Selecting an orthogonal or non-orthogonal two-level design for screening

Selecting an Orthogonal or Non-Orthogonal Two-Level Design for Screening

Robert W. Mee
Department of Business Analytics and Statistics, University of Tennessee,
Knoxville TN, 37996 (rmee@utk.edu)

Eric D. Schoen
Department of Engineering Management, University of Antwerp, Belgium
and TNO, Zeist, Netherlands (eric.schoen@uantwerpen.be)

David J. Edwards
Department of Statistical Sciences and Operations Research
Virginia Commonwealth University, Richmond VA, 23284 (dedwards7@vcu.edu)

January 18, 2016

Abstract

This article presents a comparison of criteria used to characterize two-level orthogonal arrays and nonorthogonal designs for screening purposes. To articulate the relationships between criteria, we focus on seven-factor designs with 16–32 runs and 11-factor designs with 20–48 runs. Screening based on selected designs for each of the run sizes considered is studied with simulation using a forward selection procedure and the Dantzig selector. Bayesian D-optimal designs and designs created algorithmically to optimize estimation capacity over various model spaces provide an alternative to orthogonal arrays. This article contrasts these non-orthogonal designs with the orthogonal arrays using both estimation-based criteria and simulation. In this way, we furnish both general insights regarding various design approaches, as well as a guide for how to choose among a few final candidate designs.

Keywords: Estimation capacity, Extended word length pattern, False discovery rate, G-aberration, Orthogonal array, Power, Q_B

1 INTRODUCTION

One of the possible purposes of a two-level fractional factorial design (FFD) is to screen a large number of factors to determine whether each is active or not. Initial interest in FFDs focused on *regular* designs. Assuming that the factor levels are coded with -1 and $+1$, these designs are constructed by taking a full factorial in the first few factors and defining the additional factors by multiplying subsets of the first factors. Regular FFDs and criteria to evaluate these designs are standard topics of introductory textbooks on experimental design such as Montgomery (2012).

Plackett and Burman (1946) introduced *nonregular* FFDs to the statistics literature. While regular FFDs require that the run size be a power of two, nonregular orthogonal designs are less restrictive in run size, being available for run sizes equaling any multiple of four. However, until the early 1990s experimenters were cautioned against these designs, because the aliasing between main effects and two-factor interactions and the aliasing among two-factor interactions is more complex than the aliasing in regular FFDs. For this reason nonregular FFDs were formerly recommended only when one could be sure that the two-factor interactions were negligible.

Hamada and Wu (1992), Lin and Draper (1993), Cheng (1995) and Wang and Wu (1995) showed that it is indeed possible to recover information on one or two interactions in a nonregular FFD. This feature sparked interest in developing new nonregular designs along with new criteria to evaluate these designs; references include Deng and Tang (1999, 2002); Tang and Deng (1999); Li and Nachtsheim (2000); Ingram and Tang (2005); Xu (2005); Xu and Deng (2005); Loeppky et al. (2007); Sun et al. (2008); Bulutoglu and Margot (2008), and Schoen et al. (2010).

The screening task we intend for the experimental design is to identify main effects and two-factor interactions, assuming all higher-order interactions are negligible. Since it is rare for more than a small number of two-factor interactions to be important (Li et al., 2006), we consider designs where the run size is generally too small to estimate all two-factor interactions. In addition to regular and nonregular orthogonal FFDs, several optimal design criteria have been proposed for constructing designs in this context; see DuMouchel and Jones (1994), Li and Nachtsheim (2000), and Smucker et al. (2012). In our subsequent comparisons, we will contrast these non-orthogonal FFDs with orthogonal FFDs.

Particular screening applications differ in terms of the number of factors involved and the nature of prior information available. The following describes one such application. In a recent experiment, researchers of TNO (Eindhoven, the Netherlands) were involved in making phantoms to calibrate medical devices. Phantoms are cylindrical pieces of gelatinous material that mimic human tissues; these tissues are to be investigated with the device once it is properly calibrated. A phantom is tested by exposing it to light of various wavelengths. For each of the wavelengths, the reflection is recorded, which can be affected by the absence or presence of seven ingredients. One of the ingredients scatters light of all wavelengths and so is expected to be active for all wavelengths. The other six ingredients were colorants, which mainly absorb the light around a specific wavelength. The main interest was in the size of the factorial effects. Only a few of the ingredients are expected to be active for any given wavelength. Further, optical laws suggest that main effects are much more prominent than interaction effects. Given this scenario, 16 or 20 phantoms should be a sufficient number to identify the largest main effects and interactions at each wavelength. We will revisit this example later.

Generally, the design literature lists the best few designs under some criterion for given run sizes and numbers of factors. This was understandable in the initial development of criteria. However, now that the field has matured, there is need for a comparison of the many criteria. Also, existing literature rarely addresses directly the potential of the designs to detect active effects. Simulation studies to compare designs are computationally intensive if they contemplate a wide variety of assumed models and possible designs. The results can also be highly dependent on the analysis methods used and any required tuning parameters. While simulation studies are often employed to compare analysis methods, this is not true for comparing designs. Miller and Sitter (2001, 2005) did compare designs using simulation, but they assumed that the main effects have been correctly identified in a first stage analysis, and simply simulate the probability that the true interaction model has a smaller error sum of squares than any other model the same size – thus sidestepping the problem of choosing between models of various sizes. Also, Liao and Chai (2009) compared five designs with partial or full replication with two unreplicated FFDs, yet their comparison ignores the consequences of aliasing; see Mee et al. (2009). In a study relevant to our purposes, Draguljić et al. (2014) compare Bayesian D-optimal designs for 10, 15, and 20 factors with two-stage group screening experiments, where the first stage consisted of a 2^{5-1} experiment with five groups and the second stage utilized Bayesian D-optimal designs, with the run size depending on the number of groups identified as active. The Dantzig selector (DS) (Candes and Tao, 2007) was found to be the most successful analysis method for both design strategies. Using the DS, the two-stage group screening approach performed slightly better than the one-stage designs. Since one-stage designs are simpler to implement – and there are many such designs to compare – this article will focus on screening using single-stage designs.

