Statistica Sinica Preprint No: SS-2024-0227	
Title	Collaborative Analysis for Paired A/B Testing Experiments
Manuscript ID	SS-2024-0227
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202024.0227
Complete List of Authors	Qiong Zhang, Lulu Kang and Xinwei Deng
Corresponding Authors	Xinwei Deng
E-mails	xdeng@vt.edu
Notice: Accepted author version.	

Collaborative Analysis for Paired A/B Testing Experiments

Qiong Zhang[†], Lulu Kang[‡] and Xinwei Deng[*]

[†]*School of Mathematical and Statistical Sciences, Clemson University*

[‡]*Department of Mathematics and Statistics, University of Massachusetts Amherst*

[*]*Department of Statistics, Virginia Tech, Blacksburg, VA*

Abstract: With the extensive use of digital devices, online experimental platforms are commonly used to conduct experiments to collect data for evaluating different variations of products, algorithms, and interface designs, a.k.a., A/B tests. In practice, multiple A/B testing experiments are often carried out based on a common user population on the same platform. The same user's responses to different experiments can be correlated to some extent due to the individual effect of the user. In this paper, we propose a novel framework that collaboratively analyzes the data from paired A/B tests, namely, a pair of A/B testing experiments conducted on the same set of experimental subjects. The proposed analysis approach for paired A/B tests can lead to more accurate estimates than the traditional separate analysis of each experiment. We obtain the asymptotic distribution of the proposed estimators and demonstrate that the proposed estimators are asymptotically the best linear unbiased estimators under certain assumptions. Moreover, the proposed analysis approach is computationally efficient, easy to implement, and robust to different types of responses. Both numerical simulations and numerical studies based on a real case are used to examine the performance of the proposed method.

Key words and phrases: Design and analysis of experiments; Best unbiased linear estimator; Online controlled experiments; Mixed effect models.

1. Introduction

1.1 Background and Motivation

With the global coverage of the internet, many online platforms, offering e-commerce, digital services, social media, etc., have been established by technology companies, governments, healthcare and education organizations. They have engaged a huge population of users. Massive amounts of data are recorded daily from user activities, serving as the basis of data-driven decisions and policies. This potential has prompted these organizations to collect data through online platforms more proactively by carrying out online controlled experiments (e.g., A/B testing experiments). As pointed out by Kohavi et al. (2020), A/B testing experiments are often effective in revealing new insights to generate significant impacts potentially, and thus A/B testing has become ubiquitous in these organizations.

For an online experimental platform, multiple A/B test experiments are often conducted on the same population of users within a short time frame, especially when the treatment settings involved in different experiments are not in conflict (Nassi and Jewkes, 2021). In this paper, we consider the paired A/B tests, namely, a pair of A/B testing experiments conducted on the same set of experimental subjects during the same time frame. As illustrated in Figure 1, the same set of users are participating in both experiments. Each experiment compares two different treatment settings, i.e., A1/B1 or A2/B2. In each experiment, users are divided into two groups based on the treatment settings that they are assigned to, and their responses to both experiments are recorded. Since the outcomes of the pair of experiments from the same user share common user characteristics, the analysis of experiments can be more effective if they are conducted collaboratively rather than separately for each experiment.

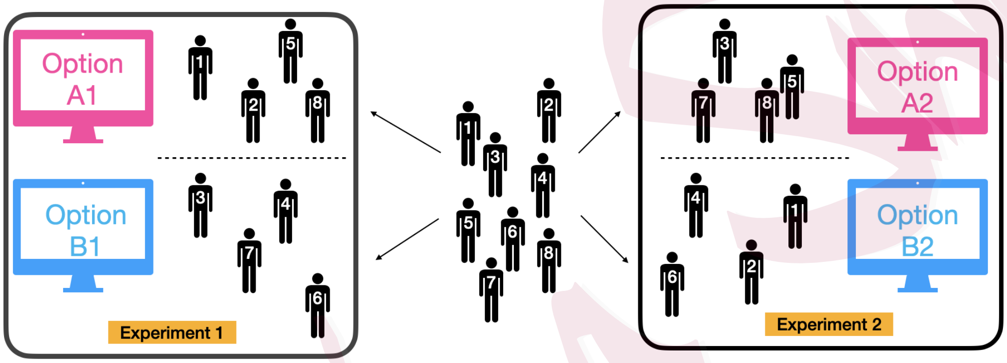

Figure 1: An illustration of paired A/B tests: A number of users (eight users here to represent any number) are participating in two experiments, each with two levels. The users' responses to both experiments are recorded.

We illustrate the motivation of paired (or multiple) A/B tests using the case study from the online math teaching platform, ASSISTments (Heffernan and Heffernan, 2014), also see `https://new.assistments.org/`. This platform supports "E-trials" for different teaching methods. Selent et al. (2016) described a set of experiments based on this platform (early version). According to Selent et al. (2016), a group of students participated in multiple ASSISTments experiments. Each experiment is associated with solving a type of math problem such as multiplying mixed numbers, equivalent expressions, and writing inequalities from situations, etc. Different types of problems can be created by different math teachers. Students in each experiment were randomly assigned to two hint conditions (treatment and control) to solve a sequence of math problems of the same type. The outcomes can be whether or not the student completed this type of problem (binary), the number of problems the students solved before completing (counts), the total time spent on this type of problem (continuous), etc. One research goal is to estimate the treatment effects for each type of math problem and find out the hint condition that leads to the best student performances for each particular problem type. One student is often involved in more than one experiment (e.q., type of math problems). In this case, the students' characteristics make an important contribution to their experimental outcomes. Therefore, combining the paired experiments that share common users can potentially enhance the analysis of the treatment effect of one individual experiment. This example also indicates that different from one 2×2 factorial experiment, paired A/B testing experiments are two experiments with different sets of control and treatment, possibly conducted by different teams on the same online experimental platform. In Section 5, we will conduct a case study using the real data from the ASSISTments experiments to elaborate the benefits of analyzing the paired A/B testing experiments.

1.2 Related Literature

Thanks to its wide application in practice, there has been a growing interest in the statistical research community to study the new challenges of A/B testing and online controlled experiments. The combined literature on this topic from statistics, machine learning, and application areas has attracted great attention. The review paper by Larsen et al. (2023) provides a comprehensive look into the new challenges and development of online controlled experiments and A/B testing in the recent literature. Due to the space limit, we only highlight a few works closely related to our work. The potential outcome framework (Neyman, 1923; Rubin, 1974; Imbens and Rubin, 2015; Ding, 2024) is the foundation for conducting causal inference of A/B testing experiments. To accommodate users' heterogeneity, covariate measurements of users can be available from prior experiments. Deng et al. (2013), Poyarkov et al. (2016) and Jin and Ba (2023) provided linear and nonlinear ways to adjust the outcome measurements based on covariates to reduce the variance of the estimate. When covariate measurements are hidden, Syrgkanis et al. (2019) and Zhao and Ding (2022) proposed different ways to estimate the treatment effect. It is also noted that recent works considered to make inferences on the treatment effect with missing outcomes (Chen et al., 2015; Shen et al., 2023; Zuo et al., 2024; Zhao et al., 2024). For multiple A/B experiments, Gupta et al. (2019) described the common scenario that users are simultaneously involved in hundreds of A/B testing experiments, which is the challenge this work aims to overcome. Despite the large literature on A/B testing experiments, few works consider how to jointly analyze multiple A/B testing experiments.

1.3 Our Contribution and Paper Outline

In this paper, we aim to fill the gap of jointly analyzing paired (or multiple) A/B testing experiments to discover new knowledge that was buried when each experiment was analyzed separately. To start, we focus on the paired A/B testing experiments. Particularly, we consider each user has unique characteristics that are not necessarily available as covariate measurements to the experimenter. The proposed collaborative estimators of the treatment effects are easy to compute and can be quickly implemented in any large-scale online platform. Our method can be extended to the cases with missing outcomes for a common practice: partially paired experiments. Theoretically and numerically, we demonstrate that the proposed estimators that combine the information of the paired experiments are better than the traditionally used separate experiment analysis approach. We also perform a robustness analysis to show the numerical performances of the proposed method when the model assumption is misspecified. The proposed collaborative analysis framework improves the design and analysis of future A/B testing practices and makes better use of all the simultaneous experiments.

