Causal inference for multiple treatments using fractional factorial designs

Nicole E. PASHLEY and Marie-Abèle C. BIND

Department of Statistics, Rutgers University, New Brunswick, NJ, U.S.A.
Biostatistics Center, Massachusetts General Hospital, Boston, MA, U.S.A.

Key words and phrases: Interactions; joint effects; multiple treatments; Neymanian inference; observational studies; potential outcomes.

MSC 2020: Primary 62K15; secondary 62P12.

Abstract: We consider the design and analysis of multi-factor experiments using fractional factorial and incomplete designs within the potential outcome framework. These designs are particularly useful when limited resources make running a full factorial design infeasible. We connect our design-based methods to standard regression methods. We further motivate the usefulness of these designs in multi-factor observational studies, where certain treatment combinations may be so rare that there are no measured outcomes in the observed data corresponding to them. Therefore, conceptualizing a hypothetical fractional factorial experiment instead of a full factorial experiment allows for appropriate analysis in those settings. We illustrate our approach using biomedical data from the 2003–2004 cycle of the National Health and Nutrition Examination Survey to examine the effects of four common pesticides on body mass index.

1. INTRODUCTION

What is the effect of exposing you to pesticide A, compared with no exposure, on your health? What about exposing you to pesticide B? Or exposing you to both pesticides at the same time? To answer these questions, we need to estimate the effects, including interactions, of multiple treatments. A randomized factorial experiment uses a random assignment of all possible treatment combinations to units to estimate these different effects. There is much interest in assessing the effects of multiple treatments, as reflected by the recent growth in the literature regarding the use of factorial designs in causal inference (e.g., Dasgupta, Pillai & Rubin, 2015; Dong, 2015;
Branson, Dasgupta & Rubin, 2016; Espinosa, Dasgupta & Rubin, 2016; Lu, 2016a,b; Lu & Deng, 2017; Mukerjee, Dasgupta & Rubin, 2018; Zhao et al., 2022; Egami & Imai, 2019).

However, a full factorial design, in which all possible combinations of treatments are randomized, may not be possible due to constraints on resources such as units or cost. Instead, a fractional factorial or an incomplete factorial design, which uses a subset of all possible treatment combinations, may be more practical. We lay out a Neyman-style randomization-based framework for the design and analysis of these designs using potential outcomes. Consideration of observational studies with multiple factors, in which certain treatment combinations may be missing from the dataset, further motivates the usefulness of these designs.

With only a subset of all treatment combinations, saturated linear regression models have underlying assumptions resulting in estimators that are not always transparent, especially with respect to the implicit imputation of the missing potential outcomes. We discuss regression estimates in this setting in Section 3.2 and assess how they connect to design-based estimates. Analyses using regression models in observational studies are even trickier, as the use of regression without a careful design phase, in which one tries to uncover or approximate some underlying randomized experiment, can lead to incorrect conclusions (e.g., see Rubin, 2008). Despite this difficulty, regression models with interaction terms are commonly used to estimate the effects of multiple treatments, in particular in observational studies (Patel, Bhattacharya & Butte, 2010; Bobb et al., 2015; Oulhote et al., 2017; Valeri et al., 2017). For instance, Bobb et al. (2015) consider a Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Therefore, exploration of design-based methods for multi-factor observational studies is an important missing piece in the literature.

The article proceeds as follows. Section 2 reviews full factorial designs within the potential outcomes framework described in Dasgupta, Pillai & Rubin (2015). Section 3 discusses extensions of this framework to fractional and incomplete factorial designs, expanding upon current inference results for variance and variance estimation. Section 4 examines how to embed an observational study into one of these experimental designs by applying our method to data from the National Health and Nutrition Examination Survey (NHANES), which is conducted by the Centers for Disease Control and Prevention (CDC). With these data, we explore the effects of four pesticides on body mass index (BMI). We illustrate our method and its challenges when working with observational data with multiple treatments. Section 5 summarizes our results and concludes with ideas for further research.

Throughout, we will use a running example of an experiment examining the impact of three pesticides (which we call A, B, and C) on weight gain in zebrafish. Our data example provides an illustration of the way we foresee these designs being useful when experimentation is not possible, such as when discussing the impact of pesticides on humans.

2. SETUP
2.1. Notation and Factorial Estimands
We closely follow the potential outcomes (Splawa-Neyman, 1990; Rubin, 1974) framework for $2^K$ factorial designs in Dasgupta, Pillai & Rubin (2015). We focus on designs with multiple two-level factors assigned in combination; that is, there are a number of distinct treatments (e.g., medications), each having two levels (e.g., receiving a certain medication or not). Let there be $K$ two-level factors, which creates $2^K = J$ total possible treatment combinations. The $j$th treatment combination is denoted $z_j$. See Table 1 for an example with $K = 3$, which illustrates the notation. For consistency, we list treatment combinations in lexicographic order, the standard ordering. Let $z_{j,k} \in \{-1,+1\}$ be the level of the $k$th factor in the $j$th treatment combination, so $z_j = (z_{j,1}, \ldots, z_{j,K})$. Let there be observations on $n$ units in the experiment. Following Rubin (1980), we assume the stable unit treatment value assumption (SUTVA): there
is no interference between units and no hidden forms of each treatment level. Then the potential outcome for unit \( i \) under treatment combination \( z_j \) can be written as \( \bar{Y}(z_j) \) and the sample average potential outcome under treatment \( z_j \) as \( \bar{Y}(z_j) = \frac{1}{n} \sum_{i=1}^n Y_i(z_j) \). The vector of mean potential outcomes for all \( 2^K \) treatment combinations is \( \bar{Y} = (\bar{Y}(z_1), \bar{Y}(z_2), \ldots, \bar{Y}(z_J)) \)

We use the framework from Dasgupta, Pillai & Rubin (2015) and denote the contrast column \( j \) in the design matrix by \( g_j \), as illustrated in Table 1. Following that paper, we can also define the contrast vector for the two-factor interaction between factors \( k \) and \( k' \) as \( g_{k,k'} = g_k \odot g_{k'} \), where \( \odot \) indicates the Hadamard (element-wise) product. Similarly, the contrast vector for three-factor interactions is \( g_{k,h,i} = g_k \odot g_h \odot g_i \), and all higher-order interaction contrast vectors can be calculated analogously.

We continue to follow Dasgupta, Pillai & Rubin (2015) and define the finite population main causal effect for factor \( k \) and the finite population interaction effect for factors \( k \) and \( k' \) as

\[
\tau(k) = \frac{1}{2^{K-1}} g_k^\top \bar{Y} \quad \text{and} \quad \tau(k, k') = \frac{1}{2^{K-1}} g_{k,k'}^\top \bar{Y},
\]

respectively, where by finite population, we mean that we are only interested in inference for the units we have in the experiment. Higher-level interaction terms are defined analogously. We can similarly define the individual-level effects as

\[
\tau_i(k) = \frac{1}{2^{K-1}} g_k^\top Y_i \quad \text{and} \quad \tau_i(k, k') = \frac{1}{2^{K-1}} g_{k,k'}^\top Y_i,
\]

where \( Y_i = (Y_i(z_1), \ldots, Y_i(z_J))^\top \). We also define the average potential outcome across treatments as \( \tau(0) = 2^{-K} g_0^\top \bar{Y} \), where \( g_0 \) is a vector of length \( 2^K \) of all \( +1 \)'s.

Consider our example of estimating the effects of pesticides A, B, and C on weight gain in zebrafish. We take the three pesticides to be three factors, each with two levels: applying the pesticide \((+1)\), or not \((-1)\). In Table 1, zebrafish assigned to \( z_6 \) would be those that get pesticides A and C, but not B. The main effect of pesticide A (\( \tau(1) \)) is the average effect of receiving pesticide A, compared with not receiving it, uniformly averaged over all combinations of receiving or not receiving pesticides B and C. The two-factor interaction between pesticides A and B would be one-half the difference in the effect of A in the presence of B versus in the absence of B, averaging uniformly over receiving or not receiving pesticide C.

