Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

RANDOMIZATION-BASED CAUSAL INFERENCE FROM SPLIT-PLOT DESIGNS*

By Anqi Zhao†, Peng Ding‡, Rahul Mukerjee§ and Tirthankar Dasgupta¶

Harvard University†, University of California at Berkeley‡ and Indian Institute of Management Calcutta §, Rutgers University ¶

Under the potential outcomes framework, we propose a randomization based estimation procedure for causal inference from split-plot designs, with special emphasis on 2^2 designs that naturally arise in many social, behavioral and biomedical experiments. Point estimators of factorial effects are obtained and their sampling variances are derived in closed form as linear combinations of the between- and within-group covariances of the potential outcomes. Results are compared to those under complete randomization as measures of design efficiency. Conservative estimators of these sampling variances are proposed. Connection of the randomization-based approach to inference based on the linear mixed effects model is explored. Results on sampling variances of point estimators and their estimators are extended to general split-plot designs. The superiority over existing model-based alternatives in frequency coverage properties is reported under a variety of simulation settings for both binary and continuous outcomes.

1. Introduction. Factorial experiments, originally developed in the context of agricultural experiments (Fisher, 1925, 1935; Yates, 1935) and later extensively used in industrial and engineering applications, are nowadays undergoing a third popularity surge among social, behavioral, and biomedical sciences, as a result of the massive trend in these areas to generalize the previous treatment-control experiments to include multiple factors. Among the plethora of possible multi-factor randomization schemes available, split-plot design, thanks to its flexibility and ease of application, has always remained a popular choice, especially when practical difficulties like economic constraints or hard-to-change factor preclude the use of simple, unrestricted

* Zhao and Dasgupta were partially supported by NSF grant number CMMI-1334178. Ding was partially supported by IES grant number R305D150040. Mukerjee was supported by the J. C. Bose National Fellowship of the Government of India and a grant from Indian Institute of Management Calcutta.

MSC 2010 subject classifications: Primary 62K15, 62K10; secondary 62K05

Keywords and phrases: Between-whole-plot additivity, Model-based inference, Neymanian inference, Potential outcomes framework, Projection matrix, Within-whole-plot additivity.

randomizations (Jones and Nachtsheim, 2009). As a motivating example, consider a simplified version of the education experiment described in Dasgupta, Pillai and Rubin (2015). The goal is to evaluate the efficacies of two interventions — A: a mid-year quality review by a team of experts, and B: a bonus scheme to teachers — on 224 schools in the state of New York. Assume two possible actions for each intervention — application or non-application, a complete randomization of the four combinations likely scatters the schools to be reviewed throughout the state. Considering the prohibitive cost of travel and amount of time cost associated with such a plan, a more practical alternative would be to divide the 224 schools by geographic proximity into sixteen 'blocks,' choose eight at random, and conduct expert quality review for all schools therein. The teacher bonus scheme can then be applied to half of the schools within each block. This exemplifies split-plot design, in which each block is considered as a larger experimental unit referred to as a *whole-plot*, and each school, the original experimental unit, a *sub-plot*. See Kirk (1982), Cochran and Cox (1957), Box, Hunter and Hunter (2005), and Wu and Hamada (2009) for formal definitions.

Most factorial experiments, like any experiment, receive regression-based methods as their default method of analysis. For those under split-plot designs, this default is either the analysis of variance (ANOVA) or the linear mixed effects model (Wu and Hamada, 2009). Despite the good intention of both methods to adjust for the group structure that defines split-plot designs, the actual variance estimation often turns out inconsistent (Gelman, 2005; Hinkelmann and Kempthorne, 2008), likely due to the required model assumptions not being satisfied. A detailed examination of this argument can be found in Freedman (2006, 2008a), who recommended randomized-based inference as the proper solution.

Despite its long tradition in the context of treatment-control experiments (Neyman, 1923/1990, 1935; Kempthorne, 1952; Imbens and Rubin, 2015; Ding and Dasgupta, 2016), randomization-based inference remains an almost uncharted field when it comes to factorial experiments. The recent works of Dasgupta, Pillai and Rubin (2015), Espinosa, Dasgupta and Rubin (2016) and Lu (2016) are, to the best of our knowledge, the only literature along this line, each documenting improvements of randomization-based analysis over existing model-based methods in the context of multi-factor completely randomized designs. Extending their methods to split-plot designs is a promising next step.

Randomization-based inference is particularly appealing when the inference has to be restricted to a finite population of experimental units that cannot be considered a random sample from a hypothetical super popula-

tion. Such scenarios mostly occur in social, behavioral and biomedical applications, where each factor typically has two levels - treatment and control. We will thus emphasize on two-level factorial experiments with a split-plot structure first, and in particular, consider a 2^2 experiment for two reasons. First, such experiments are the most simple, yet non-trivial extensions of treatment-control experiments with a multitude of applications in the social, behavioral and biomedical sciences. Second, a 2^2 experiment will make the exposition of the concepts and results more intuitive than a general case. The results and insights obtained from such a 2^2 design will then be extended to the case of more general split-plot designs.

The contribution of this paper is three-fold. First, we develop a randomization based estimation procedure for causal inference under 2^2 split-plot designs, and demonstrate its superior frequency coverage properties over existing alternatives via extensive simulations. Second, motivated by the group structure of units as a defining feature of split-plot designs, we propose a decomposition of the potential outcomes that links the difference in efficiency between a split-plot design and a complete randomization of the same size to the level of heterogeneity among blocks. Third, in an attempt to reconcile the finite-population randomization-based perspective and a hypothetical super-population model-based perspective, we offer a heuristic argument that connects the two. This connection is established by using the asymptotics of the finite-population randomization-based residual covariances to justify the block-diagonal structure assumed by the linear mixed effects model for the covariances of its super-population sampling errors. This, to the best of our knowledge, is the very first attempt that aims at reconciling the difference between finite and super-population inferences.

The article is organized as follows. We review in Section 2 the potential outcomes framework, discuss possible extensions when the experimental units exhibit a "split-plot" structure, and define the causal questions in 2^2 factorial experiments. The split-plot design, characterized by its treatment assignment mechanism, is introduced in Section 3. The point estimators of the factorial effects and their sampling variances are derived in Section 4, and their estimation addressed in Section 5. We discuss the connection and distinction between the model-based and randomization-based inferences in Section 6. Section 7 extends the results to general split-plot designs. The superiority of the proposed approach over model-based alternatives with respect to frequency coverage properties is demonstrated through simulation studies in Section 8. We conclude in Section 9. All proofs are deferred to the online supplementary material (Zhao, Ding and Dasgupta, 2017).

2. Potential outcomes and additivity assumptions. We review in this section the major concepts within the *potential outcomes framework* (Neyman, 1923/1990; Rubin, 1974, 1978, 2005), and discuss some possible extensions when the experimental units are nested within whole-plots.

2.1. *Potential outcomes framework for causal inference.* Consider an experiment in which K different *treatment*s are to be tested on N experimental *unit*s. The Stable Unit Treatment Value Assumption (Rubin, 1980) allows us to write the *potential outcome* of unit i when exposed to treatment k as $Y_i(k)$. Whereas causal effects are then defined as comparisons of such potential outcomes for a given set of units, any experiment, however well designed and implemented, allows us to observe at most one of K potential outcomes per unit, according to the treatment it receives. This poses the *fundamental problem of causal inference* (Holland, 1986). Various assumptions are introduced in this context as attempts to infer the unobserved from the observed, among which the *strict additivity assumption* is arguably the most common one.

DEFINITION 1. *The potential outcomes of N units under K treatments are 'strictly additive' if the differences between any two treatments are constant across all units, i.e., $Y_i(l) = Y_i(k) + C(k,l)$, where $C(k,l)$ are some fixed real numbers, for all $1 \le i \le N$, $1 \le k,l \le K$.*

For any positive integer p, let $\mathbf{1}_p$ be the p-dimensional vector of 1's, $\mathbf{J}_p$ be the $p \times p$ matrix of 1's, $\mathbf{I}_p$ be the $p \times p$ identity matrix, and $\mathbf{P}_p = \mathbf{I}_p - p^{-1}\mathbf{J}_p$ be the $p \times p$ projection matrix with column space orthogonal to $\mathbf{1}_p$. Given $\boldsymbol{Y}(k) = (Y_1(k), \ldots, Y_N(k))^{\mathrm{T}}$, let $\bar{Y}(k) = N^{-1}\sum_{i=1}^{N} Y_i(k) = N^{-1}\mathbf{1}_N^{\mathrm{T}}\boldsymbol{Y}(k)$ be the population average under treatment k, let $S(k,l) = (N-1)^{-1}\boldsymbol{Y}(k)^{\mathrm{T}}\mathbf{P}_N\boldsymbol{Y}(l)$ be the *finite-population covariance* of $Y_i(k)$ and $Y_i(l)$, and let

$$\mathbf{S} = (\!(S(k,l))\!)_{K\times K} = (N-1)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{P}_N\mathbf{Y} \tag{2.1}$$

be the *finite-population covariance matrix*, where $\mathbf{Y}$ is the $N \times K$ potential outcomes matrix with columns $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(K)$. Lemma 1 gives an alternative characterization of strict additivity in terms of $S(k,l)$.

LEMMA 1. *The potential outcomes $Y_i(k)$ of N units under K treatments are strictly additive if and only if the finite-population covariances $S(k,l)$ are the same for all $k,l \in \{1,\ldots,K\}$, i.e., $\mathbf{S} = S_0\mathbf{J}_K$, where S_0 is a non-negative constant.*

For simplicity, we will omit the 'finite-population' before 'covariance' in the following text when no confusion would arise. All averages and covariances over *a finite set of fixed numbers* will be finite-population in nature, and defined the same way as $\bar{Y}(k)$ and $S(k,l)$ are defined for $Y_i(k)$.

2.2. *Experimental units with a split-plot structure* . Whereas all definitions and discussion above apply universally to any K-treatment experiment with N experimental units, possible extensions arise when the experimental units in question exhibit a structure called the split-plot structure, as a result of either intrinsic characteristics like geographic proximity, or extrinsic arrangements as induced by the design.

In particular, assume the N experimental units are nested under W groups called whole-plots, each of size $M = N/W$. Index the whole-plots by w, running from 1 to W, and the units within whole-plot w by wm, running from $w1$ to wM. The *whole-plot average potential outcome*s are defined as

$$\bar{Y}_{w\cdot}(k) = M^{-1}\sum_{m=1}^{M} Y_{wm}(k), \quad (k = 1, \ldots, K).$$

These aggregated potential outcomes enable the definitions of some weaker forms of additivity as compared to that in Definition 1.

DEFINITION 2. *The potential outcomes of N units in W whole-plots under K treatments are*

- *'between-WP additive' if the corresponding whole-plot average potential outcomes $\bar{Y}_{w\cdot}(k)$ for W whole-plots under K treatments are strictly additive, i.e., $\bar{Y}_{w\cdot}(l) = \bar{Y}_{w\cdot}(k) + C(k,l)$, where $C(k,l)$ are some fixed real numbers, for all $1 \le w \le W$ and $1 \le k,l \le K$;*
- *'within-WP additive' if for each w, the potential outcomes of the M units within whole-plot w are strictly additive, i.e., $Y_{wm}(l) = Y_{wm}(k) + C_w(k,l)$, where $C_w(k,l)$ are some fixed real numbers, for all $1 \le m \le M$, $1 \le w \le W$ and $1 \le k,l \le K$.*

Strictly additive potential outcomes under a split-plot structure must be strictly additive within each whole-plot and have strictly additive whole-plot averages. Lemma 2 asserts that the converse is also true.