The paper is organized as follows. Section 2 illustrates the proposed method under the ideal situation: complete paired experiments with balanced and orthogonal designs. Section 3 extends the method to partially paired experiments with nearly balanced and orthogonal designs. Section 4 provides numerical studies and Section 5 provides a case study with a real problem background. Section 6 concludes the paper with a discussion and future directions.

2. Collaborative Analysis of Fully Paired A/B Tests

Assume that there are two experiments each with a two-level experimental factor, i.e., paired A/B tests, and they are conducted on the same group of experimental units (i.e., experimental subject). The goal is to estimate the treatment effect for each of the two experimental factors. Let $y_{i,k}$ be the experimental outcome of the i-th unit of the k-th experiment with $k = 1, 2$ and $i = 1, \ldots, n$. The underlying model of the outcome is assumed to be

$$y_{i,k} = u_i + \alpha_k + x_{i,k}\beta_k + \epsilon_{i,k}, \quad k = 1, 2, \tag{2.1}$$

where α_k is the intercept, $x_{i,k} \in \{-1, 1\}$ is the design setting for i-th unit and β_k is the treatment effect of the k-th factor, and $\epsilon_{i,k}$ is the random noise that is independent of test units, with mean zero and variance σ_k^2. The individual effect u_i can be a random effect with mean zero and variance τ^2, independent of the random noise. It represents the unique characteristics of each unit and carries the dependence of the outcomes from a paired of experiments. We do not assume any probabilistic distributions for u_i's and $\epsilon_{i,k}$'s. Note that, combining the models of the two experiments in (2.1) forms the linear mixed effect model (see for example, Andrzej and Tomasz (2012)) with individual effect u_i. In this paper, we formulate the problem of paired A/B testings under the framework of linear mixed effect models.

2.1 Analysis of Paired A/B Tests

Under the model assumption in (2.1), we illustrate the ideas of a collaborative analysis of paired A/B tests by describing three analysis approaches to estimate the treatment effects. For clarity of illustration, we consider the ideal case that the *balanced* and *orthogonal* designs are used for the two experiments, as stated in Assumption 1.

Assumption 1. Assume that the balanced and orthogonal designs are used for the two experiments, i.e.,

$$\sum_{i=1}^{n} x_{i,1} = \sum_{i=1}^{n} x_{i,2} = \sum_{i=1}^{n} x_{i,1}x_{i,2} = 0. \tag{2.2}$$

and the level combinations of the two design factors $x_{i,1}$ and $x_{i,2}$ form a random partition of the n users into four groups.

Remark 1. We are aware that the orthogonal design assumption rarely holds in practice, but it facilitates our initial introduction of the proposed estimator. Later we will discuss the situation in Section 3 where this assumption is relaxed.

Single Analysis: In this paper, single analysis refers to the case that each of the two experiments is analyzed separately. Under Assumption 1, the least squared estimator, denoted by $\hat{\beta}_k^s$, for each single experiment is given by

$$\hat{\beta}_k^s = n^{-1}\sum_{i=1}^{n} x_{i,k}y_{i,k} \quad \text{and} \quad \mathrm{Var}(\hat{\beta}_k^s) = \frac{\tau^2+\sigma_k^2}{n}, \quad \text{for} \quad k=1,2. \tag{2.3}$$

Also, notice that $\mathrm{Cov}\left(\hat{\beta}_1^s, \hat{\beta}_2^s\right) = \tau^2 n^{-2}\sum_{i=1}^n x_{i,1}x_{i,2} = 0$.

Paired Analysis: For the paired experiments with common user random effects in (2.1), one can bypass the random effects u_i by taking the differences of the two outcomes from the same user:

$$z_i \triangleq y_{i,1} - y_{i,2} = \alpha + x_{i,1}\beta_1 - x_{i,2}\beta_2 + \delta_i, \tag{2.4}$$

where $\alpha = \alpha_1 - \alpha_2$ and $\delta_i = \varepsilon_1 - \varepsilon_2$. Under this model (2.4), the estimators of the treatment effects β_1 and β_2 can be obtained by the least squared regression method. Let $\boldsymbol{x}_1 = (x_{1,1},\ldots,x_{n,1})^\top$ and $\boldsymbol{x}_2 = (x_{1,2},\ldots,x_{n,2})^\top$ be the design vectors of the paired experiments. The design matrix of this model is $X = (\mathbf{1}_n, \boldsymbol{x}_1, -\boldsymbol{x}_2)$. Then we have that the least squared estimator of $\boldsymbol{\theta} = (\alpha, \beta_1, \beta_2)^\top$ is

$$\hat{\boldsymbol{\theta}} = (\hat{\alpha}, \hat{\beta}_1^p, \hat{\beta}_2^p)^\top = (X^\top X)^{-1} X^\top \boldsymbol{z}, \tag{2.5}$$

where $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$. The analysis based on the model of the differences of the two outcomes is called *paired analysis*.

Under Assumption 1, the least square solution leads to paired analysis estimators, denoted by $\hat{\beta}_k^p$,

$$\hat{\beta}_1^p = n^{-1} \sum_{i=1}^n x_{i,1} z_i \quad \text{and} \quad \hat{\beta}_2^p = -n^{-1} \sum_{i=1}^n x_{i,2} z_i \tag{2.6}$$

and

$$\mathrm{Var}\left(\hat{\beta}_1^p\right) = \mathrm{Var}\left(\hat{\beta}_2^p\right) = \frac{\sigma_1^2 + \sigma_2^2}{n} \quad \text{and} \quad \mathrm{Cov}\left(\hat{\beta}_1^p, \hat{\beta}_2^p\right) = n^{-2}(\sigma_1^2 + \sigma_2^2) \sum_{i=1}^n x_{i,1} x_{i,2} = 0.$$

The relative efficiency between the single analysis estimator and paired analysis estimator is

$$\mathrm{RE}\left(\hat{\beta}_k^s, \hat{\beta}_k^p\right) = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_k^2 + \tau^2} \quad \text{for} \quad k = 1, 2. \tag{2.7}$$

We see that the estimators based on the paired experiments are more efficient than the estimator of the single experiment analysis if $\tau^2 > \sigma_k^2$. Since both estimators are unbiased estimators of β_k under the model assumption in (2.5), we can combine them to obtain another unbiased estimator of the treatment effect which is shown to be universally more efficient than both the single and paired analysis, as introduced next.

Proposed Collaborative Analysis: Given the two unbiased estimators $\hat{\beta}_k^s$ and $\hat{\beta}_k^p$ of β_k, we can obtain another unbiased estimator of β_k by taking a linear combination of $\hat{\beta}_k^s$ and $\hat{\beta}_k^p$. The linear weights are given by Lemma 1. In this paper, we call this analysis approach *collaborative analysis* and the resulting estimator *collaborative estimator.*

Lemma 1. *Suppose that $\boldsymbol{T}$ is a $d \times 1$ random vector with mean $\mu \mathbf{1}_d$ and variance-covariance matrix $\boldsymbol{\Sigma}$. Then the best unbiased linear estimator of μ given by $\boldsymbol{T}$ is*

$$\hat{\mu} = \frac{\mathbf{1}_d^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{T}}{\mathbf{1}_d^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_d}, \tag{2.8}$$

where the linear weights are given by $\boldsymbol{\Sigma}^{-1}\mathbf{1}_d / \mathbf{1}_d^\top \boldsymbol{\Sigma}^{-1} \mathbf{1}_d$, which is derived from solving the optimization problem:

$$\min_{\boldsymbol{a} \in \mathbb{R}^d} \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a} \quad \text{s.t.} \quad \boldsymbol{a}^\top \mathbf{1}_d = 1.$$

