. Proof of the bias and mean squared error of $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$

For general case, given $W_2 \sim N_p(\mu_2, I_{p\times p})$, the MGF of truncated normal distribution $W_2 | a \le W_2^\intercal W_2 \le b$ is [60]

$$\begin{aligned}
\alpha m(t) &= \mathbb{E}\{\exp(t^\intercal W_2)\}\\
&= (2\pi)^{-p/2}\int_{\mathbb{C}} \exp(t^\intercal W_2)\exp\left\{-\frac{1}{2}(W_2-\mu_2)^\intercal (W_2-\mu_2)\right\} dW_2\\
&= (2\pi)^{-p/2}\exp(\frac{1}{2}t^\intercal t + \mu_2^\intercal t)\\
&\times \int_{\mathbb{C}} \exp\left\{-\frac{1}{2}(W_2-\mu_2-t)^\intercal (W_2-\mu_2-t)\right\} dW_2\\
&= \exp(-\frac{1}{2}\mu_2^\intercal \mu_2)\sum_{k=0}^{\infty}\{F_{p+2k}(b) - F_{p+2k}(a)\}\{(\mu_2+t)^\intercal(\mu_2+t)/2\}^k/k!,
\end{aligned}$$

where $\alpha = F_p(b;\mu_2^\intercal\mu_2/2) - F_p(a;\mu_2^\intercal\mu_2/2)$ is the normalization constant and $F_p(a;\mu_2^\intercal\mu_2/2)$ is CDF of chi-square distribution at value a with non-central parameter $\mu_2^\intercal\mu_2/2$. The second and the third equality above are justified by

$$\begin{aligned}
&(2\pi)^{-p/2}\int_{\mathbb{C}} \exp\left\{-\frac{1}{2}(W_2-\mu_2-t)^\intercal (W_2-\mu_2-t)\right\}\\
&= \mathbb{P}\{a \le W_2^\intercal W_2 \le b \mid W_2 \sim \mathcal{N}(\mu_2+t, I_{p\times p})\}\\
&= F\{b; k=p, \lambda = (\mu_2+t)^\intercal(\mu_2+t)\} - F\{a; k=p, \lambda=(\mu_2+t)^\intercal(\mu_2+t)\}\\
&= \exp\{-\frac{1}{2}(\mu_2+t)^\intercal(\mu_2+t)\}\\
&\times \sum_{k=0}^{\infty}\{F_{p+2k}(b) - F_{p+2k}(a)\}\{(\mu_2+t)^\intercal(\mu_2+t)/2\}^k/k!.
\end{aligned}$$

To compute the first and second moment of this truncated normal distribution, we take derivative of the MGF and evaluate the function at $t=0$

$$\begin{aligned}
\alpha \left.\frac{dm(t)}{dt^\intercal}\right|_{t=0} &= (\mu_2+t)\exp(-\frac{1}{2}\mu_2^\intercal\mu_2)\\
&\times \sum_{k=0}^{\infty}\{F_{p+2k+2}(b) - F_{p+2k+2}(a)\}\{(\mu_2+t)^\intercal(\mu_2+t)\}/k!|_{t=0}\\
&= \mu_2\exp(-\frac{1}{2}\mu_2^\intercal\mu_2)\sum_{k=0}^{\infty}\{F_{p+2k+2}(b) - F_{p+2k+2}(a)\}\{\mu_2^\intercal\mu_2/2\}^k/k!\\
&= \mu_2\left\{F_{p+2}(b;\mu_2^\intercal\mu_2/2) - F_{p+2}(a;\mu_2^\intercal\mu_2/2)\right\}.
\end{aligned}$$

By the nature of MGF, we obtain the expectation of the first moment of W_2

$$\mathbb{E}(W_2 | a \le W_2^\intercal W_2 \le b) = \mu_2 \cdot \frac{F_{p+2}(b;\mu_2^\intercal\mu_2/2) - F_{p+2}(a;\mu_2^\intercal\mu_2/2)}{F_p(b;\mu_2^\intercal\mu_2/2) - F_p(a;\mu_2^\intercal\mu_2/2)}.$$

Then, taking the second derivative of the MGF follows by

$$
\begin{aligned}
&\alpha \frac{d^2 m(t)}{dtdt^\intercal}\bigg|_{t=0} \\
&= \exp(-\frac{1}{2}\mu_2^\intercal \mu_2) \left[\sum_{k=0}^{\infty} \{F_{p+2k+2}(b) - F_{p+2k+2}(a)\}\{(\mu_2+t)^\intercal(\mu_2+t)/2\}^k/k!|_{t=0} \right. \\
&+(\mu_2+t)(\mu_2+t)^\intercal \\
&\left. \times \sum_{k=0}^{\infty} \{F_{p+2k+4}(b) - F_{p+2k+4}(a)\}\{(\mu_2+t)^\intercal(\mu_2+t)/2\}^k/k!|_{t=0} \right] \\
&= \exp(-\frac{1}{2}\mu_2^\intercal \mu_2) \left[\sum_{k=0}^{\infty} \{F_{p+2k+2}(b) - F_{p+2k+2}(a)\}\{\mu_2^\intercal\mu_2/2\}^k/k! \right. \\
&\left. +\mu_2\mu_2^\intercal \sum_{k=0}^{\infty} \{F_{p+2k+4}(b) - F_{p+2k+4}(a)\}\{\mu_2^\intercal\mu_2/2\}^k/k! \right] \\
&= I_{p\times p}(F_{p+2}(b;\mu_2^\intercal\mu_2/2) - F_{p+2}(a;\mu_2^\intercal\mu_2/2)) \\
&+ \mu_2\mu_2^\intercal(F_{p+4}(b;\mu_2^\intercal\mu_2/2) - F_{p+4}(a;\mu_2^\intercal\mu_2/2)),
\end{aligned}
$$

which leads to

$$
\begin{aligned}
\mathbb{E}(W_2W_2^T | a \le W_2^T W_2 \le b) = I_{p\times p} &\frac{F_{p+2}(b;\mu_2^T\mu_2/2) - F_{p+2}(a;\mu_2^T\mu_2/2)}{F_p(b;\mu_2^T\mu_2/2) - F_p(a;\mu_2^T\mu_2/2)} \\
&+ \mu_2\mu_2^\intercal \frac{F_{p+4}(b;\mu_2^T\mu_2/2) - F_{p+4}(a;\mu_2^T\mu_2/2)}{F_p(b;\mu_2^T\mu_2/2) - F_p(a;\mu_2^T\mu_2/2)}.
\end{aligned}
$$

In our case,

$$
p = l, \quad \mu_1 = -\Sigma_S^{-1/2} \left[\mathbb{E}\{\partial\Phi_B(\mu_{g,0},\tau_0)/\partial\mu\}\right]^{-1} \eta, \quad \mu_2 = -\Sigma_T^{-1/2}\eta.
$$

Recall, for $T \le c_\gamma$, we have $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \to -V_{\text{eff}}^{1/2} W_1 + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) W_2$ $|W_2^\intercal W_2 \le c_\gamma$ with probability $\xi = F_l(c_\gamma; \mu_2^\intercal\mu_2)$, the bias would be

$$
\begin{aligned}
\text{bias}(\lambda, c_\gamma; \eta)_{T\le c_\gamma} &= -V_{\text{eff}}^{1/2}\mu_1 + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) \cdot \mathbb{E}(W_2 | W_2^\intercal W_2 \le c_\gamma) \\
&= -V_{\text{eff}}^{1/2}\mu_1 + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) \cdot \frac{F_{l+2}(c_\gamma; \mu_2^T\mu_2/2)\mu_2}{F_l(c_\gamma;\mu_2^T\mu_2/2)}.
\end{aligned}
$$

The MSE can be derived based on the known formula $\text{mse}(X+Y) = \text{var}(X+Y) + \{\mathbb{E}(X+Y)\}^{\otimes 2} = \{\text{var}(X) + \mu_X^{\otimes 2}\} + \{\text{var}(Y) + \mu_Y^{\otimes 2}\} + 2\mu_X\mu_Y^\intercal$

$$
\begin{aligned}
&\text{mse}(\lambda, c_\gamma; \eta)_{T\le c_\gamma} = V_{\text{eff}}^{1/2}(\mu_1\mu_1^\intercal + I_{l\times l})V_{\text{eff}}^{1/2} + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) \\
&\times E(W_2W_2^T | W_2^\intercal W_2 \le c_\gamma)(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) \\
&- 2V_{\text{eff}}^{1/2}\mu_1 E(W_2^\intercal | W_2^\intercal W_2 \le c_\gamma)(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) \\
&= V_{\text{eff}}^{1/2}(\mu_1\mu_1^\intercal + I_{l\times l})V_{\text{eff}}^{1/2} + (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2})
\end{aligned}
$$

$$\times \left\{ \frac{F_{l+2}(c_\gamma; \mu_2^T \mu_2/2)}{F_l(c_\gamma; \mu_2^T \mu_2/2)} I_{l\times l} + \frac{F_{l+4}(c_\gamma; \mu_2^T \mu_2/2)}{F_l(c_\gamma; \mu_2^T \mu_2/2)} \mu_2 \mu_2^\intercal \right\} (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2})$$
$$- \frac{2F_{l+2}(c_\gamma; \mu_2^T \mu_2/2)}{F_l(c_\gamma; \mu_2^T \mu_2/2)} V_{\text{eff}}^{1/2} \mu_1 \mu_2^\intercal (\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}).$$

For $T > c_\gamma$, we have $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \rightarrow -V_{\text{eff}}^{1/2} W_1 + V_{\text{A-eff}}^{1/2} W_2 | W_2^\intercal W_2 > c_\gamma$ with probability $1 - \xi = 1 - F_l(c_\gamma; \mu_2^\intercal \mu_2)$, the corresponding bias and MSE would be

$$\begin{aligned} \text{bias}(\lambda, c_\gamma; \eta)_{T>c_\gamma} &= -V_{\text{eff}}^{1/2} \mu_1 + V_{\text{A-eff}}^{1/2} \cdot E(W_2 | W_2^\intercal W_2 > c_\gamma) \\ &= -V_{\text{eff}}^{1/2} \mu_1 + V_{\text{A-eff}}^{1/2} \cdot \frac{1 - F_{l+2}(c_\gamma; \mu_2^T \mu_2/2)\mu_2}{1 - F_l(c_\gamma; \mu_2^T \mu_2/2)}, \end{aligned}$$

and

$$\begin{aligned} &\text{mse}(\lambda, c_\gamma; \eta)_{T>c_\gamma} = V_{\text{eff}}^{1/2} (\mu_1 \mu_1^\intercal + I_{l\times l}) V_{\text{eff}}^{1/2} \\ &+ V_{\text{A-eff}}^{1/2} E(W_2 W_2^T | W_2^\intercal W_2 > c_\gamma) V_{\text{A-eff}}^{1/2} \\ &- 2V_{\text{eff}}^{1/2} \mu_1 E(W_2^\intercal | W_2^\intercal W_2 > c_\gamma) V_{\text{A-eff}}^{1/2} \\ &= V_{\text{eff}}^{1/2} (\mu_1 \mu_1^\intercal + I_{l\times l}) V_{\text{eff}}^{1/2} + V_{\text{A-eff}}^{1/2} \\ &\times \left\{ \frac{1 - F_{l+2}(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_l(c_\gamma; \mu_2^T \mu_2/2)} I_{l\times l} + \frac{1 - F_{l+4}(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_l(c_\gamma; \mu_2^T \mu_2/2)} \mu_2 \mu_2^\intercal \right\} V_{\text{A-eff}}^{1/2} \\ &- 2 \left\{ \frac{1 - F_3(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_1(c_\gamma; \mu_2^T \mu_2/2)} \right\} V_{\text{eff}}^{1/2} \mu_1 \mu_2^\intercal V_{\text{A-eff}}^{1/2}. \end{aligned}$$

Overall, the bias and mean squared error for $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ can be characterized as

$$\text{bias}(\lambda, c_\gamma; \eta) = \xi \cdot \text{bias}(\lambda, c_\gamma; \eta)_{T\le c_\gamma} + (1 - \xi) \cdot \text{bias}(\lambda, c_\gamma; \eta)_{T>c_\gamma},$$
$$\text{mse}(\lambda, c_\gamma; \eta) = \xi \cdot \text{mse}(\lambda, c_\gamma; \eta)_{T\le c_\gamma} + (1 - \xi) \cdot \text{mse}(\lambda, c_\gamma; \eta)_{T>c_\gamma}.$$

