A Comparison of Supersaturated Designs and Orthogonal Arrays

Jacob Akubire Abugre*[1] and Byran J. Smucker†[1]

[1]Department of Statistics, Miami University, Oxford, OH

Abstract

The purpose of a screening experiment is to accurately and cost-efficiently identify the few most influential factors from among the many studied. Two types of screening experiments use orthogonal arrays (OAs) and supersaturated designs (SSDs), respectively. The first requires the number of runs n to be a multiple of 4 and greater than the number of factors k, while the second is a bolder approach wherein $n < k$. In this study, we compare the performance of OAs with SSDs, using both simulation and a method based on the probability of perfect sign recovery under the Lasso. Our findings indicate that for the scenarios considered, OAs that were far above saturation had similar performance to those that were only slightly above saturation, slightly supersaturated designs required effects between 1.5 and 2 times larger than those for the OAs in order to have similar screening performance, and very supersaturated designs were clearly less effective than the OAs even when detecting effects up to three times as large.

Keywords: Screening experiments; Sign recovery; Regularization; the Lasso

*current affiliation: Department of Mathematical Sciences, University of Cincinnati, Cincinnati, OH

†corresponding author, bsmucke1@hfhs.org; current affiliation: Department of Public Health Sciences, Division of Biostatistics, Henry Ford Health, Detroit, MI

1 Introduction

The goals of the design and analysis of screening experiments is the selection of active experimental factors and the dismissal of factors with no meaningful effect on the response variable. A particularly daring screening approach is supersaturated designs (SSDs), where the number of runs n is less than the number of factors k. Though these experiments have been well-studied, practitioners have been hesitant to use them (Weese et al. 2021). Orthogonal arrays (OAs), on the other hand, have a long history of implementation. The aim of this paper is to compare these two screening strategies, in order to better understand the tradeoffs between run size and screening quality.

In a screening experiment, it is customary to adopt the linear main effect model represented as:

$$\mathbf{Y} = \beta_0 \mathbf{1} + D\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$, D is the design matrix of size $n \times k$ with entries ± 1, $\boldsymbol{\beta}$ is a vector of model parameters of size $k \times 1$, and $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are each $n \times 1$ vectors. The objective of screening experiments is to cost-effectively determine which elements of $\boldsymbol{\beta}$ are non-zero, and the task of the screen designer is the selection of both the dimensions and elements of D. The use of OAs is prevalent in screening experiments (e.g. Hedayat et al. 1999), where all columns of D are mutually orthogonal, n is a multiple of 4, and $n > k$. On the other hand, SSDs are constrained to have $n \leq k$ which means D does not have full column rank and thus its columns cannot be pairwise orthogonal.

Currently, there is a notable degree of skepticism regarding the adoption of SSDs. According to an informal survey conducted by Weese et al. (2021), only 13 out of 63 respondents reported having used them in the past, with only a handful using them regularly. Some respondents were concerned that SSDs don't have sufficient power to detect effects of interest. On the other hand, orthogonal designs, such as fractional factorials and Plackett-Burman designs, were used much more regularly among the respondents.

Given this prevailing sentiment, our objective in this work is to quantify the difference in screening ability of OAs vs. SSDs, across various run sizes, effect sizes and levels of sparsity. Specifically, we seek to evaluate the amount of screening capability lost when using SSDs instead of OAs, as well as to determine conditions under which the performances of the designs are similar. We analyze the designs using the Lasso and compare them based upon simulated true positive rate (TPR, proportion of true active factors that are identified as active) and false positive rate (FPR, proportion of inactive factors that are identified as active), as well as with a measure based on the probability of perfect sign recovery under the Lasso (Stallrich et al. 2024). Our goal is to increase understanding of the performance of SSDs relative to their more accepted counterparts. Then, experimenters will be better equipped to weigh the benefits of SSDs along with their drawbacks.

The outline of this paper is as follows: Section 2 offers a review of the relevant technical material regarding SSDs and OAs. Section 3 provides a description of our simulation protocol as well as the criteria proposed by Stallrich et al. (2024), both used to assess the quality of the designs. We provide the results of our study in Section 4, and a discussion in Section 5.

2 Background

Screening experiments have been used widely in the physical sciences and manufacturing (Hedayat et al. 1999; Mee 2009), as well as computer experiments (Owen 1992). Of particular interest to this work are orthogonal arrays (Plackett and Burman 1946; Rao 1947; Bush 1952; Bose and Bush 1952; Hedayat et al. 1999; Xu et al. 2009) and supersaturated designs (Satterthwaite 1959; Lin 1993; Wu 1993; Georgiou 2014; Weese et al. 2021). With regard to the latter, there have been only a limited number of applications in the literature (Carpinteiro et al. 2004; Dejaegher and Vander Heyden 2008; Jridi et al. 2015; Zarkadas and Besseris 2023), while the former are widely accepted by practitioners.

2.1 Screening Designs

Plackett and Burman (1946) were the first to introduce something like orthogonal arrays, and this led to the development of a considerable theory (e.g., Rao 1947; Bush 1952; Bose and Bush 1952; Hedayat et al. 1999). An orthogonal array is an $n \times k$ matrix which can be described with (n, k, s, d), where k is the number of factors each at one of s levels, and n is the number of experimental runs and a subset of all possible s^k treatment combinations. This array is said to be of strength d if all s^d treatment combinations corresponding to any d factors chosen out of k occur an equal number of times. Much of the development in OAs over the years relates to how various OAs project into lower dimensions (e.g., Cheng 1995, 1998; He et al. 2022; Chen et al. 2023) and also about the performance of the designs to account for two-factor interactions (e.g., Cheng et al. 2008; Mee et al. 2017; Schoen et al. 2017; Vazquez et al. 2019; He et al. 2022; Vazquez et al. 2022; Nguyen et al. 2023). Extensive reviews of orthogonal arrays can be found in Xu et al. (2009) and Mee et al. (2017). Because these designs have orthogonal columns, they estimate the main effects with minimum variance and no bias under the assumptions of model (1). For our purposes, any OA will do since we assume no interactions are contaminating the estimation of main effects. This assumption is in line with our goal, which is to compare design screening effectiveness.

Since $n < k$ for SSDs, orthogonal designs are not possible. These experiments were first suggested by Satterthwaite (1959) who described the construction of random balance designs, in which factor levels were chosen randomly according to some distribution. After the statistical properties of these random supersaturated designs were criticized, Booth and Cox (1962) proposed a systematic way of constructing SSDs based upon the intuition that design columns be as close to orthogonal as possible. In particular, they suggested a criterion, called $E(s^2)$, based on the average of the squared off-diagonal elements of $D^T D$ and requiring that each column of D have the same number of $+1$ and -1 entries. Let $X = [\mathbf{1}|D]$, where $\mathbf{1}$ is a $n \times 1$ vector representing the intercept and indexed as column 0 of X, and let s_{ij} be the ij^{th} off-diagonal value of $X^T X$. Then we have that $E(s^2) = \frac{2}{k(k-1)} \sum_{1 \leq i < j \leq k} s_{ij}^2$.

Later, several authors (Marley and Woods 2010; Jones and Majumdar 2014; Weese et al. 2015) relaxed the column balance constraint, leading to the so-called unconditional $E(s^2)$ criterion, $UE(s^2) = \frac{2}{k(k+1)} \sum_{0 \leq i < j \leq k} s_{ij}^2$.