Our article improves the literature in three ways. First, we provide insights that will aid practitioners in understanding the relevance of various criteria to evaluate designs. Second, by simulation of the screening process we compare nonregular FFDs with three types of nonorthogonal designs. Finally, by comparing best designs over multiple run sizes, we illustrate how one may use simulation to select the run size for screening, based on achieving the desired power.

The organization of the paper is as follows. In Section 2, we give a quick survey of regular two-level FFDs, as most criteria for nonregular designs are generalizations of simpler, regular fraction criteria. In Section 3, we extend the summary to nonregular design criteria. Many of these criteria can be applied to nonorthogonal designs as well. We illustrate the criteria with the cases of 7 factors in 20 runs and 11 factors in 40 runs; the numerous orthogonal designs with $n =$ 40 runs will enable us to appreciate differences in the rankings provided by the various criteria. Next, in Section 4, we use simulation to evaluate directly the screening potential of various designs, assuming models of different sizes and applying two common analysis methods. The simulation illustrates the consequence of increasing the number of factors in an experiment by evaluating designs with 7 and 11 factors. In Section 5, we return to TNO's phantoms experiment and evaluate 22 designs ranging in size from 16 to 32 runs. The paper ends with a brief discussion.

2 CRITERIA FOR REGULAR FRACTIONS

Regular 2^{k-f} fractions were initially characterized by their resolution. For example, consider the 2^{7-4} fraction produced by augmenting a full 2^3 factorial in the three factors $\{A, B, C\}$ with the $f = 4$ generated factors $D = ABC, E = AB, F = AC$, and $G = BC$. Using columns of ± 1 for each factor, the defining relation for a regular fraction consists of all $2^f - 1$ interactions that are identically $+1$ for all treatment combinations in the fraction; these consist of f words produced directly from the generators (e.g., $ABCD, ABE, ACF$, and BCG) and all generalized interactions of these (such as $ABCD * ABE = CDE$). Resolution, defined as the defining relation's shortest word length, reflects the most critical aliasing. Resolution III designs, such as this 2^{7-4}_{III}, alias two-factor interactions with main effects.

The word length pattern (wlp) discriminates between designs more effectively than resolution does by counting how many of the interactions in the defining relation are of a given length. The wlp for the 2^{7-4} design is

$$\text{wlp} = (A_3, A_4, \ldots, A_7) = (7, 7, 0, 0, 1), \tag{1}$$

where A_j denotes the number of j-factor interactions in the defining relation (j = 3, ..., k).

All 2^{7-4}_{III} designs are *isomorphic* in that one design can be obtained from another by reordering rows, reordering columns, and/or reversing the two levels for some columns. When $k < n - 3$, there are regular fractions that are not isomorphic to one another (Mukerjee and Wu, 2006, p. 59). When non-isomorphic designs exist, the aberration criterion (Fries and Hunter, 1980), which sequentially orders the designs based on the wlp, is sufficient to identify the best resolution III design. The minimum aberration design minimizes the number of two-factor interactions aliased with main effects and, subject to this, minimizes the number of aliased two-factor interactions.

To avoid any confusion between main effects and two-factor interactions, regular fractions must have resolution of IV or more. Different criteria for comparing resolution IV designs have appeared because the aberration criterion is not sufficient to address the subtleties of different designs. Chen et al. (1993) use the 2^{9-4}_{IV} case to illustrate this deficiency. While the minimum aberration design, denoted 9-4.1, has $A_4 = 6$, the second lowest aberration design, denoted 9-4.2, has $A_4 = 7$. However, Design 9-4.2 has 15 two-factor interactions clear of aliasing with main effects and two-factor interactions (vs. only 8 clear for design 9-4.1) and 22 degrees of freedom for two-factor interactions (vs. 21 for design 9-4.1).

These results seem to challenge ranking based on aberration. However, Cheng et al. (1999) showed that the aberration criterion is a surrogate for estimation capacity as introduced by Sun (1993). While every resolution IV design can estimate a model with a single two-factor interaction, some models with multiple interactions cannot be estimated. Estimation capacity (EC) is a vector $(EC_1, EC_2, \ldots, EC_g)$ of the proportions of estimable models with all k main effects and $1, 2, \ldots, g$ two-factor interactions. For designs 9-4.1 and 9-4.2, the estimation capacities are $(1, 0.971, 0.915, 0.835, 0.737)$ and $(1, 0.967, 0.902, 0.811, 0.702)$, respectively, for 1–5 interactions. So lower aberration implies that more models with several interactions can be estimated.

The fact that the true model can be estimated does not mean that one can distinguish the true model from other models that fit the data equally well. For instance, while design 9-4.1 can estimate more of the 58,905 models with four interactions (49,206 vs. 47,775 for design 9-4.2), most estimable models involve two-factor interactions that are aliased with interactions not in the model; only 70 four-interaction models are *clear* for design 9-4.1 (vs. 1365 for design 9-4.2). Thus, even if the true model is estimable, one might still mislabel an active effect. If the true model is not estimable, then active interactions are aliased together. For some fractions, two such active effects will sum, increasing the chance that this contrast estimate will be statistically significant in an analysis; for other fractions the effects will cancel, increasing the likelihood that the linear combination will be not significant and so both interactions will be overlooked.

We review one more criterion in this brief summary of regular fractions, relevant for all screening designs. Loeppky (2004) defined the Projection Estimation Capacity (PEC) sequence as the proportion of subsets of factors of various sizes for which the two-factor interaction model is estimable. For regular FFDs, this is simply the proportion of projections which are either full factorials or which are resolution V (or higher) fractions. For design 9-4.1, $PEC = (p_3, \ldots, p_7) = (84/84, 120/126, 96/126, 32/84, 0/36)$, while for design 9-4.2, $PEC = (84/84, 119/126, 91/126, 28/84, 0/36)$. The fact that $p_7 = 0$ follows from the absence of resolution V 2^{7-2} fractions. By Loeppky's Lemma 4.3, both $1 - p_4$ and $1 - p_5$ are proportional to A_4, so a weak minimum aberration 2^{k-f}_{IV} design, which has the smallest possible value of A_4, will always have a better PEC sequence than same-sized regular fractions having a larger A_4.

