A Bayesian Perspective on the Analysis of Unreplicated Factorial Experiments Using Potential Outcomes

Valeria Espinosa, Tirthankar Dasgupta and Donald B. Rubin

Department of Statistics, Harvard University

Abstract

Unreplicated factorial designs have been widely used in scientific and industrial settings, when it is important to distinguish "active" or real factorial effects from "inactive" or noise factorial effects used to estimate residual or "error" terms. We propose a new approach to screen for active factorial effects from such experiments that utilizes the potential outcomes framework and is based on sequential posterior predictive model checks. One advantage of the proposed method is its ability to broaden the standard definition of active effects and to link their definition to the population of interest. Another important aspect of this approach is its conceptual connection to Fisherian randomization tests. Extensive simulation studies are conducted, which demonstrate the superiority of the proposed approach over existing ones in the situations considered.

1 Introduction

Two-level full factorial designs have been used extensively in engineering and industrial applications. In many situations, where the unit-to-unit variation is reasonably assumed negligible, unreplicated factorial designs are used to reduce expenses. Discriminating between active and inactive effects is the crucial goal of screening experiments, which has been broadly studied by researchers. In particular, this is a non-trivial problem for unreplicated experiments because the residual mean square error cannot be estimated in the usual way. Hamada & Balakrishnan (1998) offer an extensive review and comparison of many methods, most of which rely on the assumption that the estimated factorial effects are independently and identically distributed (iid) normal random variables (e.g., the commonly used Lenth (1989) approach and Dong's (1993) method). Loughin &

Noble (1997) proposed a method based on permutation tests, which is not included in the Hamada & Balakrishnan (1998) review. Focusing on the control of the false discovery rate (FDR) instead of the experimentwise error rate (EER), Tripolski et al. (2008) proposed and compared modifications of the Lenth (1989) and Dong (1993) methods. Bayesian approaches have also been proposed for screening factorial effects, for example Box & Meyer (1986) and Chipman et al. (1997), where active and inactive effects are distinguished by their variances: the standard deviation of active effects is assumed to be at least k times larger than the standard deviation of inactive effects, where k is a pre-set integer (both Box & Meyer (1986) and Chipman et al. (1997) suggest using $k \approx 10$). The former was included in the Hamada & Balakrishnan (1998) study but did not perform as well as Lenth's (1989) method. The approach proposed by Chipman et al. (1997) is, in principle, similar to the one proposed by Box & Meyer (1986), but has greater flexibility in terms of incorporating prior information through the effect heredity and effect hierarchy principles (Wu & Hamada (2009), Ch. 4).

In these formulations, the definition of an "active" effect is somewhat vague because it is not related to the experimental units or the population of units of interest to an experimenter but to model parameters. Consider, for example, experiments involving growth of nanostructures on substrates of silicon (Dasgupta et al., 2008), which is somewhat analogous to the yield of crops on plots of lands. Suppose that we are interested in whether a change of temperature increases the yield on one or more substrates. Because the inference made from a small population of substrates in a particular laboratory is difficult to generalize to a larger population of substrates in other laboratories, a natural question is whether the temperature affects the yield of at least one of the substrates used for experimentation. Further, if we visualize a *potential yield* of each substrate for each level of temperature, then we may be interested in a summary of the distribution of these potential yields across units, e.g., the median or a percentile, rather than the average. None of the existing methods can directly address such questions.

Our view is that assessing significance of factorial effects without first defining both (a) the population of experimental units for which the inference is made and (b) the estimand, is inappropriate for addressing *causal inference* questions, which is the sole objective when conducting screening experiments. We propose a Bayesian approach for screening active factorial effects, which addresses limitations of current procedures by utilizing the concept of potential outcomes that lies at the center stage of causal inference (Rubin, 1974, 1980). Although such a framework for single-factor

experiments with two levels is well-developed and popularly known as the Rubin Causal Model, RCM (Holland, 1986), it is not yet fully exploited for multiple-factor experiments. A theoretical framework for causal inference from two-level factorial designs has recently been proposed by Dasgupta et al. (2012), which addresses inference for factorial effects assuming replicated factorial experiments, but does not consider unreplicated experiments.

In the next section, we describe how the RCM can be applied to two-level factorial designs. In Section 3, we describe the Fisher randomization test (Fisher, 1925, 1935) using the potential outcomes framework, extend it to the setting of two-level factorial designs, and describe its connection to the permutation tests proposed by Loughin and Noble (1997). In Section 4, we show how the Fisherian approach to causal inference can be naturally extended to a Bayesian approach for screening factorial effects and propose a method based on sequential posterior predictive checks. In Section 5, we go over a simple numerical example of the procedures outlined in Sections 3 and 4. In Section 5 we demonstrate the usefulness of our method in a super population setting by first calibrating the proposed algorithm to achieve the desired experimentwise error rate (EER), and then by comparing its performance to that of existing methods for screening factorial effects. Some concluding remarks are presented in Section 6.

2 RCM for two-level factorial designs

For simplicity, an unreplicated 2^2 design is used to introduce the concepts, even though such designs are rarely used. Here, each of two treatment factors takes one of two levels, typically denoted by -1 and 1, and thus, there are four treatment combinations denoted by $\boldsymbol{z} = (-1,-1), (-1,1), (1,-1)$ and $(1,1)$, and four experimental units.

2.1 Potenital outcomes and factorial effects

Let $Y_i(\boldsymbol{z}), i = 1, \ldots, 4$, denote the potential outcome of the ith unit if exposed to treatment combination $\boldsymbol{z}$. Note that when introducing this notation, we accept the stable unit treatment value assumption (Rubin, 1980), SUTVA, which means that the potential outcome of unit i depends only on the treatment combination it is assigned, and NOT on the assignments of the other units, and that there are no hidden versions of treatments not represented by these four combinations. Thus the ith unit has four potential outcomes $Y_i(-1,-1), Y_i(-1,1), Y_i(1,-1), Y_i(1,1)$, which comprise

Unit	Potential outcome for treatment combination				Unit-level factorial effects			
(i)	$(-1,-1)$	$(-1,1)$	$(1,-1)$	$(1,1)$	$\theta_{i,0}$	$\theta_{i,1}$	$\theta_{i,2}$	$\theta_{i,3}$
1	$Y_1(-1,-1)$	$Y_1(-1,1)$	$Y_1(1,-1)$	$Y_1(1,1)$	$\theta_{1,0}$	$\theta_{1,1}$	$\theta_{1,2}$	$\theta_{1,3}$
2	$Y_2(-1,-1)$	$Y_2(-1,1)$	$Y_2(1,-1)$	$Y_2(1,1)$	$\theta_{2,0}$	$\theta_{2,1}$	$\theta_{2,2}$	$\theta_{2,3}$
3	$Y_3(-1,-1)$	$Y_3(-1,1)$	$Y_3(1,-1)$	$Y_3(1,1)$	$\theta_{3,0}$	$\theta_{3,1}$	$\theta_{3,2}$	$\theta_{3,3}$
4	$Y_4(-1,-1)$	$Y_4(-1,1)$	$Y_4(1,-1)$	$Y_4(1,1)$	$\theta_{4,0}$	$\theta_{4,1}$	$\theta_{4,2}$	$\theta_{4,3}$
Average	$\overline{Y}(-1,-1)$	$\overline{Y}(-1,1)$	$\overline{Y}(1,-1)$	$\overline{Y}(1,1)$	θ_0	θ_1	θ_2	θ_3

Table 1: The *science* (left portion) and factorial effects (right portion) for the 2^2 experiment.

the 1×4 row vector $\boldsymbol{Y}_i$. We define *the Science* as the 4×4 matrix $\boldsymbol{Y}$ of potential outcomes whose ith row is the 4-component row vector $\boldsymbol{Y}_i$ as shown in the left portion of Table 1. Only one potential outcome in each row of the Science is actually observed from an experiment, and the remaining three are missing, making the causal inference problem essentially a missing data problem.

For each unit, all levels of every factor (e.g., 1 or -1) appear in half of its potential outcomes. Therefore, at the unit-level we are generally interested in contrasting one half of the unit's potential outcomes with the other half. For example, the difference between the average of the potential outcomes when factor 1 is at its high level (1) and when at its low level (-1), is the so-called "main effect of factor 1". Of course, other definitions could be the difference in medians, or the difference in the logarithm of the averages, but the tradition in the study of factorial experiments is to deal with the difference of averages, and we adhere to this focus here. Consequently, we define three unit-level factorial effects representing the main effects of the two factors and their interactions as three contrasts (denoted by $\theta_{i,1}, \theta_{i,2}$ and $\theta_{i,3}$ respectively) of elements of the vector $\boldsymbol{Y}_i$:

$$\begin{aligned}
\theta_{i,1} &= \frac{Y_i(1,-1)+Y_i(1,1)}{2} - \frac{Y_i(-1,-1)+Y_i(-1,1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_1, \\
\theta_{i,2} &= \frac{Y_i(-1,1)+Y_i(1,1)}{2} - \frac{Y_i(-1,-1)+Y_i(1,-1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_2, \\
\theta_{i,3} &= \frac{Y_i(-1,-1)+Y_i(1,1)}{2} - \frac{Y_i(-1,1)+Y_i(1,-1)}{2} = \frac{1}{2}\boldsymbol{Y}_i\boldsymbol{g}_3, \qquad (1)
\end{aligned}$$

where $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ and $\boldsymbol{g}_3$ are the three mutually orthogonal contrast column vectors $(-1,-1,1,1)'$, $(-1,1,-1,1)'$ and $(1,-1,-1,1)'$. For completeness, we label the vector generating the ith unit's mean potential outcome as $\boldsymbol{g}_0 = (1,\ldots,1)'$, which is orthogonal to $\boldsymbol{g}_1$, $\boldsymbol{g}_2$ and $\boldsymbol{g}_3$, so that unit i's

average potential outcome is

$$\theta_{i,0} = \frac{Y_i(-1,-1) + Y_i(-1,1) + Y_i(1,-1) + Y_i(1,1)}{4} = \frac{1}{4}\boldsymbol{Y}_i\boldsymbol{g}_0.$$

Defining

$$\boldsymbol{\theta}_i = (\theta_{i,0}, \theta_{i,1}/2, \theta_{i,2}/2, \theta_{i,3}/2)$$

as the 1×4 row vector of unit-level factorial effects, as shown in the last four columns of Table 1, and denoting by $\boldsymbol{G}$ the 4×4 matrix whose columns are the vectors $\boldsymbol{g}_0, \boldsymbol{g}_1, \boldsymbol{g}_2$ and $\boldsymbol{g}_3$, from (1) it follows that $\theta_i = \frac{1}{4}\boldsymbol{Y}_i\boldsymbol{G}$. Noting that $\boldsymbol{G}$ is an orthogonal matrix satisfying $\boldsymbol{G}'\boldsymbol{G} = 4I$, we have

$$\boldsymbol{Y}_i = \boldsymbol{\theta}_i\boldsymbol{G}', \tag{2}$$

implying that $\boldsymbol{Y}_i$ and $\boldsymbol{\theta}_i$ are linear transformations of each other. The matrix $\boldsymbol{G}$ is often referred to as the model matrix (Wu & Hamada (2009), Ch. 4).