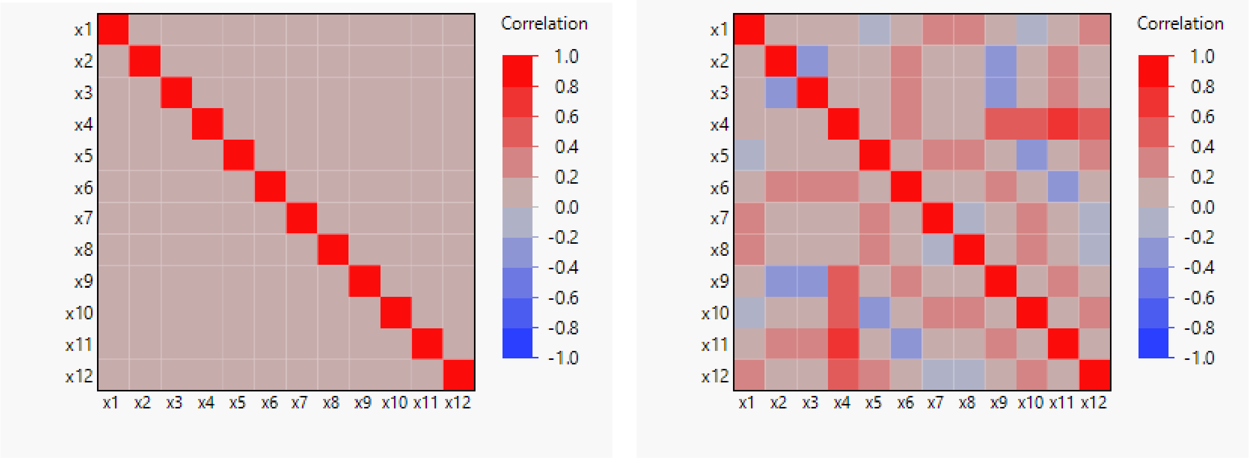
Though many other criteria have been proposed (e.g., Jones et al. 2008, 2009), none have been found to be clearly superior, in general, to $E(s^2)$-optimal designs (Marley and Woods 2010; Weese et al. 2015). However, Weese et al. (2017) developed the constrained $Var(s+)$ criterion, which was shown via simulation to be better than $E(s^2)$-type designs when effect directions are known, while similar in performance to $UE(s^2)$ when effect directions were misspecified. This and other work suggested that the Gauss-Dantzig selector be used to analyze SSDs (Phoa et al. 2009; Marley and Woods 2010; Weese et al. 2015), and Singh and Stufken (2023) used properties of the Dantzig selector to create designs that largely improved upon the $Var(s+)$ designs when effect directions were known. Recently, some theoretical results have corroborated the advantage of $Var(s+)$-type designs under a known sign vector (Stallrich et al. 2024), with an accompanying framework that can, in principle, construct designs that improve performance when analyzing the SSD using the Lasso (see Section 2.2). For the range of design sizes that we are considering, the designs of Singh and Stufken (2023) and Stallrich et al. (2024) are computationally expensive to construct. Thus, for this work, we focus primarily on $Var(s+)$ designs because they are computationally fast to construct and similar in quality to $E(s^2)$ and $UE(s^2)$ designs when effect directions are unknown (Weese et al. 2017, 2021; Singh and Stufken 2023, and also Section 3 of the Supplementary Material). Constrained $Var(s+)$ designs minimize

$$Var(s) = UE(s^2) - UE(s)^2 \tag{2}$$

$$s.t. \quad \frac{UE^*(s^2)}{UE(s^2)} > c \text{ and } UE(s) > 0, \tag{3}$$

where $UE^*(s^2)$ is the $UE(s^2)$ value for the $UE(s^2)$-optimal design. Generally, $c = 0.8$ has been found to work well.

In this paper, we compare Orthogonal Arrays with constrained $Var(s+)$ SSDs. The OAs are constructed using the `oa.design` function of `DoE.base`, an R package due to Grömping (2018), using the default "order" column option which ignores aliasing between main effects and two-factor interactions since we assume a main effects only model. We construct the constrained $Var(s+)$ designs using the algorithm described in Weese et al. (2017). In Figure 1, we show a comparison of an OA with a constrained $Var(s+)$ design. While the OA evidently has no correlation between columns, for the SSD there are many nonzero off-diagonals, and more often than not those nonzero entries are greater than 0 in keeping with the $UE(s) > 0$ part of the $Var(s+)$ criterion.



(a)

(b)

Figure 1: Correlation color plot for two 12-factor designs. (a) An OA with $n = 16$; (b) A constrained $Var(s+)$ design with $n = 10$.

2.2 Analysis of Screening Designs

Because $n < k$, a unique least squares solution for $[\beta_0|\boldsymbol{\beta}]$ does not exist in the case of SSDs. In the literature, various analysis methods have been studied, including forward selection, Bayesian methods, model averaging, and regularization (e.g., Marley and Woods 2010; Draguljić et al. 2014; Weese et al. 2015). Phoa et al. (2009) proposed to use the Dantzig selector and ensuing studies (e.g., Marley and Woods 2010; Weese et al. 2015) showed that it performed well compared to other methods. The Lasso (Tibshirani 1996) is an alternative regularization method, whose close relationship to the Dantzig selector is well-known (Meinshausen et al. 2007; Lounici 2008; Bickel et al. 2009; James et al. 2009; Asif and Romberg 2010). Indeed, Singh and Stufken (2023) remark that their designs, based on theoretical properties of the Dantzig selector, would also be effective if the Lasso was used to analyze the designs. The Lasso is more mathematically tractable than the Dantzig selector (Stallrich et al. 2024), and performs similarly in the context of screening designs (Draguljić et al. 2014). Since the two methods are similar, and because we use the probability of perfect sign recovery method of Stallrich et al. (2024), we use the Lasso to analyze our simulated data in this work. To estimate the main effect model in (1), the Lasso chooses $[\beta_0, \boldsymbol{\beta}]$ which minimizes

$$\frac{1}{2n}\|Y - \beta_0\mathbf{1} - D\boldsymbol{\beta}\|_2^2 + \lambda\|[\beta_0, \boldsymbol{\beta}]\|_1, \tag{4}$$

for a given regularization parameter λ. This parameter controls how much the parameters are shrunk toward 0; we will discuss our λ selection strategy in the next section. Equation (4) produces an OLS solution if $\lambda = 0$, while as $\lambda \to \infty$, $[\beta_0, \boldsymbol{\beta}] \to 0$. In our work, we actually solve a slightly different version of the Lasso, with centered Y, call it Y_c, and centered and scaled D, D_{cs}, where scaling D consists of dividing each column by its column length. This eliminates the need to model the intercept, and also ensures that each design column has zero mean and equal length. We denote these centered and scaled Lasso estimates as $\hat{\boldsymbol{\beta}}_{cs}$.

For the analysis of orthogonal arrays we use the Lasso as well, even though ordinary least squares are available since $n > k$. We did this in order to be consistent, in two ways. First, our goal is to compare the screening effectiveness when moving from OAs to SSDs. This seems most straightforwardly done by keeping the analysis method constant across design types. Secondly, we wanted to be consistent with the probability of perfect sign recovery measure, which is based upon the Lasso. However, it is fair to wonder whether we are shortchanging OAs by not using OLS to analyze them. In Section 2 of the Supplementary Material document, we provide some comparisons between Lasso and OLS for small OAs, and observe that the Lasso has higher TPRs than OLS based on p-values of 0.05.