A.6. Proof of the asymptotic distribution for $U(a)$

Throughout the proof, we assume that the regularity conditions in Lemma 2.1 and assumptions in Theorem 4.2 hold, we prove that the coverage probability for the adaptive projection sets is guaranteed to be larger than $1 - \alpha$, which is

$$\mathbb{P}\left\{ a^\intercal \mu_g \in \mathbb{C}_{\mu_g, 1-\alpha}^{\text{BACI}}(a) \right\} \ge 1 - \alpha + o(1),$$

where $\mathbb{C}_{\mu_g, 1-\alpha}^{\text{BACI}}(a) = \left[a^\intercal \widehat{\mu}_{\text{tap}} - \widehat{U}_{1-\alpha/2}(a)/\sqrt{n}, a^\intercal \widehat{\mu}_{\text{tap}} - \widehat{L}_{\alpha/2}(a)/\sqrt{n} \right]$. As we already know that

$$a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \le U(a), \quad a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) \ge L(a),$$

it is needed to show that $\widehat{U}(a)$ obtained by bootstrapping converges to the same asymptotic distribution as $U(a)$. Let $D_{p\times p}$ denotes the space of $p \times p$ symmetric positive-definite matrices equipped with the spectral norm. We can rewrite $U(a)$ as

$$
\begin{aligned}
U(a) &= -a^{\intercal} V_{\text{eff}}^{1/2} W_1\{\Sigma_S, n^{1/2}(\widehat{\mu}_A - \mu_g), n^{1/2}(\widehat{\mu}_B - \mu_g), \tau\} \\
&+ a^{\intercal}(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) W_2\{\Sigma_T, n^{1/2}(\widehat{\mu}_A - \mu_g), n^{1/2}(\widehat{\mu}_B - \mu_g), \tau\} \\
&+ a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})\mu^t_{[c_\gamma,\infty)} \\
&+ a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) \\
&\times \left[W_2\{\Sigma_T, n^{1/2}(\widehat{\mu}_A - \mu_g), n^{1/2}(\widehat{\mu}_B - \mu_g), \tau\}_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)}\right] \mathbf{1}_{T \geq v_n} \\
&+ a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) \\
&\times \sup_{\mu_2 \in \mathbb{R}^l} \left[W_2\{\Sigma_T, n^{1/2}(\widehat{\mu}_A - \mu_g), n^{1/2}(\widehat{\mu}_B - \mu_g), \tau\}_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)}\right] \mathbf{1}_{T < v_n}.
\end{aligned}
$$

Next, we adopt the notation for the bootstrapping to express the upper bound $\widehat{U}(a) = U^{(b)}(a)$ as

$$
\begin{aligned}
U^{(b)}(a) &= -a^{\intercal} \widehat{V}_{\text{eff}}^{1/2} W_1\{\widehat{\Sigma}_S, n^{1/2}(\widehat{\mu}_A^{(b)} - \widehat{\mu}_A), n^{1/2}(\widehat{\mu}_B^{(b)} - \widehat{\mu}_A), \widehat{\tau}\} \\
&+ a^{\intercal}(\omega_A \widehat{V}_{\text{A-eff}}^{1/2} - \omega_B \widehat{V}_{\text{B-eff}}^{1/2}) W_2\{\widehat{\Sigma}_T, n^{1/2}(\widehat{\mu}_A^{(b)} - \widehat{\mu}_A), n^{1/2}(\widehat{\mu}_B^{(b)} - \widehat{\mu}_A), \widehat{\tau}\} \\
&+ a^{\intercal}\omega_B(\widehat{V}_{\text{B-eff}}^{1/2} + \widehat{V}_{\text{A-eff}}^{1/2}) \bar{W}_2^{(b)}{}_{[c_\gamma,\infty)} \\
&+ a^{\intercal}\omega_B(\widehat{V}_{\text{B-eff}}^{1/2} + \widehat{V}_{\text{A-eff}}^{1/2}) \\
&\times \left[W_2\{\widehat{\Sigma}_T, n^{1/2}(\widehat{\mu}_A^{(b)} - \widehat{\mu}_A), n^{1/2}(\widehat{\mu}_B^{(b)} - \widehat{\mu}_A), \widehat{\tau}\}_{[c_\gamma,\infty)} - \bar{W}_2^{(b)}{}_{[c_\gamma,\infty)}\right] \mathbf{1}_{T \geq v_n} \\
&+ a^{\intercal}\omega_B(\widehat{V}_{\text{B-eff}}^{1/2} + \widehat{V}_{\text{A-eff}}^{1/2}) \\
&\times \sup_{\mu_2 \in \mathbb{R}^l} \left[W_2\{\widehat{\Sigma}_T, n^{1/2}(\widehat{\mu}_A^{(b)} - \widehat{\mu}_A), n^{1/2}(\widehat{\mu}_B^{(b)} - \widehat{\mu}_A), \widehat{\tau}\}_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)}\right] \mathbf{1}_{T < v_n},
\end{aligned}
$$

where $\bar{W}_2^{(b)} = (1/K)\sum_{b=1}^K W_2\{\widehat{\Sigma}_T, n^{1/2}(\widehat{\mu}_A^{(b)} - \widehat{\mu}_A), n^{1/2}(\widehat{\mu}_B^{(b)} - \widehat{\mu}_A), \widehat{\tau}\}$. Next, we define some functions to proceed our proof. $w_{11} : D_{l\times l} \times D_{l\times l} \times \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$, $w_{12} : D_{l\times l} \times \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^d \times \mathbb{R}^l \to \mathbb{R}$ and $\rho : D_{2l\times 2l} \times D_{l\times l} \times \mathbb{R}^l \times \mathbb{R}^l \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^l \to \mathbb{R}$ are functions defined as below

$$
\begin{aligned}
w_{11}(\Sigma_T, \Sigma_S, \mathbb{G}_A, \mathbb{G}_B, \tau, \mu_2) &= -a^{\intercal} V_{\text{eff}}^{1/2} W_1(\Sigma_S, \mathbb{G}_A, \mathbb{G}_B, \tau) \\
&+ a^{\intercal}(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) W_2(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \tau) \\
&+ a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})\mu^t_{[c_\gamma,\infty)} \\
&+ a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) \\
&\times \left\{W_2(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \tau)_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)}\right\} \mathbf{1}_{\mu_2^{\intercal}\mu_2 \in \mathbb{B}^{\complement}}, \\
w_{12}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) &= a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) \\
&\times \left\{W_2(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \tau)_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)}\right\} \mathbf{1}_{\mu_2^{\intercal}\mu_2 \in \mathbb{B}}, \\
\rho_{11}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) &= a^{\intercal}\omega_B(V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2})
\end{aligned}
$$

$$\times \left\{ W_2(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \tau)_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)} \right\} (\mathbf{1}_{T \geq v_n} - \mathbf{1}_{\mu_2^\intercal \mu_2 \in \mathbb{B}^\complement}),$$

$$\rho_{12}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) = a^\intercal \omega_B (V^{1/2}_{\text{B-eff}} + V^{1/2}_{\text{A-eff}}) \times \left\{ W_2(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \tau)_{[c_\gamma,\infty)} - \mu^t_{[c_\gamma,\infty)} \right\} (\mathbf{1}_{T < v_n} - \mathbf{1}_{\mu_2^\intercal \mu_2 \in \mathbb{B}}),$$

where $\mathbb{G}_A = n^{1/2}(\widehat{\mu}_A - \mu_g)$ and $\mathbb{G}_B = n^{1/2}(\widehat{\mu}_B - \mu_g)$. Using the functions we have defined, we could re-express the upper bound $U(a)$ in terms of

$$U(a) = w_{11}(\Sigma_T, \Sigma_S, \mathbb{G}_A, \mathbb{G}_B, \tau, \mu_2) + \rho_{11}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) + \sup_{\mu_2 \in \mathbb{R}^l} \{ w_{12}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) + \rho_{12}(\Sigma_T, \mathbb{G}_A, \mathbb{G}_B, \mu_2) \}.$$

Assume the conditions in Theorem 4.2, we can show that

1. w_{11} is continuous at points in $(\Sigma_T, \Sigma_S, \mathbb{R}^l, \mathbb{R}^l, \mathbb{R}^d, \mu_2)$ and w_{12} is continuous at points in $(\Sigma_T, \mathbb{R}^l, \mathbb{R}^l, \mu_2)$ uniformly in μ_2. That is, for any $\widehat{\Sigma}_T \to \Sigma_T$, $\widehat{\Sigma}_S \to \Sigma_S$, $\mathbb{G}^{(b)}_A = n^{1/2}(\widehat{\mu}^{(b)}_A - \widehat{\mu}_A) \to Z_1$, $\mathbb{G}^{(b)}_B = n^{1/2}(\widehat{\mu}^{(b)}_B - \widehat{\mu}_A) \to Z_2$ and $\widehat{\tau} \to \tau$, we have
$$\begin{aligned} &\sup_{\mu_2 \in \mathbb{R}^l} |w_{11}(\widehat{\Sigma}_T, \widehat{\Sigma}_S, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \widehat{\tau}, \mu_2) - w_{11}(\Sigma_T, \Sigma_S, Z_1, Z_2, \tau, \mu_2)| \to 0, \\ &\sup_{\mu_2 \in \mathbb{R}^l} |w_{12}(\widehat{\Sigma}_T, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \mu_2) - w_{12}(\Sigma_T, Z_1, Z_2, \mu_2)| \to 0. \end{aligned} \tag{A.16}$$
2. $\rho_{11}(\widehat{\Sigma}_T, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \mu_2)$ and $\rho_{12}(\widehat{\Sigma}_T, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \mu_2)$ converge to zeros with probability one as $n \to \infty$ uniformly in μ_2. That is,
$$\sup_{\mu_2 \in \mathbb{R}^l} |\rho_{11}(\widehat{\Sigma}_T, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \mu_2)| \to 0, \quad \max_{\mu_2 \in \mathbb{R}^l} |\rho_{12}(\widehat{\Sigma}_T, \mathbb{G}^{(b)}_A, \mathbb{G}^{(b)}_B, \mu_2)| \to 0. \tag{A.17}$$
See Lemma B.9. and Lemma B.11. in [32] for details.

By far, combine (A.16) and (A.17), $U(a)$ is guaranteed to be continuous, and the continuity of $L(a)$ can be derived in the same way. Based on continuous mapping theorem and Theorem 4.2 in [32], we can state that

$$\sup_M |\mathbb{E}\{L(a), U(a)\} - \mathbb{E}_M\{L^{(b)}(a), U^{(b)}(a)\}|$$

converges to zero in probability, where $\mathbb{E}_M(\cdot)$ denotes the expectation taken with respect to the bootstrap weights.

A.7. Proof of Theorem 4.2

Based on the established consistency of the bootstrapping bounds in Section A.6, the proof can be decomposed into two parts. One part is for

$$\mathbb{P}\{a^\intercal \sqrt{n}(\widehat{\mu}_{\text{tap}} - \mu_g) \leq \widehat{U}_{1-\alpha/2}(a)\} \geq \mathbb{P}\{U(a) \leq \widehat{U}_{1-\alpha/2}(a)\}$$

$$
\begin{aligned}
&= G_U\{\widehat{U}_{1-\alpha/2}(a)\} - \widehat{G}_U\{\widehat{U}_{1-\alpha/2}(a)\} \\
&+ \widehat{G}_U\{\widehat{U}_{1-\alpha/2}(a)\} \\
&= o(1) + 1 - \alpha/2,
\end{aligned}
$$

where $G_U(\cdot)$ is the cumulative distribution function for $U(a)$. Let $\widehat{G}_U(\cdot)$ be the empirical cumulative distribution function $\widehat{U}(a)$ estimated by bootstrapping. Similarly, we can show that the other part of our proof as

$$
\begin{aligned}
\mathbb{P}\{a^\intercal\sqrt{n}(\widehat{\mu}_{\text{tap}} - \mu_g) \le \widehat{L}_{\alpha/2}(a)\} &\le \mathbb{P}\{L(a) \le \widehat{L}_{\alpha/2}(a)\} \\
&= G_L\{\widehat{L}_{\alpha/2}(a)\} - \widehat{G}_L\{\widehat{L}_{\alpha/2}(a)\} \\
&+ \widehat{G}_L\{\widehat{L}_{\alpha/2}(a)\} \\
&= o(1) + \alpha/2,
\end{aligned}
$$

where $G_L(\cdot)$ is the cumulative distribution function for $L(a)$. Combine the results we have above, we can obtain that

$$
\begin{aligned}
&\mathbb{P}(\widehat{L}_{\alpha/2}(a) \le a^\intercal\sqrt{n}(\widehat{\mu}_{\text{tap}} - \mu_g) \le \widehat{U}_{1-\alpha/2}(a)) \\
&= \mathbb{P}\{a^\intercal\sqrt{n}(\widehat{\mu}_{\text{tap}} - \mu_g) \le \widehat{U}_{1-\alpha/2}(a)\} \\
&- \mathbb{P}\{a^\intercal\sqrt{n}(\widehat{\mu}_{\text{tap}} - \mu_g) \le \widehat{L}_{\alpha/2}(a)\} \\
&\ge 1 - \alpha/2 + o(1) - \alpha/2 + o(1) = 1 - \alpha.
\end{aligned}
$$

Thus, the proof is completed.