LEMMA 2. *The potential outcomes of N units in W whole-plots are strictly additive if and only if they are both between- and within-WP additive.*

Denote by $\boldsymbol{Y}_w(k) = (Y_{w1}(k), \ldots, Y_{wM}(k))^{\mathrm{T}}$ the sub-vector of $\boldsymbol{Y}(k)$ corresponding to whole-plot w, and by $\bar{\boldsymbol{Y}}(k) = (\bar{Y}_{1\cdot}(k), \ldots, \bar{Y}_{W\cdot}(k))^{\mathrm{T}}$ the vector

of whole-plot average potential outcomes under treatment k. Let '$\otimes$' be the Kronecker product. Let

$$\mathbf{P}_{\text{in}} = \mathbf{I}_W \otimes \mathbf{P}_M\,, \quad \mathbf{P}_{\text{btw}} = \mathbf{P}_W \otimes (M^{-1}\mathbf{J}_M)$$

be two mutually orthogonal $N \times N$ projection matrices that satisfy

$$\begin{aligned}\mathbf{P}_{\text{in}}\boldsymbol{Y}(k) &= \boldsymbol{Y}(k) - \bar{\boldsymbol{Y}}(k) \otimes \mathbf{1}_M = \big(\!\big(Y_{wm}(k) - \bar{Y}_{w\cdot}(k)\big)\!\big)\,,\\ \mathbf{P}_{\text{btw}}\boldsymbol{Y}(k) &= \bar{\boldsymbol{Y}}(k) \otimes \mathbf{1}_M - \bar{Y}(k) \otimes \mathbf{1}_N = \big(\!\big(\bar{Y}_{w\cdot}(k) - \bar{Y}(k)\big)\!\big)\end{aligned}$$

and thus decompose the variation of unit wm with respect to the population average into the within- and between-WP parts.

It is straightforward to verify that

$$\mathbf{P}_N\boldsymbol{Y}(k) = \mathbf{P}_{\text{in}}\boldsymbol{Y}(k) + \mathbf{P}_{\text{btw}}\boldsymbol{Y}(k)\,.$$

Multiplying both sides of the above equation by $\boldsymbol{Y}(l)^{\text{T}}$, we have

$$\boldsymbol{Y}(l)^{\text{T}}\mathbf{P}_N\boldsymbol{Y}(k) = \boldsymbol{Y}(l)^{\text{T}}\mathbf{P}_{\text{in}}\boldsymbol{Y}(k) + \boldsymbol{Y}(l)^{\text{T}}\mathbf{P}_{\text{btw}}\boldsymbol{Y}(k)\,,\ \forall 1 \le k, l \le K\,. \tag{2.2}$$

This gives the potential outcome analogue of the sum of squares decomposition pervading the ANOVA, extending the results of the observed outcomes in the presence of split-plot structure to that of the potential outcomes. Recalling the definition of the matrix $\mathbf{Y}$ with columns $\boldsymbol{Y}(1), \ldots, \boldsymbol{Y}(K)$, the K^2 identities in (2.2) can be summarized in matrix form as

$$\mathbf{Y}^{\text{T}}\mathbf{P}_N\mathbf{Y} = \mathbf{Y}^{\text{T}}\mathbf{P}_{\text{in}}\mathbf{Y} + \mathbf{Y}^{\text{T}}\mathbf{P}_{\text{btw}}\mathbf{Y}\,. \tag{2.3}$$

Analogous to the between- and within-WP variances of the observed outcomes in ANOVA, define the *between-* and *within-WP variances* of $Y_{wm}(k)$ as the between- and within-WP sums of squares normalized by their respective degrees of freedom:

$$\begin{aligned}S_{\text{btw}}(k,k) &= \frac{\|\,\big(\!\big(\bar{Y}_{w\cdot}(k) - \bar{Y}(k)\big)\!\big)\,\|^2}{W-1} = \frac{\|\mathbf{P}_{\text{btw}}\boldsymbol{Y}(k)\|^2}{W-1} = \frac{\boldsymbol{Y}(k)^{\text{T}}\mathbf{P}_{\text{btw}}\boldsymbol{Y}(k)}{W-1}\,,\\ S_{\text{in}}(k,k) &= \frac{\|\,\big(\!\big(Y_{wm}(k) - \bar{Y}_{w\cdot}(k)\big)\!\big)\,\|^2}{N-W} = \frac{\|\mathbf{P}_{\text{in}}\boldsymbol{Y}(k)\|^2}{N-W} = \frac{\boldsymbol{Y}(k)^{\text{T}}\mathbf{P}_{\text{in}}\boldsymbol{Y}(k)}{N-W}\,,\end{aligned}$$

and the *between-* and *within-WP covariances* of $Y_{wm}(k)$ and $Y_{wm}(l)$ as

$$S_{\text{btw}}(k,l) = \frac{\boldsymbol{Y}(k)^{\text{T}}\mathbf{P}_{\text{btw}}\boldsymbol{Y}(l)}{W-1}\,, \quad S_{\text{in}}(k,l) = \frac{\boldsymbol{Y}(k)^{\text{T}}\mathbf{P}_{\text{in}}\boldsymbol{Y}(l)}{N-W}\,. \tag{2.4}$$

Let

$$\mathbf{S}_{\text{btw}} = \big(\!\big(S_{\text{btw}}(k,l)\big)\!\big) = \frac{\mathbf{Y}^{\text{T}}\mathbf{P}_{\text{btw}}\mathbf{Y}}{W-1}\,, \quad \mathbf{S}_{\text{in}} = \big(\!\big(S_{\text{in}}(k,l)\big)\!\big) = \frac{\mathbf{Y}^{\text{T}}\mathbf{P}_{\text{in}}\mathbf{Y}}{N-W}\,. \tag{2.5}$$

These two covariances matrices allow us to write identity (2.3) as

$$\mathbf{S} = \frac{N-W}{N-1}\mathbf{S}_{\text{in}} + \frac{W-1}{N-1}\mathbf{S}_{\text{btw}}. \tag{2.6}$$

We further characterize the between- and within-WP additivities in Definition 2 as follows.

LEMMA 3. *Given the potential outcomes $Y_{wm}(k)$ of K treatments on N units nested within W whole-plots, we have $\mathbf{S}_{\text{btw}} = S_{\text{btw}}\mathbf{J}_K$ for some non-negative number S_{btw} if $Y_{wm}(k)$ are between-WP additive, and $\mathbf{S}_{\text{in}} = S_{\text{in}}\mathbf{J}_K$ for some non-negative number S_{in} if $Y_{wm}(k)$ are within-WP additive.*

We have so far introduced, in the context of general K-treatment experiments, all concepts about the potential outcomes framework that we consider relevant to the current topic. Specific definitions, concepts, notations and causal questions associated with potential outcomes for 2^2 factorial experiments are introduced in the next subsection.

2.3. *Potential outcomes and causal effects for 2^2 factorial experiments.* As the name suggests, 2^2 factorial experiments involve $K = 4$ different treatments as the 2^2 possible combinations of two 2-level factors of interest. Refer to the two factors as factors 'A' and 'B.' Of primary causal interest are the *main effect of factor A* (indexed by 'A'), the *main effect of factor B* (indexed by 'B'), and the *effect of interaction between A and B* (indexed by 'AB', also refer to as 'factor AB'). We set out in this section their formal definitions at unit, block, and population levels.

To start with, code the levels of factors A and B as $\{-1_A, +1_A\}$ and $\{-1_B, +1_B\}$ respectively. We represent the four treatment combinations as $(-1_A, -1_B)$, $(-1_A, +1_B)$, $(+1_A, -1_B)$, and $(+1_A, +1_B)$, and name in lexicographic order as treatments 1 to 4.

The *factorial effect of factor $F \in \mathcal{F} = \{A, B, AB\}$* on unit i is defined as

$$\tau_i(F) = 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}(Y_i(1), \ldots, Y_i(4))^{\mathrm{T}} \quad i = 1, \ldots, N, \tag{2.7}$$

where

$$\boldsymbol{g}_A = (-1, -1, +1, +1)^{\mathrm{T}},\ \boldsymbol{g}_B = (-1, +1, -1, +1)^{\mathrm{T}} \text{ and } \boldsymbol{g}_{AB} = (+1, -1, -1, +1)^{\mathrm{T}}$$

are mutually orthogonal contrasts that vectorize the levels of factors A, B, and AB in treatments 1 to 4 respectively. Let

$$\tau_F = N^{-1}\sum_{i=1}^{N}\tau_i(F) = 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}(\bar{Y}(1), \ldots, \bar{Y}(4))^{\mathrm{T}} = 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}\bar{\boldsymbol{Y}} \tag{2.8}$$

be the population average, where $\bar{\boldsymbol{Y}} = (\bar{Y}(1), \ldots, \bar{Y}(4))^{\mathrm{T}}$ is the vector of average potential outcomes for the treatments. Let $S_F = (N-1)^{-1} \sum_{i=1}^{N} \{\tau_i(F) - \tau_F\}^2$ be the variance of the unit-level factorial effects. Lemma 4 restates strict additivity in terms of S_F.

LEMMA 4. *The potential outcomes of N units in a 2^2 factorial experiment, $Y_i(k)$ ($i = 1, \ldots, N$; $k = 1, 2, 3, 4$), are strictly additive if and only if all three unit-level factorial effects are constant across all units, i.e., $\tau_i(F) = \tau_F$ for all $i \in \{1, \ldots, N\}$ and $F \in \mathcal{F}$; or equivalently, $S_F = 0$ for each $F \in \mathcal{F}$.*

When the experimental units are nested in whole-plots, we index the units by double-index wm, and extend the idea of aggregate effects to whole-plot level to define the *whole-plot average factorial effects* as

(2.9)
$$\tau_{w\cdot}(F) = M^{-1} \sum_{m=1}^{M} \tau_{wm}(F) = 2^{-1} \boldsymbol{g}_F^{\mathrm{T}} (\bar{Y}_{w\cdot}(1), \ldots, \bar{Y}_{w\cdot}(4))^{\mathrm{T}} = 2^{-1} \boldsymbol{g}_F^{\mathrm{T}} \bar{\boldsymbol{Y}}_{w\cdot},$$

interpretable as WP-level factorial effects along the lines of (2.7), where $\bar{\boldsymbol{Y}}_{w\cdot}$ is the vector of whole-plot averages for the treatments.

Let $\boldsymbol{\tau}_F = (\tau_{11}(F), \ldots, \tau_{WM}(F))^{\mathrm{T}}$. We define the *between-* and *within-WP variances* of $\tau_{wm}(F)$ the same way as (2.4) defined $S_{\mathrm{btw}}(k,k)$ and $S_{\mathrm{in}}(k,k)$:

$$S_{\mathrm{btw}}(F) = \frac{\boldsymbol{\tau}_F^{\mathrm{T}} \mathbf{P}_{\mathrm{btw}} \boldsymbol{\tau}_F}{W-1}, \quad S_{\mathrm{in}}(F) = \frac{\boldsymbol{\tau}_F^{\mathrm{T}} \mathbf{P}_{\mathrm{in}} \boldsymbol{\tau}_F}{N-W}.$$

These variances give an alternative characterization of the between- and within-WP additivities as detailed in Lemma 5.

LEMMA 5. *Given N experimental units in a 2^2 factorial experiment that are nested under W whole-plots and indexed by wm, the corresponding potential outcomes are*

- *between-WP additive if and only if all three whole-plot average factorial effects $\tau_{w\cdot}(F)$ are constant across all whole-plots, i.e., $\tau_{w\cdot}(F) = \tau_F$ for all $1 \le w \le W$ and $F \in \mathcal{F}$, or equivalently, $S_{\mathrm{btw}}(F) = 0$ for each $F \in \mathcal{F}$;*
- *within-WP additive if and only if all three unit-level factorial effects $\tau_{wm}(F)$ are constant within each whole-plot, i.e., $\tau_{wm}(F) = \tau_{w\cdot}(F)$ for all $1 \le m \le M$, $1 \le w \le W$ and $F \in \mathcal{F}$, or equivalently, $S_{\mathrm{in}}(F) = 0$ for each $F \in \mathcal{F}$.*

3. Treatment assignment for a 2^2 split-plot design. We now introduce the notion of the *treatment assignment mechanism* (Imbens and Rubin, 2015) that leads to generation of observed outcomes as a function of the potential outcomes discussed in Section 2 and a random vector of treatment assignment variables. We also highlight the key difference between the assignment variables for a completely randomized and a split-plot 2^2 design, and provide characterizations of both types of designs through the randomization distribution of these assignment variables.