3 Methodology

To quantify the difference between the screening effectiveness of orthogonal arrays and supersaturated designs, we will investigate the suite of designs shown in Table 1, across a number of effect sparsity assumptions and effect size conditions. The main comparisons of interest are between SSDs and OAs with the same number of factors, and so we investigate the difference in effectiveness when the effect sizes are the same but also when the effect sizes of the model queried by the SSD are larger than those of the OA. Though this is somewhat unconventional, one could imagine a more or less aggressive [-1,1] coding scheme in which numerical factors are transformed to result in larger or smaller effect sizes. We use both traditional simulation and a new criterion for screening design quality (Stallrich et al. 2024) based on the probability of perfect sign recovery for the Lasso.

3.1 Simulation

From model (1), we have that D is either an OA or SSD with entries ± 1, where each column represents a factor and each row a treatment. Without loss of generality, we assume $\sigma = 1$, making $\boldsymbol{\beta}$ the effect size. Let $A(\boldsymbol{\beta})$ denote the set of nonzero elements of $\boldsymbol{\beta}$, that is, $A(\boldsymbol{\beta}) = \{\beta_j | \beta_j \neq 0, j = 1, 2, \ldots, k\}$, while $I(\boldsymbol{\beta})$ represents the inactive factors, $I(\boldsymbol{\beta}) = \{\beta_j | \beta_j = 0, j = 1, 2, \ldots, k\}$. A screening design paired with an analysis method that accurately identifies the elements of the sets $A(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$, call them $\hat{A}(\boldsymbol{\beta})$ and $\hat{I}(\boldsymbol{\beta})$, is considered to be an effective screen. (Note that we suppress full notation below and refer to these sets as A, $\hat{A}$, I, and $\hat{I}$.) For the simulation study, we consider two metrics: true positive rate (TPR) and false positive rate (FPR). TPR is defined as the proportion of truly active factors correctly estimated as active, that is $n(A \cap \hat{A})/n(A)$, where $n(A)$ denotes the number of elements in set A. A large TPR indicates that the design is consistently identifying the truly active factors. On the other hand, FPR represents the proportion of truly inactive factors that are mistakenly estimated as active, $n(I \cap \hat{A})/n(I)$. An elevated FPR indicates a screening approach that isn't careful enough in its assignment of factors as active.

An important aspect of using the Lasso in 4 (or the related centered and scaled version) to estimate A and I is the choice of λ. For the analysis of a single experiment, a holistic analysis would include consideration of a profile plot; but in a simulation, we must clearly and automatically specify $\hat{A}$ and $\hat{I}$ in each iteration. Cross validation is a general approach to choose λ but traditionally is not used in small, highly structured experiments. Instead, we use the Bayesian Information Criterion (BIC) to determine λ as follows: We consider a grid of λ values uniformly placed from 0 to λ_0 in increments of 0.1, where $\lambda_0 = \max_j \left|D_{cs}^t Y_c\right|$ (see Weese et al. 2021). For each λ, we obtain the Lasso estimate and set $\hat{\beta}_{j,\lambda} = 0$ for all $\left|\hat{\beta}_{j,\lambda}\right| < \gamma$, where γ is chosen as discussed below. Then, for each λ we use OLS to refit the models using the columns of D corresponding to the nonzero $\hat{\beta}_{j,\lambda}$ and calculate the BIC of each fitted OLS model. The $\hat{A}$ and $\hat{I}$ corresponding to the λ with minimum BIC is the screen of choice.

The method we use to choose λ has an additional tuning parameter γ. Weese et al. (2017) followed an idealistic approach by setting γ equal to σ. Conversely, Phoa et al. (2009) adopted a data-driven approach, by setting $\gamma = r * (\max_j(\left|\hat{\beta}_{j,\lambda=0}\right|))$ where $r = 10\%$. Weese et al. (2021) suggested using r values of either 10%, 25% or 50%. The value of r involves a trade-off. If r is large, more factors from $A(\boldsymbol{\beta})$ will be in $\hat{A}(\boldsymbol{\beta})$ (larger TPR) but also in $\hat{I}(\boldsymbol{\beta})$ (larger FPR), while smaller r means that fewer factors will be zeroed out which lowers both TPR and FPR. In the present study, we want a value of r that produces meaningful comparisons. Specifically, we don't want comparisons where nearly all the results have TPRs of 1, TPRs quite low, or FPRs quite high. Based on our testing, we found $r = 0.5$ to produce such meaningful comparisons (see Section 1 of the Supplementary Material document for a comparison of $r = 0.1$ and $r = 0.5$).

In our simulations in Section 4, we compared the designs across a range of scenarios, defined by the combination of the following variables:

- Number of factors. We considered k as small as 7 and as large as 60 (see Table 1). For each k, we compare designs in four categories:

    1. Far from saturated (OA): $1.5k < n \leq 2k$ (e.g., $n = 40, k = 24$)
    2. Slightly above saturation (OA): $k < n \leq 1.25k$ (e.g., $n = 28, k = 24$)
    3. Slightly supersaturated (SSD): $n < k \leq 1.25n$ (e.g., $n = 20, k = 24$)
    4. Very supersaturated (SSD): $1.5n \leq k < 2n$ (e.g., $n = 14, k = 24$)

- Sparsity. Assuming ceil() rounds up to the nearest integer, we varied the number of active factors a such that $a = \{\text{ceil}(0.2k), \text{ceil}(0.25k), \text{ceil}(0.3k)\}$.

- Relative effect size. We abuse language slightly by making statements such as "the OA had effect sizes of 1," which should be interpreted to mean "the underlying true model queried by the OA had active factors with $\beta = 1$, where β is the coefficient value for all non-zero factor coefficients." We considered four effect size scenarios:

    1. For the a active factors, both the OAs and SSDs had effect sizes of 1.
    2. For the a active factors, the OAs had effect sizes of 1 while the SSDs had effect sizes of 1.5.
    3. For the a active factors, the OAs had effect sizes of 0.5 while the SSDs had effect sizes of 1.
    4. For the a active factors, the OAs had an effect sizes of 0.5 while the SSDs had effect sizes of 1.5.

As an example, consider the following simulation setting: 24 factors with SSDs $n = 14$ (very supersaturated design, $n << k$), $n = 20$ (slightly supersaturated design, $n < k$), OA

Table 1: Designs to investigate. Each design size is specified as (n, k).

Very supersaturated ($n << k$)	*Slightly supersaturated ($n < k$)*	*Slightly above saturation ($n > k$)*	*Far from saturated ($n >> k$)*	*Size of screen*
(4, 7)	(6, 7)	(8, 7)	(12, 7)	Small
(5, 8)	(7, 8)	(12, 8)	(16, 8)	
(6, 9)	(8, 9)	(12, 9)	(16, 9)	
(6, 10)	(8, 10)	(12, 10)	(16, 10)	
(8, 12)	(10, 12)	(16, 12)	(20, 12)	Medium
(14, 23)	(20, 23)	(24, 23)	(36, 23)	
(14, 24)	(20, 24)	(28, 24)	(40, 24)	
(20, 34)	(28, 34)	(36, 34)	(52, 34)	
(25, 40)	(37, 40)	(44, 40)	(64, 40)	Large
(28, 45)	(43, 45)	(48, 45)	(72, 45)	
(30, 50)	(46, 50)	(52, 50)	(80, 50)	
(35, 60)	(50, 60)	(64, 60)	(92, 60)	

$n = 28$ (slightly above saturation, $n > k$), and OA $n = 40$ (far from saturated, $n >> k$); $a = \text{ceil}(0.2k) = 5$; and SSD effect size of $\beta_{SSD} = 1$ and OA effect size of $\beta_{OA} = 0.5$. In this case we are comparing the effectiveness of 24-factor designs with run sizes ranging from 14 and 20 (SSDs) to 28 and 40 (OAs), while comparing the effectiveness of the two SSDs with the OAs when SSDs have effect sizes twice as large.