Consistent with the traditional definition of causal effects in the factorial design literature, the causal estimands at the *population level* are the averages of the unit level factorial effects. These quantities, denoted by θ_1, θ_2 and θ_3, are the population level main effects of each treatment factor and their interaction respectively, and can be expressed in terms of potential outcomes as:

$$\theta_j = \frac{\sum_{i=1}^{4}\theta_{i,j}}{4} = \frac{1}{2}\overline{\boldsymbol{Y}}\boldsymbol{g}_j, \quad j = 1, 2, 3, \tag{3}$$

where

$$\bar{\boldsymbol{Y}} = \frac{1}{4}\sum_{i=1}^{4}\boldsymbol{Y}_i. \tag{4}$$

As in the case of unit-level effects, letting $\theta_0 = \sum_{i=1}^{N}\theta_{i,0}/4 = \frac{1}{4}\overline{\boldsymbol{Y}}\boldsymbol{g}_0$, and the row vector of population-level estimands by $\boldsymbol{\theta} = (\theta_0, \theta_1/2, \theta_2/2, \theta_3/2)$, it follows that

$$\overline{\boldsymbol{Y}} = \boldsymbol{\theta}\boldsymbol{G}'. \tag{5}$$

2.2 Definition of active and inactive effects

Under the potential outcomes perspective, there are different possible definitions for *active* effects. Here we call a factorial effect *active* if it is non-zero for at least one unit of the finite population.

Thus, in the 2^2 case, we call the jth factorial effect (corresponding to the vector $\boldsymbol{g}_j$) *active* if $\theta_{i,j} \neq 0$ for some i in $\{1, \ldots, 4\}$, and *inactive* if $\theta_{i,j} = 0$ for all $i = 1, \ldots, 4$. Let $\mathcal{A}$ denote the set of active effects with cardinality a where $0 \leq a \leq 3$. Also, let $\bar{\mathcal{A}}$ denote the set of effects that are inactive with cardinality $3 - a$.

The potential outcomes framework also allows us to define active/inactive effects in terms of the finite population factorial effects and super-population factorial effects. For example, although we do not pursue these ideas here, the jth factorial effect could be called inactive at the finite population level if $\theta_j = 0$ (see (3) for definition of θ_j), and at the super population level if the expectation of $\theta_{i,j}$ is zero. In Section 5, we show that the finite population approach is robust enough to perform well in super-population settings, where inferences have more uncertainty.

2.3 Assignment mechanism and observed outcomes

The treatment assignment mechanism selects the subset of potential outcomes that will be revealed and observed; the other potential outcomes are missing. Let

$$W_i(\boldsymbol{z}) = \begin{cases} 1 & \text{if the } i\text{th unit is assigned to } \boldsymbol{z} \\ 0 & \text{otherwise.} \end{cases}$$

For an unreplicated completely randomized 2^2 factorial experiment, $Pr(W_i(\boldsymbol{z}) = 1) = 1/4$. Also, $\sum_z W_i(\boldsymbol{z}) = 1$ for $i = 1, \ldots, 4$, and $\sum_i W_i(\boldsymbol{z}) = 1$ for all $\boldsymbol{z}$. Let $w_i = \sum_{\boldsymbol{z}} \boldsymbol{z} W_i(\boldsymbol{z})$ be the treatment combination that the ith subject receives. Let $\boldsymbol{W}$ be the generic treatment assignment vector of random variables, and let $\boldsymbol{w}$ be a specific realization of $\boldsymbol{W}$, i.e., a vector comprising all the individual treatment assignments *after* randomization. Hence, each W_i is a random variable, and their joint probability distribution defines the treatment assignment mechanism of $\boldsymbol{W}$. The vector of post randomization treatment assignments, $\boldsymbol{w}$, is a draw from this distribution.

Denote the observed outcome corresponding to the ith experimental unit by $Y_i^{\text{obs}}, i = 1, \ldots, 4$, so that

$$Y_i^{\text{obs}} = \sum_{\boldsymbol{z}} W_i(\boldsymbol{z}) Y_i(\boldsymbol{z}), \tag{6}$$

and let $\boldsymbol{Y}^{\text{obs}} = (Y_i^{\text{obs}}, \ldots, Y_4^{\text{obs}})'$ be the 4×1 column vector of observed outcomes. For treatment assignment $\boldsymbol{w} = \big((-1,-1),(1,1),(1,-1),(-1,1)\big)'$, the table of *observed* potential outcomes is displayed in Table 2; the missing potential outcomes are represented by question marks.

Unit	observed outcome for treatment combination				$\boldsymbol{w}$
(i)	$(-1,-1)$	$(-1,\ \ 1)$	$(\ \ 1,-1)$	$(\ \ 1,\ \ 1)$	
1	Y_1^{obs}	?	?	?	$(-1,-1)$
2	?	?	?	Y_2^{obs}	$(\ \ 1,\ \ 1)$
3	?	?	Y_3^{obs}	?	$(\ \ 1,-1)$
4	?	Y_4^{obs}	?	?	$(-1,\ \ 1)$

Table 2: *Observed Outcomes* for the 2^2 experiment with $\boldsymbol{w} = \big((-1,-1),(1,1),(1,-1),(-1,1)\big)'$.

Let $Y^{\text{obs}}(\boldsymbol{z})$ denote the observed outcome for treatment combination $\boldsymbol{z}$, and let

$$\tilde{\boldsymbol{Y}}^{\text{obs}} = \big(Y^{\text{obs}}(-1,-1), Y^{\text{obs}}(-1,1), Y^{\text{obs}}(1,-1), Y^{\text{obs}}(1,1)\big)'$$

denote the permutation of components of $\tilde{\boldsymbol{Y}}^{\text{obs}}$ arranged according to the lexicographic ordering of the treatment combinations: $(-1,-1),(-1,1),(1,-1),(1,1)$.

2.4 Inference from data

Unbiased estimators of population-level factorial effects θ_j defined by (3) are:

$$\hat{\theta}_j = \frac{1}{2}\tilde{\boldsymbol{Y}}^{\text{obs}}\boldsymbol{g}_j, \quad j = 1,2,3. \tag{7}$$

Denoting the vector of these estimators by $\hat{\boldsymbol{\theta}} = (\bar{Y}^{\text{obs}}, \hat{\theta}_1/2, \hat{\theta}_2/2, \hat{\theta}_3/2)$, it follows that

$$(\tilde{\boldsymbol{Y}}^{\text{obs}})' = \hat{\boldsymbol{\theta}}\boldsymbol{G}', \text{ or equivalently } \tilde{\boldsymbol{Y}}^{\text{obs}} = \boldsymbol{G}\hat{\boldsymbol{\theta}}',$$

from which, using the identity $\boldsymbol{G}'\boldsymbol{G} = 4\boldsymbol{I}$, we have

$$\hat{\boldsymbol{\theta}}' = (\boldsymbol{G}'\boldsymbol{G})^{-1}\boldsymbol{G}'\tilde{\boldsymbol{Y}}^{\text{obs}}. \tag{8}$$

The unbiased *point estimator* of $\boldsymbol{\theta}$ given in (8) is the same as that obtained by ordinary least squares in the classical linear model with an additive iid error. However, to perform inference, such as interval estimation or significance tests using the least squares framework, we need to justify the assumption of additive errors and asymptotic normality of the estimators. In our current set-up, the potential outcomes are considered fixed, and the estimands are functions of potential outcomes from a finite population of units. Therefore, the only sampling distribution that arises is the one

induced by randomization.

Inferential methods such as Lenth's test, that are based on asymptotic normality, do not distinguish between the following two situations: (i) the units considered here are the only ones comprising the population of interest and the estimands θ_j's are those defined by (3), versus (ii) the units in the experiment are randomly sampled from an infinitely large super-population where the estimands are counterparts of θ_j's for this super population. Methods that rely on asymptotic normality typically aim at situation (ii).

There are two standard inferential methods that address situation (i) under the potential outcomes framework: (a) the Neymanian approach and (b) the Fisherian approach. We will not pursue the Neymanian approach here beyond the unbiasedness of estimators, such as in (8); see Dasgupta et al. (2012) for details. However, we do pursue the Fisherian approach because it has a natural connection to the permutation tests proposed by Loughin & Noble (1997) and because it has a natural Bayesian justification, which is the basis for the approach proposed in Section 4.