Assume that it is decided a priori that N_k experimental units will be assigned to treatment k ($k = 1, \ldots, 4$). In a completely randomized (C-R) design the N units are assigned to the four treatment combinations at random without any restriction. However, in a split-plot (S-P) design one factor (e.g., A) is identified as the *whole-plot factor* and the treatments are assigned in such a way that *all* units within the same whole-plot receive the same level (i.e. either -1_A or $+1_A$) of the whole-plot factor. Next, the two levels of the other factor, referred to as the sub-plot factor, are assigned to the units within each whole-plot using a C-R assignment mechanism. The name split-plot can be attributed to the design's agricultural origin.

Let T_i be the assignment variable, taking the value k if unit i is assigned to treatment k ($k = 1, \ldots, 4$). We characterize by Definitions 3 and 4 a '2^2 C-R design' and '2^2 S-P design'. Most of the quantitative derivations in this article will be based on these definitions and the randomization distributions of these assignment variables.

DEFINITION 3. *Given treatments 1 to 4 in a 2^2 factorial experiment and N experimental units, a 2^2 completely randomized design with planned treatment arm sizes N_1, N_2, N_3, and $N_4 = N - \sum_{k=1}^{3} N_k$ assigns N_k units to treatment k such that*

$$\text{pr}(T_i = t_i, i = 1, \ldots, N) = \begin{cases} \prod_{k=1}^{4} N_k!/N!, & \sum_{i=1}^{N} I_{\{t_i=k\}} = N_k, k = 1, 2, 3, 4, \\ 0, & \textit{otherwise.} \end{cases}$$

DEFINITION 4. *Given two 2-level factors of interest, whole-plot factor A and sub-plot factor B, and N experimental units nested within W whole-plots, each of size $M = N/W$, a 2^2 split-plot design with planned size parameters W_+ and M_+ consists of two separate randomizations:*

- *Whole-plot randomization that assigns W_+ of W whole-plots chosen at complete random to $+1_A$ level of whole-plot factor A, and the remaining $W_- = W - W_+$ ones to -1_A level,*
- *Sub-plot randomization that assigns M_+ of M sub-plots chosen at complete random within each whole-plot to $+1_B$ level of sub-plot factor B,*

and the remaining $M_- = M - M_+$ *ones to* -1_B *level.*

The final treatment for sub-plot wm *will be the combination of the level of factor* A *received by whole-plot* w *in the whole-plot randomization and the level of factor* B *received by itself in the sub-plot randomization. The treatment arm sizes are given by*

$$(N_1, N_2, N_3, N_4) = (W_-M_-, W_-M_+, W_+M_-, W_+M_+)\,. \tag{3.1}$$

Let $\boldsymbol{Z}(k) = (I_{\{T_1=k\}}, \ldots, I_{\{T_N=k\}})^{\mathrm{T}}$, such that the sum of its entries, $\sum_{i=1}^{N} I_{\{T_i=k\}}$, equals N_k. We define

$$\boldsymbol{Z}^* = (N_1^{-1}\boldsymbol{Z}(1)^{\mathrm{T}}, \ldots, N_K^{-1}\boldsymbol{Z}(K)^{\mathrm{T}})^{\mathrm{T}} \tag{3.2}$$

to be the *assignment vector*, in which each $\boldsymbol{Z}(k)$ is normalized by its treatment arm size to have entrywise sum one. This NK-dimensional vector gives a full representation of the randomization result, in a form that promises easier algebra than $\{T_i\}_{i=1}^{N}$.

LEMMA 6. *Under the* 2^2 *completely randomized design characterized by Definition 3, the sampling expectation and covariance matrix of the assignment vector* $\boldsymbol{Z}^*$ *given by (3.2) are*

$$E_{\text{C-R}}(\boldsymbol{Z}^*) = N^{-1}\mathbf{1}_{4N}\,, \quad \operatorname{cov}_{\text{C-R}}(\boldsymbol{Z}^*) = \mathbf{C} \otimes \mathbf{P}_N$$

where

$$\mathbf{C} = \frac{1}{N(N-1)}\left(\operatorname{diag}\left\{\frac{N}{N_1}, \frac{N}{N_2}, \frac{N}{N_3}, \frac{N}{N_4}\right\} - \mathbf{J}_4\right).$$

The whole-plot and sub-plot randomizations in Definition 4 are essentially two independent complete randomizations. The resulting 2^2 split-plot design can hence be thought of as a *restricted completely randomized design* (Bailey, 1983) in the sense that all possible assignments are equally likely. Thus Lemma 6, along with Definition 4, leads to the following result after considerable algebra.

THEOREM 1. *Under the* 2^2 *split-plot design qualified by Definition 4, the sampling expectation and covariance matrix of the assignment vector* $\boldsymbol{Z}^*$ *are*

$$E_{\text{S-P}}(\boldsymbol{Z}^*) = N^{-1}\mathbf{1}_{4N}\,, \quad \operatorname{cov}_{\text{S-P}}(\boldsymbol{Z}^*) = \mathbf{C}_{\text{btw}} \otimes \mathbf{P}_{\text{btw}} + \mathbf{C}_{\text{in}} \otimes \mathbf{P}_{\text{in}}$$

where

$$\mathbf{C}_{\text{btw}} = \frac{1}{N(W-1)} \begin{pmatrix} r_A & r_A & -1 & -1 \\ r_A & r_A & -1 & -1 \\ -1 & -1 & r_A^{-1} & r_A^{-1} \\ -1 & -1 & r_A^{-1} & r_A^{-1} \end{pmatrix},$$

$$\mathbf{C}_{\text{in}} = \frac{1}{N(N-W)} \begin{pmatrix} (1+r_A)r_B & -(1+r_A) & 0 & 0 \\ -(1+r_A) & (1+r_A)r_B^{-1} & 0 & 0 \\ 0 & 0 & (1+r_A^{-1})r_B & -(1+r_A^{-1}) \\ 0 & 0 & -(1+r_A^{-1}) & (1+r_A^{-1})r_B^{-1} \end{pmatrix}.$$

and

$$r_A = W_+/W_-, \quad r_B = M_+/M_-$$

are the ratios of factor arm sizes for the whole-plot and sub-plot randomizations, respectively.

To quantify the *effect of restriction* imposed through the split-plot randomization on the covariance matrix of $\boldsymbol{Z}^*$, we examine the relationship among the coefficient matrices $\mathbf{C}_{\text{btw}}$ and $\mathbf{C}_{\text{in}}$ of the split-plot randomization defined in Theorem 1 and the matrix $\mathbf{C}$ of the unrestricted randomization defined in Lemma 6. Straightforward algebra shows that

$$\mathbf{C} = \frac{W-1}{N-1}\mathbf{C}_{\text{btw}} + \frac{N-W}{N-1}\mathbf{C}_{\text{in}}. \tag{3.3}$$

4. Neymanian point estimation for 2^2 factorial effects. Neymanian causal inference (Imbens and Rubin, 2015, chapter 7) involves evaluation of the sampling (or randomization) distributions of estimators of the causal estimands of interest. In this section, we define unbiased point estimators of the three factorial effects τ_F ($F \in \mathcal{F}$), defined in (2.8), and derive their sampling variances. These point estimators are functions of the observed outcomes $Y_i^{\text{obs}} = Y_i(T_i)$, $(i = 1, \ldots, N)$, and the treatment indicators $(T_1, \ldots, T_N)$.

4.1. *Point estimators and their sampling variances.* Let

$$\bar{Y}^{\text{obs}}(k) = N_k^{-1} \textstyle\sum_{i:T_i=k} Y_i^{\text{obs}}$$

be the average observed outcome of treatment k. Substituting the unobservable $\bar{Y}(k)$ by $\bar{Y}^{\text{obs}}(k)$ in the definition of τ_F in (2.8) yields the *Neymanian point estimator* of this population-level factorial effect:

$$\widehat{\tau}_F \;=\; 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}(\bar{Y}^{\mathrm{obs}}(1),\ldots,\bar{Y}^{\mathrm{obs}}(4))^{\mathrm{T}} \;=\; 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}\bar{\boldsymbol{Y}}^{\mathrm{obs}}, \quad F\in\mathcal{F}, \tag{4.1}$$

where $\bar{\boldsymbol{Y}}^{\mathrm{obs}} = \big(\bar{Y}^{\mathrm{obs}}(1),\ldots,\bar{Y}^{\mathrm{obs}}(4)\big)^{\mathrm{T}}$.

Let

$$\widetilde{\mathbf{Y}} = \begin{pmatrix} \boldsymbol{Y}(1) & & & \\ & \boldsymbol{Y}(2) & & \\ & & \boldsymbol{Y}(3) & \\ & & & \boldsymbol{Y}(4) \end{pmatrix}.$$

be the $4N\times 4$ block-diagonal matrix with diagonal vectors $\boldsymbol{Y}(k)$. It follows from

$$\bar{Y}^{\mathrm{obs}}(k) \;=\; N_k^{-1}\sum_{i:T_i=k} Y_i^{\mathrm{obs}} \;=\; N_k^{-1}\sum_{i:T_i=k} Y_i(k) \;=\; \boldsymbol{Y}(k)^{\mathrm{T}}\{N_k^{-1}\boldsymbol{Z}(k)\}\,,$$

that $\bar{\boldsymbol{Y}}^{\mathrm{obs}} = \widetilde{\mathbf{Y}}^T\boldsymbol{Z}^*$, where $\boldsymbol{Z}^*$ is given by (3.2). Substituting this into (4.1) yields

$$\widehat{\tau}_F \;=\; 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}\widetilde{\mathbf{Y}}^{\mathrm{T}}\boldsymbol{Z}^*, \quad F\in\mathcal{F}\,, \tag{4.2}$$

with assignment vector $\boldsymbol{Z}^*$ being the only stochastic component on the right hand side. The randomness in $\widehat{\tau}_F$ under any arbitrary 2^2 factorial assignment mechanism (A-M) originates solely from the randomness in the assignment vector $\boldsymbol{Z}^*$. Thus, the sampling expectation and variance of $\widehat{\tau}_F$ are

$$E_{\text{A-M}}(\widehat{\tau}_F) \;=\; 2^{-1}\boldsymbol{g}_F^{\mathrm{T}}\widetilde{\mathbf{Y}}^{\mathrm{T}}E_{\text{A-M}}(\boldsymbol{Z}^*)\,, \tag{4.3}$$

$$\mathrm{var}_{\text{A-M}}(\widehat{\tau}_F) \;=\; 4^{-1}\boldsymbol{g}_F^{\mathrm{T}}\widetilde{\mathbf{Y}}^{\mathrm{T}}\mathrm{cov}_{\text{A-M}}(\boldsymbol{Z}^*)\widetilde{\mathbf{Y}}\boldsymbol{g}_F\,, \tag{4.4}$$

where $E_{\text{A-M}}$, $\mathrm{var}_{\text{A-M}}$, and $\mathrm{cov}_{\text{A-M}}$ are the expectation, variance, and covariance with respect to the sampling distribution under A-M over all possible assignments. Explicit formulae under completely randomized designs follow immediately from combining (4.3) and (4.4) with Lemma 6, and those under split-plot designs from combining (4.3) and (4.4) with Theorem 1 and are presented as the following two results.

THEOREM 2. *Under the 2^2 completely randomized design characterized by Definition 3, the Neymanian point estimator $\widehat{\tau}_F$ is unbiased for τ_F with sampling variance*

$$\mathrm{var}_{\text{C-R}}(\widehat{\tau}_F) \;=\; 4^{-1}(N-1)\boldsymbol{g}_F^{\mathrm{T}}\left(\mathbf{C}\circ\mathbf{S}\right)\boldsymbol{g}_F, \quad F\in\mathcal{F}\,. \tag{4.5}$$

Here, '$\circ$' denotes the entrywise product, $\mathbf{C}$ is the coefficient matrix defined in Lemma 6, and $\mathbf{S}$ is given by (2.1).