A.8. Proof of Remark 4.1

In this section, we construct a data-adaptive confidence interval based on the projection sets proposed in [48]. Starting from the common projection sets, we re-express the test-and-pool estimator

$$
\begin{aligned}
a^\intercal n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) = &-a^\intercal V_{\text{eff}}^{1/2} W_1 + a^\intercal(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) W_2 \\
&+ a^\intercal \omega_B (V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) W_{[c_\gamma,\infty)}^t.
\end{aligned}
$$

For given μ_2, we know that

$$
\begin{aligned}
n^{1/2}\{\widehat{\mu}_{\text{tap}}(\mu_2) - \mu_g\} = &-a^\intercal V_{\text{eff}}^{1/2} W_1 + a^\intercal(\omega_A V_{\text{A-eff}}^{1/2} - \omega_B V_{\text{B-eff}}^{1/2}) W_2(\mu_2) \\
&+ a^\intercal \omega_B (V_{\text{B-eff}}^{1/2} + V_{\text{A-eff}}^{1/2}) W_{[c_\gamma,\infty)}^t(\mu_2),
\end{aligned}
$$

where the right hand side can be approximated by empirical sample distribution as $\widehat{Q}_n(\mu; a)$ and we could construct a $(1 - \tilde{\alpha}_1) \times 100\%$ confidence interval $\mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a; \mu_2)$ of μ_g given μ_2 by the empirical quantile confidence interval as

$$
\mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a; \mu_2)
$$

$$= \left\{ \mu_g \in \mathbb{R}^l : \widehat{\mu}_{\text{tap}}(\mu_2) - \frac{\widehat{Q}_n^{-1}(1-\alpha/2;a)}{\sqrt{n}} \le \mu_g \le \widehat{\mu}_{\text{tap}}(\mu_2) - \frac{\widehat{Q}_n^{-1}(\alpha/2;a)}{\sqrt{n}} \right\},$$

where $\widehat{Q}_n^{-1}(d;a)$ is the d-th sample quantiles based on our empirical distribution.

However, the value of μ_2 is unknown, a useful approach is to form a $(1-\tilde{\alpha}_2)\times 100\%$ confidence region $\mathbb{B}_{\mu_2,1-\tilde{\alpha}_2}$ for μ_2, and thus the projection confidence interval for μ_g is the union of $\mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2)$ over all $\mu_2 \in \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2}$. Here, the confidence bounds for μ_2 can be constructed as $\mathbb{B}_{\mu_2 1-\tilde{\alpha}_2} = \widehat{\mu}_2 \pm \Phi^{-1}(1-\tilde{\alpha}_2/2)$ where

$$\widehat{\mu}_2 = n^{1/2} f_B^{1/2} \Sigma_T^{-1/2} \left\{ N^{-1} \sum_{i=1}^{N} \dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\mu}_A, \widehat{\tau}) \right\}^{-1} (\widehat{\mu}_A - \widehat{\mu}_B),$$

$\Phi^{-1}(\cdot)$ is the inverse cdf for a standard normal distribution. Thus, let $\alpha = \tilde{\alpha}_1 + \tilde{\alpha}_2$ and the union would be the data-adaptive projection $(1-\alpha)\times 100\%$ confidence interval for μ_g

$$\mathbb{C}^{\text{PCI}}_{\mu_g,1-\alpha}(a) = \cup_{\mu_2 \in \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2}} \mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2). \tag{A.18}$$

To limit conservatism, a pretest procedure is carried out while we construct the projection adaptive confidence intervals $\mathbb{C}^{\text{PACI}}_{\mu_g,1-\alpha}(a)$, and we would use the $\mathbb{C}^{\text{PCI}}_{\mu_g,1-\alpha}(a)$ if we cannot reject the $H_0 : \mu^\intercal \mu \in \mathbb{B}$. To prove the coverage for the projection adaptive confidence interval, denote for $\alpha \in (0,1)$, we have that

$$\begin{aligned}
&\mathbb{P}\left(a^\intercal \mu_g \notin \mathbb{C}^{\text{PACI}}_{\mu_g,1-\alpha}(a)\right) = \mathbb{P}\left(a^\intercal \mu_g \notin \mathbb{C}^{\text{PACI}}_{\mu_g,1-\alpha}(a) \mid T \le v_n\right) \mathbb{P}(T \le v_n) \\
&+ \mathbb{P}\left\{a^\intercal \mu_g \notin \mathbb{B}_{\mu_g,1-\alpha}(a;\widehat{\mu}_2) | T > v_n\right\} \mathbb{P}(T > v_n) \\
&= \mathbb{P}(a^\intercal \mu_g \notin \mathbb{C}^{\text{PCI}}_{\mu_g,1-\alpha}(a), \mu_2 \in \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2} \mid T \le v_n)\mathbb{P}(T \le v_n) \\
&+ \mathbb{P}(a^\intercal \mu_g \notin \mathbb{C}^{\text{PCI}}_{\mu_g,1-\alpha}(a), \mu_2 \notin \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2} \mid T \le v_n)\mathbb{P}(T \le v_n) \\
&+ \{\tilde{\alpha}_1 + o(1)\}\mathbb{P}(T > v_n) \\
&\le \mathbb{P}\{a^\intercal \mu_g \notin \mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2), \mu_2 \in \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2} \mid T \le v_n\}\mathbb{P}(T \le v_n) \\
&+ \mathbb{P}(\mu_2 \notin \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2} \mid T \le v_n)\mathbb{P}(T \le v_n) + \alpha\mathbb{P}(T > v_n) \\
&\le (\tilde{\alpha}_1 + \tilde{\alpha}_2)\mathbb{P}(T \le v_n) + \alpha\mathbb{P}(T > v_n) \\
&= \alpha,
\end{aligned}$$

where we know that $\mathbb{P}\{a^\intercal \mu_g \notin \mathbb{B}_{\mu_g,1-\tilde{\alpha}_1}(a;\mu_2), \mu_2 \in \mathbb{B}_{\mu_2,1-\tilde{\alpha}_2}\} \le \tilde{\alpha}_1$ holds for any value μ_2.

A.9. Proof of Lemma A.1

Following the similar arguments in [55], let $F(\cdot)$ and $G(\cdot)$ be the cumulative distribution function (c.d.f.) of $\mathcal{N}(\mu_g, V_1)$ and $\mathcal{N}(-\mu_{g,0}, V_2)$. Let $\Phi(t)$ be the convolution of $G(\cdot)$ and $F(\cdot)$ as $\Phi(\cdot) = (G * F)(\cdot)$, then we have

$$|\mathbb{P}\{(\widehat{\mu}_g - \mu_g) \le t\} - \Phi(t)|$$

$$\leq \left|\mathbb{E}_\zeta \left\{\sup_x \mathbb{P}(\widehat{\mu}_g \leq x \mid \mathcal{F}_N) - F(x)\right\}\right| + |\mathbb{E}_\zeta \left\{F(s) - \Phi(t)\right\}|,$$

where $s = t + \mu_g = t - (-\mu_g)$. By Lemma 3.2 in [45], $|\mathbb{P}(\widehat{\mu}_g \leq x \mid \mathcal{F}_N) - F(x)|$ converges to 0 uniformly in x. For the first term, we have

$$\lim_{N\to\infty} \left|\mathbb{E}_\zeta \left\{\sup_x \mathbb{P}(\widehat{\mu}_g \leq x \mid \mathcal{F}_N) - F(x)\right\}\right| \\ \leq \mathbb{E}_\zeta \left\{\lim_{N\to\infty} \left|\sup_x \mathbb{P}(\widehat{\mu}_g \leq x \mid \mathcal{F}_N) - F(x)\right|\right\} \to 0.$$

Since $F(\cdot)$ and $G(\cdot)$ are both bounded and continuous, by the dominated convergence theorem, the second term is

$$\lim_{N\to\infty} \mathbb{E}_\zeta\{\mathcal{F}(s)\} - \Phi(t) = \mathbb{E}_\zeta \left\{\lim_{N\to\infty} \mathcal{F}(t - (-\mu_g))\right\} - \Phi(t) \\ = \int_x G(x)F(t-x)dx - \Phi(t),$$

which also converges to 0 [55, Lemma 1]. Hence, the asymptotic c.d.f of $\widehat{\mu}_g - \mu_g$ is $\Phi(\cdot)$ and the result follows as the convolution of Gaussians is still Gaussian [1, 8].

A.10. Proof of Lemma A.2

Under Assumptions 2.1, 2.2 (iii) and Assumption A.1 f), we have

$$\begin{aligned} 0 =& N^{-1}\sum_{i=1}^N \mathbb{E}_{\text{np-p}}\left\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{B,0}, \tau_0) \mid \mathcal{F}_N\right\} \\ =& N^{-1}\sum_{i=1}^N \mathbb{E}_{\text{np-p}}\left\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0) \mid \mathcal{F}_N\right\} \\ &+ N^{-1}\sum_{i=1}^N \mathbb{E}_{\text{np-p}}\left\{\dot{\Phi}_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_B^*, \tau_0) \mid \mathcal{F}_N\right\}(\mu_{B,0} - \mu_{g,0}) \\ =& \mathbb{E}\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)\} \\ &+ \mathbb{E}\left\{\dot{\Phi}_B(V_i; \mu_B^*, \tau_0)\right\}(\mu_{B,0} - \mu_{g,0}) + O_{\text{np-p-}\zeta}(n^{-1/2}), \end{aligned}$$

for some μ_B^* between $\mu_{B,0}$ and $\mu_{g,0}$, where

$$\begin{aligned} &N^{-1}\sum_{i=1}^N \mathbb{E}_{\text{np-p}}\left\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0) \mid \mathcal{F}_N\right\} \\ &= \mathbb{E}_\zeta\left[\mathbb{E}_{\text{np-p}}\left\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_g, \tau_0) \mid \mathcal{F}_N\right\}\right] + O_{\text{np-p-}\zeta}(n^{-1/2}) && \text{(A.19)} \\ &= \mathbb{E}\{\Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu_{g,0}, \tau_0)\} + O_{\text{np-p-}\zeta}(N^{-1/2}) + O_{\text{np-p-}\zeta}(n^{-1/2}), && \text{(A.20)} \end{aligned}$$

where for (A.19), the first approximation $\mathbb{E}_{\text{np-p}}(\cdot \mid \mathcal{F}_N)$ is based on the design consistency and the non-probability sample-based Weak Law of Large Numbers under Assumption 2.2 (iii), and the second approximation $\mathbb{E}_{\zeta}(\cdot)$ is justified under Assumption A.1 f); For (A.20), it can be obtained by continuous mapping theorem as $\mu_g = \mu_{g,0} + O_{\zeta}(N^{-1/2})$ under Assumption A.1 f). By rearranging the terms under the local alternative, it follows that

$$
\begin{aligned}
&\mu_{B,0} - \mu_{g,0} \\
&= \left[\mathbb{E}\left\{\dot{\Phi}_B(V;\mu_B^*,\tau_0)\right\}\right]^{-1}\mathbb{E}\{\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu_{g,0},\tau_0)\} + O_{\text{np-p-}\zeta}(n^{-1/2}) \\
&= O(1)\times n_B^{-1/2}\eta + O_{\text{np-p-}\zeta}(n^{-1/2}) = o_{\text{np-p-}\zeta}(1).
\end{aligned}
$$

A.11. Proof of Lemma A.3

First, we show that the composite estimator $\widehat{\mu}_{\text{pool}}$ is essentially the solution to

$$
\sum_{i=1}^{N}\{\Phi_A(V_i,\delta_{A,i};\mu,\tau) + \Lambda\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu,\tau)\} = 0.
$$

Next, under the Assumption A.1 a)-d), we apply the Taylor expansion at point (μ_g,τ_0) which leads to

$$
\begin{aligned}
0 &= \sum_{i=1}^{N}\{\Phi_A(V_i,\delta_{A,i};\widehat{\mu}_{\text{pool}},\widehat{\tau}) + \Lambda\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}_{\text{pool}},\widehat{\tau})\} \\
&= \sum_{i=1}^{N}\{\Phi_A(V_i,\delta_{A,i};\mu_g,\tau_0) + \Lambda\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu_g,\tau_0)\} \\
&+ \sum_{i=1}^{N}\left\{\frac{\partial\Phi_A(V_i,\delta_{A,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)}{\partial\mu} + \Lambda\frac{\partial\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)}{\partial\mu}\right\}(\widehat{\mu}_{\text{pool}} - \mu_g) \\
&+ \sum_{i=1}^{N}\left\{\frac{\partial\Phi_A(V_i,\delta_{A,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)}{\partial\tau} + \Lambda\frac{\partial\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)}{\partial\tau}\right\}(\widehat{\tau} - \tau_0),
\end{aligned}
$$

for some $(\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)$ between $(\widehat{\mu}_{\text{pool}},\widehat{\tau})$ and (μ_g,τ_0). Given the asymptotic joint distribution for $\widehat{\mu}_A$ and $\widehat{\mu}_B$ in Lemma 3.1, we obtain