3 CRITERIA FOR NONREGULAR FRACTIONS

There are many criteria to rank nonregular FFDs. These include strength (Rao, 1947), generalized resolution and G-aberration (Deng and Tang, 1999), G_2-aberration (Tang and Deng, 1999),

Q_B (Tsai and Gilmour, 2010), generalized alias length pattern (Cheng et al., 2008; Mee, 2013), estimation capacity (Sun, 1993), information capacity (Sun, 1993; Li and Nachtsheim, 2000), projection estimation capacity (Loeppky et al., 2007), projection information capacity, model discrimination potential (Jones et al., 2007), minimal dependent sets (Miller and Sitter, 2005) and power. We now review all of these criteria, taking designs with seven or eleven factors to illustrate insights provided by these criteria.

Although the primary focus of this article regards orthogonal designs, it is instructive to compare nonregular FFDs with other designs of the same size that have been proposed for screening. Appendix A describes three classes of non-orthogonal designs: Bayesian D-optimal designs (DuMouchel and Jones (1994)); designs created to be optimal over the main-effects-plus-interactions (MEPI) family of models (Li and Nachtsheim 2000, Smucker and Drew 2015); and designs created to be optimal for estimating the two-factor interaction model for projections into subsets of factors (Smucker et al. 2012).

3.1 Strength and Generalized Resolution

Whereas resolution is a function of a regular FFD's defining relation, strength is an analogous property that pertains to both regular and nonregular FFDs. Let $\mathrm{OA}(n, k, t)$ denote an orthogonal array with n rows and k columns, where each column contains two symbols (levels); in this article we use ± 1 to denote the levels. The index t denotes the strength of the OA. A strength t array projects into an equally replicated full 2^t factorial in every subset of t columns. A regular 2^{k-f} fraction with resolution r will have strength $t = r - 1$. Thus, a resolution III design is strength 2 and a resolution IV design is strength 3. Note that a strength of t requires that the run size be divisible by 2^t. Thus, while strength 2 designs can have n a multiple of 4, strength 3 requires the number of runs to be a multiple of 8.

A j-factor interaction column for an $\mathrm{OA}(n, k, t)$ is simply the element-wise product of j main effect columns. If this column sums to $\pm n$, then this interaction forms a *full aliasing word.* For regular fractions, the only possible sums are $-n$, 0, and n. However, for nonregular fractions, the sum for interaction columns can take on values other than $-n$, 0 and n. For instance, consider the two $\mathrm{OA}(20, 7, 2)$ designs in Table 1. There are 35 three-factor interaction contrasts; for these two designs, every one sums to 4 or -4. This implies that the correlation between each main effect and two-factor interaction contrast vector involving three distinct factors is $\pm 4/20$. Let S denote the maximum sum in absolute value among all the interactions involving $t + 1$ columns. Then the generalized resolution ρ for the OA is defined as

$$\rho = t + 2 - S/n. \tag{2}$$

For regular 2^{k-f} fractions, $S = n$ and so $\rho = t + 1$. For nonregular designs, $S = n - 8s$ for some

Table 1: Two alternative OA$(20, 7, 2)$.

Design 20.7.1							Design 20.7.18						
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1	-1	-1
-1	-1	-1	-1	1	-1	1	-1	-1	-1	1	1	1	1
-1	-1	1	1	-1	-1	1	-1	-1	1	-1	1	1	-1
-1	-1	1	1	1	1	-1	-1	-1	1	1	-1	1	1
-1	1	-1	1	-1	-1	1	-1	1	-1	-1	1	-1	1
-1	1	-1	1	1	1	-1	-1	1	-1	1	1	1	-1
-1	1	1	-1	-1	1	-1	-1	1	1	-1	-1	1	1
-1	1	1	-1	1	-1	-1	-1	1	1	1	-1	-1	-1
-1	1	1	1	1	1	1	-1	1	1	1	1	-1	1
1	-1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1
1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	-1	-1	-1	1
1	-1	1	-1	1	1	1	1	-1	1	-1	1	1	1
1	-1	1	1	1	-1	1	1	-1	1	1	1	-1	-1
1	1	-1	-1	1	-1	-1	1	1	-1	-1	-1	1	1
1	1	-1	-1	1	1	1	1	1	-1	-1	1	1	-1
1	1	-1	1	-1	1	1	1	1	-1	1	-1	-1	1
1	1	1	-1	-1	-1	1	1	1	1	-1	1	-1	-1
1	1	1	1	-1	-1	-1	1	1	1	1	-1	1	-1

integer s, $0 \leq s < n/8$ (Deng and Tang, 1999); if $t = 3$, s must be even. The two OA$(20, 7, 2)$ designs from Table 1 both have $S = 4$, so their generalized resolution is $\rho = 3.8$. For every OA$(20, 11, 2)$, $S = 12$, so the generalized resolution is 3.4. Clearly, we favor designs with a larger generalized resolution.

3.2 G-aberration and G_2-aberration

For regular FFDs, aberration is measured using the wlp. That is, aberration is based on the number of interaction columns of each size that sum to $\pm n$. Deng and Tang (1999) extended this concept, defining the confounding frequency vector (cfv) to be the number of interaction columns that sum to each non-zero value. Table 1's Design 20.7.1 has

$$\text{cfv} = [F_3(4) = 35; F_4(12, 4) = (2, 33); F_5(8) = 11; F_6(8) = 1; F_7(4) = 1]. \qquad (3)$$

This cfv shows that all 35 three-factor interaction columns sum to ± 4, two four-factor interaction columns sum to ± 12 and the other 33 sum to ± 4, etc. Table 1's Design 20.7.18 has cfv $= [F_3(4) = 35; F_4(12, 4) = (5, 30); F_5(8) = 9; F_6(8) = 1; F_7(4) = 1]$. For more detail about possible values for the sums, see Proposition 1 from Deng and Tang (2002).