THEOREM 3. *Under the 2^2 split-plot design characterized by Definition 4, the Neymanian point estimator $\widehat{\tau}_F$ is unbiased for τ_F with sampling variance*

$$\begin{aligned}\text{(4.6)}\qquad \mathrm{var}_{\text{S-P}}(\widehat{\tau}_F) = {} & 4^{-1}(W-1)\,\boldsymbol{g}_F^{\mathrm{T}}(\mathbf{C}_{\mathrm{btw}} \circ \mathbf{S}_{\mathrm{btw}})\boldsymbol{g}_F \\ & + 4^{-1}(N-W)\,\boldsymbol{g}_F^{\mathrm{T}}(\mathbf{C}_{\mathrm{in}} \circ \mathbf{S}_{\mathrm{in}})\boldsymbol{g}_F, \quad F \in \mathcal{F}\,,\end{aligned}$$

where $\mathbf{C}_{\mathrm{btw}}$ and $\mathbf{C}_{\mathrm{in}}$ are defined in the statement of Theorem 1 and $\mathbf{S}_{\mathrm{btw}}$ and $\mathbf{S}_{\mathrm{in}}$ are given by (2.5).

4.2. *Comparison of precisions under strict additivity.* Simplified forms of Theorems 2 and 3 are available when the potential outcomes are strictly additive, enabling intuitive comparisons of the estimation precision.

COROLLARY 1. *For strictly additive potential outcomes, the sampling variances of $\widehat{\tau}_A$, $\widehat{\tau}_B$, and $\widehat{\tau}_{AB}$ in Theorem 3 reduce to*

$$\begin{aligned}\text{(4.7)}\qquad \mathrm{var}_{\text{S-P}}\left(\widehat{\tau}_A\right) &= N^{-1}\gamma_A S_{\mathrm{btw}} + (4N)^{-1}\gamma_A(\gamma_B-4)S_{\mathrm{in}}\,, \\ \mathrm{var}_{\text{S-P}}\left(\widehat{\tau}_B\right) &= \mathrm{var}_{\text{S-P}}\left(\widehat{\tau}_{AB}\right) = (4N)^{-1}\gamma_A\gamma_B S_{\mathrm{in}}\,,\end{aligned}$$

where $\gamma_A = r_A + r_A^{-1} + 2$, $\gamma_B = r_B + r_B^{-1} + 2$ with $r_A = W_+/W_-$ and $r_B = M_+/M_-$ as in Theorem 1 and $(S_{\mathrm{btw}}, S_{\mathrm{in}})$ are as defined in Lemma 3.

REMARK 1. We have $\min_{r_A}\gamma_A = 4$ and $\min_{r_B}\gamma_B = 4$. The increasing monotonicity of (4.7) in γ_A and γ_B suggests the three sampling variances are simultaneously minimized when γ_A and γ_B are at their respective minimums:

$$\min_{\gamma_A,\gamma_B} \mathrm{var}_{\text{S-P}}(\widehat{\tau}_A) = \mathrm{var}_{\text{S-P}}(\widehat{\tau}_A)\big|_{\gamma_A=4,\gamma_B=4} = 4S_{\mathrm{btw}}/N,$$

$$\min_{\gamma_A,\gamma_B} \mathrm{var}_{\text{S-P}}(\widehat{\tau}_B)\left(= \min_{\gamma_A,\gamma_B} \mathrm{var}_{\text{S-P}}(\widehat{\tau}_{AB})\right) = \mathrm{var}_{\text{S-P}}(\widehat{\tau}_B)\big|_{\gamma_A=4,\gamma_B=4} = 4S_{\mathrm{in}}/N,$$

where $\gamma_A = 4, \gamma_B = 4$ imply $r_A = r_B = 1$, i.e. the design is balanced. This establishes the optimality of balanced designs under the assumption of strictly additive potential outcomes.

REMARK 2. The sampling variances of $\widehat{\tau}_A$ and $\widehat{\tau}_B$ in (4.7) satisfy

$$\text{(4.8)}\qquad \mathrm{var}_{\text{S-P}}(\widehat{\tau}_A) - \mathrm{var}_{\text{S-P}}(\widehat{\tau}_B) = N^{-1}\gamma_A(S_{\mathrm{btw}} - S_{\mathrm{in}})\,.$$

This suggests more precise Neymanian estimation of the sub-plot factor B than that of the whole-plot factor A if $S_{\mathrm{btw}} - S_{\mathrm{in}} > 0$, and vice versa if $S_{\mathrm{btw}} - S_{\mathrm{in}} < 0$. An intuitive link between the discriminant $S_{\mathrm{btw}} - S_{\mathrm{in}}$ and

the whole-plot heterogeneity can be established from a super-population perspective for potential outcomes generated from linear mixed effects models. Specifically, assume that the study population in question is a random sample from some super-population such that

$$Y_{wm}(k) = \mu(k) + \eta_w + \xi_{wm} \quad (w = 1, \ldots, W; m = 1, \ldots, M) \tag{4.9}$$

follow the linear mixed effects model with fixed treatment effects $\mu(k)$, random whole-plot effects $\eta_w \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2)$, and individual sampling errors $\xi_{wm} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\xi^2)$ jointly independent of η_w. Note that Model (4.9) satisfies the strict additivity condition (1).

Straightforward derivations show that for $k, l = 1, \ldots, 4$,

$$E^*\{S_{\text{btw}}(k,l)\} = \sigma_\xi^2 + M\sigma_\eta^2, \quad E^*\{S_{\text{in}}(k,l)\} = \sigma_\xi^2, \tag{4.10}$$

where E^* denotes expectation with respect to the sampling distribution represented via model (4.9), and $S_{\text{btw}}(k,l)$ and $S_{\text{in}}(k,l)$ are given by (2.4).

Because model (4.9) satisfies the strict additivity condition (Definition 1), by Lemmas 2 and 3, $S_{\text{btw}}(k,l)$ and $S_{\text{in}}(k,l)$ in (4.10) can be replaced by constants S_{btw} and S_{in} respectively for all $k, l = 1, \ldots, 4$. It follows from (4.10) that

$$E^*(S_{\text{btw}} - S_{\text{in}}) = M\sigma_\eta^2 \geq 0\,. \tag{4.11}$$

This, coupled with (4.8), suggests the average sampling variance of the subplot estimator $\widehat{\tau}_B$ is strictly smaller than that of the whole-plot estimator $\widehat{\tau}_A$ unless $\sigma_\eta^2 = 0$, in which case (4.9) degenerates to a simple linear model that admits no random block effects.

Recall that Theorem 3 and Corollary 1 provide results on the decomposition of sampling variances of $\widehat{\tau}_F$ under split-plot designs into the between- and within-WP parts. An analogous result for completely randomized designs follows from substituting (2.6) into (4.5):

$$\begin{aligned}\text{var}_{\text{C-R}}(\widehat{\tau}_F) = {} & 4^{-1}(W-1)\,\boldsymbol{g}_F^{\text{T}}(\mathbf{C} \circ \mathbf{S}_{\text{btw}})\boldsymbol{g}_F \\ & + 4^{-1}(N-W)\,\boldsymbol{g}_F^{\text{T}}(\mathbf{C} \circ \mathbf{S}_{\text{in}})\boldsymbol{g}_F, \quad F \in \mathcal{F}\,.\end{aligned} \tag{4.12}$$

Contrasting (4.12) with Theorem 3 yields Corollary 2.

COROLLARY 2. *For a balanced design with $N_1 = N_2 = N_3 = N_4$ the sampling variance of $\widehat{\tau}_F$ under a 2^2 split-plot design* (S-P) *differs from that under a 2^2 completely randomized design* (C-R) *by*

$$\text{var}_{\text{S-P}}(\widehat{\tau}_F) - \text{var}_{\text{C-R}}(\widehat{\tau}_F) = C_0\,\boldsymbol{g}_F^{\text{T}}\{(\mathbf{C}_{\text{btw}} - \mathbf{C}_{\text{in}}) \circ (\mathbf{S}_{\text{btw}} - \mathbf{S}_{\text{in}})\}\,\boldsymbol{g}_F\,,$$

where C_0 is a positive constant.

Corollary 2 informs us of not only the difference in efficiency of split-plot designs for each $F \in \mathcal{F}$, but also the discrepancy in variance estimation when a split-plot experiment is wrongfully analyzed as a completely randomized one.

COROLLARY 3. *For strictly additive potential outcomes, the sampling variance under 2^2 completely randomized design in (4.12) reduces to*

$$\text{var}_{\text{C-R}}(\widehat{\tau}_F) = \frac{\gamma_A \gamma_B}{4(N-1)} \left(\frac{W-1}{N} S_{\text{btw}} + \frac{M-1}{M} S_{\text{in}} \right), \quad F \in \mathcal{F}. \tag{4.13}$$

COROLLARY 4. *For strictly additive potential outcomes, the differences in Corollary 2 reduce to*

$$\begin{aligned} &\text{var}_{\text{S-P}}(\widehat{\tau}_A) - \text{var}_{\text{C-R}}(\widehat{\tau}_A) = C_1(S_{\text{btw}} - S_{\text{in}}), \\ &\text{var}_{\text{S-P}}(\widehat{\tau}_B) - \text{var}_{\text{C-R}}(\widehat{\tau}_B) = \text{var}_{\text{S-P}}(\widehat{\tau}_{AB}) - \text{var}_{\text{C-R}}(\widehat{\tau}_{AB}) = -C_2(S_{\text{btw}} - S_{\text{in}}), \end{aligned}$$

where C_1 and C_2 are two positive constants.

With the same discriminant $S_{\text{btw}} - S_{\text{in}}$ as that in (4.8), Corollary 4 provides a similar intuition as in Remark 2. Under the super-population model (4.9), it follows from (4.11) and Corollary 4 that

$$E^*\{\text{var}_{\text{S-P}}(\widehat{\tau}_A)\} \geq E^*\{\text{var}_{\text{C-R}}(\widehat{\tau}_A)\}, \quad E^*\{\text{var}_{\text{S-P}}(\widehat{\tau}_B)\} \leq E^*\{\text{var}_{\text{C-R}}(\widehat{\tau}_B)\}.$$

Therefore, if there is a random block effect in the super-population model (4.9), i.e., $\sigma_\eta^2 > 0$, and a balanced split-plot design is used, then the Neymanian inference for $\widehat{\tau}_A$ is less precise and that for $\widehat{\tau}_B$ is more precise compared to their counterparts obtained from a balanced completely randomized design.

5. Estimating the sampling variances. In this section, we consider the problem of estimation of the sampling variances of estimated factorial effects $\hat{\tau}_F$. From the expression of this variance given by (4.6), it appears that its estimation requires estimation of the quantities $S_{\text{btw}}(k,l)$ and $S_{\text{in}}(k,l)$ for $k,l = 1,\ldots,4$. Clearly, $S_{\text{in}}(k,l)$ has to be estimated from observed outcomes *within* whole-plots. Because each sub-plot is exposed to only one treatment, it is not possible to construct such an estimator unless $k = l$.