In the detailed simulation protocol that follows, note that we simulate from model (1) with $\beta_0 = 0$. Each simulation scenario provides the number of factors k; a design matrix D (one of design types $n << k$, $n < k$, $n > k$, or $n >> k$); a; and a value β_{OA} or β_{SSD} for β_j's corresponding to active factors, depending on design category. To simulate from a particular design, do the following for each of n_{iter} iterations:

1. Randomly select a out of k columns from D, establishing $A(\boldsymbol{\beta})$ and $I(\boldsymbol{\beta})$.
2. Depending on the design type, assign β_{OA} or β_{SSD} to all $\beta_j \in A(\boldsymbol{\beta})$ and 0 to all $\beta_j \in I(\boldsymbol{\beta})$.
3. Assign each nonzero β_j a $+1$ or -1, with equal probability.
4. Simulate Y from model (1) and D.
5. Center Y and center and scale D.

6. Use the grid of λ described above to fit the Lasso and obtain $\hat{\boldsymbol{\beta}}_{cs,\lambda}$ for each λ. (For simplicity of presentation, we refer to the centered and scaled Lasso estimate $\hat{\boldsymbol{\beta}}_{cs,\lambda}$ as $\hat{\boldsymbol{\beta}}_{\lambda}$ in the steps below.)

7. For each $\hat{\boldsymbol{\beta}}_{\lambda}$, check if $\left|\hat{\beta}_{j,\lambda}\right| < \gamma$, where $\gamma = 0.5 * (\max_j(\left|\hat{\beta}_{j,\lambda=0}\right|))$; if $\left|\hat{\beta}_{j,\lambda}\right| < \gamma$ set $\hat{\beta}_{j,\lambda} = 0$.

8. For each λ, use the columns of D associated with the non-thresholded $\hat{\beta}$ to fit an OLS model to Y. Calculate the BIC for each model and select the one with the minimum. The factors associated with the non-zero $\hat{\beta}$'s from the chosen model are included in $\hat{A}$, while the rest are assigned to $\hat{I}$.

Based on $\hat{A}$ and $\hat{I}$, along with A and I, we compute $TPR = \frac{1}{n_{\text{iter}}} \sum_{i=1}^{n_{\text{iter}}} n(A \cap \hat{A})_i / n(A)_i$ and $FPR = \frac{1}{n_{\text{iter}}} \sum_{i=1}^{n_{\text{iter}}} n(I \cap \hat{A})_i / n(I)_i$. In our work, we take $n_{\text{iter}} = 1000$.

3.2 Probability of perfect sign recovery method

An important downside of using simulation to compare designs here is the need to specify the γ threshold. Its choice is somewhat arbitrary if the error variance is not available, though it affects TPR and FPR. To address this, Stallrich et al. (2024) introduced a method to compute the probability of perfect sign recovery for a screening design under the Lasso and the model given in (1), and Young et al. (2024) illustrate its capability to improve reproducibility in the comparison of screening designs. Given the true sign vector $\mathbf{z} = \text{sign}(\boldsymbol{\beta})$ and the estimated sign vector, $\hat{\mathbf{z}} = \text{sign}(\hat{\boldsymbol{\beta}}_{cs})$, the criterion ranks designs based on $\phi_\lambda(D|\boldsymbol{\beta}) = P(\hat{\mathbf{z}} = \mathbf{z})$. The criterion is based on the joint probability of two independent events, both a function of normal random vectors. These events are derived from the Lasso's Karush-Kuhn-Tucker conditions; details can be found in Stallrich et al. (2024).

Computing ϕ_λ requires knowledge of the entire model, including $\boldsymbol{\beta}$ along with the Lasso parameter λ. Thus, on its own it can't be used to evaluate designs. Stallrich et al. (2024) proposed a series of relaxed criteria which are more practically useful. For instance, they present a criterion $\phi_\Lambda(D|\boldsymbol{\beta})$ which integrates ϕ_λ over $\log(\lambda)$, thus allowing the computation of the criterion without knowledge of λ. Ultimately, they provide $\Phi_\Lambda(D|a, \beta)$, which is the $\log(\lambda)$-integrated measure averaged over all supports of size a (that is, all sets of a factors chosen from k) and assuming that, for all active factors, $\beta = \frac{|\beta_j|}{\sigma} \geq 1$ so that it is at least as large as the noise. Thus, as long as we can assume an effect size β and a number of active effects of interest, along with sign vector $\mathbf{z}$, Φ_Λ can be used to evaluate a design without knowledge of λ, A, or $\boldsymbol{\beta}$. For example, this criterion can be used to evaluate a design under the assumption that effect directions are known. On the other hand, if no sign information is available, another criterion $\Phi_\Lambda^{\pm}(D|a, \beta)$ can be computed which averages Φ_Λ over all possible sign vectors. Note that these criteria are computationally intensive to compute. One way to reduce computational burden is to average not over all $\binom{k}{a}$ possible supports but a subset of these supports. For this and other computationally related discussion, see Section 4 of Stallrich et al. (2024). Note also that the values of Φ_Λ and $\Phi_\Lambda^{\pm}$ don't have a clear interpretation on their own, since they are average probabilities, integrated over $\log(\lambda)$. However, they are useful to compare designs because larger values of the measures mean the design is more effective at recovering the true signs of the factors.

Because the probability method is more stringent (perfect sign recovery rather than TPR and FPR), the effect sizes necessary to give meaningful results are larger. So instead of the four effect size scenarios from Section 3.1, we use the following:

- For the a active factors, both the OAs and SSDs had effect sizes of 3.
- For the a active factors, the OAs had effect sizes of 6 while the SSDs had effect sizes of 4.
- For the a active factors, the OAs had effect sizes of 4 while the SSDs had effect sizes of 2.
- For the a active factors, the OAs had effect sizes of 6 while the SSDs had effect sizes of 2.

For this probability method, we evaluated the same set of designs (Table 1) across the same set of sparsity levels.

4 Results

In this section, in Figures 2, 3, and 4, we present the results of our simulation and probability calculations, for small, medium, and large-sized screening designs, respectively. For these main simulations, we use $Var(s+)$ designs for the supersaturated experiments (design categories: $n << k$ and $n < k$ in Table 1) and orthogonal arrays for those experiments which are not supersaturated (design categories: $n > k$ and $n >> k$ in Table 1). See Section 2.1 for more about the designs and their construction. Because the probability of perfect sign recovery is quite stringent, we have used larger effect magnitudes in the probability of perfect sign recovery comparisons compared to those in the simulation; however, the ratio of the SSD to OA effect sizes are the same across the two methods of comparison. The first columns of the figures represent SSDs and OAs evaluated under the same effect size conditions. The other columns represent SSDs evaluated assuming larger effect sizes than the OAs, in order to assess conditions under which screening effectiveness may be similar.