$$
\begin{aligned}
&n^{1/2}(\widehat{\mu}_{\text{pool}} - \mu_g) \\
&= -n^{1/2}\left[\sum_{i=1}^{N}\left\{\dot{\Phi}_A(V_i,\delta_{A,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*) + \Lambda\dot{\Phi}_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)\right\}\right]^{-1} \\
&\times \left[\sum_{i=1}^{N}\left\{\Phi_A(V_i,\delta_{A,i};\mu_g,\tau_0) + \Lambda\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\mu_g,\tau_0)\right\}\right. \\
&+ \sum_{i=1}^{N}\left(\partial\Phi_A(V_i,\delta_{A,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)/\partial\tau\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\Lambda\partial\Phi_B(V_i,\delta_{A,i},\delta_{B,i};\widehat{\mu}^*_{\text{pool}},\widehat{\tau}^*)/\partial\tau\big)\left(\widehat{\tau}-\tau_0\right)\big] \\
&\cong \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu^*_{\text{pool}},\tau_0)\right\}^{-1} \\
&\times\left[n^{1/2}\mathbb{E}\left\{\dot{\Phi}_A(V;\mu_{g,0},\tau_0)\right\}(\widehat{\mu}_A-\mu_g)+n^{1/2}\Lambda\mathbb{E}\left\{\dot{\Phi}_B(V;\mu^*_B,\tau_0)\right\}(\widehat{\mu}_B-\mu_g)\right],
\end{aligned} \tag{A.21}
$$

for some intermittent value μ^*_{pool} between $\text{plim}\widehat{\mu}_{\text{pool}}$ and $\mu_{g,0}$, where Equation (A.21) is obtained by using Equation (A.7) and (A.2) collectively. By Assumptions 2.1, 2.2 (iii) and suitable moments condition in Assumption A.1, under the local alternative, $n^{1/2}(\widehat{\mu}_{\text{pool}}-\mu_g)$ would follow the normal distribution with mean and variance as

$$
\begin{aligned}
&\mathbb{E}\left\{n^{1/2}(\widehat{\mu}_{\text{pool}}-\mu_g)\right\} = -f_B^{-1/2}\mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu_{g,0},\tau_0)\right\}^{-1}\Lambda\eta, \\
&\text{var}\left\{n^{1/2}(\widehat{\mu}_{\text{pool}}-\mu_g)\right\} = \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu_{g,0},\tau_0)\right\}^{-1} \\
&\times\left\{\begin{pmatrix} \mathbb{E}\dot{\Phi}_A(V;\mu_{g,0},\tau_0) \\ \Lambda\mathbb{E}\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\end{pmatrix}\begin{pmatrix} V_A & \Gamma \\ \Gamma^{\intercal} & V_B\end{pmatrix}\begin{pmatrix} \mathbb{E}\dot{\Phi}_A(V;\mu_{g,0},\tau_0) \\ \Lambda\mathbb{E}\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\end{pmatrix}^{\intercal}\right\} \\
&\times\left[\mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu_{g,0},\tau_0)\right\}^{-1}\right]^{\intercal},
\end{aligned}
$$

obtained by the similar arguments in (A.5). Plugging (A.11) into Equation (A.21), the asymptotic distribution of the most efficient estimator $\widehat{\mu}_{\text{eff}}$ follows

$$
\begin{aligned}
&n^{1/2}(\widehat{\mu}_{\text{eff}}-\mu_g)\cong\mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda_{\text{eff}},\mu_{g,0},\tau_0)\right\}^{-1}\times \\
&\left\{\mathbb{E}\dot{\Phi}_A(V;\mu_{g,0},\tau_0)\cdot n^{1/2}(\widehat{\mu}_A-\mu_g)+\Lambda_{\text{eff}}\mathbb{E}\dot{\Phi}_B(V;\mu_{g,0},\tau_0)\cdot n^{1/2}(\widehat{\mu}_B-\mu_g)\right\} \\
&\cong n^{1/2}\left\{\omega_A(\Lambda_{\text{eff}})(\widehat{\mu}_A-\mu_g)+\omega_B(\Lambda_{\text{eff}})(\widehat{\mu}_B-\mu_g)\right\}.
\end{aligned}
$$

It yields a similar efficient estimator as derived in [71]

$$
n^{1/2}(\widehat{\mu}_{\text{eff}}-\mu_g)\cong n^{1/2}\left\{\omega_A(\Lambda_{\text{eff}})\widehat{\mu}_A+\omega_B(\Lambda_{\text{eff}})\widehat{\mu}_B-\mu_g\right\}, \tag{A.22}
$$

with

$$
\begin{aligned}
\omega_A(\Lambda) &= \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu_{g,0},\tau_0)\right\}^{-1}\mathbb{E}\left\{\dot{\Phi}_A(V_i,\delta_{A,i};\mu_{g,0},\tau_0)\right\}, \\
\omega_B(\Lambda) &= \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda,\mu_{g,0},\tau_0)\right\}^{-1}\Lambda\mathbb{E}\left\{\dot{\Phi}_B(V_i,\delta_{A,i},\delta_{B,i};\mu_{g,0},\tau_0)\right\},
\end{aligned}
$$

where it is easy to show that $\omega_A+\omega_B=I_{l\times l}$. So that the asymptotic variance V_{eff} of this efficient estimator will become

$$
V_{\text{eff}}=\begin{pmatrix}\omega_A^{\intercal}(\Lambda_{\text{eff}}) \\ \omega_B^{\intercal}(\Lambda_{\text{eff}})\end{pmatrix}^{\intercal}\begin{pmatrix} V_A & \Gamma \\ \Gamma^{\intercal} & V_B\end{pmatrix}\begin{pmatrix}\omega_A(\Lambda_{\text{eff}}) \\ \omega_B(\Lambda_{\text{eff}})\end{pmatrix}.
$$

The expression of V_{eff} can be complicated when the dimension of the parameters of interest is greater than 1. Here, we provide the form of V_{eff} when estimating equations are (2.4) and (2.5):

$$
\omega_A(\Lambda_{\text{eff}}) = \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda_{\text{eff}},\mu_{g,0},\tau_0)\right\}^{-1}\mathbb{E}\left\{\dot{\Phi}_A(V,\delta_A;\mu_{g,0},\tau_0)\right\},
$$

$$
\begin{aligned}
&= \{I_{l\times l} + (V_A-\Gamma)(V_B-\Gamma^{\intercal})^{-1}\}^{-1}\\
&= (V_B-\Gamma^{\intercal})(V_A+V_B-\Gamma-\Gamma^{\intercal})^{-1},\\
\omega_B(\Lambda_{\text{eff}}) &= \mathbb{E}\left\{\dot{\Phi}_{A,B,n}(\Lambda_{\text{eff}},\mu_{g,0},\tau_0)\right\}^{-1}\Lambda_{\text{eff}}\mathbb{E}\left\{\dot{\Phi}_B(V,\delta_A,\delta_B;\mu_{g,0},\tau_0)\right\},\\
&= \{I_{l\times l} + (V_A-\Gamma)(V_B-\Gamma^{\intercal})^{-1}\}^{-1}(V_A-\Gamma)(V_B-\Gamma^{\intercal})^{-1}\\
&= (V_B-\Gamma^{\intercal})(V_A+V_B-\Gamma-\Gamma^{\intercal})^{-1}(V_A-\Gamma)(V_B-\Gamma^{\intercal})^{-1},
\end{aligned}
$$

and

$$
\begin{aligned}
V_{\text{eff}} = &\left\{\begin{pmatrix}1\\-1\end{pmatrix}\begin{pmatrix}V_A & \Gamma\\ \Gamma^{\intercal} & V_B\end{pmatrix}\begin{pmatrix}1\\-1\end{pmatrix}\right\}^{-2}\\
&\times\begin{pmatrix}V_B-\Gamma^{\intercal}\\ V_A-\Gamma\end{pmatrix}^{\intercal}\begin{pmatrix}V_A & \Gamma\\ \Gamma^{\intercal} & V_B\end{pmatrix}\begin{pmatrix}V_B-\Gamma^{\intercal}\\ V_A-\Gamma\end{pmatrix}\\
&= (V_A+V_B-\Gamma^{\intercal}-\Gamma)^{-2}\{(V_B-\Gamma^{\intercal})^2V_A + (V_A-\Gamma)^2V_B\\
&+ \Gamma(V_B-\Gamma^{\intercal})(V_A-\Gamma^{\intercal}) + \Gamma^{\intercal}(V_A-\Gamma)(V_B-\Gamma)\}\\
&= (V_AV_B-\Gamma^2)(V_A+V_B-2\Gamma)^{-1}\\
&= V_A - V_\Delta,
\end{aligned}
$$

with $V_\Delta = (V_A-\Gamma)^2(V_A+V_B-2\Gamma)^{-1}$ guaranteed to be non-negative definite, i.e., non-negative quantity. By Cauchy-Schwarz inequality, we have

$$\sqrt{\mathbb{E}\{(\widehat{\mu}_A-\mu_g)^2\}\times\mathbb{E}\{(\widehat{\mu}_B-\mu_g)^2\}} \geq \mathbb{E}\{(\widehat{\mu}_A-\mu_g)(\widehat{\mu}_B-\mu_g)\},$$

which leads to $\sqrt{V_AV_B}\geq\Gamma$, and therefore

$$V_A+V_B-2\Gamma \geq 2\{|V_AV_B|^{1/2}-\Gamma\}\geq 0,$$

where the two sides are equal if and only if $V_A=V_B=\Gamma$. The asymptotic variance of the efficient estimator for other multi-dimensional estimating equations can be obtained in an analogous way but with much heavier notations.

Appendix B: Simulation

B.1. A detailed illustration of simulation

Here, we will provide detailed proof for estimating the finite-population parameter $\mu_y = \mu_g = N^{-1}\sum_{i=1}^N Y_i$ and $\mu_0 = \mathbb{E}_\zeta(Y)$. First, we know the following expectation that

$$\mathbb{E}_{\text{np}}(\delta_{B,i}\mid X_i,Y_i) = \pi_B(X_i,Y_i), \quad \mathbb{E}_{\text{np}}(Y_i\mid X_i) = m(X_i).$$

To obtain the asymptotic joint distribution $\widehat{\mu}_A$ and $\widehat{\mu}_B$, the stacked estimating equation system $\Phi(V,\delta_A,\delta_B;\theta)$ is constructed with $\theta = (\mu_A^{\intercal},\mu_B^{\intercal},\tau^{\intercal})^{\intercal}$ where

$$\Phi(V,\delta_A,\delta_B;\theta) = \left\{\Phi_A(V,\delta_A;\theta)^{\intercal},\Phi_B(V,\delta_A,\delta_B;\theta)^{\intercal},\Phi_\tau(V,\delta_A,\delta_B;\theta)^{\intercal}\right\}^{\intercal}, \quad \text{(B.1)}$$

where we use μ_A and μ_B to distinguish between estimators yielded by $\Phi_A(V, \delta_A; \mu_A)$ and $\Phi_B(V, \delta_A, \delta_B; \mu_B, \tau)$. By positing a logistic regression model $\pi_B(X_i; \alpha) = \exp(X_i^\intercal \alpha)/\{1 + \exp(X_i^\intercal \alpha)\}$ and a linear model $m(X_i; \beta) = X_i^\intercal \beta$, one common choices for $\Phi_A(V, \delta_A; \mu_A)$ and $\Phi_B(V, \delta_A, \delta_B; \mu_B, \tau)$ are

$$\Phi_A(V, \delta_A; \mu_A) = \delta_A \pi_A^{-1}(Y - \mu_A),$$
$$\Phi_B(V, \delta_A, \delta_B; \mu_B, \tau) = \frac{\delta_B}{\pi_B\left(X; \alpha\right)} \left\{Y - m\left(X; \beta\right)\right\} + \frac{\delta_A}{\pi_A} m\left(X; \beta\right) - \mu_B,$$

where $\tau = (\alpha, \beta)$ and π_A is the known sample weights under probability samples accounting for sample design. There are various ways to construct the estimating functions $\Phi_\tau(V_i; \alpha, \beta)$ for (α, β). One standard approach is to use the pseudo maximum likelihood estimator $\widehat{\alpha}$ and the ordinary least square estimator $\widehat{\beta}$ [54, 26]. In usual, the maximum likelihood estimator of α can be computed by maximizing the log-likelihood function $l(\alpha)$

$$\begin{aligned}\widehat{\alpha} &= \arg\max_{\alpha} \sum_{i=1}^{N} \left[\delta_{B,i} \log \pi_B(X_i; \alpha) + (1 - \delta_{B,i}) \log\{1 - \pi_B(X_i; \alpha)\}\right] \\ &= \arg\max_{\alpha} \sum_{i=1}^{N} \delta_{B,i} \log\left\{\frac{\pi_B(X_i; \alpha)}{1 - \pi_B(X_i; \alpha)}\right\} + \sum_{i=1}^{N} \log\{1 - \pi_B(X_i; \alpha)\}.\end{aligned}$$