The cfv has been subsequently referred to as the extended word length pattern (ewlp) composed of counts denoted by $A_{j.h}$. Just as generalized resolution (2) can take on fractional values, a word of length $j.h$ is a j-factor interaction that sums to $n(1 - 0.h) > 0$. Thus, the three-factor interactions for the designs in Table 1 that sum to ±4 correspond to words of length 3.8. Using this notation, the positive counts in the ewlp for Design 20.7.1 are

$$\text{ewlp} = (A_{3.8}, A_{4.4}, A_{4.8}, A_{5.6}, A_{6.6}, A_{7.8}) = (35, 2, 33, 11, 1, 1). \tag{4}$$

Equation (4) closely resembles the notation for the usual wlp (1) and has an obvious connection with generalized resolution (2). Each ewlp subscript of the form $j.h$ denotes a correlation of $1-0.h$ between a j-factor interaction and the constant intercept column. While the cfv notation (3) shows the interaction column sums explicitly, we generally utilize the more compact ewlp (4) notation in the remainder of this article.

Generalized aberration (or G-aberration) is based on sequentially sorting designs according to entries in the ewlp (or cfv). The two designs in Table 1, along with 19 other designs, have the best possible $A_{3.8} = 35$ among all OA(20,7,2), as enumerated by Sun et al. (2008) and Schoen et al. (2010). These 21 designs are then compared with respect to 4-factor interactions, where $A_{4.0} = 0$ and $A_{4.4} = 2$, achieved by two designs, is best. Table 1's Design 20.7.1 is one of these designs; both have identical ewlp, so both are minimum G-aberration designs. Design 20.7.18 is 18th best in G-aberration.

Tang and Deng (1999) proposed G_2-aberration as a more convenient characterization of aberration in nonregular designs. G_2-aberration is based on the generalized word length pattern (gwlp) = $(B_3, B_4, \ldots, B_k)$, where B_j is the sum of the squares of all the j-factor interaction contrast sums, divided by n^2. For regular designs, these squared sums equal n^2 or 0, so $B_j = A_j$, the number of j-factor interactions in the defining relation. For nonregular designs, the elements of gwlp can be non-integers. Table 2 shows how the gwlp is a compression of the ewlp for Design 20.7.1. The gwlp for an orthogonal design with n distinct rows, such as Design 20.7.1, sums to $(2^k/n) - 1$; this result is derived using (9.10) from Xu (2015). Design 20.7.18's gwlp = (1.4, 3, 1.44, 0.16, 0.04); it's sum equals 6.04 (> 5.4) because Design 20.7.18's first two rows are identical.

Table 2: Relation between j-factor interaction entries for extended word length pattern (4) and generalized word length pattern, gwlp = (1.4, 2.04, 1.76, 0.16, 0.04), for Design 20.7.1.

.

j	ewlp	gwlp
3	$A_{3.8} = 35$	$B_3 = 35(4/20)^2 = 1.4$
4	$A_{4.4} = 2, A_{4.8} = 33$	$B_4 = 2(12/20)^2 + 33(4/20)^2 = 2.04$
5	$A_{5.6} = 11$	$B_5 = 11(8/20)^2 = 1.76$
6	$A_{6.6} = 1$	$B_6 = 1(8/20)^2 = 0.16$
7	$A_{7.8} = 1$	$B_7 = 1(4/20)^2 = 0.04$

Besides being more concise, G_2-aberration is much simpler to compute, since the gwlp can be computed easily from the moments of the row coincidence matrix $T = DD'$ (Butler, 2003), where D is the $n \times k$ design matrix with ±1 coding.

Schoen and Mee (2012) illustrate that G-aberration and G_2-aberration can produce different rankings of designs. The minimum G-aberration design in 11 factors and 32 runs has ($A_{4.0}$, $A_{4.5}$) $= (3, 90)$ and $B_4 = 25.5$, while the minimum G_2-aberration design has ($A_{4.0}$, $A_{4.5}$) $= (4, 84)$ and $B_4 = 25$.

The G_2-aberration criterion can also be applied to non-orthogonal designs. B_1 is the (uncorrected) sum of squares of the main effect column means and B_2 is the sum of squares of the two-factor interaction column means. For a balanced design, $B_1 = 0$ and for an orthogonal array, $B_2 = 0$. Hence, minimum G_2-aberration ranks orthogonal designs above non-orthogonal arrays, and balanced designs above unbalanced ones.

3.3 Q_B Criterion

Tsai et al. (2007) present the Q_B criterion for selecting designs intended to provide efficient estimation of factorial effects excluding the intercept over the class of all sub-models of a maximal model, with prior probabilities serving as weights for the possible models. Tsai and Gilmour (2010, Section 5.1) apply this criterion to two-level factorial designs where the full two-factor-interaction model is the maximal model and where all reduced models considered satisfy the marginality requirement that a main effect may be omitted only if that factor does not appear in any interactions. We assume the prior of Bingham and Chipman (2007), where the prior probability of a main effect being active is π_1, and the conditional probability that an interaction is active is π_2 if both main effects are active, π_3 if only one of the main effects is active, and zero otherwise. Under these assumptions, the Q_B criterion is

$$Q_B = \{[\xi_{10} + 2(k-1)\xi_{21}]B_1 + [2\xi_{20} + \xi_{21} + 2(k-2)\xi_{32}]B_2 + 6\xi_{31}B_3 + 6\xi_{42}B_4\}/n, \quad (5)$$

where the six ξ_{ij} coefficients in (5) depend on the prior probabilities. For simplification, we remove the prior's dependence on n as proposed by Tsai et al. (2007). Appendix B derives the coefficients under the weak effect heredity prior where $\pi_3 > 0$. However, if the prior assumes strong effect heredity (Hamada and Wu, 1992) so that $\pi_3 = 0$, then $\xi_{ij} = \pi_1^i \pi_2^j$.

Suppose $(\pi_1, \pi_2, \pi_3) = (0.5,\ 0.8,\ 0)$. Then, for $k = 7$, the expected number of main effects and interactions are $\pi_1 k = 3.5$ and $\pi_1^2 \pi_2 k(k-1)/2 = 4.2$, respectively, and $Q_B = (2.9B_1 + 1.5B_2 + 0.6B_3 + 0.24B_4)/n$. Table 3 lists the two orthogonal designs from Table 1 and the three non-orthogonal designs from Appendix A, ordered based on minimizing Q_B. Observe how Q_B ranks the MEPI and Bayes-D designs higher than the poorer of the two orthogonal designs.