Consider first the small design results in Figure 2. As an example, the top left plot in Figure 2a shows that when $a = \text{ceil}(0.2k) = 2$ and the active effect sizes are 1, the 4-run, 7-factor SSD (red line; cross-reference with Table 1) has a true positive rate well below 50%, while the 6-run, 7-factor SSD (green line) has a TPR of approximately 70%. Both of the OAs, meanwhile, have TPRs above 85%. In addition, in this example the false positive rates are also ordered by the design sizes: the $n << k$ design with the largest FPR and the $n >> k$ design with the smallest. The probability of perfect sign recovery plots in Figure 2b can be read similarly, although instead of TPR and FPR we have the integrated probability measure. Note that since this probability measure is much more stringent, even with larger effect sizes the values are much lower.

The most striking aspect of Figure 2 is how clearly inferior the $n << k$ designs are. That is, compared to the OAs, even when advantaged with effect sizes up to three times larger, $n << k$ designs are *still* less likely to identify truly active factors in the simulations of Figure 2a. In the probability calculations of Figure 2b, the $n << k$ designs fare a little better when effect sizes are two or three times larger, at least for relatively more sparse settings. On the other hand, when SSD effect sizes are twice that of the OAs, the $n < k$ designs are of similar quality (in the simulations) or even better quality (in the probability calculations) than the OAs. We also see that there is surprisingly little advantage to $n >> k$ OAs over $n > k$ OAs, except that they tend to have somewhat smaller FPRs.

For the results for the medium-sized screening designs (Figure 3), the story is similar. That is, the $n << k$ SSDs are still notably inferior in nearly all settings and $n < k$ SSDs are similar in quality to the OAs when effect sizes are doubled in the simulations and arguably slightly better when effect sizes are doubled in the probability calculations. The $n >> k$ OAs are again little better than $n > k$ OAs.

For the large design results (Figure 4), conclusions are slightly different. For the $n < k$ SSDs, overall when the SSD effects are twice that of OA effects, these designs perform better than the OAs, and for the 1.5 vs. 1 setting the $n < k$ SSDs are similar in performance to the OAs. The $n << k$ SSDs still obviously lag behind the other designs in nearly all settings; it is notable that in Figure 4b the chances of *perfect* sign recovery for these bold designs is near or equal to 0 in several scenarios. This likely reflects the strictness of the probability measure that becomes more and more severe as the number of factors increases.

5 Discussion and Conclusion

The purpose of this study was to compare the performance of supersaturated designs (SSDs) against the better-known orthogonal arrays (OAs), in terms of their ability to screen important factors. We used two methods of evaluation: traditional statistical simulation and a new probability of perfect sign recovery method for the Lasso. We compared designs with small, medium, and large numbers of factors, and for each size we examined smaller and larger SSDs as well as smaller and larger OAs. We also considered several sparsity levels and four effect size ratios. A key aspect of our scenarios is that we compared OAs with smaller effect sizes to SSDs with larger effect sizes, in order to provide an assessment of conditions under which true positive rates were similar. These differential effect sizes might conceptually correspond to, for instance, situations in which numerical factors are transformed from different intervals in their natural units to $[-1, 1]$.

As expected, we found that the OAs exhibited better screening performance than the SSDs when the underlying model had effects of the same size. When the data obtained by the SSD had effect sizes 1.5 times those of the OA, the OAs were still preferable for small and medium-sized designs, while the slightly supersaturated designs ($n < k$) had largely comparable performance to the OAs. When the effect sizes were twice or three times as large for the SSDs, in most cases the slightly supersaturated designs ($n < k$) were preferable to the OAs. Two other notable insights: (1) in most of the cases considered, no matter the size ratio, the very supersaturated designs ($n << k$) were clearly worse than all the other designs; and (2) the large OAs ($n >> k$) showed little improvement over the smaller OAs ($n > k$).

The results of this study provide guidance to practitioners by comparing the effectiveness of an established class of designs (orthogonal arrays) with a bolder type of screening design that is used less frequently (supersaturated designs). To put it another way, our study provides an opportunity for experimenters to calibrate the relatively unknown effectiveness of SSDs to their experience and knowledge of OAs. For example, in a setting in which an experimenter is comfortable using an OA because adequate effect sizes are anticipated, our study suggests that in order to reduce experimental resources by running a slightly supersaturated design, the effect size that the smaller design will reliably detect will be between 1.5 and 2 times the effect size comfortably detected by the OA.

This may seem a steep price to pay for using even slightly supersaturated designs (and our results are even less optimistic about $n << k$ supersaturated designs). However, this leads to a further insight. Orthogonal arrays have at least some capacity to screen moderate or even small effects, while supersaturated designs should typically be used in cases where effect sizes and/or sparsity is expected to be greater. Notice that in our simulations we considered effect sizes no larger than 1.5. In many cases, supersaturated designs are inadequate to reliably detect effects of that size. However, if the application is focused on identifying larger effects, or sparsity is much higher, then even very supersaturated designs ($n << k$) will be effective. We have included a set of results exploring $n << k$ SSDs for larger effect sizes in Section 4 of the Supplementary Material document; see also Kainkaryam and Woolf (2008); Ji et al. (2023); Liu et al. (2024); Smucker et al. (2024) for SSD applications in biological and chemical high-throughput screening for which sparsity is typically expected to be high and effect sizes large.

The two methods of design assessment, simulation and probability of perfect sign recovery, generally agree when making relative comparisons. That is, for a given scenario the methods generally agree which design category is preferred. However, the absolute assessments disagree substantially. We see that our simulation results suggest relatively large true positive rates for all but the very supersaturated designs, for effect sizes of 1σ or 1.5σ. In contrast, even for effect sizes as large as 4σ or 6σ, the probability of perfect sign recovery approach is much lower. There are at least two reasons for this discrepancy. First, the probabilistic framework's criterion is *perfect* sign recovery. Thus, in the presence of several active effects an analysis could have a fairly large TPR, with a small FPR, but still fail to screen perfectly. This is particularly true for larger screens. Secondly, the simulation approach uses a threshold parameter γ while the probability approach does not. Using a threshold prevents parameters with small estimates from being included as important effects.

Future work could continue to explore in more detail the relationship between the number of runs and screening effectiveness. We have studied this connection using four fairly crude categories, but additional investigation could provide better definition. We also did not account for the possibility of screening contamination due to unmodeled interactions. Though interactions and OAs have been studied extensively, little work exists which addresses the challenging problem of interactions and SSDs (one exception is Singh and Stufken 2024). An unanswered question is whether the existence of unmodelled interactions would affect SSDs more seriously than a well-chosen OA of strength 3.

Acknowledgements

The authors are grateful to Maria Weese, Steve Wright, and two anonymous reviewers for their feedback on this work. The authors have no funding support or conflicts of interest to acknowledge.

Supplementary Material

The supplementary material accompanying this paper includes:

- "A Supplementary Material document.pdf" - a document referred to throughout the main paper, with additional simulation results.
- "B Code and Data" - a folder with R code which simulates and calculates the probability of perfect sign recovery, data that is the result of the simulations and probability calculations, and R code that uses the data to produce the plots in the paper and supplementary document.
- "C Designs" - the designs used in the paper.

References

Asif, M. S. and Romberg, J. (2010), "On the lasso and dantzig selector equivalence," in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, IEEE, pp. 1–6.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous analysis of Lasso and Dantzig selector," *The Annals of Statistics*, 37, 1705–1732.