Turning next to $S_{\text{btw}}(k,l)$, define $\mathcal{W}_k$ as the set of whole-plots that are assigned to the level of the whole-plot factor A in treatment k and hence have some sub-plots assigned to k. Also, let $\mathcal{M}_k^w = \{m : T_{wm} = k\}$ denote the set of sub-plots within whole-plot $w \in \mathcal{W}_k$ that receive treatment k. Define the

average of all observed outcomes within whole-plot w that receive treatment k as

$$\bar{Y}^{\text{obs}}_{w\cdot}(k) = |\mathcal{M}^w_k|^{-1} \sum_{m:T_{wm}=k} Y^{\text{obs}}_{wm}, \tag{5.1}$$

where $|\mathcal{M}^w_k|$ denotes the cardinality of $\mathcal{M}^w_k$, and equals either M_- or M_+ depending on the level of factor B in treatment k. Then, for any k, l, which involve the same level of A, and hence have $\mathcal{W}_k = \mathcal{W}_l$, consider, in the spirit of $S_{\text{btw}}(k,l)$, the estimator

$$s_{\text{btw}}(k,l) = \frac{1}{|\mathcal{W}_k| - 1} \sum_{w \in \mathcal{W}_k} \{\bar{Y}^{\text{obs}}_{w\cdot}(k) - \bar{Y}^{\text{obs}}(k)\}\{\bar{Y}^{\text{obs}}_{w\cdot}(l) - \bar{Y}^{\text{obs}}(l)\}, \tag{5.2}$$

where $|\mathcal{W}_k|$ is the cardinality of $\mathcal{W}_k$, $\bar{Y}^{\text{obs}}_{w\cdot}(k)$ is given by (5.1), and $\bar{Y}^{\text{obs}}(k)$ is the average of $\bar{Y}^{\text{obs}}_{w\cdot}(k)$ over $w \in \mathcal{W}_k$. Note that $s_{\text{btw}}(k,l)$ is defined only for pairs (k,l) that belong to the set

$$\{(1,1),(1,2),(2,1),(2,2),(3,3),(3,4),(4,3),(4,4)\}.$$

LEMMA 7. *Under the 2^2 split-plot design characterized by Definition 4, the sampling expectations of $s_{\text{btw}}(k,l)$ satisfy*

$$\begin{aligned}
E\begin{pmatrix} s_{\text{btw}}(1,1) & s_{\text{btw}}(1,2) \\ s_{\text{btw}}(2,1) & s_{\text{btw}}(2,2) \end{pmatrix} &= M^{-1}\begin{pmatrix} S_{\text{btw}}(1,1) & S_{\text{btw}}(1,2) \\ S_{\text{btw}}(2,1) & S_{\text{btw}}(2,2) \end{pmatrix} \\
&\quad + M^{-1}\begin{pmatrix} r_B & -1 \\ -1 & r_B^{-1} \end{pmatrix} \circ \begin{pmatrix} S_{\text{in}}(1,1) & S_{\text{in}}(1,2) \\ S_{\text{in}}(2,1) & S_{\text{in}}(2,2) \end{pmatrix}, \\
E\begin{pmatrix} s_{\text{btw}}(3,3) & s_{\text{btw}}(3,4) \\ s_{\text{btw}}(4,3) & s_{\text{btw}}(4,4) \end{pmatrix} &= M^{-1}\begin{pmatrix} S_{\text{btw}}(3,3) & S_{\text{btw}}(3,4) \\ S_{\text{btw}}(4,3) & S_{\text{btw}}(4,4) \end{pmatrix} \\
&\quad + M^{-1}\begin{pmatrix} r_B & -1 \\ -1 & r_B^{-1} \end{pmatrix} \circ \begin{pmatrix} S_{\text{in}}(3,3) & S_{\text{in}}(3,4) \\ S_{\text{in}}(4,3) & S_{\text{in}}(4,4) \end{pmatrix}.
\end{aligned}$$

Lemma 7 shows that the sampling expectations of $s_{\text{btw}}(k,l)$ involve both $S_{\text{btw}}(k,l)$ and $S_{\text{in}}(k,l)$. This renders them 'self-adequate' for estimating the $\text{var}_{\text{S-P}}(\widehat{\tau}_F)$ in (4.6).

THEOREM 4. *Under the 2^2 split-plot design characterized by Definition 4, the sampling variance of $\widehat{\tau}_F$ can be conservatively estimated by*

$$\widehat{v}_F = 4^{-1}\boldsymbol{g}_F^{\mathrm{T}} \begin{pmatrix} W_-^{-1}\begin{pmatrix} s_{\text{btw}}(1,1) & s_{\text{btw}}(1,2) \\ s_{\text{btw}}(2,1) & s_{\text{btw}}(2,2) \end{pmatrix} & \mathbf{0} \\ \mathbf{0} & W_+^{-1}\begin{pmatrix} s_{\text{btw}}(3,3) & s_{\text{btw}}(3,4) \\ s_{\text{btw}}(4,3) & s_{\text{btw}}(4,4) \end{pmatrix} \end{pmatrix} \boldsymbol{g}_F,$$

in the sense that

$$\mathrm{var}_{\text{S-P}}(\widehat{\tau}_F) - E_{\text{S-P}}(\widehat{v}_F) \;=\; -N^{-1}S_{\text{btw}}(F) \;\leq\; 0\,.$$

The last inequality is strict unless the whole-plot average factorial effects $\tau_{w\cdot}(F)$ *are constant across all* $w = 1, \ldots, W$*, i.e.,* $S_{btw}(F) = 0$*.*

The estimator of $\mathrm{var}_{\text{S-P}}(\widehat{\tau}_F)$ proposed in Theorem 4 can be used for Neymanian interval estimation. Asymptotic coverage of such a procedure can be studied by application of the finite population central limit theorem (Hajek, 1960; Li and Ding, 2017) and is left for future research. Some empirical examination of coverage will be done via simulations later.

6. Randomization-based versus model-based inference . We now discuss some of the key features that set this randomization-based approach apart from existing model-based alternatives. Recall $\boldsymbol{g}_A = (-1, -1, +1, +1)$, $\boldsymbol{g}_B = (-1, +1, -1, +1)$, and $\boldsymbol{g}_{AB} = \boldsymbol{g}_A \circ \boldsymbol{g}_B$, such that the kth entry in $\boldsymbol{g}_F$ equals the level of factor F in treatment k. Let $\mathbf{D} = 2^{-1}(\mathbf{1}_4, \boldsymbol{g}_A, \boldsymbol{g}_B, \boldsymbol{g}_{AB})$ be the orthonormal design matrix, and let $g_F(k)$ be the kth entry in $\boldsymbol{g}_F$. Denoting the vector of potential outcomes for unit wm by $\boldsymbol{Y}_{wm}$, we have by (2.7) and the orthonormality of $\mathbf{D}$ that

$$\begin{aligned}
\boldsymbol{Y}_{wm} &= \mathbf{D}\mathbf{D}^{\mathrm{T}}\boldsymbol{Y}_{wm} \;=\; \mathbf{D}\left\{2^{-1}(\mathbf{1}_4, \boldsymbol{g}_A, \boldsymbol{g}_B, \boldsymbol{g}_{AB})^{\mathrm{T}}\,\boldsymbol{Y}_{wm}\right\} \\
&= \mathbf{D}\left(2^{-1}\mathbf{1}_4^{\mathrm{T}}\boldsymbol{Y}_{wm}, 2^{-1}\boldsymbol{g}_A^{\mathrm{T}}\boldsymbol{Y}_{wm}, 2^{-1}\boldsymbol{g}_B^{\mathrm{T}}\boldsymbol{Y}_{wm}, 2^{-1}\boldsymbol{g}_{AB}^{\mathrm{T}}\boldsymbol{Y}_{wm}\right)^{\mathrm{T}} \\
&= \mathbf{D}\left(2\mu_{wm}, \tau_{wm}(A), \tau_{wm}(B), \tau_{wm}(AB)\right)^{\mathrm{T}},
\end{aligned} \tag{6.1}$$

where $\mu_{wm} = 4^{-1}\sum_{k=1}^{4} Y_{wm}(k)$ is the average of all potential outcomes for unit wm. Denoting the kth entry of $\boldsymbol{g}_F$ by $g_F(k)$, we have that

$$\begin{aligned}
Y_{wm}(k) &= 2^{-1}\left(1, g_A(k), g_B(k), g_{AB}(k)\right)\left(2\mu_{wm}, \tau_{wm}(A), \tau_{wm}(B), \tau_{wm}(AB)\right)^{\mathrm{T}} \\
&= \mu_{wm} + \sum_{F\in\mathcal{F}} 2^{-1} g_F(k)\tau_{wm}(F).
\end{aligned} \tag{6.2}$$

Averaging (6.2) over all w and m yields

$$\bar{Y}(k) = \mu + \sum_{F\in\mathcal{F}} 2^{-1} g_F(k)\tau_F, \tag{6.3}$$

where μ is the average of all $4N$ potential outcomes.

Recall that the treatment indicator T_{wm} equals k if sub-plot wm is assigned to treatment k, and that the observed outcome for unit wm is $Y_{wm}^{\text{obs}} =$

$Y_{wm}(T_{wm})$. The *derived linear model* (Hinkelmann and Kempthorne, 2008) treats the population average $\bar{Y}(T_{wm})$ as the part in $Y_{wm}(T_{wm})$ explainable by the treatment, and decomposes the observed outcomes as

$$Y_{wm}^{\text{obs}} = Y_{wm}(T_{wm}) = \bar{Y}(T_{wm}) + \epsilon_{wm} = \mu + \sum_{F\in\mathcal{F}} 2^{-1} g_F(T_{wm})\tau_F + \epsilon_{wm}\,, \tag{6.4}$$

where $\epsilon_{wm} = Y_{wm}^{\text{obs}} - \bar{Y}(T_{wm})$ are the unit-level random errors, and the last equality follows from letting $k = T_{wm}$ in (6.3). Let

$$\delta_{wm}(\mu) = \mu_{wm} - \mu\,, \quad \delta_{wm}(F) = \tau_{wm}(F) - \tau_F, \quad F \in \mathcal{F}$$

be the deviations of unit-level parameters from the *finite-population* averages. Substitution of (6.2), with $k = T_{wm}$, into (6.4) yields

$$\epsilon_{wm} = \delta_{wm}(\mu) + \sum_{F\in\mathcal{F}} 2^{-1} g_F(T_{wm})\delta_{wm}(F)\,. \tag{6.5}$$

The term $g_F(T_{wm})$ in (6.4) has the straightforward interpretation as the level of factor F received by sub-plot wm. Such an interpretation, together with the functional form of (6.4), reminds us of the family of additive regression models:

$$Y_{wm}^{\text{obs}} = \beta_0 + \sum_{F\in\mathcal{F}} g_F(T_{wm})\beta_F + \epsilon_{wm}^{\text{model}}\,. \tag{6.6}$$

Despite the apparent resemblance between (6.4) and (6.6), however, their difference is fundamental, with the source of randomness being the first and foremost.

The family of additive regression models (6.6), on one hand, conditions on the treatment assignments T_{wm} for all its inference, and attributes the randomness in Y_{wm}^{obs} to the population under study being a *random sample* of some *hypothetical* super-population, reflected via $\epsilon_{wm}^{\text{model}}$ as the individual sampling errors. The regression coefficients β_F are treated as super-population causal parameters, and $\beta_0 + \sum_{F\in\mathcal{F}} g_F(T_{wm})\beta_F$ as deterministic super-population means.

The derived linear model (6.4), on the other hand, conditions on the composition of the finite population under study for all its inference, and attributes the randomness in Y_{wm}^{obs} solely to the *random assignment of treatments*, reflected via the joint distribution of treatment assignment variables T_{wm}. As a result, not only the residuals ϵ_{wm}, but also the linear combinations $\mu + \sum_{F\in\mathcal{F}} 2^{-1} g_F(T_{wm})\tau_F$ are now stochastic via their dependence on T_{wm}

(Freedman, 2008a,b,c; Lin, 2013), with coefficients τ_F, by definition (2.8), describing the finite population. See formula (6.5) for a full specification of ϵ_{wm} in terms of $g_F(T_{wm})$.

More quantitative comparison follows from the difference in residual covariance structure. Whereas the covariances of the $\epsilon_{wm}^{\text{model}}$ in (6.6) are in general specified as model assumptions, those of the ϵ_{wm} in (6.4) follow naturally from identity (6.5) and the joint distribution of T_{wm} as determined by the treatment assignment mechanism.