Since we do not have the X_i for all units in the finite population, we then instead construct the following pseudo log-likelihood function $l^*(\alpha)$

$$\begin{aligned}l^*(\alpha) &= \sum_{i=1}^{N} \delta_{B,i} \log\left\{\frac{\pi_B(X_i; \alpha)}{1 - \pi_B(X_i; \alpha)}\right\} + \sum_{i=1}^{N} \delta_{A,i} \pi_{A,i}^{-1} \log\{1 - \pi_B(X_i; \alpha)\} \\ &= \sum_{i=1}^{N} \left[\delta_{B,i} X_i^\intercal \alpha - \delta_{A,i} \pi_{A,i}^{-1} \log\{1 + \exp(X_i^\intercal \alpha)\}\right],\end{aligned}$$

where the second equality is derived under the logistic regression model for $\pi_B(X_i; \alpha)$. By taking derivative of $l^*(\alpha)$ with respect to α, the estimating functions for (α, β) can be constructed as follows:

$$\Phi_{\tau,1}(V, \delta_A, \delta_B; \alpha, \beta) = \delta_B X - \delta_A \pi_A^{-1} \pi_B(X; \alpha) X, \tag{B.2}$$
$$\Phi_{\tau,2}(V, \delta_A, \delta_B; \alpha, \beta) = \delta_B X\{Y - m(X; \beta)\}, \tag{B.3}$$

with $\Phi_\tau(V, \delta_A, \delta_B; \alpha, \beta) = (\Phi_{\tau,1}(V, \delta_A, \delta_B; \alpha, \beta)^\intercal \quad \Phi_{\tau,2}(V, \delta_A, \delta_B; \alpha, \beta)^\intercal)^\intercal$. Under our setup, both Sample A and Sample B provide information on X and Y, thus we can also consider the estimating equation based on the combined samples for β:

$$\begin{aligned}\Phi_{\tau,1}(V, \delta_A, \delta_B; \alpha, \beta) &= \delta_B X - \delta_A \pi_A^{-1} \pi_B(X; \alpha) X, \\ \Phi^*_{\tau,2}(V, \delta_A, \delta_B; \alpha, \beta) &= (\delta_A + \delta_B) X\{Y - m(X; \beta)\}.\end{aligned} \tag{B.4}$$

In addition, [29] propose a new set of estimating functions, in which $(\widehat{\alpha}, \widehat{\beta})$ are obtained by jointly solve the following estimating functions:

$$\Phi_{\tau,1}^{\mathrm{KH}}(V, \delta_A, \delta_B; \alpha, \beta) = \left\{\delta_B \pi_B^{-1}(X;\alpha) - \delta_A \pi_A^{-1}\right\} X, \tag{B.5}$$

$$\Phi_{\tau,2}^{\mathrm{KH}}(V, \delta_A, \delta_B; \alpha, \beta) = \delta_B\{\pi_B^{-1}(X;\alpha) - 1\} X\{Y - m(X;\beta)\}. \tag{B.6}$$

Denote the solution to $\sum_{i=1}^N \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta) = 0$ as $\widehat{\theta} = (\widehat{\mu}_A, \widehat{\mu}_B, \widehat{\tau}^\intercal)^\intercal$. Under Assumption A.1 a)-e), we could apply the Taylor expansion to around $\theta_y = (\mu_y, \mu_y, \tau_0^\intercal)^\intercal$ and obtain

$$\begin{aligned} 0 &= \sum_{i=1}^N \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\theta}) \\ &= \sum_{i=1}^N \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) + \left\{\sum_{i=1}^N \frac{\partial \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\theta}^*)}{\partial \theta^\intercal}\right\}(\widehat{\theta} - \theta_y), \end{aligned} \tag{B.7}$$

for some $\widehat{\theta}^* = (\widehat{\mu}_A^*, \widehat{\mu}_B^*, \widehat{\tau}^{*\intercal})^\intercal$ lying between $\widehat{\theta}$ and θ_y. Under Assumption 2.1, the consistency of $\widehat{\mu}_A$ for μ_y can be established, i.e., $\widehat{\mu}_A = \mu_y + O_{\mathrm{p}}(n^{-1/2})$. Moreover, under Assumption A.1 f), we have $\mu_y = \mu_0 + O_\zeta(N^{-1/2})$ and hence $\mathrm{plim}\widehat{\mu}_A^* = \mu_0$, i.e., $\widehat{\mu}_A^*$ converges to μ_0 in probability. Under Assumption A.1 b), $\widehat{\mu}_B$ is consistent to $\mu_{B,0}$, and $\mu_{B,0} = \mu_0 + O_{\zeta\text{-p-np}}(n^{-1/2})$ under the local alternative. Denote $\theta_0 = (\mu_0, \mu_0, \tau_0^\intercal)^\intercal$, and the following uniform convergence can be established under Assumption A.1 (a)-(c) and (e)

$$\begin{aligned} &N^{-1}\sum_{i=1}^N \frac{\partial \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\theta}^*)}{\partial \theta^\intercal} \\ &= \mathbb{E}\left\{\frac{\partial \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_0)}{\partial \theta^\intercal}\right\} + O_{\zeta\text{-p-np}}(n^{-1/2}) + O_\zeta(N^{-1/2}), \end{aligned}$$

and by Assumption A.1 (d), we have

$$\left\{N^{-1}\sum_{i=1}^N \frac{\partial \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \widehat{\theta}^*)}{\partial \theta^\intercal}\right\}^{-1} = \left[\mathbb{E}\left\{\frac{\partial \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_0)}{\partial \theta^\intercal}\right\}\right] + o_{\zeta\text{-p-np}}(1).$$

Rearrange the terms of (B.7), we then have

$$\begin{aligned} &n^{1/2}(\widehat{\theta} - \theta_y) \\ &= \left\{-N^{-1}\sum_{i=1}^N \phi(\widehat{\theta}^*)\right\}^{-1}\left\{n^{1/2}N^{-1}\sum_{i=1}^N \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y)\right\} + o_{\zeta\text{-p-np}}(1) \\ &= -\left\{\mathbb{E}\phi(\theta_0)\right\}^{-1}\left\{n^{1/2}N^{-1}\sum_{i=1}^N \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y)\right\} + o_{\zeta\text{-p-np}}(1), \end{aligned}$$

where $\phi(\theta) = \partial\Phi(V,\delta_A,\delta_B;\theta)/\partial\theta^\intercal$. For the simplicity of notation, we denote $\pi_B(X_i;\alpha) = \pi_{B,i}, m(X_i;\beta) = m_i, \dot{m}_i = \partial m\left(X_i;\beta\right)/\partial\beta$, and its expectation is given by

$$\begin{aligned}&\mathbb{E}\left\{\phi(\theta)\right\}\\&= -\mathrm{diag}\left[\begin{array}{cccc} 1 & 1 & \pi_B(X_i;\alpha)\{1-\pi_B(X_i;\alpha)\}X_iX_i^\intercal & (\pi^*_{B,i}+\Omega d_i^{-1})X_iX_i^\intercal\end{array}\right],\end{aligned} \tag{B.8}$$

where $\Omega = 0$ if $\Phi_\tau(V,\delta_A,\delta_B;\alpha,\beta)$ is constructed by (B.2) and (B.3), and $\Omega = 1$ if $\Phi_\tau(V,\delta_A,\delta_B;\alpha,\beta)$ is constructed by (B.2) and (B.4); $\pi^*_{B,i} = \mathbb{P}(\delta_{B,i} = 1 \mid X_i)$ is the true probability. In addition, if (B.5) and (B.6) are used to estimate τ, it gives us

$$\begin{aligned}&\mathbb{E}\left\{\phi_{\mathrm{KH}}(\theta)\right\}\\&= -\begin{pmatrix}\mathbb{E}(\delta_{A,i}d_i) & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & \mathbb{E}\left\{\frac{\delta_{B,i}(1-\pi_{B,i})X_iX_i^\intercal}{\pi_{B,i}}\right\} & 0\\ 0 & 0 & \mathbb{E}\left\{\frac{\delta_{B,i}(1-\pi_{B,i})(Y_i-m_i)X_iX_i^\intercal}{\pi_{B,i}}\right\} & \mathbb{E}\left\{\frac{\delta_{B,i}(1-\pi_{B,i})X_iX_i^\intercal}{\pi_{B,i}}\right\}\end{pmatrix}\\&= -\mathrm{diag}\left\{1 \quad 1 \quad (1-\pi^*_{B,i})X_iX_i^\intercal \quad (1-\pi^*_{B,i})X_iX_i^\intercal\right\}.\end{aligned} \tag{B.9}$$

Below, we focus on the asymptotic properties of $n^{1/2}(\widehat{\theta}-\theta_y)$ under (B.8), and the asymptotics under under (B.9) can be obtained in an analogous way. First, the inverse of $\mathbb{E}\left\{\phi(\theta)\right\}$ is

$$\begin{aligned}&\left[\mathbb{E}\left\{\phi(\theta)\right\}\right]^{-1}\\&= -\mathrm{diag}\left[1 \quad 1 \quad \pi_B(X_i;\alpha)\{1-\pi_B(X_i;\alpha)\}X_iX_i^\intercal \quad (\pi^*_{B,i}+\Omega d_i^{-1})X_iX_i^\intercal\right]^{-1}.\end{aligned}$$

As shown in [12] under Assumption A.1 g), the asymptotic variance of $\widehat{\mu}_B$ will not be affected by the estimated $\widehat{\beta}$. Let $\pi_{B,i,0} = \pi_B(X_i;\alpha_0)$ and $m_{i,0} = m(X_i;\beta_0)$ be the correct working model evaluated the true parameter value (α_0,β_0). Therefore, the $\sum_{i=1}^N \Phi(V_i,\delta_{A,i},\delta_{B,i};\theta_y)$ can be found by using the decomposition

$$\begin{aligned}&\sum_{i=1}^N \Phi(V_i,\delta_{A,i},\delta_{B,i};\theta_y)\\&= \begin{pmatrix} 0\\ N\left(h_N-\mu_y\right)+\sum_{i=1}^N \delta_{B,i}\left\{\pi_{B,i,0}^{-1}\left(Y_i-m_{i,0}-h_N\right)-b^\intercal X_i\right\}\\ \sum_{i=1}^N \delta_{B,i}X_i - \sum_{i=1}^N \pi_{B,i,0}X_i\\ \sum_{i=1}^N \delta_{B,i}(Y_i-X_i^\intercal\beta_0)X_i\end{pmatrix}\\&+\begin{pmatrix}\sum_{i=1}^N \delta_{A,i}d_i(Y_i-\mu_y)\\ \sum_{i=1}^N \delta_{A,i}d_it_i\\ \sum_{i=1}^N \pi_{B,i,0}X_i - \sum_{i=1}^N \delta_{A,i}d_i\pi_{B,i,0}X_i\\ 0\end{pmatrix},\end{aligned}$$

where

$$h_N = N^{-1} \sum_{i=1}^{N} (Y_i - m_{i,0}),$$

$$b^\intercal = [(1-\pi_{B,i,0})\{Y_i - m_{i,0} - h_N\} X_i^\intercal] \{N^{-1} \sum_{i=1}^{N} \pi_{B,i,0}(1-\pi_{B,i,0}) X_i X_i^\intercal\}^{-1},$$

$$t_i = \pi_{A,i} X_i^\intercal b + m_{i,0} - N^{-1} \sum_{i=1}^{N} m_{i,0}.$$

Since the probability sample is assumed to be independent of the non-probability sample [12], we could express the variance for $\sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y)$ as two components $\mathcal{V}_1$ and $\mathcal{V}_2$ under Assumption 2.1 and 2.2 (iii)

$$\begin{aligned}
&\operatorname{var}\left\{ n^{1/2} N^{-1} \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \right\} = \mathcal{V}_1 + \mathcal{V}_2 \\
&= nN^{-2} \sum_{i=1}^{N} \pi_{B,i,0}(1-\pi_{B,i,0}) \\
&\times \mathbb{E}_\zeta \left\{ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \Delta^2 & \Delta X_i^\intercal & \Delta Y_i X_i^\intercal \\ 0 & \Delta X_i & X_i X_i^\intercal & Y_i X_i X_i^\intercal \\ 0 & \Delta Y_i X_i & Y_i X_i X_i^\intercal & (Y_i - X_i^\intercal \beta_0)^2 X_i X_i^\intercal \end{pmatrix} \right\} && \text{(B.10)} \\
&+ nN^{-2} \mathbb{E}_\zeta \left\{ \begin{pmatrix} \mathcal{D}_{11} & \mathcal{D}_{12} & \mathcal{D}_{13} & 0 \\ \mathcal{D}_{12}^\intercal & \mathcal{D}_{22} & \mathcal{D}_{23} & 0 \\ \mathcal{D}_{13}^\intercal & \mathcal{D}_{23}^\intercal & \mathcal{D}_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right\} + o(1), && \text{(B.11)}
\end{aligned}$$

where

$$\mathcal{V}_1 = \operatorname{var}_{\zeta\text{-np}} \left\{ \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \right\},$$

$$\mathcal{V}_2 = \operatorname{var}_{\zeta\text{-p}} \left\{ \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \right\}$$

and

$$\Delta = \pi_{B,i,0}^{-1} \{y_i - m_{i,0} - h_N\} - b^\intercal x_i.$$

By the law of total variance, we have

$$\begin{aligned}
\operatorname{var}_{\zeta\text{-np}} \left\{ \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \right\} &= \mathbb{E}_\zeta \left[\operatorname{var}_{\text{np}} \left\{ \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \mid \mathcal{F}_N \right\} \right] \\
&+ \operatorname{var}_\zeta \left[\mathbb{E}_{\text{np}} \left\{ \sum_{i=1}^{N} \Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y) \mid \mathcal{F}_N \right\} \right],
\end{aligned}$$

Algorithm B.1: Replication-based method for estimating variance of $\widehat{\mu}_A$ and $\widehat{\mu}_B$

Input: the probability sample $\{(V_i, \delta_{A,i}) : i \in \mathcal{A}\}$, the non-probability sample $\{(V_i, \delta_{B,i}) : i \in \mathcal{B}\}$ and the number of bootstrap K.

for $b = 1, \cdots, K$ **do**

Sample n_A units from the probability sample with replacement as $\mathcal{A}^{(b)}$.