Consider $\hat{\beta}_k^s$ and $\hat{\beta}_k^p$ as a $d = 2$ random vector whose mean vector and covariance matrix of $(\hat{\beta}_k^s, \hat{\beta}_k^p)^\top$ are

$$\beta_k \mathbf{1}_2 \quad \text{and} \quad \frac{1}{n}\begin{bmatrix} \tau^2 + \sigma_k^2 & \sigma_k^2 \\ \sigma_k^2 & \sigma_1^2 + \sigma_2^2 \end{bmatrix} \quad \text{for} \quad k = 1, 2,$$

respectively. Lemma 1 gives the best linear weights to construct an unbiased estimator of β_k based on the single and paired analysis estimators. The resulting collaborative estimators, denoted by $\hat{\beta}_k^c$, are given by

$$\hat{\beta}_1^c = \frac{\tau^2\hat{\beta}_1^p + \sigma_2^2\hat{\beta}_1^s}{\tau^2 + \sigma_2^2} \quad \text{and} \quad \hat{\beta}_2^c = \frac{\tau^2\hat{\beta}_2^p + \sigma_1^2\hat{\beta}_2^s}{\tau^2 + \sigma_1^2} \tag{2.9}$$

with $\text{cov}(\hat{\beta}_1^c, \hat{\beta}_2^c) = 0$, and variances

$$\text{Var}\left(\hat{\beta}_1^c\right) = \frac{\sigma_1^2\tau^2 + \sigma_2^2\tau^2 + \sigma_1^2\sigma_2^2}{n(\sigma_2^2 + \tau^2)} \quad \text{and} \quad \text{Var}\left(\hat{\beta}_2^c\right) = \frac{\sigma_1^2\tau^2 + \sigma_2^2\tau^2 + \sigma_1^2\sigma_2^2}{n(\sigma_1^2 + \tau^2)}.$$

Thus, the relative efficiencies over the single and paired analysis estimators are

$$\text{RE}\left(\hat{\beta}_k^s, \hat{\beta}_k^c\right) = \frac{\sigma_1^2\tau^2 + \sigma_2^2\tau^2 + \sigma_1^2\sigma_2^2}{(\sigma_1^2 + \tau^2)(\sigma_2^2 + \tau^2)} \leq 1, \text{RE}\left(\hat{\beta}_k^p, \hat{\beta}_k^c\right) = \frac{(\sigma_k^2 + \tau^2)}{(\sigma_1^2 + \sigma_2^2)} \cdot \frac{\sigma_1^2\tau^2 + \sigma_2^2\tau^2 + \sigma_1^2\sigma_2^2}{(\sigma_1^2 + \tau^2)(\sigma_2^2 + \tau^2)} \leq 1,$$

which indicates that the collaborative estimator is more efficient than the single and paired analysis estimators disregarding the values of the parameters τ^2, σ_1^2, and σ_2^2.

Note that the proposed collaborative estimators contain unknown parameters τ^2, σ_1^2, and σ_2^2. We propose the following plug-in estimators for them. Let S_{k+}^2 and S_{k-}^2 be the sample variances of the k-th experiment under experimental setting 1 and -1, respectively. Let S_{++}^2, S_{+-}^2, S_{-+}^2 and S_{--}^2 be the sample variances of z_i's in (2.4) under experimental setting $x_{i,1}$ and $x_{i,2}$ with the sub-index representing the signs of $x_{i,1}$ and $x_{i,2}$. We obtain the moment estimators

$$\widehat{\sigma_k^2+\tau^2} = \frac{S_{k+}^2+S_{k-}^2}{2} \quad \text{for} \quad k=1,2, \tag{2.10}$$

and

$$\widehat{\sigma_1^2+\sigma_2^2} = \frac{S_{++}^2+S_{+-}^2+S_{-+}^2+S_{--}^2}{4}. \tag{2.11}$$

Therefore, the moment estimators of τ^2, σ_1^2 and σ_2^2 can be calculated by solving the linear systems (2.10)-(2.11):

$$\begin{bmatrix} \hat{\tau}^2 \\ \hat{\sigma}_1^2 \\ \hat{\sigma}_2^2 \end{bmatrix} = \frac{1}{2} \begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 1 \\ -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \widehat{\sigma_1^2+\tau^2} \\ \widehat{\sigma_2^2+\tau^2} \\ \widehat{\sigma_1^2+\sigma_2^2} \end{bmatrix} \tag{2.12}$$

In sum, to construct the collaborative estimators under Assumption 1, we will need to first compute $\hat{\beta}_k^s$ and $\hat{\beta}_k^p$, S_{k+}^2 and S_{k-}^2 for $k=1,2$, and S_{++}^2, S_{+-}^2, S_{-+}^2 and S_{--}^2, and then assemble them according to (2.9) and (2.12). Under the model assumption, we can also obtain maximum likelihood estimators of the variance parameters. The computation of the moment estimators is trivial and straightforward to implement.

2.2 Properties of Collaborative Estimators for Paired A/B Tests

Although the proposed collaborative estimators are the best linear combination of the single and paired analysis estimators, we have not fully demonstrated the benefit of the proposed approach compared to other potential estimators. The following proposition further states that the collaborative estimators are the best unbiased linear estimators of β_1 and β_2 with respect to the original responses vector $(y_{i,1}, y_{i,2})$'s in (2.1).

Proposition 1. *Assume that τ^2, σ_1^2 and σ_2^2 are known. Under the model assumption in* (2.1) *and the balance and orthogonal assumption of the two designs in* (2.2)*, the collaborative estimators $\hat{\beta}_1^c$ and $\hat{\beta}_2^c$ in* (2.9) *have the following properties:*

(i) They are the best unbiased linear estimators (BLUE) of β_1 and β_2.

(ii) $\hat{\beta}_1^c$ and $\hat{\beta}_2^c$ are uncorrelated, and

$$\left(\frac{\hat{\beta}_1^c - \beta_1}{\sqrt{\mathrm{Var}(\hat{\beta}_1^c)}}, \frac{\hat{\beta}_2^c - \beta_2}{\sqrt{\mathrm{Var}(\hat{\beta}_2^c)}}\right)^\top \to \mathcal{N}_2(\mathbf{0}_2, \boldsymbol{I}_2)$$

in distribution, as $n \to \infty$.

The proof of this proposition is deferred to Supplement S1. The proof also demonstrates that under Assumption 1, the proposed collaborative estimators echo the weighted least squared estimators, which are known as the BLUE under model (2.1). The proposed collaborative estimators are convenient to compute and attain clear asymptotic properties. In Sections 3, we further relax the assumptions in Section 2 and extend this approach to partially paired A/B tests.

3. Collaborative Analysis of Partially Paired A/B Tests

Often, the experimental outcomes of some experimental units may not be available. Since the missing outcomes are not planned in the experimental design stage, the orthogonality assumption 1 can not hold. In this section, we extend the collaborative analysis approach to partially paired A/B tests.

Without loss of generality, we assume that, for $i = 1, \ldots, n_0$, the outcomes of both experiments from the i-th unit are available; for unit $i = n_0 + 1, \ldots, n_0 + n_1$, the outcome of the first experiment is available, and for unit $i = n_0 + n_1 + 1, \ldots, n_0 + n_1 + n_2$, the outcome of the second experiment is available; for unit $i = n_0 + n_1 + n_2 + 1, \ldots, n$, the outcomes of both experiments are missing. The structure of collected data is illustrated in Table 1.

Given the structure of collected data in Table 1, we insert the following assumptions on the collected data.