To start with, viewing (6.5) in conjunction with Lemma 4 renders the computation of $\text{cov}_{\text{S-P}}(\epsilon_{wm}, \epsilon_{w'm'})$ almost trivial under strict additivity: With $\delta_{wm}(F) = 0$ for all wm and $F \in \mathcal{F}$, the residuals in (6.5) reduce to constants $\epsilon_{wm} = \delta_{wm}(\mu)$, and the covariance of constants is always zero, i.e., $\text{cov}_{\text{S-P}}(\epsilon_{wm}, \epsilon_{w'm'}) = 0$ for all wm and $w'm'$ under strict additivity.

Without strict additivity, the algebra becomes tedious. To avoid unnecessary complexity, we report an asymptotic result in Theorem 5. In this theorem $\delta_{wm}(F) = \tau_{wm}(F) - \tau_F$ as before, with τ_F interpreted as the limiting average of the experimental unit wise effects of factor F, as W and M approach infinity.

THEOREM 5. *Let W and M approach infinity such that $r_A = W_+/W_{-1}$ and $r_B = M_+/M_{-1}$ converge to positive constants ρ_A and ρ_B respectively. Let $e_A = (\rho_A - 1)/(\rho_A + 1)$ and $e_B = (\rho_B - 1)/(\rho_B + 1)$. Then, the residual covariance $\text{cov}_{\text{S-P}_{(W,M,r_A,r_B)}}(\epsilon_{wm}, \epsilon_{w'm'})$ for sub-plots wm and $w'm'$ converges to*

(i)

$$\frac{\rho_A}{(\rho_A+1)^2}\left(\delta_{wm}(A), \delta_{wm}(AB)\right)\begin{pmatrix}1 & e_B\\ e_B & e_B^2\end{pmatrix}\begin{pmatrix}\delta_{w'm'}(A)\\ \delta_{w'm'}(AB)\end{pmatrix}$$
$$+\quad \frac{\rho_B}{(\rho_B+1)^2}\left(\delta_{wm}(B), \delta_{wm}(AB)\right)\begin{pmatrix}1 & e_A\\ e_A & 1\end{pmatrix}\begin{pmatrix}\delta_{w'm'}(B)\\ \delta_{w'm'}(AB)\end{pmatrix},$$

if $(w, m) = (w', m')$;

(ii)

$$\frac{\rho_A}{(\rho_A+1)^2}\left\{\delta_{wm}(A) + e_B\delta_{wm}(AB)\right\}\left\{\delta_{w'm'}(A) + e_B\delta_{w'm'}(AB)\right\}$$

if $w = w'$ but $m \neq m'$; and

(iii) zero if $w \neq w'$.

REMARK 3. Theorem 5 involves *finite-population asymptotics* (Li and Ding, 2017; Hajek, 1960). The asymptotic condition '$W, M \to \infty$' can be

visualized as adding till infinity new whole-plots to the current study population, and new sub-plots to the current whole-plots. The covariance at each finite (W, M) is computed under the split-plot design characterized by Definition 4.

COROLLARY 5. *When the sequence of split-plot designs is asymptotically balanced, i.e., $\rho_A = \rho_B = 1$, the asymptotic residual covariance in Theorem 5 reduces to $4^{-1}\delta_{wm}(A)\delta_{w'm'}(A)$ for sub-plots wm and $w'm'$ in the same whole-plot.*

Theorem 5 and Corollary 5 provide an explicit account of the non-vanishing within-whole-plot correlation of ϵ_{wm} under 2^2 split-plot designs, and thereby justify heuristically the block-diagonal covariance structure that a linear mixed effects (LME) model assumes for its sampling errors. With $\epsilon^{\text{LME}}_{wm} = \eta_w + \xi_{wm}$ where $\eta_w \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\eta)$ and $\xi_{wm} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_\xi)$ are jointly independent, the covariance of $\epsilon^{\text{LME}}_{wm}$ and $\epsilon^{\text{LME}}_{w'm'}$ equals σ^2_η if $w = w'$, and 0 otherwise. To summarize the findings of this section, we note that despite the similarity in structure between the LME model (6.6) and the derived linear model (6.4), there are two major differences. First, whereas the linear mixed effects model assumes equal covariances for all pairs of residuals from the same whole-plot, those under the derived linear model, as is clear from Theorem 5 and Corollary 5, vary from pair to pair even in the asymptotics. Second, whereas the linear mixed effects model assumes independence between whole-plots at any finite (W, M), formula (6.5) suggests otherwise for the derived model.

7. Extension to a general split-plot assignment. We now consider a general factorial experiment with m_1 whole-plot factors $F_{11}, \ldots, F_{1m_1}$ (whose levels are difficult to change from unit to unit) and m_2 sub-plot factors $F_{21}, \ldots, F_{2m_2}$ for which such restrictions do not apply. Assume that each of the $m_1 + m_2$ factors has two or more levels. Let Z_k denote the set of level combinations of $F_{k1}, \ldots, F_{km_k}$ $(k = 1, 2)$, and $z = z_1 z_2$ $(z_k \in Z_k)$ denote a treatment combination. Let $Y_i(z_1 z_2)$ denote the potential outcome of unit i if exposed to treatment combination $z_1 z_2$. A typical unit-level treatment contrast for unit i is of the form

$$\tau_i = \sum_{z_1 \in Z_1} \sum_{z_2 \in Z_2} g(z_1 z_2) Y_i(z_1 z_2), \tag{7.1}$$

where $g(z_1 z_2), z_1 \in Z_1, z_2 \in Z_2$, are known, not all zeros, and sum to zero. The mean τ of the unit-level contrasts defines a treatment contrast for the

finite population and is a typical estimand of interest. From (7.1),

$$\tau = N^{-1}\sum_{i=1}^{N}\tau_i = \sum_{z_1\in Z_1}\sum_{z_2\in Z_2} g(z_1z_2)\bar{Y}(z_1z_2), \tag{7.2}$$

where $\bar{Y}(z_1z_2) = N^{-1}\sum_{i=1}^{N} Y_i(z_1z_2)$ is the average potential outcome under z_1z_2. It is important to note here that the unit-level contrasts τ_i and population-level contrast τ implicitly depend on the contrast coefficients $g(z_1z_2)$.

As before, let $N = WM$, where $W, M \geq 2$, and suppose the N experimental units are grouped into W whole-plots $\Omega_1, \ldots, \Omega_W$, each consisting of M sub-plots. Along the lines of the decomposition in Section 2.2, we define the following quantities for any $z_1, z_1^* \in Z_1$ and $z_2, z_2^* \in Z_2$:

$$\bar{Y}_w(z_1z_2) = M^{-1}\sum_{i\in\Omega_w} Y_i(z_1z_2),\ (w = 1, \ldots, W), \tag{7.3}$$

$$S_{\text{btw}}(z_1z_2, z_1^*z_2^*) = \frac{M}{W-1}\sum_{w=1}^{W}\left\{\bar{Y}_w(z_1z_2) - \bar{Y}(z_1z_2)\right\}\left\{\bar{Y}_w(z_1^*z_2^*) - \bar{Y}(z_1^*z_2^*)\right\}, \tag{7.4}$$

$$S_{\text{in}}(z_1z_2, z_1^*z_2^*) = \frac{\sum_{w=1}^{W}\sum_{i\in\Omega_w}\left\{Y_i(z_1z_2) - \bar{Y}_w(z_1z_2)\right\}\left\{Y_i(z_1^*z_2^*) - \bar{Y}_w(z_1^*z_2^*)\right\}}{W(M-1)}. \tag{7.5}$$

Expressions (7.4) and (7.5) represent, respectively, the between and within whole-plot mean squares or products in an analysis of variance/covariance decomposition of $\sum_{w=1}^{W}\sum_{i\in\Omega_w}\left\{Y_i(z_1z_2) - \bar{Y}(z_1z_2)\right\}\left\{Y_i(z_1^*z_2^*) - \bar{Y}(z_1^*z_2^*)\right\}$.

7.1. *Assignment mechanism and observed outcomes.* Consider a two-stage randomization similar to that discussed in Section 3, which assigns $N_1(z_1)$ whole-plots to level combination z_1 of $F_{11}, \ldots, F_{1m_1}$ and then, within each whole-plot, assigns $N_2(z_2)$ sub-plots to level combination z_2 of $F_{21}, \ldots, F_{2m_2}$. All assignments at each stage are equiprobable, the fixed positive integers $N_1(z_1)$, $z_1 \in Z_1$, sum to W, and the fixed positive integers $N_2(z_2)$, $z_2 \in Z_2$, sum to M.

Let $T_1(z_1)$ denote the set of indices w such that the whole-plot Ω_w is assigned to level combination z_1 of $F_{11}, \ldots, F_{1m_1}$ and $T_{w2}(z_2)$ the set of sub-plots in Ω_w that are assigned to level combination z_2 of $F_{21}, \ldots, F_{2m_2}$. Then the set of units assigned to any treatment combination z_1z_2 is

$$T(z_1z_2) = \bigcup_{w\in T_1(z_1)} T_{w2}(z_2). \tag{7.6}$$

Thus the observed outcome Y_i^{obs} for unit i equals $Y_i(z_1z_2)$ where $z_1 \in Z_1$ and $z_2 \in Z_2$ are such that the set $T(z_1z_2)$ contains index i.

7.2. *Neymanian inference of treatment contrasts.* We now explore Neymanian inference of treatment contrasts like τ in (7.2) on the basis of the observed outcomes defined in Section 7.1 . Our results will pertain to any such contrast, and hence cover, in particular, the factorial main effect and interaction contrasts that are of special interest in the present setup.

For any z_1z_2, let $\bar{Y}^{\text{obs}}(z_1z_2) = \{N_1(z_1)N_2(z_2)\}^{-1} \sum_{i \in T(z_1z_2)} Y_i(z_1z_2)$ be the mean of the observed outcomes for treatment combination z_1z_2. By (7.6),

$$\bar{Y}^{\text{obs}}(z_1z_2) = \{N_1(z_1)\}^{-1} \sum_{w \in T_1(z_1)} \bar{Y}_w^{\text{obs}}(z_1z_2), \tag{7.7}$$

where

$$\bar{Y}_w^{\text{obs}}(z_1z_2) = \{N_2(z_2)\}^{-1} \sum_{i \in T_{w2}(z_2)} Y_i(z_1z_2). \tag{7.8}$$

PROPOSITION 1. (a) *For every* z_1z_2, $E\left\{\bar{Y}^{\text{obs}}(z_1z_2)\right\} = \bar{Y}(z_1z_2)$.
(b) *For every* z_1z_2 *and* $z_1^*z_2^*$,

$$\begin{aligned}
& \operatorname{cov}\left\{\bar{Y}^{\text{obs}}(z_1z_2), \bar{Y}^{\text{obs}}(z_1^*z_2^*)\right\} \\
= \; & I\{z_1 = z_1^*\}\{MN_1(z_1)\}^{-1}\{S_{\text{btw}}(z_1z_2, z_1^*z_2^*) - S_{\text{in}}(z_1z_2, z_1^*z_2^*)\} \\
+ \; & I\{z_1 = z_1^*\}I\{z_2 = z_2^*\}\{N_1(z_1)N_2(z_2)\}^{-1} S_{\text{in}}(z_1z_2, z_1^*z_2^*) \\
- \; & (WM)^{-1} S_{\text{btw}}(z_1z_2, z_1^*z_2^*).
\end{aligned}$$

Part (a) of Proposition 1 readily yields an unbiased estimator of a treatment contrast τ defined in (7.2). The sampling variance of such an unbiased estimator follows from part (b). Theorem 6 summarizes these results.