---

**Table 1: Example of a \( 2^3 \) factorial design.**

| Treatment | Factor 1 | Factor 2 | Factor 3 | Outcomes |
|-----------|----------|----------|----------|----------|
| \( z_1 \) | -1       | -1       | -1       | \( \bar{Y}(z_1) \) |
| \( z_2 \) | -1       | -1       | +1       | \( \bar{Y}(z_2) \) |
| \( z_3 \) | -1       | +1       | -1       | \( \bar{Y}(z_3) \) |
| \( z_4 \) | -1       | +1       | +1       | \( \bar{Y}(z_4) \) |
| \( z_5 \) | +1       | -1       | -1       | \( \bar{Y}(z_5) \) |
| \( z_6 \) | +1       | -1       | +1       | \( \bar{Y}(z_6) \) |
| \( z_7 \) | +1       | +1       | -1       | \( \bar{Y}(z_7) \) |
| \( z_8 \) | +1       | +1       | +1       | \( \bar{Y}(z_8) \) |

\( g_1, g_2, g_3 \)
Consider a full factorial model matrix, $G$, that includes the mean and interactions whose rows are comprised of $g_k^\top$, where $k \in 2^K$ where $2^K$ is all subsets of $\{0, 1, \ldots, K\}$ and $g$ corresponds to $g_0$. Based on the definition of our estimands in Section 2.1, the matrix $G$ relates the mean potential outcomes and the factorial effects as follows: Let $G$ be a matrix with the $g$ vectors as rows,

$$G = \begin{bmatrix} g_0, g_1, \ldots, g_{K}, g_{1,2}, \ldots, g_{1,\ldots,K} \end{bmatrix}^\top$$

and $\tau$ be the column vector of all factorial effects,

$$\tau = \begin{bmatrix} 2\tau(0), \tau(1), \ldots, \tau(K), \tau(1,2), \ldots, \tau(K-1,K), \ldots, \tau(1,2,\ldots,K) \end{bmatrix}^\top.$$ 

Then,

$$2^{-(K-1)}GY = \tau.$$ 

Since the rows of $G$ are orthogonal with square norm $2^K$, $G^{-1} = \frac{1}{2^K}G^\top$. The mean potential outcomes can be rewritten in terms of the factorial effects as $\bar{Y} = 2^{-1}G^\top\tau$. Let the $j$th row of $G^\top$ be denoted by $h_j$. The first entry of $h_j$ is +1, corresponding to the mean, the next $K$ entries are equal to the entries of $z_j$, and the remaining entries correspond to interactions, with the order given by the order of the rows of $G$. In the $2^3$ factorial design shown in Table 1, $h_1 = (+1, -1, -1, -1, +1, +1, +1, -1)$. That is, it is the entries of the first row of Table 1, prepended with an entry for $2\tau(0)$ and appended with entries for the interactions. We have $\bar{Y}(z_j) = 2^{-1}h_j\tau$. We will use this representation in Section 3.2 when considering causal estimands in settings lacking units for some treatment combinations.

2.2. Estimation and Inference in a 2^K Factorial Design

We now review inference for $2^K$ factorial designs. Let $W_i(z_j) = 1$ if unit $i$ received treatment $z_j$ and $W_i(z_j) = 0$ otherwise. In a full $2^K$ factorial design, under complete randomization, the $n$ units are assigned completely at random to the $n_j \geq 2$, $j = 1, \ldots, 2^K$, copies of the treatment combinations. That is, there is a fixed number, $n_j \geq 2$, of units randomly assigned to treatment combination $z_j$. Let $W$ be the $n \times 2^K$ vector of $W_i(z_j)$ for all $i$ and $z_j$. Let $w$ be some $n \times 2^K$ matrix with entries in the set $\{0, 1\}$. This design imposes

$$P(W = w) = \begin{cases} \prod_{j=1}^J n_j! / n! & \text{if } w1 = 1 \text{ and } w^\top 1 = n, \\ 0 & \text{otherwise,} \end{cases}$$

where 1 is a column vector of ones of length $n$ and $n = (n_1, \ldots, n_J)^\top$. The observed potential outcome for unit $i$ is $Y_i^{\text{obs}} = \sum_{j=1}^J W_i(z_j) Y_i(z_j)$. An unbiased estimator of the average potential outcome for treatment $z_j$ is

$$\bar{Y}^{\text{obs}}(z_j) = \frac{1}{n_j} \sum_{i=1}^n W_i(z_j) Y_i(z_j) = \frac{1}{n} \sum_{i : W_i(z_j) = 1}^n Y_i^{\text{obs}}.$$ 

Denote $\bar{Y}^{\text{obs}} = (\bar{Y}^{\text{obs}}(z_1), \bar{Y}^{\text{obs}}(z_2), \ldots, \bar{Y}^{\text{obs}}(z_J))$ as the vector of observed mean potential outcomes for all $2^K$ treatment combinations.

Unbiased estimators for $\tau(k)$ and $\tau(k,k')$ are defined as

$$\hat{\tau}(k) = \frac{1}{2^{K-1}}g_k^\top \bar{Y}^{\text{obs}} \quad \text{and} \quad \hat{\tau}(k,k') = \frac{1}{2^{K-1}}g_{k,k'}^\top \bar{Y}^{\text{obs}},$$

DOI: 10.1002/cjs.11734
respectively. Estimators for higher-level interaction terms and \( \tau(0) \) are defined analogously.

The variance of a factorial effect estimator under an unbalanced design is given in Lu (2016b) (extended from balanced case in Dasgupta, Pillai & Rubin, 2015), as follows:

\[
\text{Var} (\hat{\tau}(k)) = \frac{1}{2^{2(K-1)}} \sum_{j=1}^{J} \frac{1}{n_j} S_j^2(z_j) - \frac{1}{n} S_k^2,
\]

where

\[
S_j^2(z_j) = \frac{1}{n-1} \sum_{i=1}^{n} \left( Y_i(z_j) - \bar{Y}(z_j) \right)^2 \quad \text{and} \quad S_k^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left( \tau_i(k) - \tau(k) \right)^2.
\]

An expression for the covariance between two factorial effect estimators for unbalanced factorial designs is given by Lu (2016b) (extended from Dasgupta, Pillai & Rubin, 2015) as

\[
\text{Cov} (\hat{\tau}(k), \hat{\tau}(k')) = \frac{1}{2^{2(K-1)}} \left[ \sum_{j: \delta_{kj} = \delta_{k'j}} \frac{1}{n_j} S_j^2(z_j) - \sum_{j: \delta_{kj} \neq \delta_{k'j}} \frac{1}{n_j} S_j^2(z_j) \right] - \frac{1}{n} S_{k,k'}^2,
\]

where \( S_{k,k'}^2 = \sum_{i=1}^{n} \left( \tau_i(k) - \tau(k) \right) \left( \tau_i(k') - \tau(k') \right) / (n-1) \).

Further discussion of Neymanian and Fisherian inference for full factorial designs is reviewed in Appendix A of the Supplementary Material.

3. ALTERNATE DESIGNS TO FULL FACTORIAL: FRACTIONAL AND INCOMPLETE

3.1. Fractional Factorial Designs

There are situations in which a full \( 2^K \) factorial design experiment cannot be conducted or is suboptimal. Limited resources may mean there is an insufficient number of units to randomly assign to each of the \( J = 2^K \) treatment groups, or the full factorial design may not be the most efficient allocation of resources; for example, if the experimenter believes that higher-order interactions are negligible (Wu & Hamada, 2000). Instead, an experimenter might implement a \( 2^{K-p} = J^* \) fractional factorial design in which only \( J^* \) of the total \( J \) treatment combinations are used. Here we will give a brief overview of this design, but we recommend Wu & Hamada (2000) for a more detailed review. We follow notation in Dasgupta, Pillai & Rubin (2015) and Dasgupta & Rubin (2015).

To create a \( 2^{K-p} \) design, we can write out a full factorial design for the first \( K-p \) factors, and fill in the \( p \) columns for the remaining factors using multiplicative combinations of subsets of the previous contrast columns. Factors whose columns are multiplied to get the treatment levels of another factor define a generator that equates the label of a factor to the product of other factors (Wu & Hamada, 2000). For example, to create a \( 2^{3-1} \) design with the generator \( "3 = 12" \), we write out a \( 2^2 \) design for factors 1 and 2, as in Table 2. We generate the third column, corresponding to treatment levels for factor 3, by multiplying together the contrast vectors for factors 1 and 2, as shown in Table 2. The two factors used initially are irrelevant in this case because of symmetry. Because \( 3 = 12 \), we also have \( 1 = 23 \) and \( 2 = 13 \). However, this symmetry may not hold in general. These equivalencies between effects are aliasing relationships. Factors that are aliased are confounded; that is, we cannot tease apart their effects in the experiment. One should choose which factors to use in the generator based on the final aliasing structure.
TABLE 2: Example of a $2^{3-1}$ factorial design.

| Treatment | Factor 1 | Factor 2 | Factor 3 | Outcomes  |
|-----------|----------|----------|----------|-----------|
| $z_1^*$   | $-1$     | $-1$     | $+1$     | $\bar{Y}(z_1^*)$ |
| $z_2^*$   | $-1$     | $+1$     | $-1$     | $\bar{Y}(z_2^*)$ |
| $z_3^*$   | $+1$     | $-1$     | $-1$     | $\bar{Y}(z_3^*)$ |
| $z_4^*$   | $+1$     | $+1$     | $+1$     | $\bar{Y}_*$    |

$g_1^*$ $g_2^*$ $g_3^*$ $\bar{Y}_*$

Under this design, the contrast columns, $g_k^*$, are now shortened versions or subsets of the contrast columns of a full $2^3$ factorial design, $g_k$. Denote $G^*$ as the matrix with $g_k^*$ as rows, analogous to $G$ in the full factorial setting, which relates the treatment combinations in the fractional experiment to the treatment effect estimands. The treatment combinations in the fractional design, $z_j^*$, are a subset of the full set of treatment combinations, $z_j$, for the $2^3$ design.