2.5 Generalization to 2^K experiments when $K > 2$

We conclude this section by noting that the framework just illustrated can easily be extended to a 2^K design for $K > 2$. For such a design, we have K factors labeled $1, \ldots, K$; $N = 2^K$ experimental units; and $N - 1$ mutually orthogonal contrast vectors $\boldsymbol{g}_j$, each representing a factorial effect θ_j. The left portion of Table 1 - the science - is generalized by following the *lexicographic order* of the treatment combinations specified by the levels of factors 1 through K, starting with the value -1 and then 1; that is, the first $N/2$ entries in the first column are -1's and the last $N/2$ entries are 1's. In the second column, the sign is switched after every $N/4$ entries, and so on, until the last column switches sign with every entry. Hence, for $j = 1, \ldots, K$, the construction of the vector $\boldsymbol{g}_j$ corresponding to the levels of factor j for each treatment combination consists of defining the first $N/2^j$ entries to be -1, the next $N/2^j$ entries to be 1, and repeating 2^{j-1} times (until the N entries of $\boldsymbol{g}_j$ are defined). The first treatment combination (and first potential outcome listed in a generalized science table) corresponds to the vector $\boldsymbol{z}$ defined by the first entry of each of these K $\boldsymbol{g}_j$'s, the second treatment combination corresponds to the second entries, and so on. After defining the first K $\boldsymbol{g}_j$ vectors (and the corresponding θ_j's) that represent the K main effects, the interaction terms are defined as products of $\boldsymbol{g}_1, \ldots, \boldsymbol{g}_K$, such that the next $\binom{K}{2}$ $\boldsymbol{g}_j$ vectors correspond to the two-factor interactions, and eventually the $\boldsymbol{g}_{N-1}$ vector represents the K-factor interaction. Thus,

for example, for a 2^4 design, the factorial effect θ_3 represents the main effect of factor 3, θ_5 the interaction between factors 1 and 2, θ_{10} the two-factor interaction between factors 3 and 4, and θ_{15} the four factor interaction. The $\boldsymbol{g}_j, j = 1, \ldots, N-1$ vectors define the model matrix $\boldsymbol{G}$, an $N \times N$ orthogonal matrix with the N-vector $\boldsymbol{g}_0 = (1, \ldots, 1)'$ as its first column, and such that $\boldsymbol{G}'\boldsymbol{G} = 1/2^K \boldsymbol{I}_{2^K} = 1/N \boldsymbol{I}_{2^K}$ where $\boldsymbol{I}_{2^K}$ is the 2^K identity matrix. Thus, in the generalized case, the divisors on the right hand side of (1) and (3) change from 2 to 2^{K-1}. The jth factorial effect (corresponding to the vector $\boldsymbol{g}_j$) is *active* if $\theta_{i,j} \neq 0$ for some i in $\{1, \ldots, N\}$, and *inactive* if $\theta_{i,j} = 0$ for all $i = 1, \ldots, N$. The cardinality a of the set of active effects $\mathcal{A}$ is such that $0 \leq a \leq N-1$, and the cardinality of the inactive set $\bar{\mathcal{A}}$ is $N - 1 - a$. The assignment mechanism is defined as in (2.3), where $Pr(W_i(\boldsymbol{z}) = 1) = 1/N, \quad \sum_z W_i(\boldsymbol{z}) = 1$ for $i = 1, \ldots, N$, and $\sum_i W_i(\boldsymbol{z}) = 1$ for all $\boldsymbol{z}$. The vector of observed outcomes $\boldsymbol{Y}^{\text{obs}}$ defined in (6) consists of N entries. The vector $\tilde{\boldsymbol{Y}}^{\text{obs}}$ is the permutation of $\boldsymbol{Y}^{\text{obs}}$ that follows the *lexicographic order* of the treatment combinations described above. The point estimates described in Section 2.4 are generalized by changing in the denominator of 7) from 2 to 2^{K-1} and the generalized version of the matrix $\boldsymbol{G}$. The unbiasedness of the estimators $\hat{\theta}_j$ under its randomization distribution for a general 2^K factorial design has established by Dasgupta et al. (2012).

3 Fisherian approach: randomization tests for a sharp null hypothesis and extensions to "data-dependent sharp" null hypotheses.

Randomization tests (Fisher 1925, 1935) are useful tools because they assess statistical significance of treatment effects from randomized experiments using a test statistic without making any assumption whatsoever about its distribution. Such tests can be used to test Fisher's *sharp null hypothesis* (see Rubin (1980)) of no factorial effects at the unit levels, which is a much stronger hypothesis than the traditional one of no average factorial effects. Randomization tests apparently have not been studied in the context of factorial experiments, except for the work by Loughin & Noble (1997), who studied such tests in the framework of a linear regression model for the observed response with additive error (in other words, invoking the assumption of strict additivity).

3.1 Randomization test for the sharp null hypothesis of no treatment effect

The plausibility of the sharp null hypothesis can be assessed using a randomization test of a suitable test statistic T of choice. That is, assuming the sharp null hypothesis of no treatment effect and keeping the units fixed, as well as their potential outcomes, we compare the observed value of T, T^{obs}, to the values we would have observed under hypothetical replications of the experiment. The collection of such values is also called the statistic's sampling distribution induced by the randomization under the sharp null hypothesis. To obtain this distribution we enumerate all possible treatment assignments under the actual assignment mechanism (if the number of such assignments is very large, a sample can be considered). The assumption that the sharp null hypothesis is true permits us to complete the table of all missing potential outcomes using only the observed data. For example, if the sharp null hypothesis of no treatment effect on unit i is true, then all the missing potential outcomes for the ith unit are equal to the observed one, Y_i^{obs}. Given the complete table of potential outcomes determined by the observed outcomes and the sharp null hypothesis, for each new draw of $\boldsymbol{W}$, $\boldsymbol{w}^{rep}$, we calculate the value of T that would have been observed, say T^{rep}. The proportion of such values of T (out of the total number of possible randomizations) that are as extreme or more extreme than T^{obs} is the p-value (i.e., significance level) of the test statistic under the null hypothesis. The smaller the p-value, the greater is the degree of belief that the null hypothesis is not true, because the probability of that one observed result, even when the probability of that observed event is combined with the probabilities of all of the more extreme results, would still be a rare event.

Many possible test statistics could be used. It might be of interest to use each estimated factorial effect as a test statistic because each defines a specific type of deviation from the null hypothesis. Therefore, a Fisher test can be obtained by comparing each estimated factorial effect against its randomization distribution resulting from the sharp null hypothesis of no treatment effects (note that due to the symmetry of the definition of all factorial effects, the reference distribution under the sharp null hypothesis of no effects is the same for all estimates of factorial effects). In this approach, one can adjust for multiple comparisons, for example, by using the Bonferroni correction, which simply modifies the significance value for each test to $\alpha/(N-1)$ for an EER less than or equal to α. However, if interest is in assessing if there is a deviation from the null without a particular direction in mind, a commonly used statistic in the presence of replicates is the F-statistic associated with the decomposition of the total sum of squares of the observed outcomes. Another reasonable option

is to use the estimated factorial effect $\hat{\theta}_{\max}$ that has the largest absolute value among all the $N-1$ estimated effects $\hat{\theta}_j, \quad j = 1, \ldots, N-1$ defined by (7). A version of $\hat{\theta}_{\max}$ that is scaled by Lenth's (1989) pseudo-standard error (see Section 4.3 for the definition) can also be considered.

When screening for active effects, these last two test statistics are natural choices. Once the sharp null hypothesis described above is rejected (e.g., the factorial effect that corresponds to $\hat{\theta}_{\max}$ is considered active), adopting a sequential approach for further exploration of factorial effects is also natural.

3.2 Single imputation in a sequence of "data-dependent sharp" null hypotheses

The problem with applying a strict Fisher test in the screening context is that as soon as we declare an effect active, we no longer have a sharp null hypothesis, and therefore can no longer impute the missing potential outcomes using just that null hypothesis and the observed potential outcomes. To generalize the test of the sharp null hypothesis described in Section 3.1, at every step a new data-driven null hypothesis that takes into consideration the effects already identified as active can be defined. Specifically, as in Loughin & Noble (1997), the first step can be the same as Fisher's sharp null test described in Section 3.1. The subsequent steps are designed to allow imputation of unobserved potential outcomes incorporating the existence of the identified factorial effect. One approach is to assume the estimated active effects are the actual ones, for example, using the unbiased point estimate of the factorial effect that corresponds to $\hat{\theta}_{\max}$, treating it as known with certainty, and then imputing the missing potential outcomes followed by a randomization test. Alternatively, we could modify this process and, assuming that each unit's residual is the same for all treatment combinations, use randomization tests of residuals as in Loughin & Noble (1997). We believe a more principled approach is to draw repeatedly the active effects from their joint posterior distribution and thereby multiply-impute the missing potential outcomes, which is what we propose in Section 4. However, to help clarify the use of the potential outcomes in the unreplicated experiments context, we now describe the first approach, which assumes that the estimated active effects are the actual ones.

Let $\dot{\boldsymbol{\theta}}_i$ be the $(N-1)$-vector of unit level factorial effects $\boldsymbol{\theta}_i$ without its first element θ_{i0}. At step s, let $\mathcal{A}_s$ and $\bar{\mathcal{A}}_s$ denote the sets of active effects and inactive effects, respectively (note that here $a = s$, and $s = 0$ is equivalent to the sharp null hypothesis of no treatment effects). Let $\dot{\boldsymbol{\theta}}^{(s)}$ be a $(N-1)$-component row vector with s non-zero entries that correspond to the s factorial effects in

$\mathcal{A}_s$, and zeros for the $N-s-1$ factorial effects in $\bar{\mathcal{A}}_s$. When $s = N-1$, all factorial effects take non zero values, and the largest estimated absolute effect among those effects assumed inactive is not defined. Then we can express the s-th step of a sequence of $N-2$ data-dependent sharp null hypotheses by $H_{0s} : \dot{\boldsymbol{\theta}}_i = \dot{\boldsymbol{\theta}}^{(s)} \ \ \forall \ \ i = 1, \ldots, N$.