THEOREM 6. (a) *An unbiased estimator of* τ *is given by*

$$\widehat{\tau} = \sum_{z_1 \in Z_1} \sum_{z_2 \in Z_2} g(z_1z_2)\bar{Y}^{\text{obs}}(z_1z_2).$$

(b) *The sampling variance of* $\widehat{\tau}$ *defined in* (a) *is given by*

$$\begin{aligned}
\operatorname{var}_{\text{S-P}}(\widehat{\tau}) \; = \; & \sum_{z_1 \in Z_1} \sum_{z_2 \in Z_2} \sum_{z_2^* \in Z_2} \frac{g(z_1z_2)g(z_1z_2^*)\{S_{\text{btw}}(z_1z_2, z_1z_2^*) - S_{\text{in}}(z_1z_2, z_1z_2^*)\}}{MN_1(z_1)} \\
+ \; & \sum_{z_1 \in Z_1} \sum_{z_2 \in Z_2} \frac{\{g(z_1z_2)\}^2 S_{\text{in}}(z_1z_2, z_1z_2)}{N_1(z_1)N_2(z_2)} - \frac{\sum_{w=1}^{W}(\bar{\tau}_w - \tau)^2}{W(W-1)},
\end{aligned}$$

where

$$\bar{\tau}_w = M^{-1}\sum_{i\in\Omega_w}\tau_i = \sum_{z_1\in Z_1}\sum_{z_2\in Z_2} g(z_1z_2)\bar{Y}_w(z_1z_2),\ w=1,\ldots,W.$$

We now consider the estimation of $\mathrm{var}_{\text{S-P}}(\widehat{\tau})$. Assume that $N_1(z_1)\geq 2$ for each $z_1\in Z_1$, and let

$$\widehat{v}(\widehat{\tau}) = \sum_{z_1\in Z_1}\sum_{z_2\in Z_2}\sum_{z_2^*\in Z_2}\frac{g(z_1z_2)g(z_1z_2^*)s\left(z_1z_2,z_1z_2^*\right)}{N_1(z_1)}, \tag{7.9}$$

where

$$s\left(z_1z_2,z_1z_2^*\right) = \sum_{w\in T_1(z_1)}\frac{\left\{\bar{Y}_w^{\text{obs}}(z_1z_2)-\bar{Y}^{\text{obs}}(z_1z_2)\right\}\left\{\bar{Y}_w^{\text{obs}}(z_1z_2^*)-\bar{Y}^{\text{obs}}(z_1z_2^*)\right\}}{N_1(z_1)-1}. \tag{7.10}$$

Comparing with (7.4), we note that $s\left(z_1z_2,z_1z_2^*\right)$ is a sample counterpart of $S_{\text{btw}}\left(z_1z_2,z_1z_2^*\right)$ without the multiplier M. The following result justifies using $\widehat{v}(\widehat{\tau})$ defined in (7.9) as an estimator of $\mathrm{var}_{\text{S-P}}(\widehat{\tau})$.

THEOREM 7. (a) $E\left\{\widehat{v}(\widehat{\tau})\right\}\geq \mathrm{var}_{\text{S-P}}(\widehat{\tau})$.

(b) *Equality holds in* (a) *above for every treatment contrast* τ *if and only if between-WP additivity holds in the sense of Definition 2, i.e., constancy of* $\bar{Y}_w(z_1z_2)-\bar{Y}_w(z_1^*z_2^*)$ *over* $w=1,\ldots,W$, *for every pair of treatment combinations* z_1z_2 *and* $z_1^*z_2^*$.

REMARK 4. The results stated in Proposition 1, Theorem 6 and Theorem 7 have the following implications:

(a) Theorem 7 establishes that $\widehat{v}(\widehat{\tau})$ is a conservative estimator of $\mathrm{var}_{\text{S-P}}(\widehat{\tau})$, potentially leading to overestimation on an average, and $\widehat{v}(\widehat{\tau})$ becomes unbiased for $\mathrm{var}_{\text{S-P}}(\widehat{\tau})$ for every treatment contrast τ if and only if between-WP additivity holds. It is satisfying to note that $\widehat{v}(\widehat{\tau})$ is nonnegative because, by (7.9) and (7.10), it can be expressed as:

$$\widehat{v}(\widehat{\tau}) = \sum_{z_1\in Z_1}\sum_{w\in T_1(z_1)}\frac{\left[\sum_{z_2\in Z_2} g(z_1z_2)\left\{\bar{Y}_w^{\text{obs}}(z_1z_2)-\bar{Y}^{\text{obs}}(z_1z_2)\right\}\right]^2}{N_1(z_1)\left\{N_1(z_1)-1\right\}}.$$

(b) From Proposition 1(b), which leads to the expression for $\mathrm{var}_{\text{S-P}}(\widehat{\tau})$ in Theorem 6(b), one can check that this variance formula is in agreement with its counterpart in Theorem 3 for the 2^2 factorial. Moreover, from (7.9) and (7.10), it can be seen that the expression $\widehat{v}(\widehat{\tau})$ matches the expression for $\widehat{v}_F$ defined in Theorem 4 for the 2^2 factorial.

8. Simulations. We evaluate in this section, via simulation, the frequency coverage property of the Neymanian split-plot interval estimators indicated at the end of Section 5. For ease of presentation, this is done with reference to the 2^2 factorial.

8.1. *Generative models for potential outcome matrices (*POM*s).* In what follows, the potential outcomes are said to be *without WP effect* if they arise from a generative model such that the distribution of $Y_{wm}(k)$ depends possibly on m and k but not w; else, they are *with WP effect*. As an extreme case of the latter, the potential outcomes are said to have *ultimate WP effect* if $Y_{wm}(k)$ equals $\bar{Y}_{w\cdot}(k)$ with probability 1, for all w, m, and k. Clearly, ultimate WP effect implies within-WP additivity, and hence strict additivity, if, in addition, between-WP additivity holds. We consider here five types of potential outcomes:

(I) binary potential outcomes without WP effect,
(II) binary potential outcomes with ultimate WP effect,
(III) continuous potential outcomes without WP effect,
(IV) continuous potential outcomes with WP effect,
(V) continuous potential outcomes with ultimate WP effect

in combination with three types of additivity assumption:

(i) strict additivity,
(ii) between-WP additivity,
(iii) no assumption about additivity.

This gives a total of $5 \times 3 = 15$ types of POM, from which specific POMs are generated in two steps:

1. Generate $\boldsymbol{Y}(1)$ according to the designated potential outcomes type. See Table 1 for details about the generative models.
2. Conditional on $\boldsymbol{Y}(1)$, generate $\boldsymbol{Y}(k)$ $(k = 2, 3, 4)$ according to the designated additivity type. See Table 2 for details about the generative models.

Strict additivity for all five potential outcomes types is imposed by letting $\boldsymbol{Y}(k) = \boldsymbol{Y}(1)$ $(k = 2, 3, 4)$, and between-WP additivity by letting

$$\bar{Y}_{w\cdot}(k) = \bar{Y}_{w\cdot}(1) \quad (k = 2, 3, 4; w = 1, \ldots, W), \tag{8.1}$$

such that the resulting POMs satisfy Definitions 1 and 2 respectively with all differential constants being zero. No generality is lost so far as the coverage rate is concerned.

TABLE 1
Generative models for $\boldsymbol{Y}(1)$ under potential outcomes (PO) types (I)–(V).

PO Type	Generative Model for $\boldsymbol{Y}(1) = (Y_{11}(1), \ldots, Y_{WM}(1))^T$
(I)	$Y_{wm}(1) \overset{\text{iid}}{\sim} \text{Bern}(0.5)$.
(II)	$\bar{Y}_{w\cdot}(1) \overset{\text{iid}}{\sim} \text{Bern}(0.5)$, and $Y_{wm}(1) = \bar{Y}_{w\cdot}(1)$.
(III)	$Y_{wm}(1)$ are independent normals with means $\mu_{wm} = 2(-1)^{I\{m \le M/2\}}$ and variances $(\sigma^2_{11}, \ldots, \sigma^2_{WM})$ being a random permutation of $2(\mathbf{1}^{\mathrm{T}}_{N/2}, \mathbf{0}^{\mathrm{T}}_{N/2})$. This makes half of the potential outcomes constant.
(IV)	$Y_{wm}(1) = \eta_w + \epsilon_{wm}$, where η_w and ϵ_{wm} are iid standard normals.
(V)	$\bar{Y}_{w\cdot}(1) \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, and $Y_{wm}(1) = \bar{Y}_{w\cdot}(1)$.

8.2. *Interval estimates and their coverage rates.* For each realized POM, coverage rates of the Neymanian split-plot (S-P) interval estimators, as indicated at the end of Section 5, are summarized over 1,000 independent split-plot randomizations and compared to those of the following three alternatives:

- GLM interval estimators.

  The $100(1-\alpha)\%$ confidence intervals under the standard generalized linear model (GLM) with the levels of factors A and B and their interaction as explanatory variables.

- GLME interval estimators.

  The $100(1-\alpha)\%$ confidence intervals under the standard generalized linear mixed effects model (GLME) that includes also whole-plot dummy, in addition to the levels of factors A and B and their interaction, as explanatory variable.

- C-R interval estimators.

  The $100(1-\alpha)\%$ Neymanian interval estimators for 2^2 completely randomized (C-R) design discussed by Dasgupta, Pillai and Rubin (2015).

All GLMs are fitted by the standard R function 'glm,' and all GLMEs by 'glmer,' both with 'binomial' link for binary potential outcomes types (I)–(II) and 'identity' link for continuous potential outcomes types (III)–(V). We abbreviate 'GLM' to 'LM,' and 'GLME' to 'LME' in the latter case as the identity link reduces the two generalized models to linear and linear mixed effects models, respectively.

TABLE 2

Generative models for $\boldsymbol{Y}(k)$ $(k = 2, 3, 4)$ under the 15 POM types as combinations of the five potential outcomes (PO) types (I)–(V) in Table 1, and the three additivity (ADT) types: (i) strict, (ii) between-WP , and (iii) no assumption about additivity.

ADT Type	PO Type	Generative Model for $\boldsymbol{Y}(k)$ $(k = 2, 3, 4)$
(i)	(I) – (V)	$\boldsymbol{Y}(k) = \boldsymbol{Y}(1)$.
(ii)	(I)	$\boldsymbol{Y}(k)$ are independent WP-wise permutations of $\boldsymbol{Y}(1)$, such that the numbers of 1's within each WP are the same for $\boldsymbol{Y}(k)$ and $\boldsymbol{Y}(1)$. This ensures (8.1).
	(II), (V)	$\boldsymbol{Y}(k) = \boldsymbol{Y}(1)$. Under ultimate WP effect, we have $Y_{wm}(1) = \bar{Y}_{w\cdot}(1)$ and $Y_{wm}(k) = \bar{Y}_{w\cdot}(k)$; (8.1) holds if and only if $Y_{wm}(k) = Y_{wm}(1)$.
	(III), (IV)	$Y_{wm}(k) = Y'_{wm}(k) - \{\bar{Y}'_{w\cdot}(k) - \bar{Y}_{w\cdot}(1)\}$, where $\boldsymbol{Y}'(k)$ are iid as $\boldsymbol{Y}(1)$. Subtracting $\bar{Y}'_{w\cdot}(k) - \bar{Y}_{w\cdot}(1)$ ensures (8.1).
(iii)	(I) – (V)	$\boldsymbol{Y}(k)$ are iid as $\boldsymbol{Y}(1)$.

8.3. *Results.* We realize each of the 15 POM types at two sizes: $(W, M) = (40, 40)$ and $(80, 80)$, and construct the intervals at confidence level $1 - \alpha = 0.95$. Results for the 15 POMs at $(W, M) = (40, 40)$ are shown in Figure 1; the overall superiority of S-P interval is evident. Results at $(W, M) = (80, 80)$ exhibit quite similar patterns, and are thus not included here to avoid redundancy.