Referring to Table 2, $g_3^* = g_1^* \odot g_2^*$. The generator $3 = 12$ indicates that the main effect of factor 3 is aliased with the two-factor interaction 12, which means that we cannot distinguish these effects—they are confounded and combined in our estimators. The full aliasing in this design is as follows: $I = 123$, $3 = 12$, $2 = 13$, $1 = 23$, where $I$ corresponds to a vector of all $+1$’s ($g_0$). The relation $I = 123$ is called the defining relation, which characterizes the aliasing and how to generate the rest of the columns. The main effects, as defined in Section 2.1, are aliased with the two-factor interactions. If the two-factor interactions are negligible, the $2^{3-1}$ design is a parsimonious design to estimate the main effects and we can construct unbiased estimators for the main effects (see Section 3.1.1). The resulting design is also orthogonal (i.e., all pairs of columns are orthogonal) and balanced in the sense that each $g_k^*$ has an equal number of $+1$’s and $-1$’s (Wu & Hamada, 2000). These properties simplify the aliasing structure.

We typically choose the generator or defining relation based on the maximum resolution criterion for the design. Resolution is defined as the word length (i.e., the number of factors) in the shortest word of the defining relation (see Wu & Hamada, 2000, for more details) and indicates the aliasing structure and with which levels of effects the main effects are aliased. In our example of a $2^{3-1}$ design, we have only one defining relation, $I = 123$, so the word length, and hence resolution, is 3, and main effects are aliased with two-factor interactions. Aliasing some main effects with other main effects is an alternate aliasing structure.

The effect hierarchy principle assumes that lower-order effects have larger magnitude than higher-order interaction effects (Wu & Hamada, 2000). Therefore, one generally chooses a fractional design where main effects and some other lower-order interaction effects are aliased with higher-order terms, i.e., the main effects and two-factor interactions are clear. This assumption goes along with the assumption of effect sparsity (the number of important effects is small), as a justification for the use of fractional designs over the full factorial (Wu & Hamada, 2000). It is possible that there may be multiple defining relations that lead to the same resolution. In addition to the maximum resolution criterion, it is standard to consider minimum aberration, which says we want the design with the fewest short words in its defining relations.

3.1.1. Estimators

In this section, we review estimators for the fractional design, which are similar to the full factorial case, and follow the framework laid out in Dasgupta, Pillai & Rubin (2015) and...
Dasgupta & Rubin (2015). We assume units are assigned using complete randomization to the \( J^* \) treatments. Randomization is the same as in the full factorial case, but with fewer treatment groups. All of the following inference results are derived under complete randomization.

The estimator for \( \tau(k) \) is defined as
\[
\hat{\tau}^*(k) = \frac{1}{2^{K-p-1}} g_k^T \bar{Y}_s^{obs}.
\]

The estimator for \( \tau(k, k') \) is defined as
\[
\hat{\tau}^*(k, k') = \frac{1}{2^{K-p-1}} g_{k,k'}^T \bar{Y}_s^{obs}.
\]

These estimators are no longer unbiased. Let \( A_k \) represent the set consisting of the main effect of factor \( k \) and all effects aliased with it. The number of effects aliased with the main effect of factor \( k \) is \( 2^p - 1 \) (Montgomery, 2017, Chapter 8). So, \( A_k \) has \( 2^p \) elements. The main effect of factor \( k \) may be aliased with the negative of a main effect or interaction. Let \( A_{k,j} \) be the indicator for whether effect \( j \) is negatively aliased with the main effect of factor \( k \) \( (A_{k,j} = 1) \) or positively aliased \( (A_{k,j} = -1) \).

**Result 1.** The expectation of the estimator of the main effect \( k \) is
\[
E[\hat{\tau}^*(k)] = \frac{1}{2^{K-p-1}} \{ \sum_{j \in A_k} A_{k,j} \tau(j) = \tau(k) + \sum_{j \in A_k \setminus \{k\}} A_{k,j} \tau(j) \}.
\]

Result 1 can be shown using the orthogonality of the \( g_k \) vectors. Hence, by aliasing these effects, they get combined in our estimator.

Returning to our running example, let us say we used the design in Table 2. In that design, the main effect of pesticide A is aliased with the interaction between pesticides B and C. The expectation of the main effect for pesticide A would be \( E[\hat{\tau}^*(1)] = \tau(1) + \tau(2, 3) \). Our estimate of the effect of pesticide A will be biased by the size of the interaction. If the interaction effect between pesticides B and C is small, our estimator for the main effect of pesticide A will have a small bias. As an example, if the main effect of pesticide A is \( \tau(1) = -2 \) and the interaction between pesticides B and C is \( \tau(2, 3) = 2 \), the expectation of our estimator will be 0 even though the true average effect of pesticide A is nonzero.

More generally, the estimator for \( \tau(k) \) is unbiased if the effects aliased with factorial effect \( k \) are zero, and will be approximately unbiased if the aliased effects are negligible, which may be justified by the effect hierarchy principle. That is, we expect higher-order effects to be smaller in magnitude. When aliasing occurs such that main effects are aliased with low-order interaction terms, such as two-factor interactions, this assumption may be unrealistic. However, if we have a large number of factors and are only aliasing main effects with higher-order effects, this may be reasonable. For example, in a \( 2^{6-1} \) fractional design, main effects can be aliased with five-factor interactions and two-factor interactions can be aliased with four-factor interactions, where both sets of higher-order interactions may be assumed to be small.

We extend the variance and variance estimator expressions from Dasgupta, Pillai & Rubin (2015) (see Section 2) to this setting. Recall \( J^* = 2^{K-p} \) and define \( \bar{\tau}_j(k) = \tau_j(k) + \sum_{j \in A_k \setminus \{k\}} A_{k,j} \tau_j(j) \) and \( \bar{\tau}(k) = \frac{1}{N} \sum_{i=1}^N \bar{\tau}_i(k) = E[\hat{\tau}^*(k)] \). Further define the analog of \( S_k^2 \),
\[
S_k^2 = \frac{1}{n-1} \sum_{i=1}^n (\bar{\tau}_i(k) - \bar{\tau}(k))^2.
\]
which is the variation in our newly defined aliased effects, $\tilde{\tau}_i(k)$. Let $n_j^*$ be the number of units assigned to treatment combination $z_j^*$.

**Result 2.** The variance of the estimator $\hat{\tau}^*(k)$ is

$$\text{Var} (\hat{\tau}^*(k)) = \frac{1}{2^{2(k-p-1)}} \sum_{j=1}^{J^*} \frac{1}{n_j^*} S^2(z_j^*) - \frac{1}{n} \tilde{S}_k^2. \tag{1}$$

See Appendix D.1 of the Supplementary Material for the proof of Equation (1).

We can compare the variance under a full factorial design to one under a fractional factorial design easily in a simplified setting. Take the sample size to be $n$ for both designs, and both designs to be balanced ($n_j = n/J$, $n_j^* = n/J^*$). Further, assume we have additive effects (no effect heterogeneity), which implies $S^2(z_j) = S^2$ for all $z_j$ and $S^2_k = \tilde{S}_k^2 = 0$. In this setting, the variance of the estimator for $\hat{\tau}(k)$ for the full factorial design will be

$$\text{Var} (\hat{\tau}(k)) = \frac{1}{2^{2(k-1)}} \frac{J^2}{n} S^2 = \frac{4}{n} S^2,$$

and for the fractional factorial design will be

$$\text{Var} (\hat{\tau}^*(k)) = \frac{1}{2^{2(k-p-1)}} \frac{(J^*)^2}{n} S^2 = \frac{4}{n} S^2,$$

the same as under a full factorial design. However, as the fractional factorial design has a bias when the interaction effects being aliased are nonzero, the full factorial is preferred when possible. The case without additive effects, even with a balanced design, is more complicated, especially if the treatments included in the fractional factorial design happen to be those that produce less variable outcomes. However, this would be difficult to predict before running the experiment. Unbalanced designs will further change this result, potentially increasing or decreasing the weight on more variable groups in the two designs. When a full factorial design is not feasible or practical, inducing some (hopefully negligible) bias by using a fractional factorial design may be acceptable.

We can obtain also the covariance between two fractional factorial effect estimators. We define an altered version of $S^2_{k,k'}$:

$$\tilde{S}_{k,k'}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tilde{\tau}_i(k) - \bar{\tilde{\tau}}(k)) \left( \tilde{\tau}_i(k') - \bar{\tilde{\tau}}(k') \right).$$

**Result 3.** The covariance of $\hat{\tau}^*(k)$ and $\hat{\tau}^*(k')$ is

$$\text{Cov} (\hat{\tau}^*(k), \hat{\tau}^*(k')) = \frac{1}{2^{2(k-p-1)}} \left[ \sum_{j: z_{kj} \neq z_{kj}^*} \frac{1}{n_j^*} S^2(z_j^*) - \sum_{j: z_{kj} = z_{kj}^*} \frac{1}{n_j^*} S^2(z_j^*) \right] - \frac{1}{n} \tilde{S}_{k,k'}^2.$$

These variance and covariance expressions result from an application of Theorem 3 in Li & Ding (2017), as explained in Appendix D.1 in the Supplementary Material. The variance expressions for fractional factorial designs are similar to the full factorial case, but are defined over aliased effects. In particular, the treatment effect heterogeneity term has the interpretation of the heterogeneity of a linear combination of aliased effects.
3.1.2. Statistical inference

It is straightforward to extend the analysis for factorial designs in Dasgupta, Pillai & Rubin (2015) and reviewed in Section 2 to fractional factorial designs. We can use a conservative Neyman-style variance estimator, similar to the full factorial setting (see Appendix A in the Supplementary Material):

$$\hat{\text{Var}} (\tilde{\tau}^*(k)) = \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J^*} \frac{1}{n_j^*} s^2(z_j^*),$$

where $s^2(z_j^*)$ is the observed variance of outcomes among units assigned to $z_j^*$.