Recall that the experimenter observes only one potential outcome for the ith experimental unit, Y_i^{obs}. Let $\boldsymbol{Y}_i^{\text{mis}}$ denote the $(N-1)$-component row vector of unit i's missing potential outcomes. Denote by $\boldsymbol{g}_i^{'\text{obs}}$ the column of $\boldsymbol{G}'$ that contains the treatment combination $\boldsymbol{z}$ assigned to unit i, and let the submatrix formed by the remaining $N-1$ columns be $\boldsymbol{G}_i^{'\text{mis}}$. Then from (2) we can write

$$(Y_i^{\text{obs}}, \boldsymbol{Y}_i^{\text{mis}}) = \left(\theta_{i,0}, \dot{\boldsymbol{\theta}}^{(s)}\right) (\boldsymbol{g}_i^{'\text{obs}} : \boldsymbol{G}_i^{'\text{mis}}) \tag{9}$$

Also, let $\tilde{\boldsymbol{g}}_i^{'\text{obs}}$ be the column vector $\boldsymbol{g}_i^{'\text{obs}}$ without its first element (which is unity). Imputation of the vector of missing potential outcomes $\boldsymbol{Y}_i^{\text{mis}}$ under H_{0s} requires two simple steps. Similarly, let $\dot{\hat{\boldsymbol{\theta}}}^{(s)}$ be $\hat{\boldsymbol{\theta}}$, defined in (8), without its first element. First, from (9), it follows that θ_{i0} can be estimated as $\hat{\theta}_{i0} = Y_i^{\text{obs}} - \dot{\hat{\boldsymbol{\theta}}}^{(s)} \tilde{\boldsymbol{g}}_i^{'\text{obs}}$. Then, the missing potential outcomes for the ith unit are singly imputed using $Y_i^{\text{mis}} = \left(\hat{\theta}_{i0}, \dot{\hat{\boldsymbol{\theta}}}^{(s)}\right) \boldsymbol{G}_i^{'\text{mis}}$, which again follows from (9).

Given the observed potential outcomes and the imputed missing potential outcomes, the values of the test statistic that would have been observed for each of the possible treatment assignments are computed. This process generates the randomization distribution of the test statistic under the data-dependent sharp null hypothesis H_{0s}. The p-value is calculated as the proportion of values of the test statistic that are equal to or greater than its observed value.

Thus, at each step, the sequential single imputation approach uses the point estimates of the factorial effects currently labelled active to fill in the missing potential outcomes, and uses the largest absolute value of the estimates of the effects currently labelled inactive as the test statistic. In contrast, from the potential outcomes perspective the Loughin & Noble (1997) approach assumes that for each unit the residuals for all the missing potential outcomes are the same as the observed residual, and at every step assesses the largest absolute value of all the estimated factorial effects obtained for the corresponding vector of residuals.

Although the procedure described above is more general and flexible because it permits generating the randomization distribution of *any* test statistic under *any* treatment assignment mechanism, under *any* sharp null hypothesis, it has two important drawbacks for testing a sequence of sharp-null hypotheses: first, it assumes a constant additive effect; and second, it ignores uncertainty of the

estimates. Rubin (1984) provided the following Bayesian justification for the Fisherian approach to inference: *it gives the posterior predictive distribution of the estimand of interest under a model of constant treatment effects and fixed units with fixed responses.* Thus, a natural extension of this approach is the Bayesian inferential procedure described next.

4 A Bayesian approach to screening factorial effects using Sequential Posterior Predictive Checks

We now propose a Bayesian approach for screening active factorial effects from an unreplicated 2^K design that overcomes the limitations of the Fisherian and the Loughin and Noble (1997) approaches. Our method extends the Bayesian framework proposed by Dasgupta et al. (2012) to a sequential screening procedure using posterior predictive checks (PPC) proposed by Rubin (1984) and investigated by Meng (1994), Gelman et al. (1996) and Rubin (1998). The use of PPCs is motivated by an additional intuitive appeal: When hypothesis testing, it is common practice to stop once we have a p-value that is "significant enough"; ironically, we stop when we find a model that does *not* fit (the null model). We believe that a better, cleaner and more principled strategy is one that stops when we find a model that *does* fit the data. Rubin (1984), wrote "*Although the frequentist can stop with a rejection of the null hypothesis, I believe that the Bayesian is obliged to seek and build a model that is acceptable to condition on*". The proposed Bayesian procedure does use distributional assumptions in contrast to the Fisherian approach, making it more general than the strict randomization-based approach, which can be derived as a special case of the former by putting point-mass prior distributions on the potential outcomes and the unit-level factorial effects.

The key steps in the proposed approach are: (i) postulating a suitable "null model" (a probabilistic model specifying the active effects) for the potential outcomes; (ii) obtaining an imputation model for the missing potential outcomes $\boldsymbol{Y}^{\text{mis}}$, conditional on the observed outcomes $\boldsymbol{Y}^{\text{obs}}$ and the observed assignment vector $\boldsymbol{w}$; and (iii) using the imputation model to generate the posterior predictive distribution of a suitable test statistic (or discrepancy measure, as in Gelman et al. (1996)) T, and consequently to compute the posterior predictive p-value under hypothetical replications of the same experiment.

To implement the sequential posterior predictive checks (S-PPC), the aforementioned three steps are conducted either through a "step-in" or a "step-out" approach. The former approach

starts with the sharp null model of no active effects, $\mathcal{A}_0$ (i.e., Fisher's sharp null hypothesis of no treatment effects), and then creates a sequence of non-sharp null models by including effects into the set of postulated active effects $\mathcal{A}_s$ one by one, starting with the largest estimated effect. We stop when we find a parsimonious model that is consistent with the data. In this case s is equal to the cardinality of $\mathcal{A}_s, a$. In contrast, the step-out procedure starts with the saturated model (all effects active) as a default, and then tests whether more parsimonious models are consistent with the data by eliminating factorial effects one by one from the active set, starting with the one with the smallest estimated effect. We then stop when we find a model that is inconsistent with the data, keeping the last model that seemed adequate. In this case $a \neq s$. In the following three subsections, we describe this procedure in detail.

4.1 Useful partitions related to active and inactive effects

As before, we assume that the N experimental units are *fixed*. At step s, it is possible to partition the unit-level vector of factorial effects $\boldsymbol{\theta}_i$ as $(\boldsymbol{\theta}_i^{\mathcal{A}_s} : \mathbf{0})$, where $\boldsymbol{\theta}_i^{\mathcal{A}_s}$ is an $(s+1)$-component row vector that includes the mean term, and $\mathbf{0}$ is a null vector with $N-1-s$ components. Therefore $(\boldsymbol{\theta}_i^{\mathcal{A}_s} : \mathbf{0})$ is a permutation of $\boldsymbol{\theta}_i$. Each row of $\boldsymbol{G}'$ represents a factorial effect; hence $\boldsymbol{G}'$ can also be rearranged to form the matrix

$$\begin{pmatrix} \boldsymbol{G}'^{\mathcal{A}_s} \\ \boldsymbol{G}'^{\bar{\mathcal{A}}_s} \end{pmatrix},$$

so that $\boldsymbol{G}'^{\mathcal{A}_s}$ and $\boldsymbol{G}'^{\bar{\mathcal{A}}_s}$ are matrices of dimension $(s+1) \times N$ and $(N-s-1) \times N$ corresponding to the active and inactive effects respectively.

Consequently, it follows from (2) that

$$\boldsymbol{Y}_i = (\boldsymbol{\theta}_i^{\mathcal{A}_s} : \mathbf{0}) \begin{pmatrix} \boldsymbol{G}'^{\mathcal{A}_s} \\ \boldsymbol{G}'^{\bar{\mathcal{A}}_s} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}_s} \boldsymbol{G}'^{\mathcal{A}_s}. \tag{10}$$

To express the observed and missing outcomes in terms of the active effects, we partition the vector $\boldsymbol{g}_i'^{\text{obs}}$ and the matrix $\boldsymbol{G}_i'^{\text{mis}}$, defined in Section 3, as

$$\begin{pmatrix} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}_s} \\ \boldsymbol{g}_i'^{\text{obs},\bar{\mathcal{A}}_s} \end{pmatrix} \text{ and } \begin{pmatrix} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}_s} \\ \boldsymbol{G}_i'^{\text{mis},\bar{\mathcal{A}}_s} \end{pmatrix}$$

respectively, just as we partitioned $\boldsymbol{G}'$. Therefore, from (9) we can write

$$Y_i^{\text{obs}} = \boldsymbol{\theta}_i \boldsymbol{g}_i'^{\text{obs}} = (\boldsymbol{\theta}_i^{\mathcal{A}_s} : \mathbf{0}) \begin{pmatrix} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}_s} \\ \boldsymbol{g}_i'^{\text{obs},\bar{\mathcal{A}}_s} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}_s} \boldsymbol{g}_i'^{\text{obs},\mathcal{A}_s}, \tag{11}$$

$$\boldsymbol{Y}_i^{\text{mis}} = \boldsymbol{\theta}_i \boldsymbol{G}_i'^{\text{mis}} = (\boldsymbol{\theta}_i^{\mathcal{A}_s} : \mathbf{0}) \begin{pmatrix} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}_s} \\ \boldsymbol{G}_i'^{\text{mis},\bar{\mathcal{A}}_s} \end{pmatrix} = \boldsymbol{\theta}_i^{\mathcal{A}_s} \boldsymbol{G}_i'^{\text{mis},\mathcal{A}_s}. \tag{12}$$

4.2 The imputation model and computation of the posterior predictive distribution of T

Throughout this section, we assume that the units are fixed, but the potential outcomes are generally random variables. Also, we assume that the units are exchangeable, and thus the potential outcomes can be modeled as conditionally independent across units. This assumption is reasonable in many realistic situations. Moreover, in many cases conditioning on unit level covariates increases the plausibility of this assumption, and our method can easily be extended to incorporate this information.