The intended 'approximate exact-coverage under between-WP or strict additivity and over-coverage if otherwise' is fulfilled by the S-P interval for all but potential outcomes types (II) and (V) under strict additivity. Despite its undue conservativeness towards τ_B and τ_{AB} in these two cases, the S-P interval remains to be the only interval that 'does not under-cover' — see Table 3 for the untruncated statistics regarding the severe under-coverage of τ_A by LM and LME intervals. The fact that $\widehat{\tau}_B$ and $\widehat{\tau}_{AB}$ in these two cases are virtually constant at their respective true values τ_B and τ_{AB} over all possible assignments, as a result of the ultimate WP effect, may render even S-P's undue conservativeness excusable.

For potential outcomes type (IV) in particular, S-P markedly outperforms LM (C-R) in all three factorial effects, matches LME in the main effect of

whole-plot factor A, and beats the latter in all other cases. The fact of potential outcomes type (IV) being actually generated from LME model accentuates S-P's victory even further.

The general inadequacy of C-R, LM, and GLM intervals for potential outcomes types (II), (IV), and (V), on the other hand, exemplifies the possible severe under-coverage when split-plot experiments are wrongfully analyzed as completely randomized ones, even when the preferred randomization-based perspective is adopted.

An additional set of simulations was conducted with binary potential outcomes with WP effect, and the outcomes were found similar to those reported in Table 3 and Figure 1. Details are given in the supplemental article (Zhao, Ding and Dasgupta, 2017).

TABLE 3

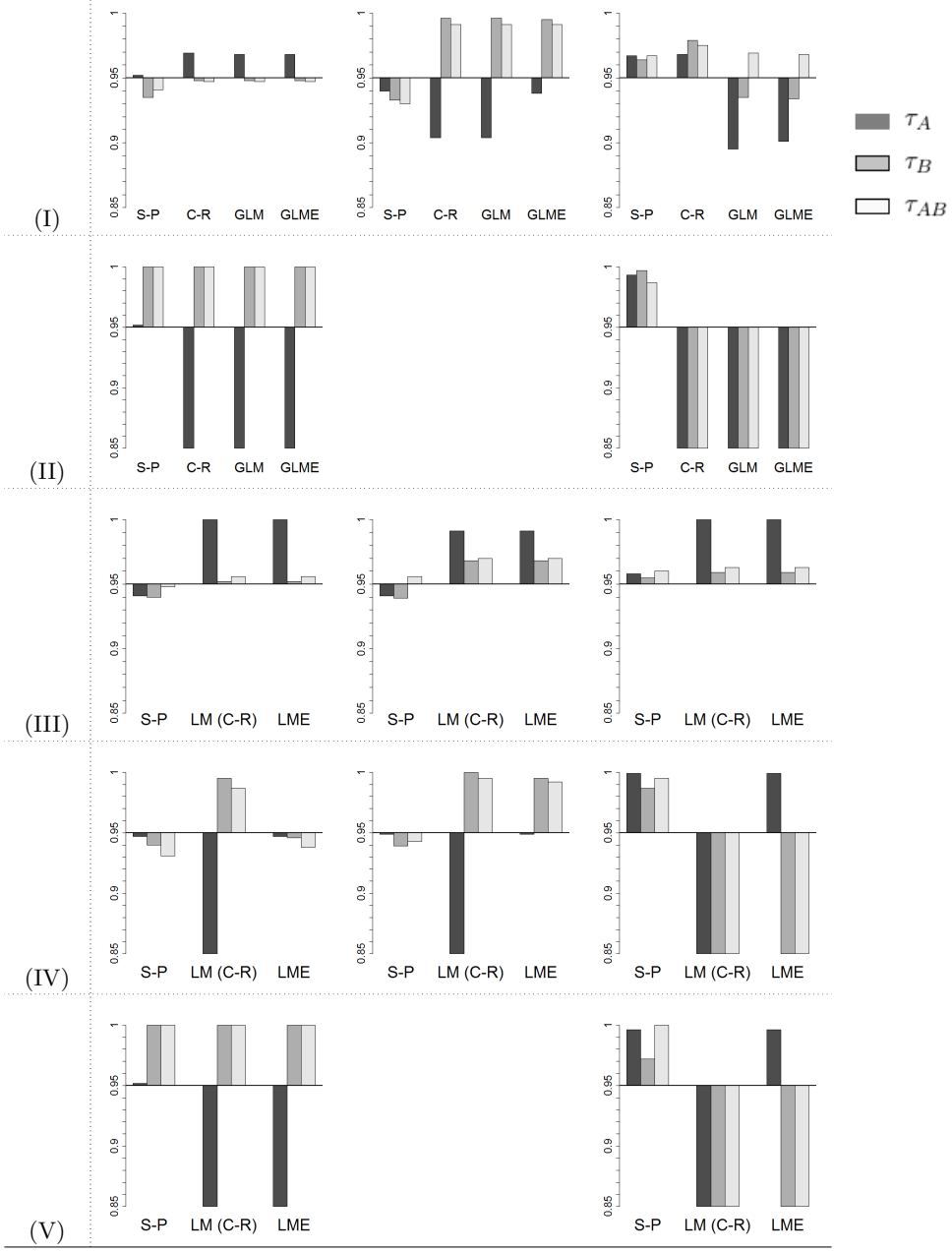
Coverage rates (%) averaged over 1,000 independent split-plot randomizations with $r_A = r_B = 1$ *at* $(W, M) = (40, 40)$ *for potential outcomes (PO) types (II) and (V).*

		PO Type (II)				PO Type (V)		
		S-P	C-R	GLM	GLME	S-P	LM (C-R)	LME
Strict Additivity	τ_A	95.0	0.0	0.0	32.5	95.0	22.9	25.9
	τ_B	100.0	100.0	100.0	100.0	100.0	100.0	100.0
	τ_{AB}	100.0	100.0	100.0	100.0	100.0	100.0	100.0
No Assumption about Additivity	τ_A	99.3	33.9	34.4	84.3	99.6	45.4	99.6
	τ_B	99.7	42.3	44.0	33.2	97.2	39.6	27.1
	τ_{AB}	98.7	36.8	47.3	26.2	100.0	65.0	43.8

9. Discussion. Randomization-based causal inference, originally developed by Neyman (1923/1990) and Neyman (1935) in the context of completely randomized, randomized block, and Latin square designs, (a) attributes the randomness in experimental data to the actual physical randomization of the experiments, (b) allows for the definition of causal effects over a finite population of interest, and (c) extends the super-population notions of 'unbiased' point estimators and 'conservative' interval estimators to the finite-population settings. Under this inferential framework, we proposed a new procedure for analyzing 2^2 and general split-plot designs, and demonstrated its superior frequency coverage property over existing model-based alternatives.

Whereas the length limit restrains us from going any further, the interested reader may find the following two directions, among others, worthy of future exploration. First, Rubin (1978) and Dasgupta, Pillai and Ru-

Fig 1: Coverage rates summarized over 1,000 independent split-plot randomizations with $r_A = r_B = 1$ at $(W, M) = (40, 40)$ $(1 - \alpha = 0.95)$. All bars start from the nominal coverage rate 0.95 and grow upwards/downwards to the actual values, truncated at 0.85. Results of C-R and LM are combined for potential outcomes (PO) types (III)–(V), since the procedure by which Dasgupta, Pillai and Rubin (2015) constructed the C-R renders it numerically identical to the LM.



bin (2015) discussed Bayesian causal inference for completely randomized designs in the context of treatment-control and 2^K factorial experiments respectively. How to extend the same framework to split-plot designs in a way that also guarantees frequency properties is yet unclear. Second, Fisher (1935) proposed the use of randomization test for sharp null hypotheses regarding the treatment effects at unit level. Extension of such framework to split-plot designs should complement the current Neymanian framework's focus on the population-level parameters.

Acknowledgments. Special thanks go to Professor Richard Tuck, Professor Joseph Blitzstein and Steven Finch for their inspirations and inputs that helped transform this manuscript. We are thankful to the two reviewers and an Associate Editor for their careful reading of the manuscript and numerous insightful comments and constructive suggestions that significantly improved the contents and presentation of this manuscript.

SUPPLEMENTARY MATERIAL

Supplement to "Randomization-based causal inference from split-plot designs"
(doi: COMPLETED BY THE TYPESETTER; .pdf). We give proofs of the theorems and provide additional simulation studies.

References.

Bailey, R. A. (1983). Restricted randomization. *Biometrika* **70** 183-198.

Box, G. E. P., Hunter, J. S. and Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd ed. John Wiley & Sons, Hoboken, New Jersey.

Cochran, W. G. and Cox, G. M. (1957). *Experimental Designs*, 2nd ed. John Wiley & Sons, Hoboken, New Jersey.

Dasgupta, T., Pillai, N. S. and Rubin, D. B. (2015). Causal inference for 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society, Series B* **77** 727-753.

Ding, P. and Dasgupta, T. (2016). A potential tale of two by two tables from completely randomized experiments. *Journal of the American Statistical Association* **111** 157-168.

Espinosa, V., Dasgupta, T. and Rubin, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. *Technometrics* **58** 62-73.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver & Boyd, Edinburgh, Scotland.

Fisher, R. A. (1935). *The Design of Experiments*, 1st ed. Oliver & Boyd, Oxford, England.

Freedman, D. A. (2006). Statistical models for causation: What inferential leverage do they provide? *Evaluation Review* **30** 691-713.

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

FREEDMAN, D. A. (2008b). On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics* **2** 176-196.

FREEDMAN, D. A. (2008c). Randomization does not justify logistic regression. *Statistical Science* **23** 237-249.

GELMAN, A. (2005). Analysis of variance — why it is more important than ever. *The Annals of Statistics* **33** 1-53.

HAJEK, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of Mathematical Institute of Hungarian Academy of Sciences, Series A* **5** 361-374.

HINKELMANN, K. and KEMPTHORNE, O. (2008). *Design and Analysis of Experiments: Introduction to Experimental Design*, 4th ed. John Wiley & Sons, Hoboken, New Jersey.

HOLLAND, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81** 945-970.

IMBENS, G. W. and RUBIN, D. B. (2015). *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction*. Cambridge University Press, New York, NY.

JONES, B. and NACHTSHEIM, C. J. (2009). Split-plot designs: What, why, and how. *Journal of Quality Technology* **41** 340-361.

KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments*, 1st ed. John Wiley & Sons, New York.

KIRK, R. E. (1982). *Experimental Design: Procedures for the Behavioral Sciences*. Brooks/Cole, Monterey, CA.

LI, X. and DING, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* in press.

LIN, W. (2013). Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics* **7** 295-318.

LU, J. (2016). On randomization-based and regression-based inferences for 2^K factorial designs. *Statistics and Probability Letters* **112** 72-78.

NEYMAN, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles (with discussion). Section 9 (translated). *Statistical Science* **5** 465–480.

NEYMAN, J. (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society* **2** 107-180.

RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* **66** 688-701.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6** 34-58.

RUBIN, D. B. (1980). Comment on "Randomization analysis of experimental data: The Fisher randomization test" by D. Basu. *Journal of the American Statistical Association* **75** 591-593.

RUBIN, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322-331.

WU, C. F. J. and HAMADA, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*, 2nd ed. John Wiley & Sons, Hoboken, New Jersey.

YATES, F. (1935). Complex experiments. *Supplement to the Journal of the Royal Statistical Society* **2** 181-247.

ZHAO, A., DING, P. and DASGUPTA, T. (2017). Supplement to "Randomization-based causal inference from split-plot designs". *DOI:xxx*.

Department of Statistics
Harvard University
Science Center, 1 Oxford Street
Cambridge, MA
E-mail: anqizhao@fas.harvard.edu

Department of Statistics
University of California, Berkeley
365 Evans Hall
Berkeley, CA
E-mail: pengdingpku@gmail.com

Indian Institute of Management Calcutta
Joka, Diamond Harbor Road
Kolkata 700 104
India
E-mail: rmuk0902@gmail.com

Department of Statistics and Biostatistics
Rutgers University
110 Frelinghuysen Road, Hill Center
Piscataway, NJ
E-mail: tirthankar.dasgupta@rutgers.edu