Result 4. The expectation of the fractional factorial variance estimator is

$$E \left[ \hat{\text{Var}} (\tilde{\tau}^*(k)) \right] = \frac{1}{2^{2(K-p-1)}} \sum_{j=1}^{J^*} \frac{1}{n_j^*} S^2(z_j^*).$$

Thus, $\hat{\text{Var}} (\tilde{\tau}^*(k))$ is a conservative estimator and unbiased if and only if $\hat{S}_k^2 = 0$, which occur if the $\tilde{\tau}_i(k)$ are constant. This condition holds when all effects aliased with factor $k$ are constant additive effects.

We can build confidence regions and confidence intervals analogously to what was reviewed in Section A. To be more rigorous in this section, we show how to use the results of Li & Ding (2017) to build asymptotic Wald-type confidence regions. For the $2^{3-1}$ example, recalling the aliasing structure, which gives $g_0^* = g_{123}^*$ and $\tilde{\tau}(2) = \tilde{\tau}(1, 3)$, we have

$$2^{-(K-p-1)} \begin{bmatrix} g_{0}^{*\top} \\ g_{1}^{*\top} \\ g_{2}^{*\top} \\ g_{3}^{*\top} \end{bmatrix} \begin{bmatrix} \tilde{Y}(z_1^*) \\ \tilde{Y}(z_2^*) \\ \tilde{Y}(z_3^*) \end{bmatrix} = \begin{bmatrix} 2\tilde{\tau}(0) \\ \tilde{\tau}(1) \\ \tilde{\tau}(2) \\ \tilde{\tau}(3) \end{bmatrix}.$$

Define $\hat{\tau}^*$ as the $2^{K-p}$ vector of unique treatment effect estimators that corresponds to estimands in $\hat{\tau}$.

Result 5. Suppose that all $S^2(z_j^*)$ and $\hat{S}_k^2$ have limiting values as $n \to \infty$, that the $n_j^*/n$ have positive limiting values, and $\max_{1 \leq j \leq J^*} \max_{1 \leq i \leq n} \left( Y_i(z_j^*) - \tilde{Y} (z_j^*) \right)^2 / n \to 0$. Then, according to Theorem 5 of Li & Ding (2017), $n\hat{\text{Var}}(\hat{\tau}^*)$ has a limiting value, which, following their notation, we denote by $V$ and

$$\sqrt{n} (\hat{\tau}^* - \hat{\tau}) \overset{d}{\to} N(0, V).$$

Result 6. Further, let $\hat{V} = \sum_{j=1}^{J^*} n_j^* G_j^* s^2(z_j^*) G_j^{*\top}$, where $G_j^*$ is the $j$th column of $G^*$. In addition to the assumptions required for Result 5, we require that the limit of $\sum_{j=1}^{J^*} n_j^* G_j^* S^2(z_j^*) G_j^{*\top}$ is nonsingular. Under Proposition 3 of Li & Ding (2017) and following their notation, the Wald-type confidence region is

$$\{ \mu : (\hat{\tau}^* - \mu)^\top \hat{V} (\hat{\tau}^* - \mu) \leq q_{p,1-a} \}.$$
where $q_{J,1-\alpha}$ corresponds to the $1-\alpha$ quantile of the $\chi^2_J$ distribution and has at least $1-\alpha$ asymptotic coverage.

We assume that the number of treatments is constant as $n \to \infty$, but Li & Ding (2017) discuss the case where the number of treatment combinations grows as well, through factors or levels being added to the experiment.

As in the full factorial setting (see Lu, 2016b; Zhao & Ding, 2021; and Appendix A in the Supplementary Material), linear regression on the full vector of observed outcomes, using a saturated model with factor levels coded as $-1$ and $+1$, yields the same point estimates (divided by 2), and the HC2 variance estimator (MacKinnon & White, 1985) yields the same variance estimate (divided by 4) as the Neymanian estimates. For the proof, see Appendix E in the Supplementary Material. Zhao & Ding (2021) show that weighted-least squares can also be used to obtain similar results in an unsaturated regression in the full factorial setting.

3.2. Incomplete Factorial Designs

In this section, we discuss an alternative experimental design called the incomplete factorial design (Byar, Herzberg & Tan, 1993), which we consider (novelly, to our knowledge) from the potential outcome and design-based perspective. This design uses a subset of data from a full factorial design but a different subset than the fractional design. In particular, the estimators we discuss will not use all of the available data. We can therefore consider each estimator to be associated with a particular “design” that corresponds to an experiment that randomizes units only to treatment combinations used in the estimator. Different estimators may give nonzero weight to different treatment combinations, and so the designs may be estimand-specific. Due to this inherent linking between the design and the estimators, we discuss the design and corresponding estimators together in the next section.

3.2.1. Design and estimators

Consider if we ran an experiment with $K$ binary factors where, whether due to human error or lack of resources, some treatment combinations from the full $2^K$ factorial design were not included in the experiment. We may have a factorial structure as in Table 3, but no outcome measurements for treatment $z_7$. How should we analyze this?

Focus on estimation of $\tau(1)$ first. We could reduce the data to a fractional factorial design, which requires removing multiple treatment groups. If we recreate a fractional factorial design

| Treatment | 1 | 2 | 3 | Observed outcomes |
|-----------|---|---|---|------------------|
| $z_1$     | $-1$ | $-1$ | $-1$ | $\bar{Y}_{\text{obs}}(z_1)$ |
| $z_2$     | $-1$ | $-1$ | $+1$ | $\bar{Y}_{\text{obs}}(z_2)$ |
| $z_3$     | $-1$ | $+1$ | $-1$ | $\bar{Y}_{\text{obs}}(z_3)$ |
| $z_4$     | $-1$ | $+1$ | $+1$ | $\bar{Y}_{\text{obs}}(z_4)$ |
| $z_5$     | $+1$ | $-1$ | $-1$ | $\bar{Y}_{\text{obs}}(z_5)$ |
| $z_6$     | $+1$ | $-1$ | $+1$ | $\bar{Y}_{\text{obs}}(z_6)$ |
| $z_7$     | $+1$ | $+1$ | $-1$ | $\bar{Y}_{\text{obs}}(z_7)$ |
| $z_8$     | $+1$ | $+1$ | $+1$ | $\bar{Y}_{\text{obs}}(z_8)$ |
| $g_1$     |     |     |     | \hspace{1cm} |
| $g_2$     |     |     |     | \hspace{1cm} |
| $g_3$     |     |     |     | \hspace{1cm} |

Note: Treatment $z_7$, highlighted in red, has no units assigned to it. Treatment $z_3$, highlighted in blue, is identical to $z_7$ except for the level of factor 1.
that aliases the main effects with the two-way interactions, we estimate \( \tau(1) + \tau(2, 3) \), which involves using outcome data from units assigned to treatment combinations \( z_2, z_3, z_5, \) and \( z_8 \), but not from units assigned to treatment combinations \( z_1, z_4, \) and \( z_6 \).

Instead, we might consider building an estimator using all of the treatment combinations except for \( z_3 \), which compared to \( z_7 \) has the same levels for factors 2 and 3 but the opposite level for factor 1. Thus, in some sense, removing \( z_3 \) “balances” the remaining treatment combinations and gives essentially the “naïve estimator” discussed in Byar, Herzberg & Tan (1993). This strategy creates a different hypothetical experimental design with a different aliasing structure. In this case, we would be estimating

\[
\hat{\tau}(1) = \frac{1}{3} \left[ \bar{Y}(z_4) + \bar{Y}(z_6) + \bar{Y}(z_8) \right] - \frac{1}{3} \left[ \bar{Y}(z_1) + \bar{Y}(z_2) + \bar{Y}(z_4) \right]
\]

\[
= \frac{1}{3} \left[ \tau(1|F_2 = -1, F_3 = -1) + \tau(1|F_2 = -1, F_3 = +1) + \tau(1|F_2 = +1, F_3 = +1) \right],
\]

where we let \( F_k \) denote the level of the \( k \)th factor so that \( \tau(k|F_j = x, F_i = y) \) is the main effect of factor \( k \) conditional on level \( x \) of factor \( j \) and level \( y \) of factor \( i \), as in Dasgupta, Pillai & Rubin (2015). In other words, we estimate the average of the conditional effects of factor 1 given the combinations \((-1, -1), (-1, +1), \) and \((+1, +1)\) for factors 2 and 3.