Booth, K. H. and Cox, D. R. (1962), "Some systematic supersaturated designs," *Technometrics*, 4, 489–495.

Bose, R. C. and Bush, K. A. (1952), "Orthogonal arrays of strength two and three," *The Annals of Mathematical Statistics*, 23, 508–524.

Bush, K. A. (1952), "Orthogonal arrays of index unity," *The Annals of Mathematical Statistics*, 426–434.

Carpinteiro, J., Quintana, J., Martınez, E., Rodrıguez, I., Carro, A., Lorenzo, R., and Cela, R. (2004), "Application of strategic sample composition to the screening of anti-inflammatory drugs in water samples using solid-phase microextraction," *Analytica chimica acta*, 524, 63–71.

Chen, G., Shi, C., and Tang, B. (2023), "Nonregular designs from Paley's Hadamard matrices: Generalized resolution, projectivity and hidden projection property," *Electronic Journal of Statistics*, 17, 2120–2138.

Cheng, C.-S. (1995), "Some projection properties of orthogonal arrays," *The Annals of Statistics*, 23, 1223–1233.

— (1998), "Some hidden projection properties of orthogonal arrays with strength three," *Biometrika*, 85, 491–495.

Cheng, C.-S., Mee, R. W., and Yee, O. (2008), "Second order saturated orthogonal arrays of strength three," *Statistica Sinica*, 105–119.

Dejaegher, B. and Vander Heyden, Y. (2008), "Supersaturated designs: set-ups, data interpretation, and analytical applications," *Analytical and bioanalytical chemistry*, 390, 1227–1240.

Draguljić, D., Woods, D. C., Dean, A. M., Lewis, S. M., and Vine, A.-J. E. (2014), "Screening strategies in the presence of interactions," *Technometrics*, 56, 1–1.

Georgiou, S. D. (2014), "Supersaturated designs: A review of their construction and analysis," *Journal of Statistical Planning and Inference*, 144, 92–109.

Grömping, U. (2018), "R package DoE. base for factorial experiments," *Journal of Statistical Software*, 85, 1–41.

He, Y., Lin, C. D., and Sun, F. (2022), "A new and flexible design construction for orthogonal arrays for modern applications," *The Annals of Statistics*, 50, 1473–1489.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999), *Orthogonal arrays: theory and applications*, Springer Science & Business Media.

James, G. M., Radchenko, P., and Lv, J. (2009), "DASSO: connections between the Dantzig selector and lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71, 127–142.

Ji, H., Lu, X., Zhao, S., Wang, Q., Liao, B., Bauer, L. G., Huber, K. V., Luo, R., Tian, R., and Tan, C. S. H. (2023), "Target deconvolution with matrix-augmented pooling strategy reveals cell-specific drug-protein interactions," *Cell Chemical Biology*, 30, 1478–1487.

Jones, B., Lin, D. K., and Nachtsheim, C. J. (2008), "Bayesian D-optimal supersaturated designs," *Journal of Statistical Planning and Inference*, 138, 86–92.

Jones, B. and Majumdar, D. (2014), "Optimal supersaturated designs," *Journal of the American Statistical Association*, 109, 1592–1600.

Jones, B. A., Li, W., Nachtsheim, C. J., and Kenny, Q. Y. (2009), "Model-robust supersaturated and partially supersaturated designs," *Journal of Statistical Planning and Inference*, 139, 45–53.

Jridi, M., Lassoued, I., Kammoun, A., Nasri, R., Nasri, M., Souissi, N., et al. (2015), "Screening of factors influencing the extraction of gelatin from the skin of cuttlefish using supersaturated design," *Food and Bioproducts Processing*, 94, 525–535.

Kainkaryam, R. M. and Woolf, P. J. (2008), "poolHiTS: A Shifted Transversal Design based pooling strategy for high-throughput drug screening," *BMC bioinformatics*, 9, 1–11.

Lin, D. K. (1993), "A new class of supersaturated designs," *Technometrics*, 35, 28–31.

Liu, N., Kattan, W. E., Mead, B. E., Kummerlowe, C., Cheng, T., Ingabire, S., Cheah, J. H., Soule, C. K., Vrcic, A., McIninch, J. K., et al. (2024), "Scalable, compressed phenotypic screening using pooled perturbations," *Nature Biotechnology*, 1–13.

Lounici, K. (2008), "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators," *Electronic Journal of Statistics*, 2, 90–102.

Marley, C. J. and Woods, D. C. (2010), "A comparison of design and model selection methods for supersaturated experiments," *Computational Statistics & Data Analysis*, 54, 3158–3167.

Mee, R. (2009), *A comprehensive guide to factorial two-level experimentation*, Springer Science & Business Media.

Mee, R. W., Schoen, E. D., and Edwards, D. J. (2017), "Selecting an orthogonal or nonorthogonal two-level design for screening," *Technometrics*, 59, 305–318.

Meinshausen, N., Rocha, G., and Yu, B. (2007), "Discussion: A tale of three cousins: Lasso, L2Boosting and Dantzig," *The Annals of Statistics*, 35, 2373–2384.

Nguyen, N.-K., Pham, T.-D., and Vuong, M. P. (2023), "A Catalog of 2-Level Orthogonal Minimally Aliased Designs with Small Runs," *Journal of Statistical Theory and Practice*, 17, 26.

Owen, A. B. (1992), "Orthogonal arrays for computer experiments, integration and visualization," *Statistica Sinica*, 439–452.

Phoa, F. K., Pan, Y.-H., and Xu, H. (2009), "Analysis of supersaturated designs via the Dantzig selector," *Journal of Statistical Planning and Inference*, 139, 2362–2372.

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

Rao, C. R. (1947), "Factorial experiments derivable from combinatorial arrangements of arrays," *Supplement to the Journal of the Royal Statistical Society*, 9, 128–139.

Satterthwaite, F. (1959), "Random balance experimentation," *Technometrics*, 1, 111–137.

Schoen, E. D., Vo-Thanh, N., and Goos, P. (2017), "Two-level orthogonal screening designs with 24, 28, 32, and 36 runs," *Journal of the American Statistical Association*, 112, 1354–1369.

Singh, R. and Stufken, J. (2023), "Selection of two-level supersaturated designs for main effects models," *Technometrics*, 65, 96–104.

— (2024), "Factor selection in screening experiments by aggregation over random models," *Computational Statistics & Data Analysis*, 194, 107940.

Smucker, B. J., Wright, S. E., Williams, I., Page, R. C., Kiss, A. J., Silwal, S. B., Weese, M., and Edwards, D. J. (2024), "Large Row-Constrained Supersaturated Designs for High-throughput Screening," *arXiv preprint arXiv:2407.06173*.

Stallrich, J. W., Young, K., Weese, M. L., Smucker, B. J., and Edwards, D. J. (2024), "An Optimal Design Framework for Lasso Sign Recovery," `https://arxiv.org/abs/2303.16843`, version 2, revised 2024-03-15.

Tibshirani, R. (1996), "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58, 267–288.

Vazquez, A. R., Goos, P., and Schoen, E. D. (2019), "Constructing two-level designs by concatenation of strength-3 orthogonal arrays," *Technometrics*, 61, 219–232.

Vazquez, A. R., Schoen, E. D., and Goos, P. (2022), "Two-level orthogonal screening designs with 80, 96, and 112 runs, and up to 29 factors," *Journal of Quality Technology*, 54, 338–358.