Table 1: The structure of data collected from paired experiments

	Experiment 1		Experiment 2	
Unit	Outcome	Design	Outcome	Design
1	$y_{1,1}$	$x_{1,1}$	$y_{1,2}$	$x_{1,2}$
2	$y_{2,1}$	$x_{2,1}$	$y_{2,2}$	$x_{2,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_0	$y_{n_0,1}$	$x_{n_0,1}$	$y_{n_0,2}$	$x_{n_0,2}$
n_0+1	$y_{n_0+1,1}$	$x_{n_0+1,1}$	NA	$x_{n_0+1,2}$
n_0+2	$y_{n_0+2,1}$	$x_{n_0+2,1}$	NA	$x_{n_0+2,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_0+n_1	$y_{n_0+n_1,1}$	$x_{n_0+n_1,1}$	NA	$x_{n_0+n_1,2}$
n_0+n_1+1	NA	$x_{n_0+n_1+1,1}$	$y_{n_0+n_1+1,2}$	$x_{n_0+n_1+1,2}$
n_0+n_1+2	NA	$x_{n_0+n_1+2,1}$	$y_{n_0+n_1+2,2}$	$x_{n_0+n_1+2,2}$
$\vdots$	NA	$\vdots$	$\vdots$	$\vdots$
$n_0+n_1+n_2$	NA	$x_{n_0+n_1+n_2,1}$	$y_{n_0+n_1+n_2,2}$	$x_{n_0+n_1+n_2,2}$
$n_0+n_1+n_2+1$	NA	$x_{n_0+n_1+n_2+1,1}$	NA	$x_{n_0+n_1+n_2+1,2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	NA	$x_{n,2}$	NA	$x_{n,2}$

Assumption 2. As $n \to \infty$,

$$\frac{n_0}{n} \to r_0 \in (0,1], \quad \frac{n_1}{n} \to r_1 \in [0,1), \quad \frac{n_2}{n} \to r_2 \in [0,1),$$

with r_0, r_1 and r_2 being constants satisfying that $r_0 + r_1 \in (0,1]$ and $r_0 + r_2 \in (0,1]$.

Assumption 2 ensures that there is a sufficient number of observations for each experiment. Therefore, the corresponding treatment effects are estimable. Assumption 2 also guarantees that the number of units in the fully paired portion of the experiment is sufficiently large to ensure the necessity of collaborative estimation. A few special cases are: 1) if $r_0 = 1$, the two experiments are exactly overlapped, which is the same as the situation described in Section 2; 2) if $r_1 = 0$ or $r_2 = 0$, one experiment is "nested" in the other experiment, but the following proposed collaborative analysis still works.

Assumption 3. The designs of the paired experiments are nearly balanced and orthogonal, namely,

$$\sum_{i=1}^{n_0} x_{i,1} = \sum_{i=1}^{n_0} x_{i,2} = \sum_{i=n_0+1}^{n_0+n_1} x_{i,1} = \sum_{i=n_0+n_1+1}^{n_0+n_1+n_2} x_{i,2} = \sum_{i=1}^{n_0} x_{i,1}x_{i,2} = o(n).$$

Assumption 3 can be easily satisfied if the original design of experiments can be well controlled under Assumption 1 or each assignment of $x_{i,k}$ is completely randomized and independent of each other for $k = 1, 2$ plus that the responses are missing at random and independent with the design allocation.

We now construct the collaborative estimators based on partially paired data in Table 1. Based on the fully paired outcomes for units 1 to n_0 from the top panel in Table 1, we can construct the paired estimator of β_1 as in (2.6):

$$n_0^{-1} \sum_{i=1}^{n_0} x_{i,1} z_i \quad \text{with} \quad z_i = y_{i,1} - y_{i,2}.$$

Also, we can obtain two single analysis estimators from the top panel and the second panel respectively. In sum, we have that the following three estimators for the estimand β_1:

$$n_0^{-1}\sum_{i=1}^{n_0} x_{i,1}z_i, \quad n_0^{-1}\sum_{i=1}^{n_0} x_{i,1}y_{i,1}, \quad \text{and} \quad n_1^{-1}\sum_{i=n_0+1}^{n_0+n_1} x_{i,1}y_{i,1}. \tag{3.13}$$

Under the model assumption in (2.1) and Assumptions 2 and 3, all three estimators are asymptotic unbiased estimators of β_1. Also, the covariance matrix of three estimators in (3.13) is proportional to

$$\begin{bmatrix} n_0^{-1}(\sigma_1^2+\sigma_2^2) & n_0^{-1}\sigma_1^2 & 0 \\ n_0^{-1}\sigma_1^2 & n_0^{-1}(\sigma_1^2+\tau^2) & 0 \\ 0 & 0 & n_1^{-1}(\sigma_1^2+\tau^2) \end{bmatrix}.$$

The covariance matrix also indicates that the weights of the two single analysis estimators are different. So we should separate the two single analysis estimators as two estimators in collaborative analysis. Based on Lemma 1, we can combine three estimators to obtain the *proposed collaborative estimator* of β_1 for the partially paired case:

$$\hat{\beta}_1^c = \frac{\sum_{i=1}^{n_0} x_{i,1}z_i \Big/ \left(\sigma_1^2+\sigma_2^2+\frac{\sigma_1^2\sigma_2^2}{\tau^2}\right) + \sum_{i=1}^{n_0} x_{i,1}y_{i,1} \Big/ \left(\tau^2+\sigma_1^2+\frac{\sigma_1^2\tau^2}{\sigma_2^2}\right) + \sum_{i=n_0+1}^{n_0+n_1} x_{i,1}y_{i,1} \Big/ (\tau^2+\sigma_1^2)}{\left(\sigma_1^2+\sigma_2^2+\frac{\sigma_1^2\sigma_2^2}{\tau^2}\right)^{-1} n_0 + \left(\tau^2+\sigma_1^2+\frac{\sigma_1^2\tau^2}{\sigma_2^2}\right)^{-1} n_0 + (\tau^2+\sigma_1^2)^{-1}n_1}, \tag{3.14}$$

where the weights are derived according to Lemma 1 to ensure that $\hat{\beta}_1^c$ is the best linear combination of the three estimators in (3.13) under the model assumption in (2.1). We also have that

$$\mathrm{Var}\left(\hat{\beta}_1^c\right) = \left[\left(\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\tau^2}\right)^{-1} n_0 + \left(\tau^2 + \sigma_1^2 + \frac{\sigma_1^2\tau^2}{\sigma_2^2}\right)^{-1} n_0 + (\tau^2 + \sigma_1^2)^{-1} n_1\right]^{-1}. \tag{3.15}$$

For comparison purposes, we provide the results for the single analysis estimator and the paired analysis estimators under partially paired experiments as follows:

$$\hat{\beta}_1^s = \frac{\sum_{i=1}^{n_0} x_{i,1}y_{i,1} + \sum_{i=n_0+1}^{n_0+n_1} x_{i,1}y_{i,1}}{n_0 + n_1} \quad \text{and} \quad \mathrm{Var}\left(\hat{\beta}_1^s\right) = \frac{\tau^2 + \sigma_1^2}{n_0 + n_1},$$

$$\hat{\beta}_1^p = n_0^{-1} \sum_{i=1}^{n_0} x_{i,1}z_i \quad \text{and} \quad \mathrm{Var}\left(\hat{\beta}_1^p\right) = \frac{\sigma_1^2 + \sigma_2^2}{n_0}.$$

We see that if $\tau^2 \to \infty$, the estimator $\hat{\beta}_1^c$ becomes the paired analysis estimator $\hat{\beta}_1^p$, and if $\tau^2 \to 0$, this estimator becomes the single analysis estimator $\hat{\beta}_1^s$. Similar to $\hat{\beta}_1^c$, the collaborative estimator of β_2 is

$$\hat{\beta}_2^c = \frac{\sum\limits_{i=1}^{n_0} x_{i,2}z_i \Big/ \left(\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\tau^2}\right) + \sum\limits_{i=1}^{n_0} x_{i,2}y_{i,2} \Big/ \left(\tau^2 + \sigma_2^2 + \frac{\sigma_2^2\tau^2}{\sigma_1^2}\right) + \sum\limits_{i=n_0+n_1+1}^{n_0+n_1+n_2} x_{i,2}y_{i,2} \Big/ (\tau^2 + \sigma_2^2)}{\left(\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\tau^2}\right)^{-1} n_0 + \left(\tau^2 + \sigma_2^2 + \frac{\sigma_2^2\tau^2}{\sigma_1^2}\right)^{-1} n_0 + (\tau^2 + \sigma_2^2)^{-1} n_2}, \tag{3.16}$$

with

$$\mathrm{Var}(\hat{\beta}_2^c) = \left[\left(\sigma_1^2 + \sigma_2^2 + \frac{\sigma_1^2\sigma_2^2}{\tau^2}\right)^{-1} n_0 + \left(\tau^2 + \sigma_2^2 + \frac{\sigma_2^2\tau^2}{\sigma_1^2}\right)^{-1} n_0 + (\tau^2 + \sigma_2^2)^{-1} n_2\right]^{-1}.$$

Different from the fully paired A/B tests under Assumption 1 in Section 2, the collaborative estimators for partially paired A/B tests are no longer the BLUEs under model (2.1). However, the collaborative estimators are still the asymptotically best linear unbiased estimators under model (2.1) as shown in the following proposition.