To find the aliasing structure, we can refer to the matrix \( G^T \) in Section 2.1 to rewrite the estimand above as

\[
\hat{\tau}(1) = \frac{1}{6} \left( h_8 + h_6 + h_5 - h_4 - h_2 - h_1 \right) \tau
\]

\[
= \tau(1) + \frac{1}{3} \left[ -\tau(1, 3) + \tau(2, 3) + \tau(1, 2, 3) \right].
\]

We have partially aliased the main effect for factor 1 with the two-factor interactions for factors 1 and 3 and factors 2 and 3, as well as the three-factor interaction, all divided by three. This is partial aliasing because the factors are neither fully aliased nor completely clear of each other (Wu & Hamada, 2000, Chapter 7). Now we shall return to our pesticide example. If \( \tau(1, 3) = 2, \tau(2, 3) = 2.5, \) and \( \tau(1, 2, 3) = 1 \), the main effect of pesticide A will be positively biased by \( 1/2 \). Whether this is preferable to aliasing the main effect with just the two-factor interaction between factors 2 and 3, as in the fractional design, depends on context. If we know that the three-factor interaction is negligible (or at least smaller than the two-factor interactions, as in the numerical example just given), this new estimator may have a lower bias, as we are dividing both two-factor interactions by 3. However, even if we had knowledge that the two-factor interactions were of the same sign, we would not know the direction of the bias in this case without knowing the relative magnitudes of the two-factor interactions.

When estimating the main effect for factor 2, we would naturally approximate a different design. Using the same logic as before, we would remove treatment combination \( z_5 \).

As an alternative design, we may alias our main effects with the highest-order interaction possible, allowing for a different design for each main effect estimator. So, when estimating the main effect of factor 1, we would use a fractional factorial design that aliases factor 1 with the three-factor interaction. This leads to a design using treatment combinations \( z_1, z_4, z_5, \) and \( z_8 \) for which we can build the following estimand:

\[
\hat{\tau}(1) = \frac{1}{2} \left[ \bar{Y}(z_5) + \bar{Y}(z_8) \right] - \frac{1}{2} \left[ \bar{Y}(z_1) + \bar{Y}(z_4) \right]
\]

\[
= \frac{1}{4} \left( h_8 + h_5 - h_4 - h_1 \right) \tau
\]

\[
= \tau(1) + \tau(1, 2, 3).
\]
Based on the hierarchy principle, an estimator with this aliasing structure should be a superior estimator to the original fractional factorial estimator mentioned because the three-factor interaction is more likely to be negligible than the two-factor interactions. In particular, using the previous numbers from our empirical example, the bias for the main effect of pesticide A would be $1/3$. In general, based on the hierarchy principle, we want to choose a design that aliases the effect we are looking at with the highest-order interaction possible (which may be limited by the missing data structure). For example, if we are missing data only for one treatment combination, then when estimating main effects, we can always choose the $2^{K-1}$ fractional factorial design that aliases the main effect with the negative of the highest-order interaction. This is not a design we would want to use to collect data, because it would alias our main effects with each other. However, differing from the fractional factorial setting, we have more data and would use a different design for estimating each factorial effect.

Denote by $\hat{g}_k$ the analogue of $g_k$ with zeros where outcome data are missing or excluded in a given estimator for factor $k$. If there is a single treatment combination with no outcome measured, we can choose the aliasing such that factor $k$ is aliased with the $K$-factor interaction. When more rows are missing, the pattern of missingness will dictate what aliasing structure is possible. For example, if we are missing two treatment combinations but they are missing from the same design that aliases factor $k$ with the positive of the $K$-factor interaction, we can still recreate the design that aliases factor $k$ with the negative of the $K$-factor interaction. But if we are missing a row from each of these designs, neither option is possible and we must choose a different aliasing structure. If this method is continued for each factor, one ends up with a set of $\hat{g}_k$, each with zeros for different treatment combinations.

Section 4.5 of Wu & Hamada (2000) gives a general strategy to design experiments while attempting to reduce aliasing for certain main effects. There are other designs we can construct, not considered here, such as nonregular designs that have partial aliasing (see Wu & Hamada, 2000, Chapter 7 for more details on nonregular designs).

3.2.2. Statistical inference

Once we pick which subsets of treatments we will use (based on the aliasing structure and practical constraints), we assume units are assigned using complete randomization to the $\tilde{J}$ treatments in the incomplete design. Assignment is similar to the full factorial or fractional factorial, just for a different subset of treatments. The major differences between a fractional factorial design and an incomplete design are therefore not in the way that the treatments are randomized but in (i) that the treatment subset used is not (in general) a fractional subset of the original $2^K$ treatments and (ii) the way in which the analysis is proposed to proceed. All of the following inference results are derived under the assumption of complete randomization.

More generally, we denote our estimator for $k$th factor under one of these alternative incomplete factorial designs as $\hat{\tau}(k) = \hat{g}_k^\top \tilde{Y}^{\text{obs}}$. It is straightforward to extend the variance expression $\text{Var}(\hat{\tau}(k))$ and variance estimator $\hat{\text{Var}}(\hat{\tau}(k))$, as well as the covariance of $\hat{\tau}(k)$ and $\hat{\tau}(k')$, from Section 3.1.2; in fact, it is easy to similarly extend these results to any linear combinations of interest on the $2^K$ treatment combinations. However, terms for some treatment combinations will be set to zero in these expressions because they are present in one estimator but not the other. See Appendix F in the Supplementary Material for the specification and derivation.

If we run a regression with all interactions on a dataset with missing treatment combinations, it is ambiguous what design the resulting estimators correspond to and, as discussed above, multiple designs may be plausible to estimate effects well. If we specify a regression of the outcome on the fully interacted treatments, but some treatment combinations are not present in the data, the regression will not be able to estimate all interactions. Not including a set of interactions implies that the linear model assumes that those interaction effects are zero, and we
can assess how reasonable this assumption is. However, the specific aliasing structure between the effects included in the model and those dropped by the regression is not obvious from the usual computer output alone, though can be discerned from the design matrix. Byar, Herzberg & Tan (1993) and Byar, Freedman & Herzberg (1995) give more discussion of these types of estimators, and Appendix F.2 in the Supplementary Material has further discussion.

3.3. Fractional Factorial Designs versus Incomplete Factorial Designs

Sometimes an experimenter may have a choice between running a fractional factorial design or an incomplete factorial design. However, in many cases this decision will be dictated in part by the ability to assign at least 2 units to each treatment in the design. The fractional factorial design strictly uses $2^{K-p}$ of the original treatments, whereas an incomplete design may use more treatment groups.

When either design may be run, we expect there to be an advantage in terms of bias by using an incomplete factorial design. In the incomplete factorial example given in Table 3, we were able to show that we can alias the main effect of factor 1 with the three-factor interaction. In our numerical example with the three pesticides, this implied a bias for our main effect estimator of pesticide A of 1/3. On the other hand, the fractional factorial design of Table 2 resulted in factor 1 being aliased by the interaction between factors 2 and 3. In our example, this meant the estimator for main effect of pesticide A was biased by 2. We expect this larger bias by aliasing with two-factor interactions over three-factor interactions in general due to the hierarchy principle.

In terms of variance, consider the case with additive effects where all designs are balanced and have the same total sample size. The fractional factorial design uses all $2^{K-p}$ treatment groups in its estimators of effects, and thus all $n$ units. The incomplete design will use only a subset of the treatment combinations for estimating each effect, and therefore fewer than the total $n$ units. It should be clear that we expect lower variance under the fractional design, in which more units are used in estimation. Therefore, there is a bias-variance trade-off comparing the fractional factorial to the incomplete factorial design.

If we are instead comparing estimators corresponding to different hypothetical incomplete designs as in Table 3, it should be clear that approximately balanced designs should yield similar variances for estimators using the same number of treatments.

4. DISCUSSION: EMBEDDING OBSERVATIONAL STUDIES IN FRACTIONAL FACTORIAL DESIGNS

4.1. General Issues

To address causality in a nonrandomized study, it has been argued that one needs to conceptualize a hypothetical randomized experiment that could correspond to the observational data (Rosenbaum, 2002; Rubin, 2008; Stuart, 2010; Bind & Rubin, 2019). The hypothetical randomization is plausible if the treatment groups are “similar” with respect to confounding variables (Rubin, 2007, 2008). With multiple treatments of interest, we would aim to recreate a hypothetical factorial randomized experiment. However, in observational studies, there may be no units that received certain treatment combinations. Therefore, we propose to recreate a hypothetical fractional factorial or incomplete factorial experiment, which further motivates the usefulness of these designs beyond randomized experiments.

For the fractional factorial design, we must decide which fraction to use. This decision should be based on criteria such as the maximum resolution criterion (see Wu & Hamada, 2000, for more). To reduce the amount of aliasing, it is desirable to use a small $p$, i.e., a larger fraction of the total design is retained. In practice, we may not be able to control the aliasing structure. We must choose a $2^{K-p}$ design such that the treatment combination(s) with no observations is not used. However, this strategy usually results in the removal of units assigned to treatment
combinations that were present in the observational dataset but not used in the design. If only one treatment combination is not present in the dataset, we could use a $2^{K-1}$ design, but in that case we are not using $2^{K-1} - 1$ treatment combinations for which we have data.

An alternative strategy to using a subset of the data would be to use an unsaturated regression model. Zhao & Ding (2021) discuss this model from the potential outcome perspective, though not in the case where there are entire treatment groups missing from the data or containing only a single unit. Those authors found that the unsaturated model will result in bias, in general settings, even with all treatment groups present.