3.4 Generalized Alias Length Pattern

The generalized alias length pattern (galp) was first proposed by Cheng et al. (2008) as a simple measure of aliasing among two-factor interactions for strength 3 arrays. Mee (2013) extends galp

Table 3: Generalized word length pattern for five (n = 20, k = 7) designs, ranked by Q_B, with prior (π_1, π_2, π_3) = (0.5, 0.8, 0)

.

Design	B_1	B_2	B_3	B_4	Q_B
20.7.1	0.00	0.00	1.40	2.04	0.066
20.7.MEPI	0.04	0.16	0.48	3.16	0.070
20.7.Bayes-D	0.00	0.04	1.68	1.64	0.073
20.7.18	0.00	0.00	1.40	3.00	0.078
20.7.PEC	0.10	0.18	1.00	2.00	0.082

to strength 2 arrays as follows. Let X denote the model matrix for the full two-factor interaction model, and compute galp as the main diagonal of the matrix $(X'X/n)^2$. Note that the ith element of this diagonal is the sum of squares for the elements in the ith column of $X'X/n$. The minimum value for the ith element is 1, which would indicate that the ith column is uncorrelated with every other column. As with Q_B, galp may omit the (1,1) element of $(X'X/n)^2$, which corresponds to the intercept.

For Design 20.7.1 in Table 1, the diagonal elements equal 2.24 for the AD interaction, 1.92 for 10 interactions, and 1.6 for the seven main effects and the other 10 interactions. Figure 1 shows the galp distributions for the minimum G-aberration design 20.7.1 and two non-orthogonal designs from Appendix A: the Bayesian D-optimal design and the MEPI design. The two non-orthogonal designs have many distinct values. The median value is smallest for Design 20.7.1 and largest for the MEPI design. The Bayesian D-optimal and minimum G-aberration designs have similar galp distributions, with rather even aliasing. The MEPI design has lower aliasing for the main effects but higher aliasing for two-factor interactions, and its galp sum is larger than for the other two designs. This suggests that the MEPI design may do well identifying active main effects, but have more difficulty identifying which interactions are active.

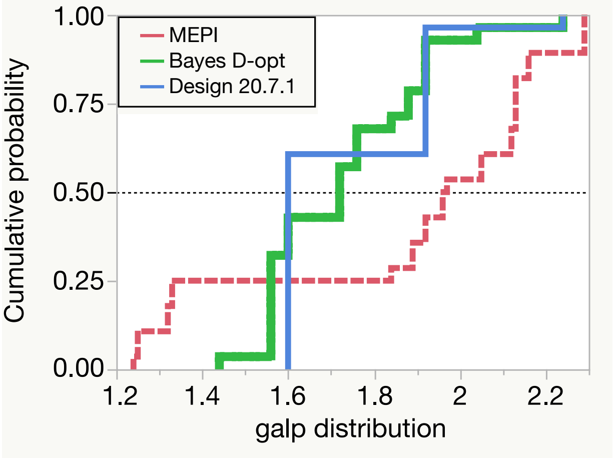


Figure 1: galp distributions for Design 20.7.1 and two non-orthogonal designs.

Following Section 2 of Tsai et al. (2000), when the elements of galp are divided by n, one obtains a first-order approximation to the variances of the coefficient estimators, assuming the model to be estimable. In this way, trace[(1/n)$(X'X/n)^2$] roughly approximates an A-optimality criterion for the full two-factor interaction model. Now

$$\text{sum(galp)} = 0.5k(k+1) + (2k-1)(B_1 + B_2) + 6(B_3 + B_4). \tag{6}$$

Thus, when the sole model under consideration is the full two-factor interaction model, B_3 and B_4 are treated with equal weight. This contrasts with the Q_B results just seen, where consideration of a family of possible simpler models gives more weight to B_1 than B_2, and to B_3 than B_4.

3.5 Estimation Capacity and Information Capacity

For Design 20.7.1 in Table 1, the estimation capacity for models with 1–7 interactions is (1, 1, 1, 1, 1, 0.99994, 0.9996). Every possible model with 7 main effects and 5 two-factor interactions is estimable; EC_6 is less than 1 because 3 of the 54,264 possible models with 7 main effects and 6 interactions are not estimable. This EC is impressive. Assuming that at most 1/3rd of the 21 possible two-factor interactions are active and no higher-order interactions are present, one is essentially guaranteed that the true model will be estimable. However, estimability ignores precision, model discrimination, and power. We discuss variance efficiency first, as this directly impacts precision of effect estimates. Model discrimination is discussed in Section 3.7, and power in Section 3.10.

For regular 2^{k-f} fractions, the model matrix is either diagonal (and, so, D-optimal) or singular (i.e., the model is not estimable). For a nonregular FFD, models may be estimable but with D-efficiency < 100%, so estimation capacity alone disregards relevant information. Sun (1993) and Li and Nachtsheim (2000) noted this deficiency and so compared nonregular designs by augmenting EC with the mean D-efficiency across all models of equal size; they refer to this measure as the information capacity (IC). The EC and IC sequences for the two designs in Table 1 and the three non-orthogonal designs from Appendix A are given in Table 4. The estimation capacity is excellent for all these designs. Given the similarity of the EC vectors, IC is a more useful criterion. The lower aberration for Design 20.7.1 vs. Design 20.7.18 is reflected in better IC. The MEPI design has the highest average efficiency for models with 3-7 interactions; it was constructed to optimize (EC, IC) for $g = 6$ (see Section A.2).

Let X_g denote the $n \times (1 + k + g)$ matrix for the first-order model augmented with g two-factor interactions. Cheng et al. (2002) reason that minimizing the average trace$[(X_g'X_g/n)^2]$ across all possible models with g interactions is a good surrogate for maximizing the average determinant of $X_g'X_g/n$. In this way, Cheng et al. (2002) relates the gwlp of a design to its information capacity. Extending their Proposition 1 to include designs with $B_1 > 0$, the average

Table 4: EC and IC for designs in Table 1 and Appendix A.

.