At step s, let $p(\boldsymbol{\theta}_i|\eta^{\mathcal{A}_s})$ denote a prior probabilistic model for $\boldsymbol{\theta}_i$, where $\boldsymbol{\eta}^{\mathcal{A}_s}$ is a vector of parameters having a suitable prior distribution for the set of active effects $\mathcal{A}_s$. Because of the identity $\boldsymbol{Y}_i = \boldsymbol{\theta}_i \boldsymbol{G}$, the model can also be specified through $\boldsymbol{Y}_i$. Then, following Rubin (1978), for a completely randomized factorial experiment, we have the following lemma:

Lemma 1. *The conditional distribution of* $\boldsymbol{Y}_i^{\text{mis}}$ *given* $\boldsymbol{Y}^{\text{obs}}$ *and* $\boldsymbol{w}$ *is given by*

$$p(\boldsymbol{Y}_i^{\text{mis}}|\boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}) \propto \int\int p(Y_i^{\text{mis}}|\boldsymbol{\eta}^{\mathcal{A}_s}, \boldsymbol{\theta}_i, \boldsymbol{Y}^{\text{obs}}, \boldsymbol{W}) p(\boldsymbol{\theta}_i|\boldsymbol{\eta}^{\mathcal{A}_s}, \boldsymbol{Y}^{\text{obs}}) p(\boldsymbol{\eta}^{\mathcal{A}_s}|\boldsymbol{Y}^{\text{obs}}) d\boldsymbol{\eta}^{\mathcal{A}_s} d\boldsymbol{\theta}_i. \tag{13}$$

The observed treatment assignment vector $\boldsymbol{w}$ does not appear in the second and third factors on the right hand side of (13) because in a randomized experiment, the treatment assignment mechanism is *ignorable* (Rubin 1978). Hence, the posterior distributions of $\boldsymbol{\theta}_i$ or $\boldsymbol{\eta}_{\mathcal{A}_s}$ explicitly depend on $\boldsymbol{w}$ only through $\boldsymbol{Y}^{\text{obs}}$. However, we need $\boldsymbol{w}$ to identify the missing observations Y_i^{mis}. Obtaining the posterior predictive p-value for the null model involves the following steps:

i. Obtain the imputation model $p(\boldsymbol{Y}_i^{\text{mis}}|\boldsymbol{Y}^{\text{obs}}, \boldsymbol{w})$ using Lemma 1.

ii. Repeat the following steps M times:

   (a) Impute a draw of the missing potential outcomes using the model obtained in step i.

(b) Re-randomize the N units to the N treatment combinations. In other words, generate a draw, $\boldsymbol{w}^{rep}$, from $p(\boldsymbol{W})$ and thereby obtain a new set of observed data.

(c) Given this specific imputation and re-randomization, compute the test statistic (T^{rep}).

iii. Compute the posterior predictive p-value for the null model as the proportion of the M draws in which T^{rep} equals or exceeds T^{obs}.

In principle, the steps described above can be carried out for any model. However, here we restrict ourselves to a simple normal hierarchical model described in (14), where δ denotes an indicator function such that $\delta(A) = 1$ if A is true and zero otherwise.

$$\begin{aligned} p(\boldsymbol{Y}_i|\boldsymbol{\theta}_i) &= \delta(\boldsymbol{Y}_i = \boldsymbol{\theta}_i^{\mathcal{A}_s}\boldsymbol{G}'^{\mathcal{A}_s}), \quad i = 1, \ldots, N, \\ p(\boldsymbol{\theta}_i^{\mathcal{A}_s}|\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2) &= N(\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2\boldsymbol{I}_s), \quad i = 1, \ldots, N, \\ p(\boldsymbol{\theta}_i^{\bar{\mathcal{A}}_s}) &= \delta(\boldsymbol{\theta}_i^{\bar{\mathcal{A}}_s} = \boldsymbol{0}), \quad i = 1, \ldots, N, \\ p(\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2) &\propto \frac{1}{\sigma^2}. \end{aligned} \tag{14}$$

Also, we assume conditional independence of the $\boldsymbol{\theta}_i$'s, and hence that of $\boldsymbol{Y}_i$'s for $i = 1, \ldots, N$. The hierarchical normal model specified by (14) permits a fair comparison of the proposed approach with the standard approaches that make analogous assumptions, which is conducted in Section 5. Note that model (14) assumes the potential outcomes are continuous with additive, iid normal residuals. Consider, for example, the 2^2 experiment with two active main effects and an inactive interaction. Then, the potential outcomes for the ith unit under treatment combination $\boldsymbol{z} = (z_1, z_2)$ can be written as $Y_i(z_1, z_2) = \mu_{i0} + (\mu_1/2)z_1 + (\mu_2/2)z_2 + \epsilon_i$, where ϵ_i's are iid with a common variance.

We now discuss how Steps i and ii described earlier can be implemented under model (14). Substituting $(\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2)$ for $\boldsymbol{\eta}^{\mathcal{A}_s}$ in (13), and after some minor manipulations, we obtain the imputation model for missing outcomes as:

$$\int\int\int p(Y_i^{\text{mis}}|\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2, \boldsymbol{\theta}_i, \boldsymbol{Y}^{\text{obs}}, \boldsymbol{w}) p(\boldsymbol{\theta}_i|\boldsymbol{\mu}^{\mathcal{A}_s}, \sigma^2, \boldsymbol{Y}^{\text{obs}}) p(\boldsymbol{\mu}^{\mathcal{A}_s}|\sigma^2, \boldsymbol{Y}^{\text{obs}}) p(\sigma^2|\boldsymbol{Y}^{\text{obs}}) d\boldsymbol{\mu}^{\mathcal{A}_s} d\sigma^2 d\boldsymbol{\theta}_i. \tag{15}$$

As shown in Gelman et al. (2003) (page 356) the joint posterior distribution of the hyperparameters

σ^2 and $\boldsymbol{\mu}^{\mathcal{A}_s}$ is the product of

$$p(\sigma^2|\boldsymbol{Y}^{\text{obs}}) = \text{Inv}\boldsymbol{\chi}^2(N-s-1, MS_{res}^{\mathcal{A}_s}) \quad \text{and} \quad p(\boldsymbol{\mu}^{\mathcal{A}_s}|\sigma^2, \boldsymbol{Y}^{\text{obs}}) = N(\hat{\boldsymbol{\theta}}^{\mathcal{A}_s}, \sigma^2\mathbf{I}_s/N), \tag{16}$$

where $MS_{res}^{\mathcal{A}_s}$ is the residual mean square under the model that corresponds to the regression of $\boldsymbol{Y}^{obs}$ on $\boldsymbol{G}'^{\mathcal{A}_s}$ (which is well defined except for the saturated model in which all factorial effects are active), and $\hat{\boldsymbol{\theta}}^{\mathcal{A}_s}$ is the posterior mean of both $\boldsymbol{\mu}^{\mathcal{A}_s}$ and $\boldsymbol{\theta}_i^{\mathcal{A}_s}$ (given by its unbiased OLS estimator). The imputation of missing potential outcomes, i.e., step ii(a), can be executed using the following simulation procedure based on (15) and (16):

1. Draw σ_*^2 from $p(\sigma^2|\boldsymbol{Y}^{\text{obs}})$.
2. With σ^2 set to σ_*^2, draw $\boldsymbol{\mu}_*^{\mathcal{A}_s}$ from $p(\boldsymbol{\mu}^{\mathcal{A}_s}|\sigma_*^2, \boldsymbol{Y}^{\text{obs}})$.
3. With $\sigma^2 = \sigma_*^2$ and $\boldsymbol{\mu}^{\mathcal{A}_s} = \boldsymbol{\mu}_*^{\mathcal{A}_s}$, for unit i draw $\boldsymbol{\theta}_{i,*}$ from $p(\boldsymbol{\theta}_i|\boldsymbol{\mu}_*^{\mathcal{A}_s}, \sigma_*^2)$.
   Specifically, for unit i
   - Every $\theta_{i,j} \in \bar{\mathcal{A}}_s$ is set to zero: $p(\theta_{i,j}|\boldsymbol{\mu}, \sigma) = \delta(\theta_{i,j} = 0)$.
   - Draw $\dot{\boldsymbol{\theta}}_{i,*}^{\mathcal{A}_s}$ from $N(\boldsymbol{\mu}_*^{\mathcal{A}_s}, \sigma_*^2\mathbf{I}_s)$.
   - Complete $\boldsymbol{\theta}_{i,*}^{\mathcal{A}_s}$ by calculating $\theta_{i,0} = Y_i^{obs} - \dot{\boldsymbol{\theta}}_{i,*}^{\mathcal{A}_s}\dot{\boldsymbol{g}}_i'^{\text{obs},\mathcal{A}_s}$, where $\dot{\boldsymbol{g}}_i'^{\text{obs},\mathcal{A}_s}$ excludes the first entry of $\boldsymbol{g}_i'^{\text{obs},\mathcal{A}_s}$. Note that the prior distribution on θ_{i0} is dominated by the observed potential outcome, the draw of $\dot{\boldsymbol{\theta}}_{i,*}^{\mathcal{A}_s}$ and the imputed zeros for $\boldsymbol{\theta}_i^{\bar{\mathcal{A}}_s}$ because of the deterministic relationship between $\boldsymbol{\theta}_i$ and $\boldsymbol{Y}_i$.
   - The missing potential outcomes for this unit are imputed as $\boldsymbol{Y}_{i,*}^{\text{mis}} = \boldsymbol{\theta}_{i,*}^{\mathcal{A}_s}\boldsymbol{G}_i'^{\text{mis},\mathcal{A}_s}$.

4.3 Test statistics and definitions of extremeness

Lenth (1989) proposed the following estimate of the standard deviation of the $\hat{\theta}_j$'s, which he called the *pseudo standard error*, PSE:

$$PSE = 1.5 \cdot \text{median}_{\{|\hat{\theta}_j| \leq 2.5 s_0\}}|\hat{\theta}_j|$$

where $s_0 = 1.5 \cdot \text{median}|\hat{\theta}_j|$, and $\hat{\theta}_j$ is the estimate of the j-th factorial effect. This definition can be adapted to take into account the sequential nature of the proposed methodology; let $PSE_{\bar{\mathcal{A}}_s}$ denote the pseudo standard error of the factorial effects in $\bar{\mathcal{A}}_s$ at step s.

We consider three test statistics: the maximum absolute value of the effects assumed to be in $\bar{\mathcal{A}}_s$ by the current prior distribution, with and without standardizing by the pseudo standard error, (i.e., $\max_{j\in\bar{\mathcal{A}}_s}\left|\frac{\hat{\theta}_j}{PSE_{\bar{\mathcal{A}}_s}}\right|$ and $\max_{j\in\bar{\mathcal{A}}_s}\left|\hat{\theta}_j\right|$), and $PSE_{\bar{\mathcal{A}}_s}$. The maximum is a useful statistic for sequential screening in unreplicated experiments because, in addition to being less sensitive to the normality assumption, it helps identify outliers among factorial effects assumed inactive. Its effectiveness in Loughin & Noble (1997) motivated the inclusion of test statistics involving $\max_{j\in\bar{\mathcal{A}}_s}\left|\hat{\theta}_j\right|$ in our study. For these statistics, extremeness is defined by the right tail of the distribution of T because, given the assumed prior distribution, we are interested in how unlikely it is that the inactive effects would lead to a value of T as large or larger than the observed value, T^{obs}.