After a design is chosen, a strategy to balance covariates should be used to ensure that the units across treatment groups are similar. We assume strong ignorability, i.e., conditional on measured covariates, the assignment mechanism is individualistic, probabilistic, and unconfounded (Rosenbaum & Rubin, 1983). Then we can obtain unbiased causal estimates by analyzing the data as if it arose from a hypothetical randomized experiment after balancing.

Importantly, these assumptions apply to all treatment combinations in the final experimental design used. We must also assume that the treatment assignment is, at least hypothetically, manipulable such that all potential outcomes are well defined. This ensures that our estimands of interest are defined. See Appendix G in the Supplementary Material for more discussion. A similar argument must hold for the unconfoundedness assumption.

4.2. Covariate Balance

An important stage when estimating the causal effect of nonrandomized treatments is the design phase (Rubin, 2007, 2008). At this stage, we attempt to obtain a subset of units for which we can assume unconfoundedness of the treatment assignment. That is, units for which $P(W_i|Y_i(z),X_i) = P(W_i|X_i)$, where $X_i$ is an $m$-dimensional vector of covariates (Imbens & Rubin, 2015). Matching strategies, as reviewed by Stuart (2010), are often used to ensure no evidence of covariate imbalance between treatment groups.

With multiple treatments, matching can be difficult and methods for obtaining covariate balance specifically for factorial type designs require further exploration. There are extensions of propensity score to multiple treatments, notably the generalized propensity score (GPS) (Hirano & Imbens, 2004) and the covariate-balancing propensity score (CBPS) (Imai & Ratkovic, 2014). For dealing with observational studies with multiple treatments, Lopez & Gutman (2017) review techniques, including matching, and Bennett, Vielma & Zubizarreta (2020) use template matching, which matches units to a “template” population of units. Although neither of those works explores factorial (full or fractional) designs, it may be straightforward to extend their methods to this setting. Nilsson (2013) discusses matching in the $2^2$ design. In our data illustration in Section 4.4, we demonstrate a method of sequential trimming and checks on covariate balance. Tests for covariate balance are reviewed in Appendix B of the Supplementary Material.

One issue likely to occur with matching is the difficulty in achieving good balance across many groups. Matching could result in dropping a large number of units, leading to lower power to detect effects. Weighting techniques can avoid the issue of completely dropping units, but the weights may be unstable if there is little overlap among the treatment groups.

Due to challenges in obtaining full balance, a first step might be to obtain covariate balance between only two treatment groups, a task commonly done in the causal inference literature, and compare outcomes for these groups. To start, we could estimate the difference in outcomes between the units that were assigned level $+1$ for all factors and the units that were assigned $-1$ for all factors. Under certain assumptions, testing the difference in these groups can act as a global test for whether any effects of interest are significant. We discuss the test and assumptions that are needed for this simple comparison to be meaningful in Appendix C of the Supplementary Material.
4.3. Comparing Designs

There may be complications in the observational setting that make attempting to recreate one design easier or more desirable than another. If we use a different design for each estimator, as in Section 3.2, we are able to use more of the data than a fractional factorial design, but this can also incur a cost. If we have no observations for only one treatment combination, we would use all \(2^K - 1\) treatment combinations if we did an analysis for all effects and used a different design for each. This approach would require us to either first obtain balance among all \(2^K - 1\) treatment groups or to obtain balance among the treatment groups within each design separately. The former option may be difficult; as the number of treatment combinations grows, obtaining covariate balance across all treatment groups becomes increasingly difficult and may result in smaller and smaller sample sizes, especially if trimming is used. The latter option will make joint inferences more challenging because different units would be used in each analysis. Therefore, although these incomplete factorial designs may improve the bias of our estimators, the fractional factorial design in which we are using the same \(2^K-p\) rows may be more attractive in terms of obtaining covariate balance.

Further, estimators for different designs result in different types of bias due to aliasing of effects. In order to believe that the biases of main effect estimators are negligible, we must believe that the interactions aliased with that main effect are negligible. Researchers need to carefully unpack the aliasing structure for the design they choose in the observational setting in order to understand the assumptions necessary for unbiased estimation. This requires even more care and thought than a standard observational study.

4.4. Data Illustration

4.4.1. Data description

We now give an illustration of the implementation of our methods using data on pesticide exposure and BMI, the ratio between weight and height-squared. Previous studies have found an association between pesticide exposures and BMI (Buser, Murray & Scinicariello, 2014; Ranjbar et al., 2015). We use the 2003–2004 cycle of the National Health and Nutrition Examination Survey (NHANES) collected by the Centers for Disease Control and Prevention (CDC & NCHS, 2003-2004). We access the data via the \texttt{R} (R Core Team, 2017) package \texttt{RNHANES} (Susmann, 2016). We focus on four organochlorine pesticides, measured via a blood serum test and then dichotomized based on whether they were above (+1) or below (−1) the detection limit, as given in the NHANES dataset, as factors. Organochlorine pesticides are persistent in the environment and adverse health effects have been reported by the CDC (2009), making them an interesting group of pesticides to study.

The four pesticides we consider are \(\beta\)-hexachlorocyclohexane (beta-Hex), heptachlor epoxide (Hept Epox), mirex, and p,p'-dichlorodiphenyltrichloroethane (p,p'-DDT), chosen primarily on the basis of data availability and exposure rates. We removed 271 units with missing values of pesticide or BMI, noting that because this is an illustration and not intended to draw causal conclusions we simply drop those units. We decided to study a nonfarmer population as farmers are more likely to be exposed to pesticides than the general population and may also differ on other unobserved covariates that affect health outcomes. This is our first step to achieving covariate balance, leaving a dataset with 1259 observations (see Figure H.1 in the Supplementary Material). Importantly, as shown in Table 4, we have one treatment group with only one unit. Thus, using the full dataset can lead to unstable estimates and an inability to calculate Neyman variance.

4.4.2. Design phase

To show the process of how estimands change as we adjust our design phase, we consider three different hypothetical “experiment” designs: (I) a \(2^4\) factorial design, without adjusting
### Table 4: Design matrix and counts.

|     | Factor levels | Number of Obs. |
|-----|---------------|----------------|
|     | beta-Hex      | Hept Epox      | Mirex p,p'-DDT |
| $z_1$ | +1            | +1             | +1            | +1          | 426 | - |
| $z_2$ | -1            | +1             | +1            | +1          | 12  | 6  |
| $z_3$ | +1            | -1             | +1            | +1          | 70  | 22 |
| $z_4$ | -1            | -1             | +1            | +1          | 51  | -  |
| $z_5$ | +1            | +1             | -1            | +1          | 291 | 102|
| $z_6$ | -1            | +1             | -1            | +1          | 25  | -  |
| $z_7$ | +1            | -1             | -1            | +1          | 94  | -  |
| $z_8$ | -1            | -1             | -1            | +1          | 54  | 10 |
| $z_9$ | +1            | +1             | +1            | -1          | 21  | 8  |
| $z_{10}$ | -1           | +1             | +1            | -1          | 1   | -  |
| $z_{11}$ | +1           | -1             | +1            | -1          | 19  | -  |
| $z_{12}$ | -1           | -1             | +1            | -1          | 19  | 10 |
| $z_{13}$ | +1           | +1             | -1            | -1          | 42  | -  |
| $z_{14}$ | -1           | +1             | -1            | -1          | 19  | 3  |
| $z_{15}$ | +1           | -1             | -1            | -1          | 37  | 8  |
| $z_{16}$ | -1           | -1             | -1            | -1          | 78  | -  |
| $g_1$ |              | $g_2$          | $g_3$         | $g_4$     | 1259 | 169|

*Note:* $+1$ refers to exposure to the pesticide, $-1$ refers to no detectable exposure to the pesticide. Red treatments are used when recreating a fractional factorial design. “Original” column on the right gives the counts for each treatment for the full dataset. “Trimmed” column on the right gives counts for those treatments used in the fractional factorial design, after trimming.

the sample to balance for covariates; (II) a fractional factorial $2^{4-1}$ design, without adjusting the sample to balance for covariates; (III) a fractional factorial $2^{4-1}$, with trimming to obtain covariate overlap and no evidence of covariate imbalance with respect to gender (recorded as male vs. female), smoking status (smoker vs. nonsmoker), and age at the time of survey (in years). Comparing designs (I) and (II) will show how estimates change between a full factorial and a fractional factorial design. Comparing designs (II) and (III) will show how estimates change when the design phase induces covariate balance. We use simple trimming, because the focus of this article is not to compare different methods of balancing for covariates. However, exploration of different balancing methods in this setting would be worthwhile. We consider gender, age, and smoking status as our first tier of most important covariates. For smoking status, we use the question “Have you smoked at least 100 cigarettes in your life time?” and categorize “Yes” and “Don’t know” as smokers. Only one observation with value “Don’t know” was recorded.