Weese, M. L., Edwards, D. J., and Smucker, B. J. (2017), "A criterion for constructing powerful supersaturated designs when effect directions are known," *Journal of Quality Technology*, 49, 265–277.

Weese, M. L., Smucker, B. J., and Edwards, D. J. (2015), "Searching for powerful supersaturated designs," *Journal of Quality Technology*, 47, 66–84.

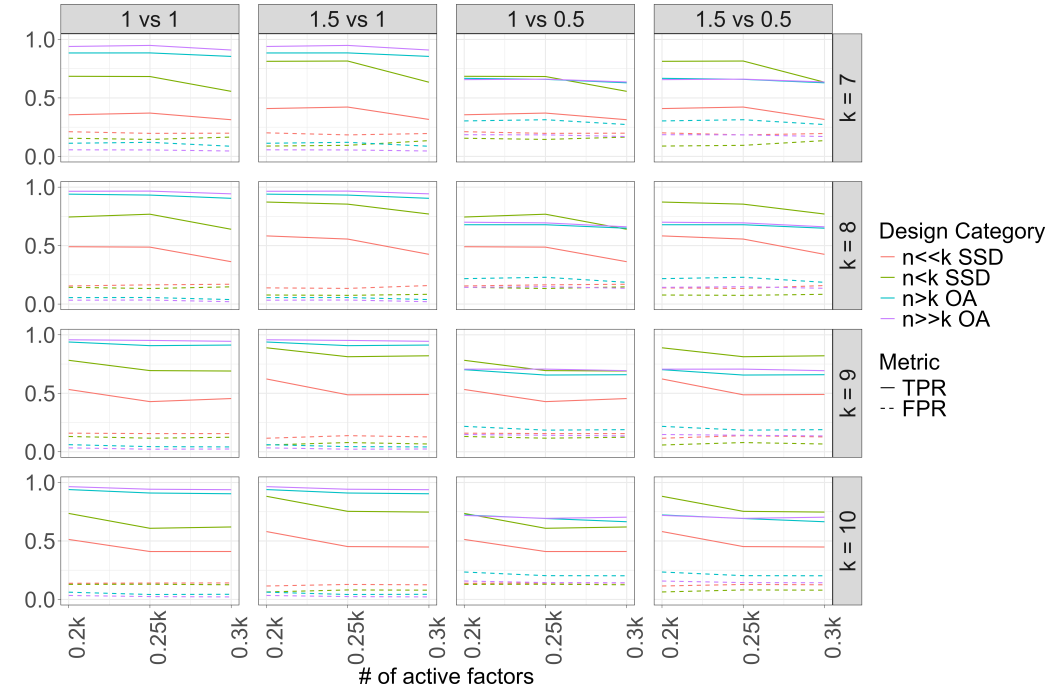
Weese, M. L., Stallrich, J. W., Smucker, B. J., and Edwards, D. J. (2021), "Strategies for supersaturated screening: Group orthogonal and constrained var (s) designs," *Technometrics*, 63, 443–455.

Wu, C. (1993), "Construction of supersaturated designs through partially aliased interactions," *Biometrika*, 80, 661–669.

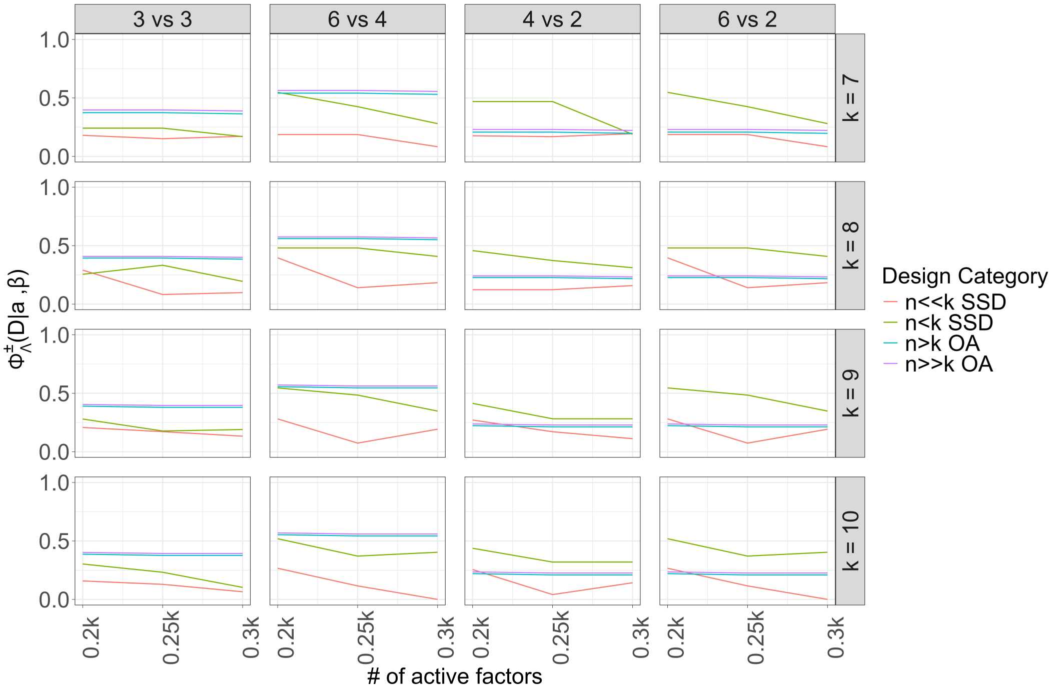
Xu, H., Phoa, F. K., and Wong, W. K. (2009), "Recent developments in nonregular fractional factorial designs," .

Young, K., Weese, M. L., Stallrich, J. W., Smucker, B. J., and Edwards, D. J. (2024), "A graphical comparison of screening designs using support recovery probabilities," *Journal of Quality Technology*, 56, 355–368.

Zarkadas, S. and Besseris, G. (2023), "Using Lean-and-Green Supersaturated Poly-Factorial Mini Datasets to Profile Energy Consumption Performance for an Apartment Unit," *Processes*, 11, 1825.