Design	# interactions g						
	1	2	3	4	5	6	7
			Estimation Capacity (EC)				
20.7.1	1	1	1	1	1	0.9999	0.9996
20.7.18	1	1	1	1	1	0.9997	0.9980
20.7.MEPI	1	1	1	1	1	0.9998	0.9987
20.7.Bayes-D	1	1	1	1	1	1	0.9999
20.7.PEC	1	1	1	1	1	1	1
			Information Capacity (IC)				
20.7.1	0.9755	0.9512	0.9266	0.9011	0.8745	0.8461	0.8153
20.7.18	0.9755	0.9491	0.9207	0.8902	0.8572	0.8212	0.7808
MEPI	0.9670	0.9546	0.9378	0.9170	0.8926	0.8644	0.8318
BayesD	0.9637	0.9345	0.9064	0.8785	0.8503	0.8211	0.7902
PEC	0.9412	0.9207	0.8990	0.8758	0.8511	0.8245	0.7956

trace$[(X_g'X_g/n)^2]$ over all subsets of g two-factor interactions equals

$$1+k+g+2[1+\frac{g}{G}(k-1)]B_1+2[1+\frac{g}{G}+\frac{g(g-1)}{G(G-1)}(k-2)]B_2+6\frac{g}{G}B_3+6\frac{g(g-1)}{G(G-1)}B_4, \quad (7)$$

where $G = 0.5k(k-1)$. In contrast to (5) and (6), this derivation includes the intercept. Apart from this minor difference, when $g = G$, then $X_g = X$ and (7) simplifies to (6). If (7) is a good surrogate for IC_g, as Cheng et al. (2002) suggest, this would be quite a computational shortcut in the search for IC-optimal designs, since computing the gwlp for a design is very quick, whereas computing IC_g can be time consuming. For the five designs in Table 4, the correlation between IC_g and the surrogate (7) ranges from -0.9915 to -0.9982 for $1 \leq g \leq 7$. Ranking by the surrogate identifies the best and second best designs in every case, and perfectly ranks all five designs for $g = 2, 3, 4, 7$. Dividing the coefficients in (7) by their sum, we obtain weights that sum to 1. Figure 2 shows how increasing weight is given to B_3 and B_4 as the number of interactions considered grows. Note that $w_1 = w_2$ and $w_3 = w_4$ when $g = 21$. The main insight from the lower portion of Table 4 and the surrogate reflected in Figure 2 is that the fewer number of interactions expected, the more justification one has for using an orthogonal array. That is, orthogonal arrays will often be preferred for screening, but not necessarily for estimation of the full two-factor interaction model.

3.6 Projection Estimation Capacity

EC_j and IC_j measures in Section 3.5 pertained to the family of models with all main effects and j two-factor interactions. Based on the assumption of *factor* sparsity, Loeppky et al. (2007) defined projection estimation capacity (PEC) based on the proportion of subsets of the k factors for which the two-factor interaction model is estimable. Especially when the analysis is based

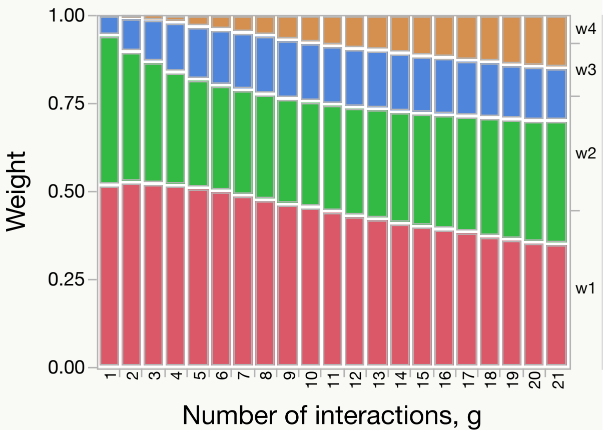


Figure 2: Weights for B_1,B_2,B_3,B_4 in (7) for 7 factors and $g = 1, \ldots, 21$ interactions

on first identifying which factors are active, PEC is well-motivated. For Design 20.7.1, the two-factor interaction model can be estimated for any subset of four factors and for 19 out of the 21 five-factor projections; hence $p_4 = 1$ and $p_5 = 19/21$. Design 20.7.18 is slightly worse, with $p_5 = 18/21$, while all three non-orthogonal designs from Appendix A have $p_5 = 21/21$. Due to insufficient sample size, $p_6 = 0$.

Although Loeppky et al. (2007) did not compute information capacity for the same sets of models, it is a straightforward extension to do so. For 3-, 4- and 5-factor projections, the projection information capacity (PIC) for Designs 20.7.1 and 20.7.18, respectively, are (0.9827, 0.9328, 0.7584) and (0.9827, 0.9226, 0.6737). Thus, for projections into 4 or 5 factors, the efficiency is better for the minimum G-aberration design. Note also that IC_3 and IC_6 for Design 20.7.1 (Table 4), are lower than the PIC values for models with 3 and 4 factors, even though these two-factor interaction models also have 3 and 6 interactions, respectively.

Table 5 compares the PIC sequence for the three non-orthogonal designs with those of the orthogonal designs. We note that the average D-efficiencies for projections are less than 1 for the non-orthogonal designs. This criterion suggests that the non-orthogonal designs are better for projections into five factors, but the nonregular FFDs are better for projections into 4 or fewer factors. The PEC design was optimized in terms of maximizing the minimum D-efficiency, which is not captured in Table 5. The PEC-optimal design's minimum D-efficiency equals 0.741 across the 21 five-factor projections. This value is indeed larger than for the other four designs; the minimum D-efficiency for the Bayes D-optimal design is 0.694.

3.7 Model Discrimination

Jones et al. (2007) observed that designs having the same estimation capacity may differ considerably with respect to model-discrimination capabilities. To identify the correct model, in

Table 5: *PIC* sequences for designs in Table 1 and Appendix A.

.