Proposition 2. *Suppose that τ^2, σ_1^2 and σ_2^2 are known. Under the model assumption in* (2.1)*, data structure in Table 1 and Assumptions 2-3, the collaborative estimators in* (3.14) *and* (3.16) *have the following properties:*

(i) They are asymptotically the best linear unbiased estimators of the treatment effects β_1 and β_2 under the true model assumption (2.1)*;*

(ii) The asymptotic distribution of $\hat{\beta}_1^c$ and $\hat{\beta}_2^c$ in Proposition 1 also holds under the partially paired case.

The proof of this proposition is deferred to Supplement S2.

The collaborative estimators for partially paired A/B tests also contain unknown parameters τ^2, σ_1^2 and σ_2^2, which can be estimated as in (2.10), (2.11) and (2.12). However, different from the fully paired case in Section 2, the sample variances S_{k+}^2

and S^2_{k-} for $k = 1, 2$, and S^2_{++}, S^2_{+-}, S^2_{-+} and S^2_{--} are computed only based on available outcomes. We summarize the steps of the collaborative analysis procedure for partially paired experiments as follows.

Step 1. For $k = 1, 2$, obtain sample variances S^2_{k+} and S^2_{k-} using available outcomes $y_{i,k}$ from experiment k with design value equal to 1 and -1, respectively.

Step 2. Split the differences $z_i = y_{i,1} - y_{i,2}$'s from the paired parts, i.e., $i = 1, \ldots, n_0$ in Table 1 according to the designs $x_{i,1}$ and $x_{i,2}$ into four groups and obtain sample variances S^2_{++}, S^2_{+-}, S^2_{-+} and S^2_{--} of each group, respectively.

Step 3. Compute the moment estimators: $\widehat{\sigma_1^2 + \tau^2}$, $\widehat{\sigma_2^2 + \tau^2}$, and $\widehat{\sigma_1^2 + \sigma_2^2}$ according to (2.10) and (2.11).

Step 4. Compute the variance estimators $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ and $\hat{\tau}^2$ according to (2.12).

Step 5. Obtain the collaborative estimators $\hat{\beta}^c_k$ for $k = 1, 2$ according to (3.14) and (3.16) with the variance estimators from Step 4.

The computational complexity of the whole procedure is $\mathcal{O}(n)$, which is the same as the estimators of the single and paired analysis.

4. Numerical Study

We provide numerical studies to compare the proposed method with other alternatives. All the methods involved in the comparison are described below:

1. **SINGLE:** The single experiment estimator in (2.3) with available outcomes from a single experiment.

2. **PAIRED:** The least squared estimator based on the difference model in (2.6) with fully paired outcomes.

3. **COE:** The proposed collaborative estimator in (3.14) with all available outcomes.

4. **LME:** The estimators are given by fitting a linear mixed-effects model using the `R` function "lmer" in the package `lme4` (Bates et al., 2015). The model is assumed to be (2.1) with the normal random effects u_i's and normal random errors $\varepsilon_{i,k}$'s. Using this function, the unknown variance parameters are given by optimizing the restricted maximum likelihood (REML) criterion in this package (Bates et al., 2015).

It is worth noting that, the estimators generated by LME are equivalent to the weighted least squared estimators with variance parameters τ^2, σ_1^2 and σ_2^2 given by maximum likelihood type approaches. Therefore, using the same set of plug-in variance estimators, the difference of the estimators given by LME and COE converges to zero as $n \to \infty$ given by the proof of Proposition 2 in the Supplement S2.

We describe the data generation scheme for the simulation study as follows.

Assume that there are n test units. The outcomes of the paired experiments are simulated from the model

$$y_{i,k}(x) = 1 + x\beta_k + \epsilon_{i,k} + u_i, \quad \text{for} \quad i = 1, \ldots, n, \quad k = 1, 2, \quad \text{and} \quad x \in \{-1, 1\}, \tag{4.17}$$

where $x\beta_k$ is a linear function of x, $\epsilon_{i,k}$ is a mean-zero normal random error with variance $\sigma_k^2 = 1$ and u_i is the individual random effect. We use some different ways to generate u_i's. If u_i's are iid samples from $N(0, \tau^2)$, the outcomes are exactly generated under the model assumption in (2.1). We generate two $n \times 1$ design vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ satisfying the balanced and orthogonal assumption in Assumption 1. For $k = 1, 2$, we generate the responses $y_{i,k}(x)$ for each entry in $\boldsymbol{x}_k$. We specify the missing rate of outcomes r and randomly mark nr responses as missing for each experiment.

We use mean squared error (MSE) to evaluate the accuracy of the estimators. Without loss of generality, we only report the MSE of β_1 for the first experiment. The results and conclusion for β_2 should be similar to β_1. For an estimator, we replicate the data generation and estimation procedure under the same setting 100 times and obtain 100 estimators $\hat{\beta}_1^1, \ldots, \hat{\beta}_1^{100}$. For a clear comparison of different estimators, we compute the ratio between the MSEs of one estimator and the single experiment estimator:

$$\text{MSE.ratio} = \frac{100^{-1} \sum_{l=1}^{100} (\hat{\beta}_1^l - \beta_1)^2}{100^{-1} \sum_{l=1}^{100} (\hat{\beta}_1^{s,l} - \beta_1)^2}, \tag{4.18}$$

where $\hat{\beta}_1^{s,l}$'s are the single experiment estimators of β_1 in (2.3). The value of MSE.ratio is the smaller the better. If this value is less than 1, it indicates that the estimator is more accurate than the single analysis estimator.

In the remaining of this section, we conduct numerical studies to investigate three aspects: (1) the computational advantage of COE compared with LME; (2) the robustness of COE under different assumptions of the user effect u_i in (4.17); (3) the robustness of COE under different types of responses.

4.1 Computational Advantage

We demonstrate the computational advantage of COE compared with LME. The responses are generated following model in (4.17) with $u_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ with $\tau = 2$ and $\beta_k = 1$ for $k = 1, 2$. Therefore, the data is exactly generated by the true model in (2.1). Since the purpose of this study is to demonstrate the computational advantage, we assume that the data are exactly paired without missing values. To compute the MSE.ratio in (4.18), $\beta_1 = 1$. We show the MSE.ratio and average computational time (i.e., average CPU time over 100 replications) in Figure 2. According to the results in Figure 2, COE outperforms LME in computational time, and the advantage of COE is more distinct as n increases. Also, COE and LME have the same level of accuracy, and both outperform the single experimental estimator since the values of MSE.ratio are less than one.

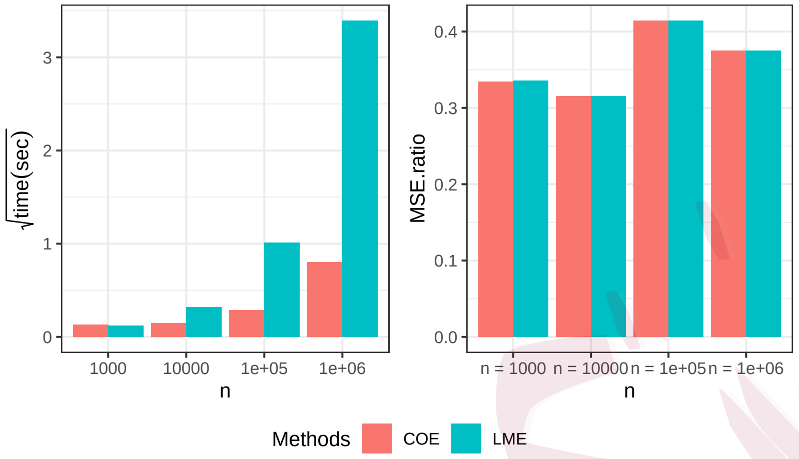

Figure 2: Comparison between COE and LME on the average computational time over 100 replication and the MSE.ratio in (4.18).