Sample n_B units from the non-probability sample with replacement as $\mathcal{B}^{(b)}$.

Compute the bootstrap replicates $\widehat{\mu}_A^{(b)}$ and $\widehat{\mu}_B^{(b)}$ by solving

$$\sum_{i\in\mathcal{A}^{(b)}} \Phi_A(V_i, \delta_{A,i}; \mu) = 0, \qquad \sum_{i\in\mathcal{A}^{(b)}\cup\mathcal{B}^{(b)}} \Phi_B(V_i, \delta_{A,i}, \delta_{B,i}; \mu, \widehat{\tau}) = 0.$$

Calculate the variance estimator $\widehat{V}_A, \widehat{\Gamma}$ and $\widehat{V}_B$

$$\begin{aligned}\widehat{\Gamma} &= n(K-1)^{-1}\sum_{b=1}^{K}(\widehat{\mu}_A^{(b)} - \overline{\widehat{\mu}}_A)(\widehat{\mu}_B^{(b)} - \overline{\widehat{\mu}}_B)^\intercal, \\ \widehat{V}_D &= n(K-1)^{-1}\sum_{b=1}^{K}(\widehat{\mu}_D^{(b)} - \overline{\widehat{\mu}}_D)(\widehat{\mu}_D^{(b)} - \overline{\widehat{\mu}}_D)^\intercal, \quad D = A, B,\end{aligned}$$

where $\overline{\widehat{\mu}}_D = K^{-1}\sum_{b=1}^{K}\widehat{\mu}_D^{(b)}$ for $D = A, B$.

where the second term will be negligible under Assumption A.1 g) and h). Similar arguments hold for $\text{var}_{\zeta\text{-p}}\left\{\sum_{i=1}^{N}\Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y)\right\}$, therefore, (B.10) and (B.11) follow. The sub-matrices $\mathcal{D}_{kl}, k = 1, \cdots, 3, l = 1, \cdots, 3$ are all design-based variance-covariance matrices under the probability sampling design, and can be obtained using standard plug-in approach.

Alternatively, a with-replacement bootstrap variance estimation can also be used here [43]. To illustrate, we consider a single-stage probability proportional to size sampling with negligible sampling ratios. Following [57], the bootstrap procedures in Algorithm B.1 are conducted.

Under Assumptions 2.1 and A.1, $\widehat{\theta} - \theta_y \mid \mathcal{F}_N$ and θ_y are both approximately normal, which leads to the asymptotic normality of the unconditional distribution over all the finite populations by Lemma A.1:

$$\begin{aligned}&n^{1/2}(\widehat{\theta} - \theta_y) \to \\ &\mathcal{N}\left(\theta^*, \{\mathbb{E}\phi(\theta_0)\}^{-1}\,\text{var}\left\{n^{1/2}N^{-1}\sum_{i=1}^{N}\Phi(V_i, \delta_{A,i}, \delta_{B,i}; \theta_y)\right\}\{\mathbb{E}\phi(\theta_0)^\intercal\}^{-1}\right),\end{aligned}$$

where $\theta^* = (0 \quad -f_B^{-1/2}\left[\mathbb{E}\{\partial\Phi_B(\mu_0, \tau_0)/\partial\mu\}\right]^{-1}\eta \quad 0)^\intercal$. Thus, the asymptotic variance for the joint distribution $n^{1/2}(\widehat{\mu}_A - \mu_y, \widehat{\mu}_B - \mu_y)^\intercal$ is obtain by the 2×2 submatrix corresponding as

$$\begin{aligned}&\text{var}\{n^{1/2}(\widehat{\mu}_A - \mu_y, \widehat{\mu}_B - \mu_y)^\intercal\} \\ &= nN^{-2}\begin{pmatrix}\mathcal{D}_{11} & \mathcal{D}_{12} \\ \mathcal{D}_{21} & \sum_{i=1}^{N}(1-\pi_{B,i,0})\pi_{B,i,0}\Delta^2 + \mathcal{D}_{22}\end{pmatrix} + o(1)\end{aligned}$$

$$= \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} + o(1),$$

and

$$n^{1/2} \begin{pmatrix} \widehat{\mu}_A - \mu_y \\ \widehat{\mu}_B - \mu_y \end{pmatrix}$$
$$\to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ -f_B^{-1/2} \left[\mathbb{E}\{\partial\Phi_B(\mu_0, \tau_0)/\partial\mu\}\right]^{-1} \eta \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \right\}$$
$$\to \mathcal{N}\left\{ \begin{pmatrix} 0 \\ f_B^{-1/2}\eta \end{pmatrix}, \begin{pmatrix} V_A & \Gamma \\ \Gamma^\intercal & V_B \end{pmatrix} \right\},$$

where $\mathbb{E}\{\partial\Phi_B(\mu_0, \tau_0)/\partial\mu\} = -1$.

B.2. A detailed illustration of bias and mean squared error

Here, we take $\Phi_A(V, \delta_A; \mu)$ as Equation (2.4) and $\Phi_B(V, \delta_A, \delta_B; \mu, \tau)$ as Equation (2.5) for an illustration. For $T \le c_\gamma$, we have

$$n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$$
$$= -\left(\frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma}\right)^{1/2} W_1 + \frac{(\Gamma - V_A) - \lambda(\Gamma - V_B)}{(1+\lambda)(V_A + V_B - 2\Gamma)^{1/2}} W_2 | W_2^2 \le c_\gamma,$$

with probability $\xi = F_1(c_\gamma; \mu_2^2)$, which leads to

$$\text{bias}(\lambda, c_\gamma; \eta)_{T \le c_\gamma} = -\left(\frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma}\right)^{1/2} \mu_1$$
$$+ \frac{(\Gamma - V_A) - \lambda(\Gamma - V_B)}{(1+\lambda)(V_A + V_B - 2\Gamma)^{1/2}} \cdot E(W_2 | W_2^2 \le c_\gamma)$$
$$= -\left(\frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma}\right)^{1/2} \mu_1$$
$$+ \frac{(\Gamma - V_A) - \lambda(\Gamma - V_B)}{(1+\lambda)(V_A + V_B - 2\Gamma)^{1/2}} \cdot \mu_2 \frac{F_3(c_\gamma; \mu_2^T \mu_2/2)}{F_1(c_\gamma; \mu_2^T \mu_2/2)}$$
$$= \frac{-\eta f_B^{-1/2}(\Gamma - V_A)}{V_A + V_B - 2\Gamma} + \frac{\eta f_B^{-1/2}\{(\Gamma - V_A) - \lambda(\Gamma - V_B)\}}{(1+\lambda)(V_A + V_B - 2\Gamma)} \frac{F_3(c_\gamma; \mu_2^T \mu_2/2)}{F_1(c_\gamma; \mu_2^T \mu_2/2)},$$

and

$$\text{mse}(\lambda, c_\gamma; \eta)_{T \le c_\gamma} = \frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} \cdot (\mu_1^2 + 1) + \left\{ \frac{(\Gamma - V_A) - \lambda(\Gamma - V_B)}{(1+\lambda)(V_A + V_B - 2\Gamma)^{1/2}} \right\}^2$$
$$\times E(W_2^2 | W_2^2 \le c_\gamma)$$
$$- 2 \frac{\left(V_A V_B - \Gamma^2\right)^{1/2} \{(\Gamma - V_A) - \lambda(\Gamma - V_B)\}}{(1+\lambda)(V_A + V_B - 2\Gamma)} \mu_1 \cdot \mu_2 \frac{F_3(c_\gamma; \mu_2^T \mu_2/2)}{F_1(c_\gamma; \mu_2^T \mu_2/2)}$$

$$= \frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} \cdot (\mu_1^2 + 1) + \left\{ \frac{\lambda(\Gamma - V_B) - (\Gamma - V_A)}{(1+\lambda)(V_A + V_B - 2\Gamma)^{1/2}} \right\}^2$$

$$\times \left\{ \frac{F_3(c_\gamma; \mu_2^2/2)}{F_1(c_\gamma; \mu_2^2/2)} + \mu_2^2 \frac{F_5(c_\gamma; \mu_2^2/2)}{F_1(c_\gamma; \mu_2^2/2)} \right\}$$

$$- 2 \frac{\left(V_A V_B - \Gamma^2\right)^{1/2} \{(\Gamma - V_A) - \lambda(\Gamma - V_B)\}}{(1+\lambda)(V_A + V_B - 2\Gamma)} \mu_1 \cdot \mu_2 \frac{F_3(c_\gamma; \mu_2^T \mu_2/2)}{F_1(c_\gamma; \mu_2^T \mu_2/2)}.$$

For $T > c_\gamma$, we have

$$n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g) = -\left(\frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} \right)^{1/2} W_1 + \frac{-(\Gamma - V_A)}{(V_A + V_B - 2\Gamma)^{1/2}} W_2 | W_2^2 > c_\gamma,$$

with probability $1 - \xi = 1 - F_1(c_\gamma; \mu_2^2)$, the corresponding bias and mean squared error would be

$$\text{bias}(\lambda, c_\gamma; \eta)_{T > c_\gamma} = -\left(\frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} \right)^{1/2} \mu_1$$

$$+ \frac{(\Gamma - V_A)}{(V_A + V_B - 2\Gamma)^{1/2}} \cdot \mu_2 \frac{1 - F_3(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_1(c_\gamma; \mu_2^T \mu_2/2)}$$

$$= \frac{-\eta f_B^{-1/2}(\Gamma - V_A)}{V_A + V_B - 2\Gamma} + \frac{\eta f_B^{-1/2}(\Gamma - V_A)}{V_A + V_B - 2\Gamma} \frac{1 - F_3(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_1(c_\gamma; \mu_2^T \mu_2/2)},$$

and

$$\text{mse}(\lambda, c_\gamma; \eta)_{T > c_\gamma} = \frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} \cdot (\mu_1^2 + 1) + \frac{(\Gamma - V_A)^2}{V_A + V_B - 2\Gamma}$$

$$\times E(W_2^2 | W_2^2 > c_\gamma)$$

$$- 2 \frac{\left(V_A V_B - \Gamma^2\right)^{1/2} (\Gamma - V_A)}{V_A + V_B - 2\Gamma} \mu_1 \cdot \mu_2 \frac{1 - F_3(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_1(c_\gamma; \mu_2^T \mu_2/2)}$$

$$= \frac{V_A V_B - \Gamma^2}{V_A + V_B - 2\Gamma} + \frac{(\Gamma - V_A)^2}{V + V_B - 2\Gamma}$$

$$\times \left\{ \frac{1 - F_3(c_\gamma; \mu_2^2/2)}{1 - F_1(c_\gamma; \mu_2^2/2)} + \mu_2^2 \frac{1 - F_5(c_\gamma; \mu_2^2/2)}{1 - F_1(c_\gamma; \mu_2^2/2)} \right\}$$

$$- 2 \frac{\left(V_A V_B - \Gamma^2\right)^{1/2} (\Gamma - V_A)}{V_A + V_B - 2\Gamma} \mu_1 \cdot \mu_2 \frac{1 - F_3(c_\gamma; \mu_2^T \mu_2/2)}{1 - F_1(c_\gamma; \mu_2^T \mu_2/2)}.$$