Design	# factors			
	2	3	4	5
20.7.1	1	0.9827	0.9328	0.7584
20.7.18	1	0.9827	0.9226	0.6737
20.7.MEPI	0.9903	0.9766	0.9201	0.7790
20.7.Bayes-D	0.9990	0.9759	0.9198	0.8111
20.7.PEC	0.9801	0.9507	0.8886	0.7756

addition to being estimable it must stand out as a superior fit to the data. When the number of models considered is vast, there are generally models so similar in fit to one another that the given data will not permit the analyst to distinguish them. Jones et al. (2007) propose three measures to assess the ability of a design to discriminate between a pair of models; these are the subspace angle (SA), the maximum prediction difference (MPD), and the expected prediction difference (EPD), with EPD being the simplest to compute, since EPD $=$ trace$[H_1 - H_2]^2/n$, where H_1 and H_2 are the hat matrices corresponding to the two models being compared. EPD is also the most intuitive discrimination measure. Jones et al. (2007) proposed computing both the average and the minimum EPD over all pairs of models being considered, which we denote by AvgEPD and MinEPD, respectively. Jones et al. (2007) and Androulakis et al. (2014) applied these criteria to three-level designs for model spaces with 1 or 2 interactions; we are not aware if the criteria have been previously applied to two-level designs.

We compute AvgEPD and MinEPD for all pairs of models with all main effects and g two-factor interactions for $g \leq 4$. For each g, the average and minimum are computed across $0.5M(M-1)$ pairs of models, where $M = \binom{0.5k(k-1)}{g}$. For $g = 4$, this is nearly 18 million model pairs. See Table 6. The Bayesian D-optimal design is best in terms of average EPD, but the minimum G-aberration design is better with respect to the minimum EPD for models with two or three interactions. The MEPI design, which performed well in terms of estimation capacity is near the bottom in terms of discrimination for the main effect plus up to 4 interactions. It is striking that the MEPI designs perform poorly, since Table 6 is based on pairs of models from the MEPI family.

Two observations are clear. First, good estimation capacity does not imply better model discrimination. This is evident in the MEPI design here and again in Section 3.9 for $k = 11$. Second, model discrimination is assessed for much simpler models, i.e., ones with few interactions. However, the ranking of designs based on AvgEPD was at least stable over the number of interactions.

Table 6: Model discrimination for designs in Table 1 and Appendix A, ordered from best to worst for average expected prediction difference. Each column maximum is underlined.

Design	# interactions				# interactions			
	1	2	3	4	1	2	3	4
	AvgEPD				MinEPD			
20.7.Bayes-D	0.0952	0.1640	0.2187	0.2574	0.0666	0.0293	0.0217	0.0132
20.7.1	0.0951	0.1638	0.2182	0.2563	0.0510	0.0339	0.0265	0.0118
20.7.PEC	0.0947	0.1626	0.2161	0.2538	0.0548	0.0343	0.0245	0.0149
20.7.MEPI	0.0940	0.1599	0.2108	0.2455	0.0494	0.0330	0.0199	0.0088
20.7.18	0.0937	0.1588	0.2079	0.2396	0.0360	0.0238	0.0093	0.0048

3.8 Minimal Dependent Sets

Whenever a particular model is not estimable, there exist one or more linear dependencies among the columns of its model matrix. Miller and Sitter (2005) and Lin et al. (2008) define the concept of minimal dependent set (MDS) as any set of two-factor interactions that creates a linear dependency when added to a model with all main effects, where the linear dependency disappears if any one of the interactions is removed. The size of the smallest MDS so defined is the smallest value of g such that $EC_g < 1$. For instance, we know from Section 3.5 that $EC_5 = 1$ and $EC_6 < 1$ for both designs in Table 1. The minimum G-aberration design 20.7.1 can estimate all but three six-interaction models; the three linear dependencies corresponding to these MDS of size $d = 6$ are:

$$\begin{aligned} A - B - C + D - 2AD + AG + DG + 2BC - BG - CG &= 0 \\ A - D + E + F + 2AD + DG + 2EF - AG - EG - FG &= 0 \\ D + G + AD + AG + BE - BF - CE + CF &= 0. \end{aligned} \tag{8}$$

AD, the interaction with the largest galp value for this design, shows up in every MDS of size 6 (and with the largest coefficient). Design 20.7.18 has 15 MDS of size 6. Table 4 showed that the PEC design is the only one that can estimate all models with 7 interactions. It has just three MDS of size 8, and all three have at least two terms involving each factor, whereas two of MDS in (8) involve just five factors.

3.9 Summary of Criteria

In terms of computational ease, strength, G_2-aberration, Q_B, and galp are the simplest, being quick functions of the row coincidence matrix DD' or the two-factor interaction model matrix. Next easiest to compute is generalized resolution and initial terms of the ewlp. Computing PEC and PIC are the next easiest, with EC and IC involving a more taxing computation, though for large g and k one can sample rather than examine all possible models with g interactions, as we

Table 7: Four of 260 alternative $\mathrm{OA}(40, 11, 3)$.

Design ID	Optimality	$A_{4.4}$	B_4	galp rank	EC_{15}	IC_{15}	p_6	PIC$_6$
40.11.1a	G, G_2, PEC	18	18.96	245/260	0.977	0.756	0.831	0.699
40.11.1b	G, G_2, galp	18	18.96	1/260	0.984	0.762	0.799	0.672
40.11.1c	G, G_2, EC	18	18.96	7/260	0.992	0.769	0.701	0.596
40.11.260	none	30	22.80	260/260	0.961	0.718	0.468	0.393

illustrate in this section. Computing the average EPD is more difficult still, since it involves all pairs of models of a given size. Other model discrimination criteria are more computationally intensive, exceeded only by that required for power computations, which we discuss in the next subsection.

Regarding the five designs we have compared, the criteria lead to different rankings. Design 20.7.1 is best (or tied for best) in terms of strength, generalized resolution, G-aberration, G_2-aberration, and Q_B, assuming our strong effect heredity prior. The Bayesian D-optimal design measured best in terms of model discrimination and in estimation of the two-factor interaction models in 5-factor projections. The MEPI design measured best in terms of D-efficiency for models with 2 to 7 interactions. The PEC design fared best in EC and very well in minEPD. Design 20.7.18, though having the maximum possible generalized resolution, was never the preferred design. We investigate which of the other four designs performs best for screening in Section 4, where we evaluate the designs with respect to power via simulation.