4.2 Robustness to Random Effect

We consider different settings of the random effect u_i's. For $i = 1, \ldots, n$, we generate $\boldsymbol{w}_i \overset{\text{iid}}{\sim} MVN(0, I_{10})$, which is fixed for each user and used as the user's ten-dimensional latent covariates.

(a) We generate $u_i \overset{\text{iid}}{\sim} N(0, \tau^2)$ under the true model assumption in (2.1). Under this setting, the random effect is not associated with the latent covariates $\boldsymbol{w}_i$. We expect that COE has the best performance as stated in the theoretical results. Also, as demonstrated in (2.7), PAIRED outperforms SINGLE when the value of τ (e.q., τ/σ_2 since $\sigma_2 = 1$ in the simulation) is greater than one, whereas SINGLE outperforms PAIRED when the value of τ is less than one.

(b) We generate

$$u_i = \boldsymbol{w}_i^\top \boldsymbol{\gamma} \quad \text{with} \quad \boldsymbol{\gamma} \sim MVN(0, \tau^2 I_{10}).$$

Under this setting, the variances (conditional on $\boldsymbol{w}_i$) of the random effects are different for different test units. Although the data under this case is not generated as the true model in (2.1), we expect that taking the difference between the outcomes of two experiments can still remove the user effects. Therefore, both COE and PAIRED can show some advantages especially when the value of τ is larger than 1.

(c) We generate

$$u_i = \boldsymbol{w}_i^\top \boldsymbol{\gamma}_k \quad \text{with} \quad \boldsymbol{\gamma}_k \sim MVN(0, \tau^2 I_{10}) \quad \text{for} \quad k = 1, 2.$$

Under this setting, the user effects of the same individual are different for different experiments. Therefore, taking the difference between the outcomes of the two experiments can not remove the user effects. We expect that PAIRED can not outperform SINGLE. However, the performances of COE remain robust and should at least be similar to SINGLE.

(d) We first generate $\boldsymbol{\gamma}_k \sim MVN(0, \tau^2 I_{10})$ for $k = 1, 2$, then generate

$$u_i = \boldsymbol{w}_i^\top \boldsymbol{\gamma}_1 I(x_{i,k} = 1) + \boldsymbol{w}_i^\top \boldsymbol{\gamma}_2 I(x_{i,k} = -1) \quad \text{for} \quad k = 1, 2.$$

where x_{ik} is the treatment allocation of i-th user in the k-th experiments. Under this setting, there are interactions between the user and treatment effects. Taking the difference between the outcomes of the two experiments may have a small contribution in removing the user effects due to the interaction. Therefore, we expect that PAIRED can not outperform SINGLE, but COE can outperform SINGLE slightly.

For all four settings, we specify $\beta_k = 1$ for $k = 1, 2$. We vary the standard deviation of the user effect $\tau \in \{0.5, 1, 2, 3, 4, 5\}$, the missing rate $r \in \{0.1, 0.3\}$ and sample size $n \in \{1000, 10000\}$. The results of MSE.ratio are shown in Figure 3. Similarly to what we expected when introducing each case above, the results show that the MSE.ratio of COE is consistently the best among the three methods for all four random effects settings u_i. Compared to SINGLE, the advantage of COE is larger for larger variance parameter τ^2. Compared to PAIRED, the advantage of COE is larger for smaller variance parameter τ^2 or larger missing rate. Also, we can see that COE is more robust than PAIRED with respect to misspecified models (i.e., case (c) and (d)). Also, the advantage of COE in terms of MSE.ratio does not diminish as

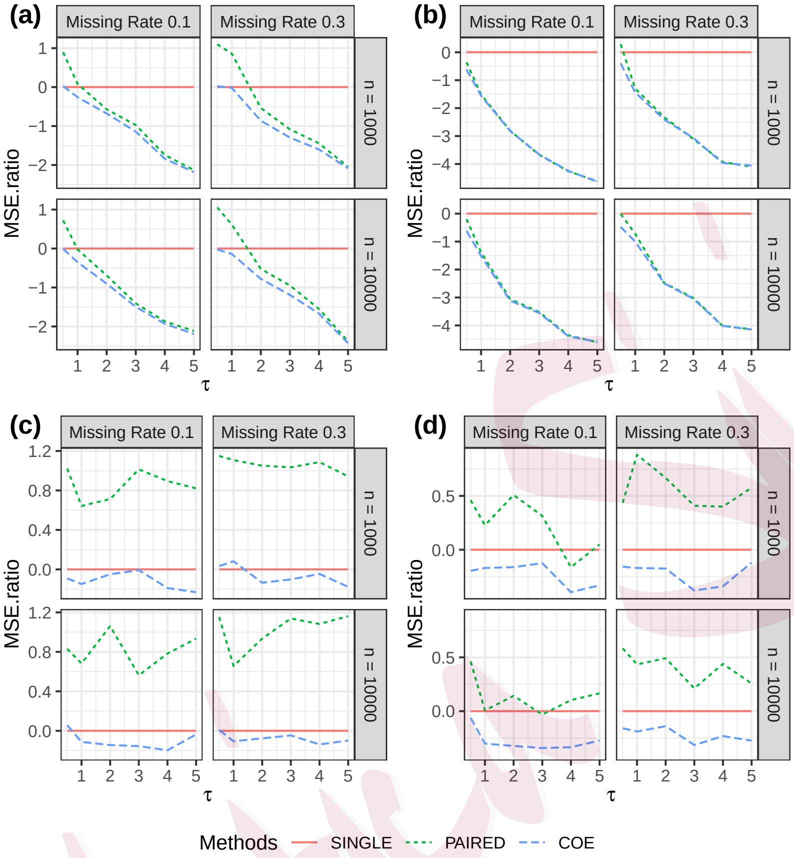

Figure 3: MSE.ratio in (4.18) under user effect settings (a)-(d)

we increase the sample size n. Note that we have include additional results, such as bias and variance, and coverage rate of confidence intervals, in Sections S3 and S4 of Supplementary Materials. In Section S6, we also provide a theoretical justification on why the COE shows the robustness with respect to model misspecification, especially for the data generating models (a)–(d).

4.3 Robustness to Different Types of Outcomes

Although the derivation of the COE is for continuous outcome data, we show here that the COE outperforms other estimators for discrete outcome data as well. In this part, we consider binary and integer types of outcomes. For convenience, we generate the discrete outcomes based on some latent continuous response that follows the model in (2.1). Specifically, for a binary outcome, we define it based on the continuous outcome $y_{i,k}(x)$ from (2.1) as

$$\tilde{y}_{i,k}(x) = I\left\{y_{i,k}(x) > \text{median}\right\}, \tag{4.19}$$

where "median" is the median outcomes of all $y_{i1}(1)$ for $i = 1, \ldots, n$. For an integer or count outcome, we define it based the continuous outcome $y_{i,k}(x)$ from (2.1) as

$$\tilde{y}_{i,k}(x) = \lfloor y_{i,k}(x) - \text{min} \rfloor, \tag{4.20}$$

where $\lfloor t \rfloor$ gives the largest integer that is less than or equal to t and "min" is the minimum of $y_{i,k}$ for $i = 1, \ldots, n$, $k = 1, 2$ and $x \in \{-1, 1\}$.

For the discrete outcomes, the estimand in such cases is the average treatment effect, defined as $\tilde{\beta}_1$ below, which is usually the default parameter of interest in causal inference (Imbens and Rubin, 2015)

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} \{\tilde{y}_{i,1}(1) - \tilde{y}_{i,1}(-1)\}}{2n},$$

which is computed for each replication. Then the MSE.ratio in (4.18) is modified by

$$\text{MSE.ratio} = \frac{100^{-1} \sum_{l=1}^{100} (\hat{\beta}_1^l - \tilde{\beta}_1^l)^2}{100^{-1} \sum_{l=1}^{100} (\hat{\beta}_1^{s,l} - \tilde{\beta}_1^l)^2}. \tag{4.21}$$

Under the model for continuous outcome in (2.1), the expected value of $\tilde{\beta}_1^l$ coincides with β_1. Therefore, our previous theoretical and numerical results still hold when the average treatment effect is the parameter of interest, provided the underlying model is (2.1).