$PSE_{\bar{\mathcal{A}}_s}$ is a measure of the size of what is not explained by the model motivated its use as a test statistic in this paper. Thus, a small value of the $PSE_{\bar{\mathcal{A}}_s}$ indicates that the assumed model is not consistent with the data because there is a smaller variability between the factorial effects assumed null than what would be expected under the assumed prior model. Hence, extremeness for this test statistic is based on a one sided assessment focusing on the lower tail of the posterior predictive distribution of T.

5 Simulation Study

The general framework presented in Section 4 offers flexibility for estimands, test statistics and sequences of models to be assessed other than those explored in this paper which may be more relevant to the study at hand. However, we believe it is fundamental to compare its performance relative to standard approaches in the literature to evaluate what benefits this method can have, even in the setting where the traditional assumptions are met. As already mentioned in 4.2, the specific hierarchical normal model given by (14) was proposed with this goal in mind. To agree with the traditional setting, the simulation is done using the superpopulation definition of an active effect, which in our procedure can be understood as $\boldsymbol{\mu}^{\bar{\mathcal{A}}_s} = \mathbf{0}$ and $\boldsymbol{\mu}^{\mathcal{A}_s}$ consists of non-zero elements. Nevertheless, as explained in Section 4.2, at step s the S-PPC method fills in the missing data assuming $\theta_{i,j} = 0$ for effects in $\bar{\mathcal{A}}_s$ and all units. We believe this simulation allows for a fair comparison of the frequency properties of the different methods.

Tukey (1953) first coined the term *experimentwise error rate* (EER) to refer to the probability of making one or more false discoveries when testing multiple factors. It is standard practice in the literature to calibrate the proposed methods to thresholds that are most frequently used in

screening. For comparison purposes, we calibrate our proposals to satisfy an EER of 0.05, and use the cutoff values for EER = 0.05 reported in the papers proposing the other methods. Two other error rates that are commonly used in the literature are compared here but not calibrated. One is the *individual error rate* (IER), which is the probability of incorrectly identifying an inactive effect as active, but does not account for the other effects being assumed inactive. The other one is the *false discovery rate* (FDR), which corresponds to the expected proportion of false positives among all the effects declared active. For the null model ($\mathcal{A}_s = \emptyset$), the FDR is equivalent to the EER.

The calibration study was performed under the null hypothesis of all individual level factorial effects being inactive in a 2^4 full factorial design (i.e. $\mu_j = 0$ for all $i = 1, \ldots, 16$ and $j > 1, \ldots, 15$, where $\mu_j = E(\theta_{i,j})$. We simulated 1000 different science tables assuming the null hypothesis. Each science table was obtained by simulating $\theta_{i,j}$'s from a $N(0, 1)$ distribution, and transforming them to the potential outcomes $Y_i(\boldsymbol{z})$. The cutoffs obtained are: 0.050 and 0.043 for the $\max_{\{j:j\in\bar{\mathcal{A}}_s\}} \left|\hat{\theta}_j\right|$ test statistic with the step-in and step-out procedures respectively, 0.048 for both procedures using the $\max_{\{j:j\in\bar{\mathcal{A}}_s\}} \left|\frac{\hat{\theta}_j}{PSE_{\bar{\mathcal{A}}_s}}\right|$ measure, and 0.049 for test statistic PSE_s using the step-in procedure. The step-out procedure was inadequate for this measure because the cutoff necessary to achieve an EER of 0.05 was relatively small (0.035), having a negative effect on its overall performance.

A similar simulation set-up was used to compare the performance of the proposed S-PPC to the randomization test under Fisher's sharp null hypothesis (described in Section 3) with and without the Bonferroni correction, the Loughin and Noble approach (L&N) (described in Loughin & Noble (1997)), the Lenth method (described in Lenth (1989)), and the Step Down Lenth method (described in Ye et al. (2001)).

The Bayesian methods of Box & Meyer (1986) and Chipman et al. (1997) are not included in this study. We decided to exclude the former because it performs poorly for large numbers of active effects, and, although it performs well for low numbers of active effects, on average, it is less powerful than Lenth's method (Hamada & Balakrishnan, 1998). We excluded the latter because of the sensitivity of the results to the choice of some hyperparameter values (Wu & Hamada, 2009).

We simulated potential outcomes repeatedly (1000 times) from 36 different alternative settings defined by fixing $\max_{\{j:j\in\mathcal{A}_s\}} |\mu_j|$ at 4, and specifying different levels of noise ($\sigma = 0.5, 1, 2$), number of active effects ($a = 1, 2, 4, 6$), and the range r between active effects (defined as $r = \max_{\{j:j\in\mathcal{A}_s\}} |\mu_j| - \min_{\{j:j\in\mathcal{A}_s\}} |\mu_j| = 1, 2, 3$). Each combination of these simulation factors determines a true value of $\boldsymbol{\mu}$ by selecting a factorial effects at random to be active, and letting the

corresponding μ_j's assume the value $4-r$ if $a=1$, or the values $4-r\left(1-\frac{t}{a-1}\right)$ for $t=0,\ldots,a-1$ if $a>1$. The remaining μ_j's are set to zero. The potential outcomes for each unit are drawn from $\boldsymbol{Y}_i \sim N(\boldsymbol{\mu}\boldsymbol{G}', \sigma^2\mathbf{I})$.

We compare these methods based on five summary measurements based on averages across simulated data sets. Three of these measures were defined in Section 5 and are related to error rates: **IER** (average proportion of false positives), **EER** (proportion of data sets with at least one false positive), and **FDR** (average proportion of false positives among all the effects declared active). The other two measures are related to the power of the procedures. The first is the rejection rate (**RR**) that measures the proportion of active effects correctly identified as such (average proportion of true positives). The second is the average number of positives (**ANP**) that corresponds to the average number of effects declared active. The **ANP** should ideally be close to zero under the null hypothesis and close to 3.25 (i.e., the average number of active effects across all 36 simulation settings) under the alternative hypotheses.

The results of the simulation study are shown in Table 3. Using our notation leads to a different expression of the test statistic used for the Step Down Lenth method than that given in Ye et al. (2001). The values obtained in the calibration study are used as threshold p-values in the proposed S-PPC methods. For the existing methods we used the values reported in the literature, which were also determined via calibration studies. We include the results for the randomization test, which does not account for multiple comparisons, as a reference point where it achieves an IER of 0.05 but a very high EER of 0.661. The Step Down Lenth method has a very high EER under the null hypothesis (0.109). The rest of the methods are reasonably close to the intended 0.05.

Table 3 shows that the *step-out* S-PPC with test statistic $\max_{\{j:j\in\bar{\mathcal{A}}_s\}}\left|\hat{\theta}_j\right|$ has the best performance overall and hence will be referred to as the best S-PPC (BS-PPC) henceforth. It has an EER below 0.049 under the null hypothesis, as well as low IER and ANP values relative to the other methods. Averaging across the alternative hypotheses, it has the highest RR of 0.637 and the highest ANP of 2.146. The closest competitors L&N, Step Down Lenth and Lenth have much lower RR and ANP values. This S-PPC procedure also has an EER of 0.028 which is more or less comparable to the EER values for L&N (0.023) and Lenth (0.031) that appear to control the EER best. In Figure 1 we display the RR, FDR, and EER, across the different simulation alternative hypothesis settings for the previously established methods and compare them to the BS-PPC. This figure clearly shows that the BS-PPC is much better than the rest with clear and distinct modes

			null hypothesis			average across alternative hypotheses				
Method	**Discrepancy Measure**	**Screening Rule**	**IER** (SE)	**EER** (SE)	**ANP** (SE)	**RR** (SE)	**IER** (SE)	**EER** (SE)	**FDR** (SE)	**ANP** (SE)
Randomization Test	$\lvert\hat{\theta}\rvert_j$	-	0.050 (0.001)	0.661 (0.015)	0.750 (0.019)	0.487 (0.006)	0.004 (<0.001)	0.049 (0.006)	0.042 (0.005)	1.200 (0.014)
Bonferroni (on randomization test)	$\lvert\hat{\theta}\rvert_j$	-	0.003 (<0.001)	0.048 (0.007)	0.048 (0.007)	0.251 (0.006)	0.000 (<0.001)	0.002 (0.001)	0.002 (0.001)	0.400 (0.009)
Lenth	$\lvert\hat{\theta}_j\rvert/PSE$	-	0.007 (0.001)	0.056 (0.007)	0.098 (0.016)	0.551 (0.008)	0.004 (0.001)	0.031 (0.005)	0.014 (0.003)	1.862 (0.030)
Step Down Lenth	$\max_{\{j:j\in\bar{A}_s\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\bar{A}_s}}\right\rvert$	step-in	0.015 (0.002)	0.109 (0.010)	0.226 (0.026)	0.560 (0.008)	0.016 (0.002)	0.095 (0.009)	0.052 (0.006)	2.050 (0.037)
L&N	$\left(\frac{N-1}{N-\lvert\bar{A}_s\rvert}\right)^{1/2}\max_{\{j:j\in\bar{A}_s\}}\left\lvert\hat{\theta}_j\right\rvert$	step-out	0.006 (0.001)	0.053 (0.007)	0.086 (0.015)	0.560 (0.009)	0.002 (0.001)	0.023 (0.005)	0.011 (0.002)	1.744 (0.034)
S-PPC	$PSE_{\bar{A}_s}$	step-in	0.008 (0.002)	0.052 (0.007)	0.121 (0.029)	0.436 (0.008)	0.013 (0.003)	0.029 (0.005)	0.018 (0.003)	1.664 (0.046)
	$\max_{\{j:j\in\bar{A}_s\}}\left\lvert\hat{\theta}_j\right\rvert$	step-in	0.003 (0.002)	0.050 (0.007)	0.050 (0.029)	0.327 (0.008)	0.001 (0.003)	0.009 (0.005)	0.005 (0.003)	0.594 (0.046)
		step-out	0.004 (0.001)	0.049 (0.007)	0.058 (0.009)	0.637 (0.009)	0.003 (0.001)	0.028 (0.005)	0.012 (0.002)	2.146 (0.035)
	$\max_{\{j:j\in\bar{A}_s\}}\left\lvert\frac{\hat{\theta}_j}{PSE_{\bar{A}_s}}\right\rvert$	step-in	0.006 (0.001)	0.049 (0.007)	0.085 (0.015)	0.527 (0.008)	0.003 (0.001)	0.026 (0.005)	0.012 (0.003)	1.768 (0.029)
		step-out	0.006 (0.001)	0.049 (0.007)	0.085 (0.015)	0.531 (0.008)	0.003 (0.001)	0.026 (0.005)	0.012 (0.003)	1.788 (0.029)