Then, the bias and mean squared error for $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ would be

$$\text{bias}(\lambda, c_\gamma; \eta) = \text{bias}(\lambda, c_\gamma; \eta)_{T \le c_\gamma} \cdot \xi + \text{bias}(\lambda, c_\gamma; \eta)_{T > c_\gamma} \cdot (1 - \xi)$$

$$= \frac{-\eta f_B^{-1/2}(\Gamma - V_A)}{V_A + V_B - 2\Gamma} + \frac{\eta f_B^{-1/2}\{-\lambda(\Gamma - V_B) + (\Gamma - V_A)\}}{(1+\lambda)(V_A + V_B - 2\Gamma)} F_3(c_\gamma; \mu_2^T \mu_2/2)$$

$$+ \frac{\eta f_B^{-1/2}(\Gamma - V_A)}{V_A + V_B - 2\Gamma} \left\{ 1 - F_3(c_\gamma; \mu_2^T \mu_2/2) \right\}$$

$$= \frac{-\lambda\eta f_B^{-1/2}}{1+\lambda}\left(\frac{\Gamma - V_B}{V_A+V_B-2\Gamma} + \frac{\Gamma - V_A}{V_A+V_B-2\Gamma}\right) F_3(c_\gamma; \mu_2^T\mu_2/2)$$
$$= \eta d_0, \tag{B.12}$$

with

$$d_0 = -\lambda f_B^{-1/2}(1+\lambda)^{-1}(\omega_A+\omega_B)F_3(c_\gamma; \mu_2^T\mu_2/2),$$

and

$$\begin{aligned}
\mathrm{mse}(\lambda, c_\gamma; \eta) &= \frac{V_AV_B - \Gamma^2}{V_A+V_B-2\Gamma}\cdot(\mu_1^2+1) \\
&+ \frac{\{\lambda(\Gamma - V_B) - (\Gamma - V_A)\}^2}{(1+\lambda)^2(V_A+V_B-2\Gamma)} \times \left\{F_3(c_\gamma;\mu_2^2/2) + \mu_2^2 F_5(c_\gamma;\mu_2^2/2)\right\} \\
&+ \frac{(\Gamma - V_A)^2}{V_A+V_B-2\Gamma} \times \left\{1 - F_3(c_\gamma;\mu_2^2/2) + \mu_2^2 - \mu_2^2 F_5(c_\gamma;\mu_2^2/2)\right\} \\
&- 2\frac{\left(V_AV_B-\Gamma^2\right)^{1/2}(\Gamma - V_A)}{V_A+V_B-2\Gamma}\left\{1 - F_3(c_\gamma;\mu_2^T\mu_2/2)\right\}\mu_1\mu_2 \\
&- 2\frac{\left(V_AV_B-\Gamma^2\right)^{1/2}\{(\Gamma - V_A) - \lambda(\Gamma - V_B)\}}{(1+\lambda)(V_A+V_B-2\Gamma)} F_3(c_\gamma;\mu_2^T\mu_2/2)\mu_1\mu_2 \\
&= V_{\mathrm{eff}}d_1 + V_{\text{B-eff}}d_2 + V_{\text{A-eff}}d_3 + V_{\mathrm{eff}}^{1/2}(V_{\text{B-eff}}^{1/2}d_4 + V_{\text{A-eff}}^{1/2}d_5),
\end{aligned} \tag{B.13}$$

with

$$\begin{aligned}
d_1 &= \mu_1^2 + 1, \\
d_2 &= \lambda(1+\lambda)^{-2}\left\{F_3(c_\gamma;\mu_2^2/2) + \mu_2^2F_5(c_\gamma;\mu_2^2/2)\right\}\{\lambda - 2\omega_B/\omega_A\}, \\
d_3 &= 1 - F_3(c_\gamma;\mu_2^2/2) + \mu_2^2\left\{1 - F_5(c_\gamma;\mu_2^2/2)\right\} \\
&\quad +(1+\lambda)^{-2}\left\{F_3(c_\gamma;\mu_2^2/2) + \mu_2^2F_5(c_\gamma;\mu_2^2/2)\right\}, \\
d_4 &= 2\lambda(1+\lambda)^{-1}\mu_1\mu_2F_3(c_\gamma;\mu_2^2/2), \\
d_5 &= -2\mu_1\mu_2\left\{1 - F_3(c_\gamma;\mu_2^2/2) + F_3(c_\gamma;\mu_2^2/2)(1+\lambda)^{-1}\right\}.
\end{aligned}$$

Let $V_A = 2, V_B = 1, \Gamma = 0.5$, and $\eta = 0, 0.5$ and 1.5 (encoding zero, weak, and strong violation of H_0) in (B.12) and (B.13). Figure B.1 shows three mean squared error surfaces as functions of (Λ, c_γ) with three values of η.

a) In the leftmost plot, where H_0 holds, for a given Λ, the mean squared error decreases drastically and then flattens out as c_γ increases. Moreover, for a given c_γ, there exists a minimizer Λ^* such that the mean squared error achieves the minimum. These observations justify our strategy by viewing Λ and c_γ jointly as tuning parameters since both of them are playing important roles when searching for the minimum value of mean squared error.
b) In the middle plot, where H_0 is weakly violated, the pattern of the mean squared error retains the similar features for c_γ as shown in (A). In addition, the optimal choice Λ^* leads to a sharp decline of the mean squared

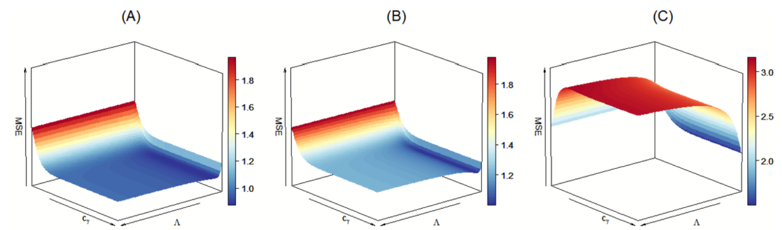

FIG B.1. *The plots for the mean squared errors in a synthetic example. Leftmost (A) plots the mean square error mse*$(\Lambda, c_\gamma; \eta)$ *of* $n^{1/2}(\widehat{\mu}_{\text{tap}} - \mu_g)$ *as function of* Λ *and* c_γ *when the null hypothesis* H_0 *holds true* $(\eta = 0)$*; Middle (B) plots mse*$(\Lambda, c_\gamma; \eta)$ *when the null hypothesis* H_0 *is weakly violated* $(\eta = 0.5)$*; Rightmost (C) plots mse*$(\Lambda, c_\gamma; \eta)$ *when the null hypothesis* H_0 *is strongly violated* $(\eta = 1.5)$.

error compared to other choices of Λ. These findings imply that despite the bias due to accepting the non-probability sample, the impact would be less compared to the increased variance due to rejecting the non-probability sample. But care is needed to determine the amount of information borrowed from the non-probability sample since a small deviation from the optimal value Λ^* can lead to a non-ignorable increase of the mean squared error. Once the optimal mean squared error is reached at (Λ^*, c_γ^*), the further increment of c_γ will not be influential.

c) In the rightmost plot, where H_0 is strongly violated, the mean squared error behaves differently as in (A) and (B). It is advisable to choose both Λ and c_γ close to zero (the low probability of combining the non-probability sample with the probability sample) to minimize the mean squared error. As above, keeping increasing c_γ after the mean squared error flattens out is of no importance.

B.3. Additional simulation results

Table B.1 provides the Monte Carlo averages and standard errors of the data-adaptive tuned parameters (Λ, c_γ) and the Monte Carlo proportion of combining the probability and non-probability samples. Figure B.2 presents the plots of Monte Carlo biases, variances and mean squared errors of the $\widehat{\mu}_A$, $\widehat{\mu}_{\text{dr}}$, $\widehat{\mu}_{\text{eff}}$, $\widehat{\mu}_{\text{tap}}$ and $\widehat{\mu}_{\text{tap:fix}}$ based on 2000 replicated datasets. For the fixed threshold strategy $\widehat{\mu}_{\text{tap:fix}}$, the threshold c_γ is held fixed to be the 95th quantile of a χ^2_1 distribution (i.e., 3.84) and the tuning parameter Λ is selected by minimizing the asymptotic mean square error at the fixed c_γ.

In Table B.1, we find that the adaptive procedure tends to select smaller values of Λ and c_γ as b increases. As a result, the Monte Carlo proportions of combining the probability and non-probability samples together are decreasing, which is desired for down-weighting the biased non-probability sample. Moreover, we compare the adaptive tuning strategy of c_γ with a fixed thresholding strategy, and Figure B.2 shows that the strategy with pre-defined cutoff cannot

TABLE B.1
Simulation results of Monte Carlo averages of the tuning parameters (Λ, c_γ) and the proportion $\mathbb{P}(comb)$ of combining the probability and non-probability samples.

H_0		Λ		c_γ		$\mathbb{P}$(comb)	
		EST	SE	EST	SE	EST	SE
holds	$\widehat{\mu}_{\text{tap}}$	3.02	4.26	35.06	9.45	0.95	0.22
	$\widehat{\mu}_{\text{tap}:B}$	3.05	4.62	35.06	9.44	0.95	0.22
	$\widehat{\mu}_{\text{tap:KH}}$	3.06	4.66	35.06	9.44	0.95	0.22
slightly violated	$\widehat{\mu}_{\text{tap}}$	2.21	3.39	31.60	13.76	0.86	0.35
	$\widehat{\mu}_{\text{tap}:B}$	2.22	3.47	31.60	13.75	0.86	0.35
	$\widehat{\mu}_{\text{tap:KH}}$	2.23	3.60	31.60	13.75	0.86	0.35
strongly violated	$\widehat{\mu}_{\text{tap}}$	0.16	0.28	1.40	1.97	0.00	0.06
	$\widehat{\mu}_{\text{tap}:B}$	0.16	0.28	1.40	1.97	0.00	0.06
	$\widehat{\mu}_{\text{tap:KH}}$	0.16	0.28	1.40	1.98	0.00	0.06

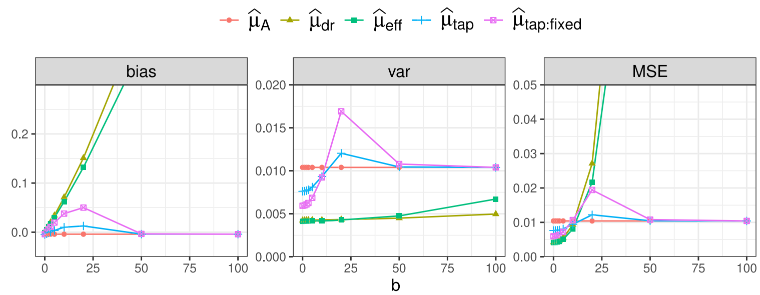

FIG B.2. *Summary statistics plots of estimators of μ_y with respect to the strength of violation, labeled by b. Each column of the plots corresponds to a different metric: "bias" for bias, "var" for variance, "MSE" for mean square error.*

satisfactorily control the mean squared error when H_0 is slightly or strongly violated.

B.4. Double-bootstrap procedure for v_n selection

Following the algorithm mentioned by [10], where optimal v_n is selected to ensure the coverage probability, we need to retain the K bootstrapped samples, called $V^{(1)}, V^{(2)}, \cdots, V^{(K)}$ where $V^{(b)} = \{V_i = (X_i^{(b)\intercal}, Y_i^{(b)})^\intercal : i \in 1, \cdots, n\}, b = 1, \cdots, K$ with $n = n_A + n_B$. The reason it is called double bootstrap is that each bootstrap sample spawns itself to a set of K' second-order bootstrap samples. Next, we set up the candidates for v_n. Under the assumption (A2), we let v_n be the form of $\kappa \log\log n$ with $\kappa \in \{2, 4, 10, 20, 30\}$, and construct the bound-based adaptive confidence intervals for each given κ at $1-\alpha$ confidence level, denoted as $\mathbb{C}^{\text{PACI},\kappa}_{\mu_g,1-\alpha}(a)$. Given each κ, we compute the coverage probability for the associated adaptive confidence intervals regarding these K' second-ordered simulated datasets. Then, choose the smallest κ that ensures the actual coverage probability larger than $1-\alpha$. Specifically, we use the estimator $\widehat{\mu}_A^{(b)}$ for μ_A in each bootstrapped dataset as the ground truth and count the number of datasets in which the adaptive confidence interval covers the ground truth, say $c(\kappa)=\sum_{b=1}^{K'}\mathbf{1}\{\widehat{\mu}_A^{(b)}\in\mathbb{C}_{\mu_g,1-\alpha}^{\mathrm{PACI},\kappa,(b)}(a)\}$ and therefore the v_n can be determined by using $v_n=\inf\{\kappa: c(\kappa)/K'>1-\alpha\}\times\log\log n$. In our simulation, K' is set to be 100.

B.5. Details of the Bayesian method

In this section, we provide the details of the Bayesian approaches proposed by [52] to combine the probability and non-probability samples as follows.

1. Solve the score function for β by using the non-probability sample:

$$\widehat{\beta}_{\mathrm{NPR}}=\arg\min_{\beta}\sum_{i=1}^{N}\delta_{B,i}X_i(Y_i-X_i^{\intercal}\beta)=0.$$

2. Construct the informative prior with three choices:

Prior 1: Choose a weakly informative parameterization of the prior as

$$\beta\sim\mathcal{N}(0,10^6),$$

which can be treated as a reference for comparison.