### 4.4.3. Statistical analysis

We analyzed the three datasets described in the design stage using: (1) a multiple linear regression that regresses BMI on treatment factors; (2) a multiple linear regression that regresses BMI on treatment factors, in addition to the following covariate factors: race and ethnicity, income, income, income.
gender, and smoking status; (3) Fisher-randomization tests of the sharp null hypothesis of no treatment effects, approximated with the Monte-Carlo method. We will refer to the “race and ethnicity” covariate as simply ethnicity, as a short hand and to make clear that this is one categorical variable in the NHANES dataset. For the first analysis, recall that regression estimates when including all factors and interactions, but not covariates, correspond to the Neymanian estimates divided by 2. The second analysis is motivated by Figures H.2 and H.3 in Appendix H.3 of the Supplementary Material, which show that further adjustments for ethnicity and income are needed, even after trimming. The second analysis also adjusts for our second tier of covariates: ethnicity and income. Income is defined as annual household income. In the second analysis for each design, units with missing covariate values were removed. Due to the positive skewness of the weight variable, BMI exhibits some degree of positive skewness. Therefore, we use log-transformed BMI as the outcome in all of these analyses. Additional data descriptions and the full analyses are available in Appendix H of the Supplementary Material.

4.4.4. Resulting designs

Table 4 provides the counts of observations for each treatment $z_j$. Factor combination $z_{10}$ has only one observation. Relying on only one observation for a treatment will lead to unstable estimates and we will not be able to use Neymanian variance estimates. Hence, the fractional factorial design aims to avoid this issue by embedding the observational study in a fractional factorial hypothetical experiment.

To exclude $z_{10}$ from our analysis, we impose a $2^{4-1}$ fractional factorial design with the defining relation $I = -1234$ (recalling that $I$ corresponds to the intercept) instead of the more conventional $I = 1234$. The dataset in this hypothetical experiment consists of 523 observations. In this design, aliasing is as follows: $I = -1234, 4 = -123, 3 = -124, 2 = -134, 1 = -234, 12 = -34, 13 = -24, 14 = -23$. The main effects are aliased with the negative of the three-factor interactions and the two-factor interactions are aliased with each other with reversed signs. In order to identify main effects, we will assume that the three-factor interactions are negligible. In practice, researchers should assess whether this aliasing assumption is realistic.

For the third design, we attempted to achieve balance on the covariates of gender, smoking status, and age. Figure 1 shows covariate imbalance with respect to gender, smoking, and age in the fractional factorial design before trimming. Recall that these form our first tier of covariates, whereas ethnicity and income constitute the second tier, for which we adjust by linear regression. In practice, balancing a large set of covariates is not always possible and so choosing the most important (first tier) covariates should be done using subject matter knowledge. We used a rejection approach that sequentially pruned observations from the fractional factorial dataset until we found no evidence of imbalance across exposure groups with respect to our first tier of covariates, based on a MANOVA using the Wilks’ statistic (Wilks, 1932, see also Appendix B of the Supplementary Material). Figure 2 shows the covariate distribution for our first tier after trimming. Gender and smoking status are still unbalanced even after trimming, and hence their effects were accommodated by linear regression on the covariates in all “designs”. The first dataset that resulted in no evidence of covariate imbalance consisted of 169 observations, and the new treatment counts are presented in Table 4. The number of units has been drastically reduced in our attempt to achieve covariate balance, a major challenge in this setting. In fact, one treatment group has only three observations.

4.4.5. Results comparison across different conceptualized experiments and statistical approaches

Figure 3 shows a comparison of the regression estimates of the main effects across designs and statistical analyses. To compare the different methods, we present univariate analyses. That is, we utilize individual tests for each main effect rather than joint tests, for better illustration of
Figure 1: Comparing covariates across treatment combinations in the $2^{4-1}$ fractional factorial design for the 2003–2004 cycle of NHANES data (CDC & NCHS, 2003-2004). Text labels give number of observations per group. For bar plots, the $y$-axis gives the percent within each treatment combination for each category of the covariate. For the boxplots, individuals older than 85 were set to 85 on the graph. Note that all individuals older than 85 were dropped in the covariate balance stage.

DOI: 10.1002/cjs.11734 The Canadian Journal of Statistics / La revue canadienne de statistique
Figure 2: Comparing covariates across treatment combinations in the $2^{4-1}$ fractional factorial design after trimming for the 2003–2004 cycle of NHANES data (CDC & NCHS, 2003-2004). Text labels give number of observations per group. For bar plots, the $y$-axis gives the percent within each treatment combination for each category of the covariate. For the boxplots, individuals older than 85 were set to 85 on the graph. Note that all individuals older than 85 were dropped in the covariate balance stage.
FIGURE 3: Plots of estimates of factorial effects (associations), on the log BMI scale. Bars indicate two standard errors (using standard OLS estimates because the Neymanian variance estimates are not available for the full factorial design due to having only a single unit in one treatment group) above and below the point estimate.

the different methods. In practice, adjustment for multiple comparisons should be considered. All methods and designs estimated a positive “effect” of heptachlor epoxide and a negative “effect” of mirex on BMI. Although the full factorial estimates generally agree with the estimates of the two fractional factorial designs, differences in estimates of beta-Hexachlorocyclohexane (Beta-Hex) and \( p,p' \)-DDT may be due to the aliasing of the three-factor interactions with the main effects. However, it is also plausible that we have reduced our data in the fractional design to a subset of individuals with different average main effects than in the full dataset.

Figure 4 shows a comparison of the significance of these estimates. The Fisher \( P \)-value is the \( P \)-value for the randomization test of no effects of any pesticides, based on effect estimates for a given pesticide, which is suggested as a screening stage in Espinosa, Dasgupta & Rubin (2016). We obtained low \( P \)-values for the main effect of mirex on BMI across all methods and designs. However, the \( P \)-values disagree for the other pesticides, especially the \( P \)-values testing the effects of \( p,p' \)-DDT and beta-Hexachlorocyclohexane. The HC2/Neyman \( P \)-value is the significance based on the HC2 variance estimate (or Neyman variance estimate as we have shown this estimator to be equivalent in settings with no covariates) and the normal approximation. This \( P \)-value was calculated only for the regression analysis without covariates and was unavailable for the full factorial model due to limited data.

4.4.6. Discussion of data illustration

We have performed an illustration with data to show the benefits and challenges of using our method, as well as working more generally with observational data involving multiple treatments.

First, it is important to note that simplifications were made in the statistical analyses to focus on illustrating how researchers can capitalize on using fractional factorial designs to estimate the main and interactive effects of multiple treatments in observational studies. We did not adjust for all important hypothetical covariates, such as diet. Another consideration is that we...
log-transformed BMI to address the fact that BMI is a ratio and so its distribution tends to have positive skewness. However, other distributions could have been considered.

Applying our method to these data led to some major challenges related to sample size. Trimming helps mitigate bias but greatly reduced the sample size, potentially leading to decreased power and precision. There was also a great reduction in sample size from using a fractional factorial design. We could have used an incomplete-factorial-type design, but this likely would have made the covariate balancing even more challenging, as discussed previously.

For p,p'-DDT, the full factorial design resulted in a lower $P$-value than the other designs, which could be due to aliased interactions in the fractional design watering down the effect, as a result of the bias noted in Result 1. It could also have occurred because the populations are different in these two designs. For $\beta$-hexachlorocyclohexane, the covariate-balance-adjusted fractional factorial design has a lower $P$-value compared with the other designs. We are more inclined to trust the balanced design as it should have reduced bias. This result may indicate that there was some confounding that made the effect appear less significant before trimming. In fact, the stark contrast between the unbalanced and balanced fractional design suggests that confounding may be to blame. Alternatively, by trimming we may have reduced our sample to a subpopulation where $\beta$-hexachlorocyclohexane has a larger effect on BMI than the rest of the population.

It is important to acknowledge that using the small subset of the NHANES dataset, we do not intend to provide policy recommendations on pesticide use. In the general population, organochlorine pesticide exposure primarily occurs through diet (excluding those with farm-related jobs), particularly eating foods such as dairy products and fatty fish [Centers for Disease Control and Prevention (CDC), 2009]. Without further adjustments for diet, we are not able to disentangle the causal effect of diet and pesticides. For instance, in our study, individuals are likely to have been exposed to mirex largely through fish consumption [Agency for Toxic Substances and Disease Registry (ATSDR), 1995]. Further studies could investigate BMI differences between a group of fish consumers with high levels of mirex and a similar group of fish consumers with low levels of mirex, where similarity is based on important confounding variables. Indeed,
it could be that eating fish causes individuals to have both high levels of mirex and also lower BMI.

5. EPILOGUE

In this article, we have laid out how to design and analyze fractional factorial and incomplete factorial designs using the potential outcomes framework. These designs are particularly useful when running a full factorial design is infeasible or impractical, but the factorial estimands are still of interest. Our work includes extensions of some of the known factorial design results for variance and regression to fractional and incomplete factorial designs. In doing so, we have expanded the tool kit available to researchers wishing to perform causal inference.

Further, we have proposed to embed observational studies with multiple treatments in hypothetical fractional factorial experiments. This type of design is useful in settings with many treatments, especially when some treatment combinations have few or no observations and the aliasing assumptions are plausible. Once we recreate a factorial or fractional factorial experiment in the design phase, we can use standard methods, extended as in Sections 2 and 3, to estimate causal effects of interest. We first reviewed the basic setup for factorial and fractional factorial designs. We have also discussed covariate balance complications that may arise when dealing with multiple nonrandomized treatments in practice and illustrated these methods on a dataset with pesticide exposure and BMI.