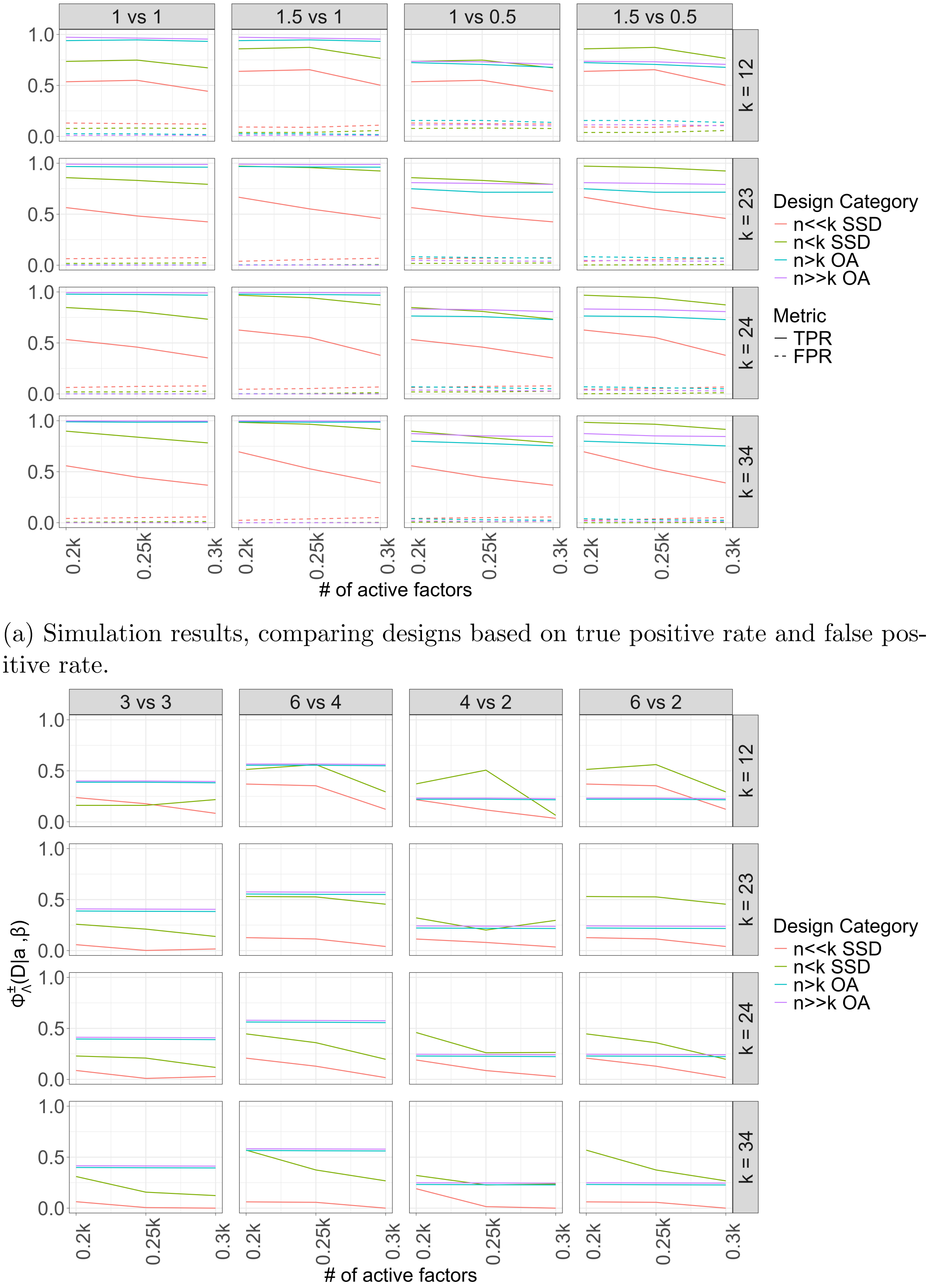


(a) Simulation results, comparing designs based on true positive rate and false positive rate.



(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}\big|\, a, \beta)$.

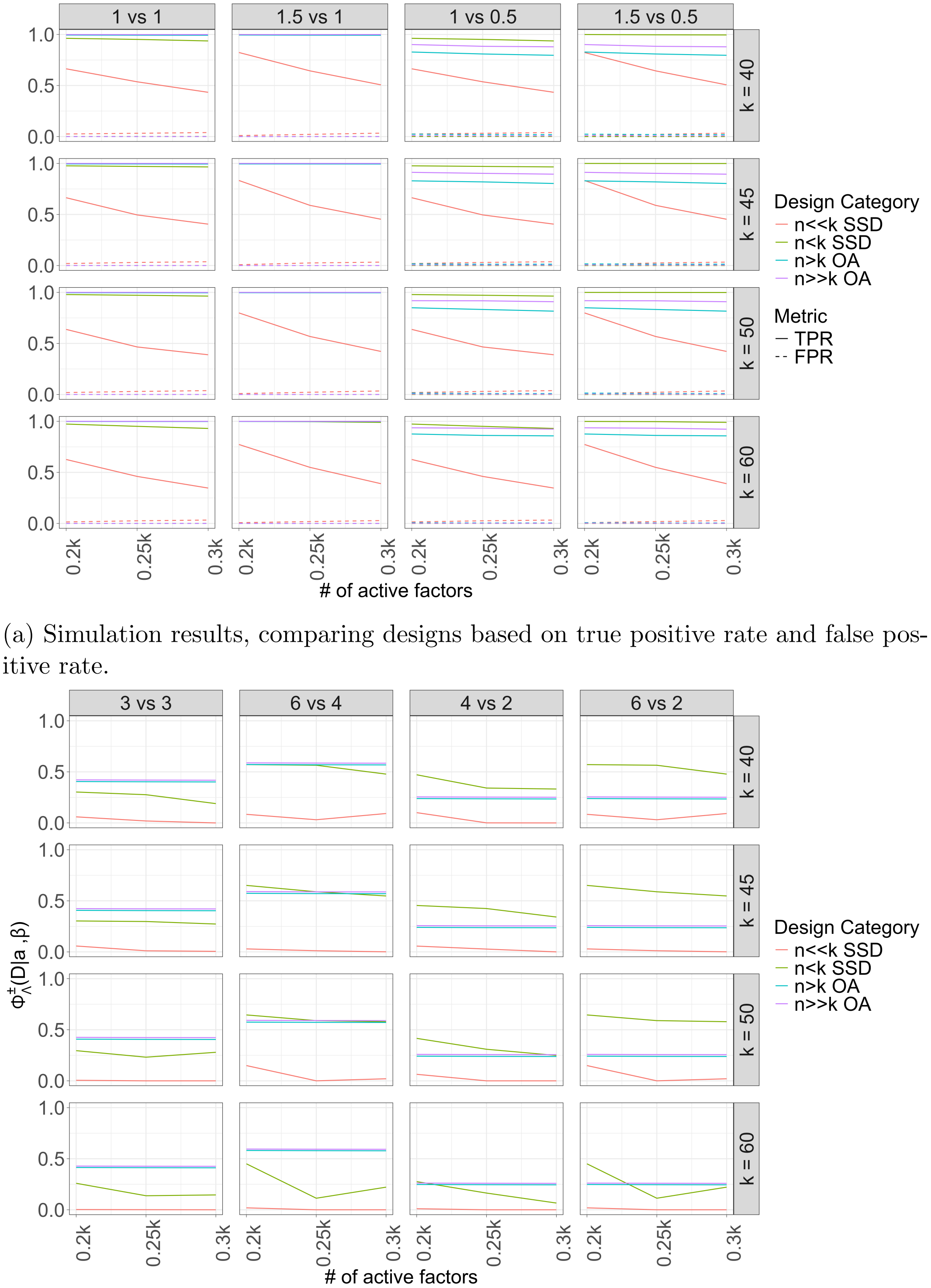
Figure 2: For small screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: β_{SSD} vs. β_{OA}. The number of factors are shown along the right-hand side, while the Design Category indicates the specific run sizes for each design as displayed in Table 1.



(a) Simulation results, comparing designs based on true positive rate and false positive rate.

(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}\big|\, a, \beta)$.

Figure 3: For medium-sized screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: β_{SSD} vs. β_{OA}. The number of factors are shown along the right-hand side, while the Design Category indicates the specific run sizes for each design as displayed in Table 1.



(a) Simulation results, comparing designs based on true positive rate and false positive rate.

(b) Probability of perfect sign recovery results, comparing designs based on $\Phi^{\pm}_{\Lambda}(\mathrm{D}\big|\, a, \beta)$.

Figure 4: For large screening designs. Both approaches assume unknown sign vector. The effect sizes are displayed along the top: β_{SSD} vs. β_{OA}. The number of factors are shown along the right-hand side, while the Design Category indicates the specific run sizes for each design as displayed in Table 1.