Next we evaluate different methods using MSE.ratio in (4.21) with β_1 replaced by the actual estimand $\tilde{\beta}_1^l$, and both $\tilde{y}_{i1}(1)$ and $\tilde{y}_{i1}(-1)$ can be generated from the simulation model. The results for binary and integer/count outcomes are shown in Figures 4 and 5, respectively. The comparison conclusion is similar to the continuous outcome case in Figure 3. The advantage of COE is robust over different types of responses. What is more, the advantage of COE over PAIRED is more significant than the continuous outcome case. Moreover, we have conducted additional analysis results, including investigation of bias and variance and the coverage rate of confidence intervals, and report them in Sections S3 and S4 of Supplementary Materials.

We would like to remark that the outcomes of a pair of experiments may differ in scale or type in practice. The proposed COE works best when the outcomes from both experiments are of the same type and have the same scale. However, when the outcomes of a pair of experiments are at difference scales, the COE could still offer certain advantages by partially removing common user effect components. An example of mixed types of outcomes is provided in Section S5 of the Supplementary Materials.

5. Case Studies

To comprehensively evaluate the performance of the proposed method, we conduct two case studies in this section. The first case study is pseudo-study based on real data on customer campaign. The details of this case study and results are in Section S7 of Supplementary Materials. The second case study is about the ASSISTments Experiments described in this section below.

Following the motivating example in Section 1.1, we provide the analysis results for three selected ASSISTments experiments. The information of the three experiments are given in Table 2. The table shows that the problem types of the three experiments are all related to introductory topics in probability. For each experiment, we provide the number of users participated the experiment, the number of users allocated to the treatment group (e.q., the remaining users are allocated to the

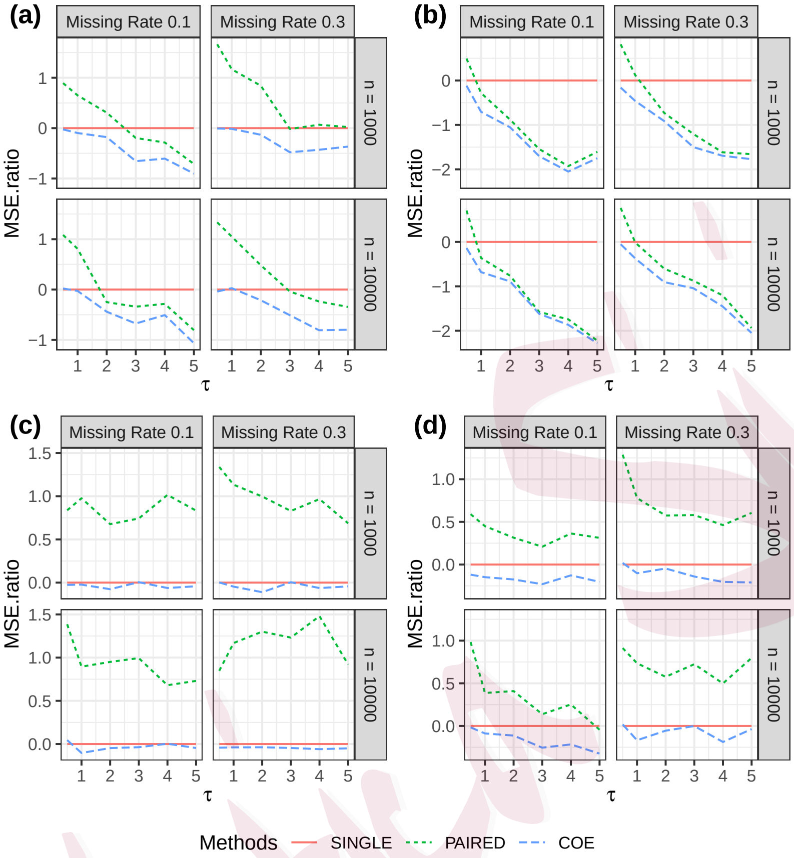

Figure 4: Modified MSE.ratio in (4.21) under user effect settings (a)-(d) for binary responses.

control group), and the number of users completed the problem set. As noted earlier, the treatment and control in each experiment are the two hint conditions tailored for each problem type, and the complete status will be recorded as 1 or 0 and serve as the binary response of the experiments.

The three experiments induced three pairs of experiments as in Table 3. This table shows the number of overlapping users for each pair of experiments, along with

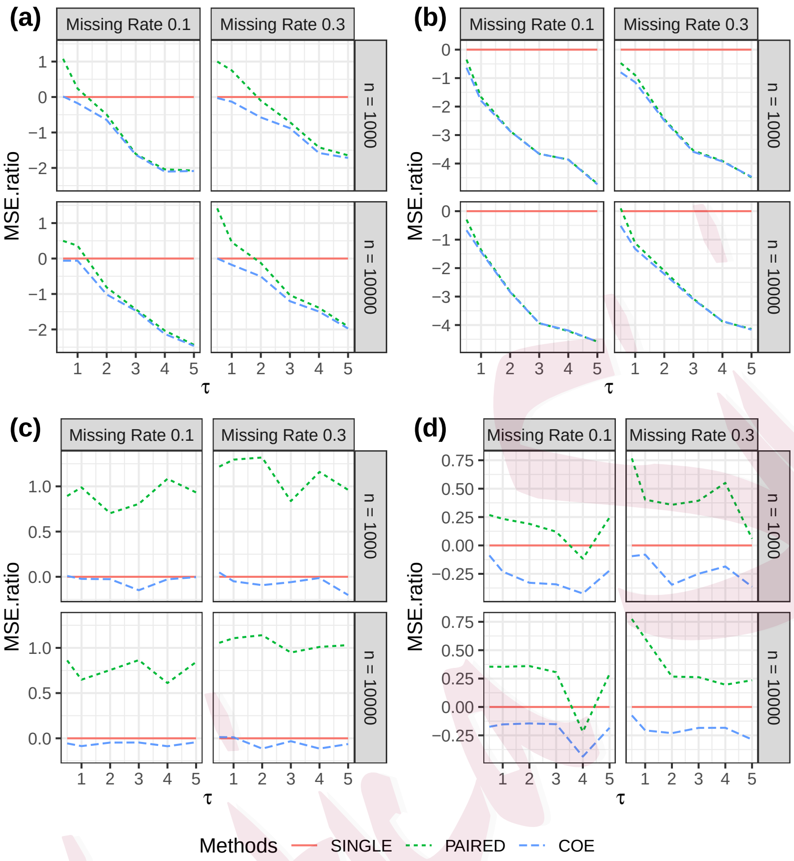

Figure 5: Modified MSE.ratio in (4.21) under user effect settings (a)-(d) for integer/count responses.

Table 2: Information and descriptive statistics of the three experiments

ID	Problem Type	# Users	# Treatment	# Complete
a	Conditional Probability	514	235	365
b	Finding Expected Value	457	232	337
c	Permutations and Combinations	538	264	455

the total number of users who participated in either experiment in each pair. As shown in Tables 2 and 3, the number of users allocated to the treatment or control group is roughly balanced across all three experiments, and there is substantial user overlap. Therefore, it is a reasonable approach to use COE to estimate the treatment effect.

Table 3: Information and descriptive statistics of the three experiments

Pair	Experiment 1	Experiment 2	# of Overlapped Users	# of Total Users
1	a	b	300	671
2	b	c	277	718
3	a	c	385	667

We use COE, PAIRED and SINGLE to estimate the treatment effects for the three pairs of experiments. The approach LME introduced in Section 4 is used as the benchmark to compute estimation error. For an estimator $\hat{\beta}_k$ of the k-th experiment, the "Error" is computed by

$$\text{Error} = |\hat{\beta}_k - \hat{\beta}_k^{\text{LME}}| \quad \text{for} \quad k = 1, 2, \tag{5.22}$$

which is the smaller the better. The Errors for three pairs of experiments are depicted in Figure 6. The results show that COE gives the best estimator compared to PAIRED and SINGLE.