We have given a short overview of factorial, fractional factorial, and incomplete factorial designs, in the potential outcomes framework. However, there are many related designs that we did not cover (see Wu & Hamada, 2000), such as nonregular design types (e.g., Plackett–Burman designs, Plackett & Burman, 1946). Practitioners may also use variable selection and the principle of effect heredity to select their model for estimating factorial effects.

We see many avenues of future exploration connected to our approach. For instance, coupling fractional factorial designs with a Bayesian framework would provide more statistical tools and would potentially offer different methodology for dealing with missing data. Additionally, development of observational tools, especially covariate-balancing techniques for factorial designs with many treatment combinations should be an area of future exploration. A particular challenge is that as we increase the number of treatment combinations, matching becomes more and more difficult because of increased dimensionality, whereas weighting methods may produce unstable estimators. Another direction would choose the design of the observational study based upon the ability to balance different treatment groups. Random allocation designs (Dempster, 1960, 1961), in which randomness of the design is incorporated, could also be utilized in this framework. Finally, we could explore other causal estimands that may be of interest in observational studies with multiple treatments, such as those in Egami & Imai (2019) and De la Cuesta, Egami & Imai (2022).

ACKNOWLEDGEMENTS

The authors thank Zach Branson, Kristen Hunter, Kosuke Imai, Xinran Li, Luke Miratrix, and Alice Sommer for their comments. They also thank Donald B. Rubin and Tirthankar Dasgupta for their insights and prior work that made this article possible. Additionally, they thank members of Marie-Abèle Bind’s research lab and Luke Miratrix’s CARES lab, as well as participants of STAT 300, the Harvard IQSS Workshop, and the Yale Quantitative Research Methods Workshop for their feedback on this project. Marie-Abèle Bind was supported by the John Harvard Distinguished Science Fellows Program within the FAS Division of Science of Harvard University and is supported by the Office of the Director, National Institutes of Health under Award Number DP5OD021412. Nicole Pashley was supported by the National Science...
Foundation Graduate Research Fellowship under Grant No. DGE1745303. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the National Science Foundation.

REFERENCES

Agency for Toxic Substances and Disease Registry (ATSDR). (1995). Public Health Statement for Mirex and Chlordecone, U.S. Department of Health and Human Services, Public Health Service, Atlanta, GA. https://www.atsdr.cdc.gov/phs/phs.asp?id=1189&ttd=276

Bennett, M., Vielma, J. P., & Zubizarreta, J. R. (2020). Building representative matched samples with multi-valued treatments in large observational studies. Journal of Computational and Graphical Statistics, 29(4), 744–757.

Bind, M.-A. C. & Rubin, D. B. (2019). Bridging observational studies and randomized experiments by embedding the former in the latter. Statistical Methods in Medical Research, 28(7), 1958–1978.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., & Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics, 16(3), 493–508.

Branson, Z., Dasgupta, T., & Rubin, D. B. (2016). Improving covariate balance in $2^K$ factorial designs via rerandomization with an application to a New York City Department of Education High School Study. The Annals of Applied Statistics, 10(4), 1958–1976.

Buser, M. C., Murray, H. E., & Scinicariello, F. (2014). Association of urinary phenols with increased body weight measures and obesity in children and adolescents. The Journal of Pediatrics, 165(4), 744–749.

Byar, D. P., Freedman, L. S., & Herzberg, A. M. (1995). Identifying which sets of parameters are simultaneously estimable in an incomplete factorial design. Journal of the Royal Statistical Society, Series D, 44(4), 451–456.

Byar, D. P., Herzberg, A. M., & Tan, W.-Y. (1993). Incomplete factorial designs for randomized clinical trials. Statistics in Medicine, 12(17), 1629–1641.

Centers for Disease Control and Prevention (CDC). (2009). Fourth Report on Human Exposure to Environmental Chemicals, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Atlanta, GA. https://www.cdc.gov/exposurereport/

Centers for Disease Control and Prevention (CDC), National Center for Health Statistics (NCHS) (2003-2004). National Health and Nutrition Examination Survey Data, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, Hyattsville, MD. Accessed Oct 1, 2018, https://www.cdc.gov/nchs/nhanes/nhanes/

Dasgupta, T., Pillai, N. S., & Rubin, D. B. (2015). Causal inference from $2^K$ factorial designs by using potential outcomes. Journal of the Royal Statistical Society, Series B, 77(4), 727–753.

Dasgupta, T. and Rubin, D. B. (2015). Harvard University STAT 240: Matched Sampling and Study Design Lecture Notes and Draft Textbook, Fall 2015.

De la Cuesta, B., Egami, N., & Imai, K. (2022). Improving the external validity of conjoint analysis: The essential role of profile distribution. Political Analysis, 30(1), 19–45. https://doi.org/10.1017/pan.2020.40

Dempster, A. (1960). Random allocation designs I: On general classes of estimation methods. The Annals of Mathematical Statistics, 31(4), 885–905.

Dempster, A. (1961). Random allocation designs II: Approximate theory for simple random allocation. The Annals of Mathematical Statistics, 32(2), 387–405.

Dong, N. (2015). Using propensity score methods to approximate factorial experimental designs to analyze the relationship between two variables and an outcome. American Journal of Evaluation, 36(1), 42–66.

Egami, N. & Imai, K. (2019). Causal interaction in factorial experiments: Application to conjoint analysis. Journal of the American Statistical Association, 114(526), 529–540.

Espinosa, V., Dasgupta, T., & Rubin, D. B. (2016). A Bayesian perspective on the analysis of unreplicated factorial experiments using potential outcomes. Technometrics, 58(1), 62–73.

Hirano, K. & Imbens, G. W. (2004). The propensity score with continuous treatments. In Gelman, A. & Meng, X. (Eds.), Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives, Wiley, Hoboken, NJ, 73–84.

The Canadian Journal of Statistics / La revue canadienne de statistique DOI: 10.1002/cjs.11734
Imai, K. & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society, Series B*, 76(1), 243–263.

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

Li, X. & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520), 1759–1769.

Lopez, M. J. & Gutman, R. (2017). Estimation of causal effects with multiple treatments: A review and new ideas. *Statistical Science*, 32(3), 432–454.

Lu, J. (2016a). Covariate adjustment in randomization-based causal inference for $2^K$ factorial designs. *Statistics & Probability Letters*, 119, 11–20.

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

Lu, J. & Deng, A. (2017). On randomization-based causal inference for matched-pair factorial designs. *Statistics & Probability Letters*, 125, 99–103.

MacKinnon, J. G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29(3), 305–325.

Montgomery, D. C. (2017). *Design and Analysis of Experiments*, Wiley, New York.

Mukerjee, R., Dasgupta, T., & Rubin, D. B. (2018). Using standard tools from finite population sampling to improve causal inference for complex experiments. *Journal of the American Statistical Association*, 113(522), 868–881.

Nilsson, M. (2013). *Causal Inference in a $2^2$ Factorial Design Using Generalized Propensity Score*. Master’s thesis, Uppsala University, Sweden.

Oulhote, Y., Bind, M.-A. C., Coull, B., Patel, C. J., and Grandjean, P. (2017). Combining ensemble learning techniques and G-computation to investigate chemical mixtures in environmental epidemiology studies. *bioRxiv*. https://doi.org/10.1101/147413

Patel, C. J., Bhatacharya, J., & Butte, A. J. (2010). An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. *PLoS ONE*, 5(5), e10746.

Plackett, R. L. & Burman, J. P. (1946). The design of optimum multifactorial experiments. *Biometrika*, 33(4), 305–325.

R Core Team. (2017). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Ranjbar, M., Rotondi, M. A., Ardern, C. I., & Kuk, J. L. (2015). The influence of urinary concentrations of organophosphate metabolites on the relationship between BMI and cardiometabolic health risk. *Journal of Obesity*, 2015, 687914.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.

Rosenbaum, P. R. (2002). *Observational Studies*, 2nd ed., Springer, New York.

Rosenbaum, P. R. & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1), 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688–701.

Rubin, D. B. (1980). Comments on “Randomization analysis of experimental data: The Fisher randomization test” by D. Basu. *Journal of the American Statistical Association*, 75(371), 591–593.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840.

Susmann, H. (2016). *RNHANES: Facilitates Analysis of CDC NHANES Data*. R package version 1.1.0. http://github.com/silentspringinstitute/RNHANES

Valeri, L., Mazumdar, M. M., Bobb, J. F., Claus Henn, B., Rodrigues, E., Sharif, O. I. A., Kile, M. L. et al. (2017). The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes.
at 20-40 months of age: Evidence from rural Bangladesh. *Environmental Health Perspectives*, 125(6), 067015.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24(3/4), 471–494.

Wu, C. F. J. & Hamada, M. S. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*, Wiley, New York.

Zhao, A. & Ding, P. (2021). Regression-based causal inference with factorial experiments: Estimands, model specifications, and design-based properties. *Biometrika*, 109(3), 799–815. https://doi.org/10.1093/biomet/asab051

Zhao, A., Ding, P., Mukerjee, R., & Dasgupta, T. (2022). Randomization-based causal inference from split-plot designs. *The Annals of Statistics*, 46(5), 1876–1903.

---

*Received 19 March 2021
Accepted 26 January 2022*