Table 3: Summary of results of the simulation study. Average rates and number of effects declared active, as well as the standard errors of these quantities are displayed for all methods and screening rules under the null ($\sigma = 1$) and across alternative hypotheses.

around 1 for the RR, and 0 for the FDR and EER. For the BS-PPC, there are no combinations with an FDR above 0.05. However, there are eight combinations with EER above 0.05. In contrast, the L&N and Lenth methods have none above this threshold, at the cost of lower rejection rates.

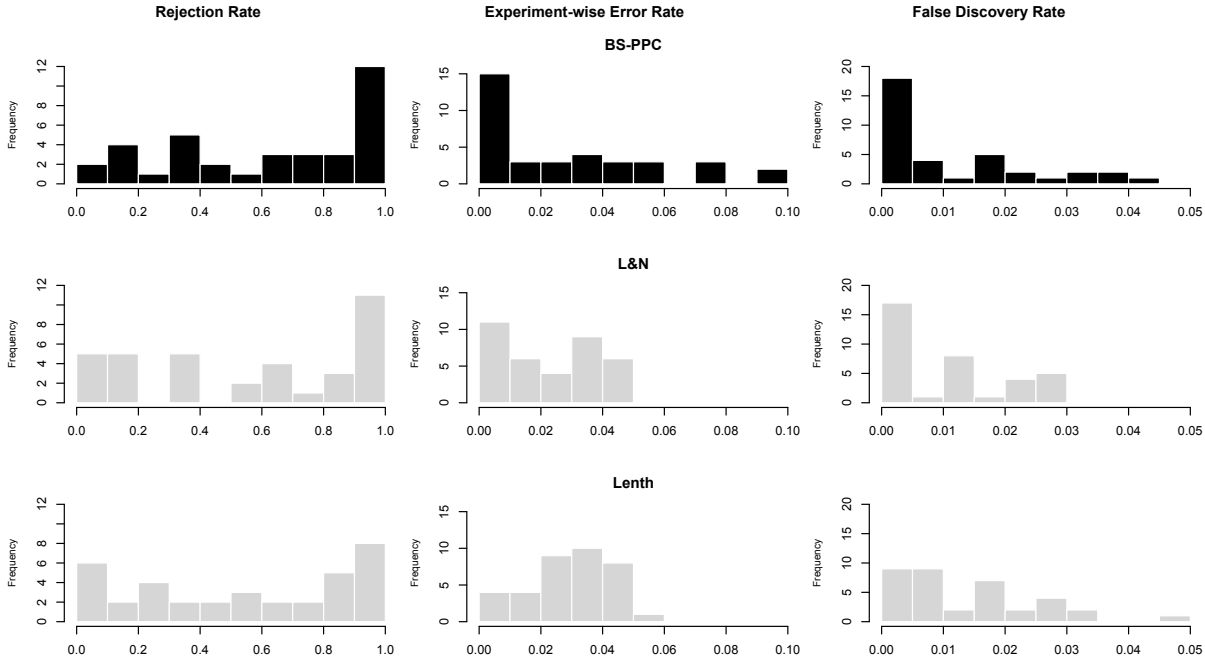

Figure 1: Distributions of the Rejection (RR), the False Discovery (FDR), and Experiment-wise Error (EER) Rates across the different alternative hypotheses.

Figures 2, 3, and 4 display the marginal effects of (i) the number of active effects, a, (ii) the standard deviation, σ, and (iii) the range of mean active effects, r, respectively, on average RR, EER, FDR, and ANP. Figure 2 shows the average number of effects declared active for each of the true numbers of active effects in the set of alternative hypotheses explored (i.e., $1, 2, 4, 6$). Again, the BS-PPC approach is the one with best overall performance. This figure agrees with the finding of Tripolski et al. (2008), that the L&N approach has a good performance for a low number of active effects (1 and 2), but that its performance deteriorates for higher numbers.

Figure 3 shows that across different values of σ the RR is highest for BS-PPC. Based on the two error rate plots, Figure 3 reveals that σ has a bigger impact on BS-PPC than on all other methods, for which both error rates appear to be quite stable across different values of σ.

Figure 4 shows that with respect to RR and ANP, BS-PPC is still sensitive to variation in r. However, with respect to EER and FDR, it appears to be less sensitive to variation in r than to σ. Although the other methods exhibit fairly stable performance for all measures with respect to

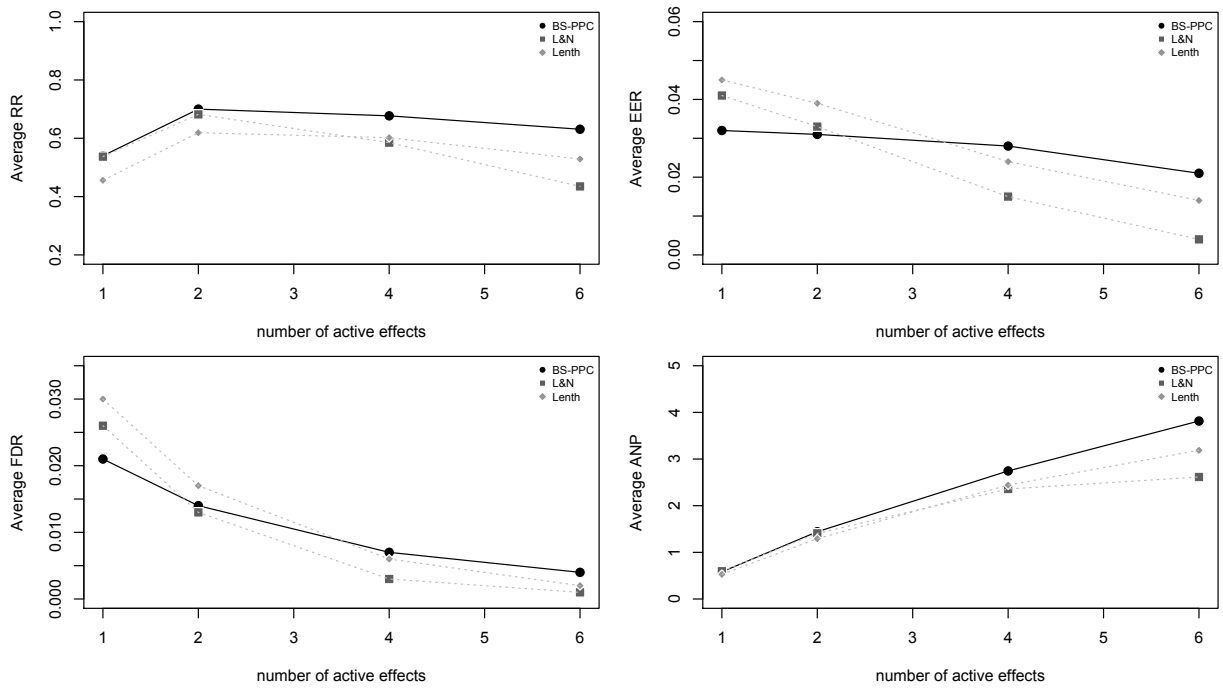

Figure 2: Performance measures (RR, EER, FDR and ANP) versus actual number of active effects

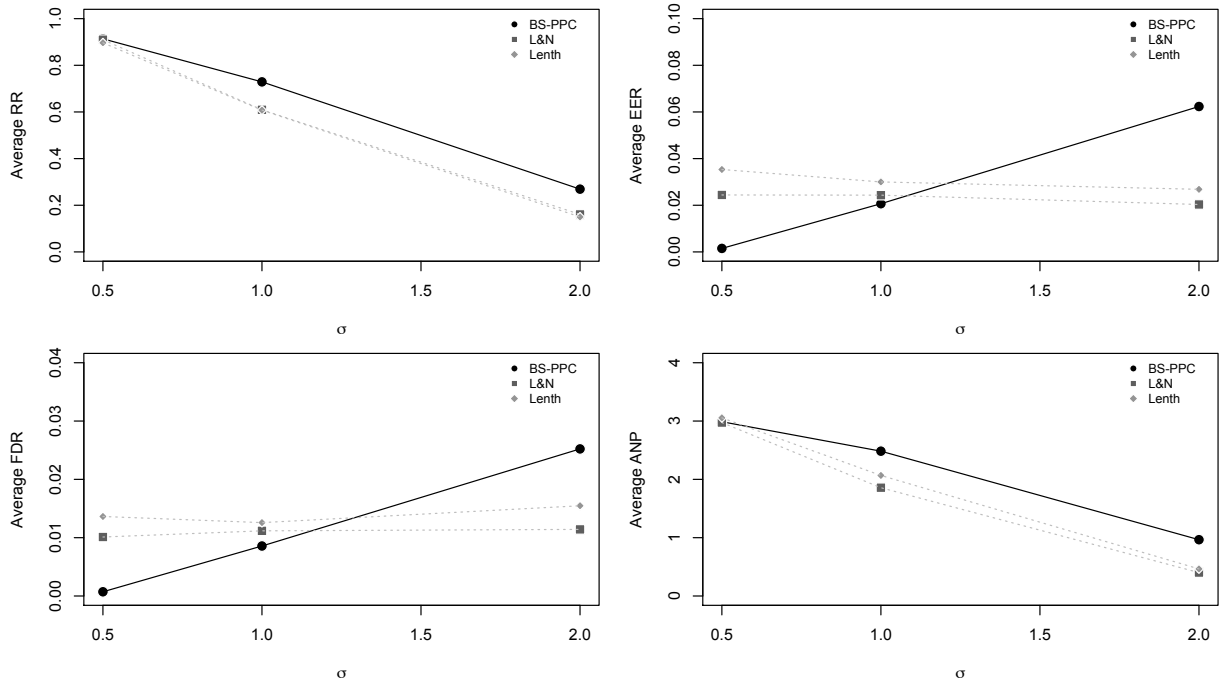

Figure 3: Performance measures (RR, EER, FDR and ANP) versus level of noise (σ): comparison of five methods

different range values, the BS-PPC is preferred because of the high RR for all levels of r, and all error rates being below 0.05.