Prior 2: Let $\widehat{\beta}_{\mathrm{PR}}$ be the solution to the score function based on the probability sample

$$\widehat{\beta}_{\mathrm{PR}}=\arg\min_{\beta}\sum_{i=1}^{N}\delta_{A,i}X_i(Y_i-X_i^{\intercal}\beta)=0.$$

Then consider the squared Euclidean distance between $\widehat{\beta}_{\mathrm{PR}}$ and $\widehat{\beta}_{\mathrm{NPR}}$ as the hyper-parameter σ_β^2 for the variance of β:

$$\beta\sim\mathcal{N}\left\{\widehat{\beta}_{\mathrm{NPR}},\mathrm{diag}(\|\widehat{\beta}_{\mathrm{PR}}-\widehat{\beta}_{\mathrm{NPR}}\|_2^2)\right\}.$$

Prior 3: In lieu of using the squared distance to extract information on σ_β^2, a nonparametric with-replacement bootstrap procedure can be implemented ($B=1000$). After estimating the coefficient in each of them, denoted by $\widehat{\beta}_{\mathrm{NPR}}^{(i)}$, one replication-based variance estimator can be obtained, $\widehat{\sigma}^2_{\beta_{\mathrm{NPR}}}=\sum_{i=1}^{B}(\widehat{\beta}_{\mathrm{NPR}}^{(i)}-\bar{\widehat{\beta}}_{\mathrm{NPR}})^2/(B-1)$ with $\bar{\widehat{\beta}}_{\mathrm{NPR}}=1/B\sum_{i=1}^{B}\widehat{\beta}_{\mathrm{NPR}}^{(i)}$. Then, the informative prior can be constructed

$$\beta\sim\mathcal{N}(\widehat{\beta}_{\mathrm{NPR}},I_{p\times p}\cdot\widehat{\sigma}^2_{\beta_{\mathrm{NPR}}}).$$

3. Assume that the model for the observed probability sample is

$$Y_i\mid\delta_{A,i}=1\sim\mathcal{N}(X_i^{\intercal}\beta,\sigma^2).$$

By imposing an informative non-probability-based prior, the resulting posterior estimates are expected to be more efficient. Specifically, these priors are:

$$\beta \sim \mathcal{N}(\beta_0, \sigma_\beta^2), \quad \sigma^{-2} \sim \Gamma(r, m), \quad r = m = 10^{-3},$$

where

$$\begin{aligned}
&\text{Prior 1: } \beta_0 = 0, \quad \sigma_\beta^2 = 10^6, \\
&\text{Prior 2: } \beta_0 = \widehat{\beta}_{\text{NPR}}, \quad \sigma_\beta^2 = \operatorname{diag}(\|\widehat{\beta}_{\text{PR}} - \widehat{\beta}_{\text{NPR}}\|_2^2), \\
&\text{Prior 3: } \beta_0 = \widehat{\beta}_{\text{NPR}}, \quad \sigma_\beta^2 = I_{p\times p} \cdot \widehat{\sigma}^2_{\beta_{\text{NPR}}}.
\end{aligned}$$

The posterior Markov chain Monte Carlo (MCMC) samples of β and Y_i are obtained by drawing 2000 samples from the posterior distributions and discarding the first 500 samples as the burn-in procedures. The Bayesian estimator is

$$\widehat{\mu}_{\text{Bayes}} = 1/\widehat{N} \sum_{i=1}^{n_A} d_i \bar{Y}_i \text{ with } \widehat{N} = \sum_{i=1}^{n_A} d_i,$$

where $\bar{Y}_i$ is the posterior mean calculated by $\bar{Y}_i = 1/(2000-500)\sum_{k=501}^{2000} Y_{i,k}$. Borrowed from Bayes' Theorem, its variance and 95% highest posterior density intervals can be estimated via the MCMC posterior samples. Denote $\widehat{\mu}_{\text{Bayes},k} = 1/\widehat{N}\sum_{i=1}^{n_A} d_i Y_{i,k}, k = 501, \cdots, 2000$. Then, we have

$$\begin{aligned}
\operatorname{var}(\widehat{\mu}_{\text{Bayes}}) &= \frac{1}{2000-500-1}\sum_{k=501}^{2000}(\widehat{\mu}_{\text{Bayes},k} - \widehat{\mu}_{\text{Bayes}})^2, \\
\text{HPDI} &= \left\{Q(\widehat{\mu}_{\text{Bayes},k}; \alpha/2), Q(\widehat{\mu}_{\text{Bayes},k}; 1-\alpha/2)\right\},
\end{aligned}$$

where $Q(\widehat{\mu}_{\text{Bayes},k}; \alpha_0)$ represents the α_0-th sample quantile of the posterior samples $\widehat{\mu}_{\text{Bayes},k}$, $k = 501, \cdots, 2000$ after burn-in.

References

[1] Abramowitz, M., Stegun, I. A. and Romer, R. H. (1988). Handbook of mathematical functions with formulas, graphs, and mathematical tables.

[2] Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J. and Tourangeau, R. (2013). Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* **1** 90–143.

[3] Baltagi, B. H., Bresson, G. and Pirotte, A. (2003). Fixed effects, random effects or Hausman–Taylor?: A pretest estimator. *Economics Letters* **79** 361–369.

[4] Barr, D. R. and Sherrill, E. T. (1999). Mean and variance of truncated normal distributions. *The American Statistician* **53** 357–361.

[5] Beaumont, J.-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* **46** 1–28.

[6] Bethlehem, J. (2016). Solving the nonresponse problem with sample matching? *Social Science Computer Review* **34** 59–77.

[7] Binder, D. A. and Roberts, G. R. (2003). Design-based and model-based methods for estimating model parameters. *Analysis of Survey Data* **29** 33–54. MR1978842

[8] Boas, M. L. (2006). *Mathematical Methods in the Physical Sciences*. John Wiley & Sons.

[9] Boos, D. D. and Stefanski, L. A. (2013). *Essential Statistical Inference: Theory and Methods* **591**. Springer. MR3024617

[10] Chakraborty, B., Laber, E. B. and Zhao, Y. (2013). Inference for optimal dynamic treatment regimes using an adaptive m-out-of-n bootstrap scheme. *Biometrics* **69** 714–723. MR3106599

[11] Chen, S., Yang, S. and Kim, J. K. (2022). Nonparametric mass imputation for data integration. *Journal of survey statistics and methodology* **10** 1–24.

[12] Chen, Y., Li, P. and Wu, C. (2019). Doubly Robust Inference With Non-probability Survey Samples. *Journal of the American Statistical Association* **115** 2011–2021. MR4189773

[13] Cheng, X. (2008). Robust confidence intervals in nonlinear regression under weak identification. *Manuscript, Department of Economics, Yale University*.

[14] Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology* **40** 137–161.

[15] Cochran, W. G. (2007). *Sampling Techniques*, 3 ed. New York: John Wiley & Sons, Inc. MR0054199

[16] Colnet, B., Mayer, I., Chen, G., Dieng, A., Li, R., Varoquaux, G., Vert, J.-P., Josse, J. and Yang, S. (2020). Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*.

[17] Couper, M. P. (2000). Web surveys: A review of issues and approaches. *The Public Opinion Quarterly* **64** 464–494.

[18] Couper, M. P. (2013). Is the sky falling? New technology, changing media, and the future of surveys. *Survey Research Methods* **7** 145–156.

[19] Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association* **87** 376–382. MR1173804

[20] Elliot, M. R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice* **2** 2982.

[21] Elliott, M. N. and Haviland, A. (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology* **33** 211–215.

[22] Elliott, M. R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology* **33** 23–34.

[23] Elliott, M. R., Valliant, R. et al. (2017). Inference for nonprobability

samples. *Statistical Science* **32** 249–264. MR3648958
[24] Fuller, W. A. (2009). *Sampling Statistics*. Wiley, Hoboken, NJ.
[25] Gao, C., Yang, S. and Kim, J. K. (2023). Soft calibration for selection bias problems under mixed-effects models. *Biometrika* doi.org/10.1093/biomet/asad016.
[26] Haziza, D. and Rao, J. N. (2006). A nonresponse model approach to inference under imputation for missing survey data. *Survey Methodology* **32** 53–64. MR2193025
[27] Kalton, G. (1983). Models in the practice of survey sampling. *International Statistical Review/Revue Internationale de Statistique* **51** 175–188.
[28] Kalton, G. (2019). Developments in survey research over the past 60 years: A personal perspective. *International Statistical Review* **87** S10–S30. MR3957341
[29] Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica* **24** 375–394. MR3183689
[30] Kim, J. K. and Wang, Z. (2019). Sampling techniques for big data analysis. *International Statistical Review* **87** S177–S191. MR3957350
[31] Kott, P. S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology* **32** 133–142.
[32] Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E. and Murphy, S. A. (2014). Dynamic treatment regimes: Technical challenges and applications. *Electronic Journal of Statistics* **8** 1225–1272. MR3263118
[33] Laber, E. B. and Murphy, S. A. (2011). Adaptive confidence intervals for the test error in classification. *Journal of the American Statistical Association* **106** 904–913. MR2894746
[34] Little, R. J. (1982). Models for nonresponse in sample surveys. *Journal of the American statistical Association* **77** 237–250. MR0664675
[35] Mashreghi, Z., Léger, C. and Haziza, D. (2014). Bootstrap methods for imputed data from regression, ratio and hot-deck imputation. *Canadian Journal of Statistics* **42** 142–167. MR3181587
[36] McRoberts, R. E., Tomppo, E. O. and Næsset, E. (2010). Advances and emerging issues in national forest inventories. *Scandinavian Journal of Forest Research* **25** 368–381.
[37] Molina, E., Smith, T. and Sugden, R. (2001). Modelling overdispersion for complex survey data. *International Statistical Review* **69** 373–384.
[38] Mosteller, F. (1948). On pooling data. *Journal of the American Statistical Association* **43** 231–242.
[39] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal* **7** 308–313. MR3363409
[40] Palmer, J. R., Espenshade, T. J., Bartumeus, F., Chung, C. Y., Ozgencil, N. E. and Li, K. (2013). New approaches to human mobility: Using mobile phones for demographic research. *Demography* **50** 1105–1128.
[41] Pfeffermann, D., Eltinge, J. L., Brown, L. D. and Pfeffermann, D. (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *Journal of Survey Statistics and Methodology* **3** 425–483.

[42] Rao, J. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya B* **83** 242–272. MR4256318
[43] Rao, J., Wu, C. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology* **18** 209–217.
[44] Rao, J. N. (2014). Small-area estimation. *Wiley StatsRef: Statistics Reference Online*. MR1953089
[45] Rao, R. R. (1962). Relations between weak and uniform convergence of measures with applications. *The Annals of Mathematical Statistics* **33** 659–680. MR0137809
[46] Rivers, D. (2007). Sample Matching for Web Surveys: Theory and Application. In *Joint Statistical Meetings*.
[47] Robbins, M. W., Ghosh-Dastidar, B. and Ramchand, R. (2021). Blending of Probability and Non-Probability Samples: Applications to a Survey of Military Caregivers. *Journal of Survey Statistics and Methodology* **9** 1114–1145. MR4417203
[48] Robins, J. M. (2004). Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics* **179** 189–326. Springer. MR2129402
[49] Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89** 846–866. MR1294730
[50] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974
[51] Rothwell, P. M. (2005). Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *The Lancet* **365** 176–186.
[52] Sakshaug, J. W., Wiśniowski, A., Ruiz, D. A. P. and Blom, A. G. (2019). Supplementing Small Probability Samples with Nonprobability Samples: A Bayesian Approach. *Journal of Official Statistics* **35** 653–681.
[53] Särndal, C.-E., Swensson, B. and Wretman, J. (2003). *Model Assisted Survey Sampling*. New York: Springer-Verlag. MR1140409
[54] Scharfstein, D. O., Rotnitzky, A. and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association* **94** 1096–1120. MR1731478
[55] Schenker, N. and Welsh, A. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16** 1550–1566. MR0964938
[56] Shao, J. (1994). Bootstrap sample size in nonregular cases. *Proceedings of the American Mathematical Society* **122** 1251–1262. MR1227529
[57] Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer, New York. MR1351010
[58] Skinner, C. et al. (1992). Pseudo-likelihood and quasi-likelihood estimation for complex sampling schemes. *Computational Statistics & Data Analysis* **13** 395–405. MR1173330
[59] Staiger, D. and Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica* **65** 557–586. MR1445622
[60] Tallis, G. (1963). Elliptical and radial truncation in normal populations.

The Annals of Mathematical Statistics **34** 940–944. MR0152081

[61] Tam, S.-M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *International Statistical Review* **83** 436–448.

[62] Tourangeau, R., Conrad, F. G. and Couper, M. P. (2013). *The Science of Web Surveys*. Oxford University Press: New York.

[63] Toyoda, T. and Wallace, T. D. (1979). Pre-testing on part of the data. *Journal of Econometrics* **10** 119–123. MR0567944

[64] Tsiatis, A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York. MR2233926

[65] van der Vaart (2000). *Asymptotic Statistics* **3**. Cambridge university press, Cambridge: Cambridge University Press. MR1652247

[66] Vavreck, L. and Rivers, D. (2008). The 2006 cooperative congressional election study. *Journal of Elections, Public Opinion and Parties* **18** 355–366.

[67] Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association* **110** 1024–1036. MR3420681

[68] Wallace, T. D. (1977). Pretest estimation in regression: A survey. *American Journal of Agricultural Economics* **59** 431–443.

[69] Williams, D. and Brick, J. M. (2018). Trends in US face-to-face household survey nonresponse and level of effort. *Journal of Survey Statistics and Methodology* **6** 186–211.

[70] Xu, C., Chen, J. and Harold, M. (2013). Pseudo-likelihood-based Bayesian information criterion for variable selection in survey data. *Survey Methodology* **39** 303–322.

[71] Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association* **115** 1540–1554. MR4143484

[72] Yang, S., Gao, C., Zeng, D. and Wang, X. (2022). Elastic integrative analysis of randomized trial and real-world data for treatment heterogeneity estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), In press.*

[73] Yang, S. and Kim, J. K. (2020). Statistical data integration in survey sampling: A review. *Japanese Journal of Statistics and Data Science* **3** 625–650. MR4181993

[74] Yang, S., Kim, J. K. and Hwang, Y. (2021). Integration of survey data and big observational data for finite population inference using mass imputation. *Survey Methodology* **47** 29–58.

[75] Yang, S., Kim, J. K. and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82** 445–465. MR4084171