6. Conclusion and Discussion

This paper proposed a collaborative analysis approach for a pair of A/B testing experiments carried out on the same set of users. The proposed approach can work well in practical situations where partially paired A/B testing experiments are commonly

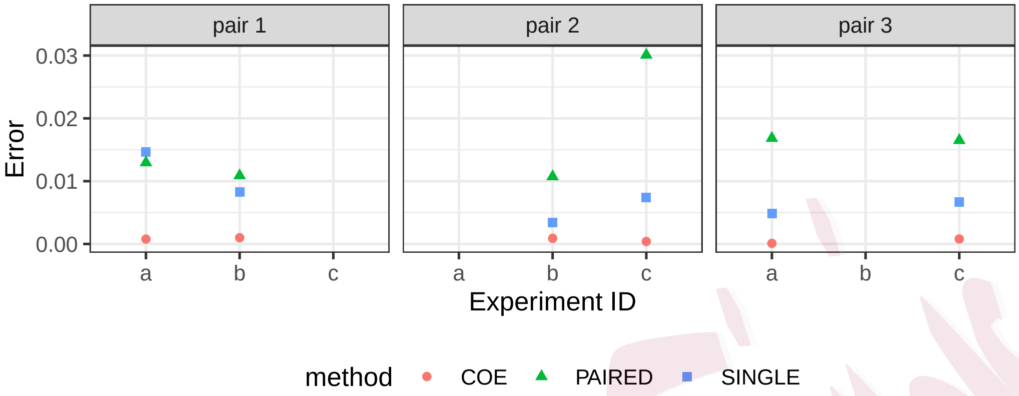

Figure 6: The Error in (5.22) for three pairs of experiments in Table 3.

encountered. Overall, compared to the linear mixed effect model, the proposed collaborative analysis approach is computationally efficient and easy to implement over online experimental platforms. Compared to single experiment analysis and paired analysis approaches, the proposed collaborative analysis approach consistently gives more accurate estimators over different scenarios.

Although this work focuses on paired A/B testing experiments, the proposed concept of collaborative analysis can be extended to the collaborative analysis of multiple A/B testing experiments that involve the same set of users. In addition, it will be interesting to consider multivariate mixed responses in each of the multiple A/B testing experiments. One can leverage some techniques developed in Chen et al. (2023) to facilitate the model estimation and inference, or the linear adjustment method developed by Freedman (2008) to reduce the variance when the working linear model is misspecified. Also, the experimental design issue in the presence of user covariates (e.g., Li et al. (2021)) and network information (e.g., Zhang and Kang (2022)) can be interesting to explore in the future. One can further extend the framework to personalized preference learning via collaborative experiments (e.g., Zhang et al. (2022); Li et al. (2023)).

References

Andrzej, G. and Tomasz, B. (2012), “Linear Mixed Effects Models Using R: A Step-by-step Approach,” .

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015), “Fitting Linear Mixed-Effects Models Using lme4,” *Journal of Statistical Software*, 67, 1–48.

Chen, H., Ding, P., Geng, Z., and Zhou, X.-H. (2015), “Semiparametric Inference of the Complier Average Causal Effect with Nonignorable Missing Outcomes,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7, 1–15.

Chen, X., Kang, X., Jin, R., and Deng, X. (2023), “Bayesian Sparse regression for mixed multi-responses with application to runtime metrics prediction in fog manufacturing,” *Technometrics*, 65, 206–219.

Deng, A., Xu, Y., Kohavi, R., and Walker, T. (2013), “Improving the sensitivity of online controlled experiments by utilizing pre-experiment data,” in *Proceedings* *of the sixth ACM international conference on Web search and data mining*, pp. 123–132.

Ding, P. (2024), *A first course in causal inference*, CRC Press.

Freedman, D. A. (2008), “On regression adjustments to experimental data,” *Advances in Applied Mathematics*, 40, 180–193.

Gupta, S., Kohavi, R., Tang, D., Xu, Y., Andersen, R., Bakshy, E., Cardin, N., Chandran, S., Chen, N., Coey, D., et al. (2019), “Top challenges from the first practical online controlled experiments summit,” *ACM SIGKDD Explorations Newsletter*, 21, 20–35.

Heffernan, N. T. and Heffernan, C. L. (2014), “The ASSISTments ecosystem: Building a platform that brings scientists and teachers together for minimally invasive research on human learning and teaching,” *International Journal of Artificial Intelligence in Education*, 24, 470–497.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Jin, Y. and Ba, S. (2023), “Toward optimal variance reduction in online controlled experiments,” *Technometrics*, 65, 231–242.

Kohavi, R., Tang, D., and Xu, Y. (2020), *Trustworthy online controlled experiments: A practical guide to a/b testing*, Cambridge University Press.

Larsen, N., Stallrich, J., Sengupta, S., Deng, A., Kohavi, R., and Stevens, N. T. (2023), "Statistical Challenges in Online Controlled Experiments: A Review of A/B Testing Methodology," *The American Statistician*, 0, 1–15.

Li, Y., Kang, L., and Huang, X. (2021), "Covariate balancing based on kernel density estimates for controlled experiments," *Statistical Theory and Related Fields*, 5, 102–113.

Li, Y., Zhang, Q., Khademi, A., and Yang, B. (2023), "Optimal Design of Controlled Experiments for Personalized Decision Making in the Presence of Observational Covariates," *The New England Journal of Statistics in Data Science*, 1, 386–393.

Nassi, T. and Jewkes, H. (2021), "Simultaneous Experimentation: Run Multiple A/B Tests Concurrently," `https://www.split.io/blog/simultaneous-experiments/`.

Neyman, J. (1923), "On the application of probability theory to agricultural experiments. Essay on principles," *Ann. Agricultural Sciences*, 1–51.

Poyarkov, A., Drutsa, A., Khalyavin, A., Gusev, G., and Serdyukov, P. (2016), "Boosted decision tree regression adjustment for variance reduction in online controlled experiments," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 235–244.

Rubin, D. B. (1974), "Estimating causal effects of treatments in randomized and nonrandomized studies," *Journal of educational Psychology*, 66, 688.

Selent, D., Patikorn, T., and Heffernan, N. (2016), "Assistments dataset from multiple randomized controlled experiments," in *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pp. 181–184.

Shen, S., Mao, H., Zhang, Z., Chen, Z., Nie, K., and Deng, X. (2023), "Clustering-Based Imputation for Dropout Buyers in Large-Scale Online Experimentation," *The New England Journal of Statistics in Data Science*, 1, 415–425.

Syrgkanis, V., Lei, V., Oprescu, M., Hei, M., Battocchi, K., and Lewis, G. (2019), "Machine learning estimation of heterogeneous treatment effects with instruments," *Advances in Neural Information Processing Systems*, 32.

Zhang, Q. and Kang, L. (2022), "Locally optimal design for a/b tests in the presence of covariates and network dependence," *Technometrics*, 64, 358–369.

Zhang, Q., Khademi, A., and Song, Y. (2022), "Min-max optimal design of two-armed trials with side information," *INFORMS Journal on Computing*, 34, 165–182.

Zhao, A. and Ding, P. (2022), "To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates," *Journal of the American Statistical Association*, 1–11.

Zhao, A., Ding, P., and Li, F. (2024), "Covariate adjustment in randomized experiments with missing outcomes and covariates," *Biometrika*, asae017.

Zuo, S., Ghosh, D., Ding, P., and Yang, F. (2024), "Mediation analysis with the mediator and outcome missing not at random," *Journal of the American Statistical Association*, 1–21.

School of Mathematical and Statistical Sciences, Clemson University, Clemson, SC

E-mail: (qiongz@clemson.edu)

Department of Mathematics and Statistics, University of Massachusetts, Amherst

E-mail: (lulukang@umass.edu)

Department of Statistics, Virginia Tech, Blacksburg, VA

E-mail: (xdeng@vt.edu)