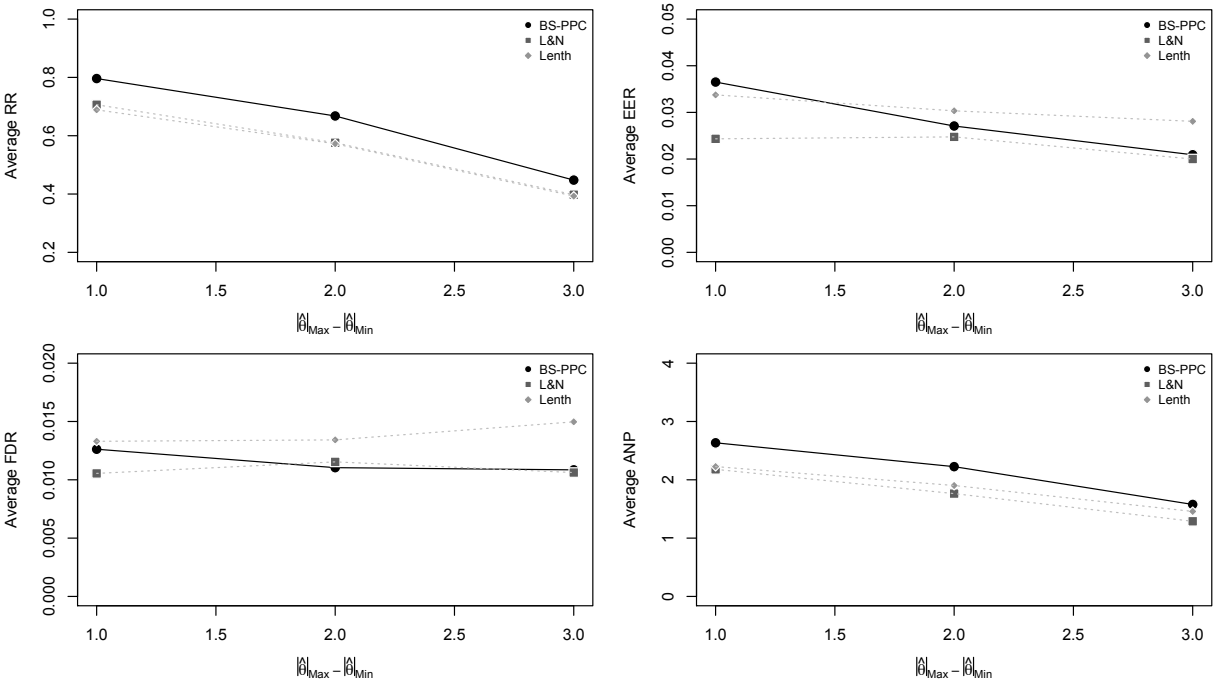

Figure 4: Performance measures (RR, EER, FDR and ANP) versus range of the absolute values of the factorial effects (r)

We decided not to include the detailed comparison of the proposed method with the FDR-corrected Lenth method due to an apparent confusion about the actual method used in Tripolski et al. (2008). The procedure outlined on page 35 of Tripolski et al. (2008) states: "For controlling the FDR, arrange the p values in ascending order and declare as active all effects that have a p-value *smaller than* the largest p-value that upholds the inequality $p_{(i)} \leq iq/k$," where $p_{(i)}$ denotes the ith ordered p-value, q the proposed cut-off value and k the total number of tests. This is not the same definition used in the original Benjamini & Hochberg (1995) procedure where the active effects are identified by p-values *smaller or equal to* the largest p-value that upholds the inequality. After conducting simulation studies with both methods (referred to as FDR-corrected Lenth method ($<$) and FDR-corrected Length method ($\leq$)), the results indicated that whereas the FDR-corrected Lenth method ($<$) controlled the EER well under the null hypothesis, it appeared to have a poor RR for the alternative settings. On the contrary, the FDR-corrected Lenth method ($\leq$) appeared to have high RR and ANP, comparable to BS-PPC, but did not appear to have the correct EER under the null hypothesis, an observation that raised some doubts about its calibration.

6 Conclusion

In order to account for the exploration of multiple factors in unreplicated screening experiments, we propose a Bayesian sequential method based on posterior predictive checks, which requires the modeling of the response variable, here assumed Normal to fit the traditional assumptions of classical screening methods. We recommend the use of step-out posterior predictive checks with $\max_{\{j:j\in\bar{\mathcal{A}}_s\}}\left|\hat{\theta}_j\right|$ as the test statistic. Our simulation results show that our approach balances rejection rates (the finding of true positives) and the error rates in an appealing way.

We believe that our method gives a more intuitive approach into screening by selecting a parsimonious model that is consistent with the data, instead of finding a null model that is not consistent with the observed data, as does the traditional p-value approach. However, a natural question to ask is, Why does the proposed approach appear to perform better than the existing ones? An intuitive explanation is as follows: the foundation of the proposed methodology is the Fisherian approach of using randomization tests that generates the randomization distribution of the test statistic under a sequence of sharp null hypotheses. Although this approach is similar (in fact, exactly identical at the first step of the sequential procedure) to the Loughin & Noble (1997) approach, our procedure generates a larger number of support points of the distribution of the test statistic by allowing each unit to have its own intercept. The proposed Bayesian approach, which uses the Fisherian approach as its foundation, incorporates additional flexibility into the analysis by (i) accounting for the uncertainty of estimation, thereby improving upon the "plug-in" Fisherian approach (based on a single imputation of missing outcomes) and (ii) not assuming a constant treatment effect. These aspects appear to give it an advantage over the competing procedures.

Extension to the fractional factorial designs are a topic for future work. In this setting, further assumptions are required to fill in all missing potential outcomes when an aliased group is deemed inactive. The approach proposed in this paper is a viable option with the additional assumption that all factorial effects in an aliased group deemed null, are themselves null. However, performance for fractional replication designs still needs to be evaluated.

Our approach is the only one that has been proposed that includes the consideration of the finite population. Nevertheless, it has a better performance in the classical super population setting than currently available methods proposed with the superpopulation in mind. Further study of its relative performance in the finite population setting is left for future work. Another topic for future work are extensions to non-linear estimands, such as median effects and other percentiles.

Acknowledgments

We are thankful to two referees and an Associate Editor whose comments were helpful in substantial improvement of the contents and presentation of this paper. This research was supported by National Foundation Grant number DMS-1107004.

References

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* **57**, 289–300.

Box, G. E. & Meyer, R. D. (1986). An analysis for unreplicated fractional factorials. *Technometrics* **28**, 11–18.

Chipman, H., Hamada, M. & Wu, C. (1997). A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics* **39**, 372–381.

Dasgupta, T., Ma, C., Joseph, V. R., Wang, Z. L. & Wu, C. F. J. (2008). Statistical modeling and analysis for robust synthesis of nanostructures. *Journal of the American Statistical Association* **103**, 594–603.

Dasgupta, T., Pillai, N. S. & Rubin, D. B. (2012). Causal inference from 2^k factorial designs using the potential outcomes model. *arXiv preprint arXiv:1211.2481* .

Dong, F. (1993). On the identification of active contrasts in unreplicated fractional factorials. *Statistica Sinica* **3**, 209–217.

Fisher, R. A. (1925). *Statistical methods for research workers*. Genesis Publishing.

Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (2003). *Bayesian Data Analysis*. CRC Press.

Gelman, A., Meng, X. L. & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica* **6**, 733–807.

Hamada, M. & Balakrishnan, N. (1998). Analyzing unreplicated factorial experiments: a review with some new proposals. *Statistica Sinica* **8**, 1–41.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association* **81**, 945–960.

Lenth, R. V. (1989). Quick and easy analysis of unreplicated factorials. *Technometrics* **31**, 469–473.

Loughin, T. M. & Noble, W. (1997). A permutation test for effects in an unreplicated factorial design. *Technometrics* **39**, 180–190.

Meng, X.-L. (1994). Posterior predictive p-values. *The Annals of Statistics* **22**, 1142–1160.

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* , 34–58.

Rubin, D. B. (1980). Randomization Analysis of Experimental Data: The Fisher Randomization Test: Comment. *Journal of the American Statistical Association* **75**, 591–593.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applies statistician. *The Annals of Statistics* **12**, 1151–1172.

Rubin, D. B. (1998). More powerful randomization-based p-values in double-blind trials with non-compliance. *Statistics in medicine* **17**, 371–385.

Tripolski, M., Benjamini, Y. & Steinberg, D. M. (2008). The false discovery rate for multiple testing in factorial experiments. *Technometrics* **50**, 32–38.

Tukey, J. (1953). The problem of multiple comparisons. *Unpublished manuscript. In The Collected Works of John W Tukey VIII. Multiple Comparisons: 1948-1983* .

Wu, C. & Hamada, M. S. (2009). *Experiments: Planning, Analysis and Optimization*. New Jersey: Wiley.

Ye, K. Q., Hamada, M. & Wu, C. (2001). A step-down lenth method for analyzing unreplicated factorial designs. *Journal of Quality Technology* **33**, 140–152.