We have used 7-factor designs with 20 runs to illustrate the criteria in this section. We now examine designs with 11 factors in 40 runs, as the numerous minimum G-aberration designs of this case will again enable us to appreciate differences in the rankings provided by other criteria. There are 260 $\mathrm{OA}(40, 11, 3)$, all with generalized resolution 4.4; $A_{4.4}$ ranges between 18 and 30 and B_4 ranges from 18.96 to 22.8. There are 48 minimum G-aberration designs, which have an ewlp of $(A_{4.4}, A_{4.8}, A_{6.6}, A_{8.4}, A_{8.8}, A_{10.6}) = (18, 312, 142, 4, 161, 4)$. These 48 designs are also minimum G_2-aberration designs and they all produce the optimal PEC for projections into five factors ($p_5 = 1$). Differences appear when considering PEC for projections into six or more factors. Each of the 260 $\mathrm{OA}(40, 11, 3)$ permit estimation of up to 19 two-factor interactions. Since the PEC into six factors involves models with 15 interactions and because differences in EC$_g$ are slight for smaller g, we chose to evaluate EC and IC for $g = 15$. With $k = 11$, there are 11.9 trillion models with 15 interactions; EC$_{15}$ and IC$_{15}$ were estimated, based on sampling 2 million models.

Table 7 highlights three of the minimum G-aberration designs and contrasts them with the less attractive maximum aberration OA(40,11,3). Design 40.11.1a corresponds with the 40-run design 11.68 in Schoen and Mee (2012) with the highest PEC ($p_6 = 0.831$). We also consider here Design 40.11.1b having the best galp, Design 40.11.1c (having the best EC among these 260 designs), and Design 40.11.260, which is the OA(40,11,3) with the worst $A_{4.4} = 30$. Note that criteria (5), (6), and (7), which are simple functions of $B_1, \ldots, B_4$, cannot distinguish designs such as 40.11.1a-c that have identical gwlp.

Figure 3 displays all 260 OA(40,11,3), denoting Designs 40.11.1a-c with filled circles; 'Z' denotes the worst aberration OA(40.11.3). The vertical axis in this scatterplot is PIC_6, the average D-efficiency across the 462 full two-factor interaction models in subsets of six factors. Design 40.11.1a has the highest value (0.699). The horizontal axis IC_{15} is the average D-efficiency across models with all 11 main effects and 15 of the 55 possible interactions. This average was estimated by sampling 100,000 subsets of interactions for most designs, with sampling an additional 2 million subsets among the best few to ensure that the design that maximized IC_{15} was correctly identified. Design 40.11.1c has the largest IC_{15} (0.769), but many designs do nearly as well. Design 40.11.1b, with the best galp, is the third design denoted by a filled circle and is a compromise between maximizing PIC_6 and IC_{15}.

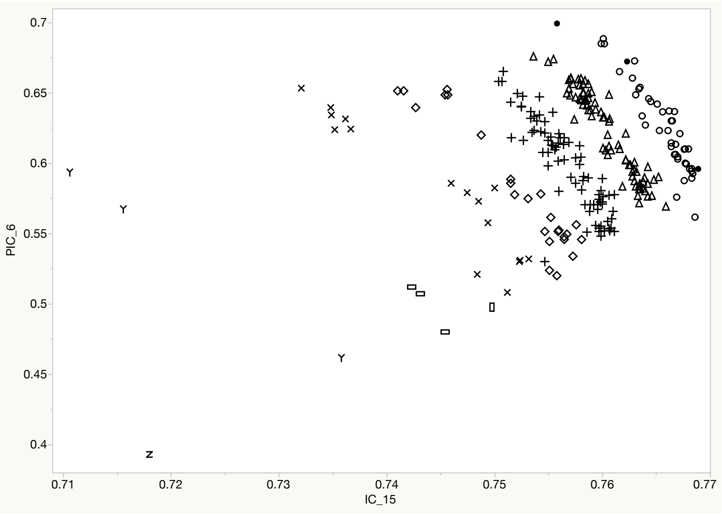


Figure 3: Comparison of 260 OA(40, 11, 3): average D-efficiency for two-factor interaction models across all 462 6-factor projections (PIC_6) vs. average D-efficiency for models with 11 main effects and 15 interactions (IC_{15}). Different symbols indicate different values of B_4, with empty and filled circles marking the 48 minimum G-aberration designs and z marking the maximum G-aberration design.

Figure 3 reveals a strong negative correlation between PIC_6 and IC_{15} for designs with the same B_4. The correlations between the variables displayed in the figure range from -0.874 to -1.000 for the seven different values of B_4 having multiple designs. The pair of variables (PIC_6, IC_{15}) results in more distinct clusters with different B_4 than would be obtained using either PIC or IC alone. The smallest B_4 corresponds to designs on or near the optimal frontier in the upper right corner (displayed using circles), and the clusters of designs move away from this frontier as B_4 increases. We have found this pattern for some other size OAs when there are several OAs tied on G- or G_2-aberration.

It is surprising that the all-main-effect-plus-15-interactions models have IC (and EC) larger than do the six-factor projection models with their 15 interactions, which have five fewer main effects. Differences in PIC_6 are much larger than differences in IC_{15}. Thus, we are inclined to select the design with the highest PIC_6.

3.10 The Bottom Line Criterion: Power

Several of the OA(40,11,3) show only small differences in the design criteria. It is not obvious whether such differences matter. When a final comparison is to be made between several designs, it is informative to speculate a set of possible models and to use simulation to estimate each design's power for detecting active effects. While this criterion would be too cumbersome to distinguish between thousands of designs, it is worthwhile as a final criterion, which is why it is listed last here. In Section 4, we use estimated power both to compare the top designs of a given size and to see what is gained by increasing the sample size.

4 POWER

Power, defined as the probability of detecting active effects for a specified set of effect sizes, can be considered the ultimate measure of a screening design's effectiveness. As just mentioned, when a final comparison is to be made between several competing designs, it is useful to speculate a set of possible models and employ simulation to compare power to detect active effects. A measure of false positive results should be included as well to check whether good power is not mitigated by a high error rate. Accordingly, we measure false discovery rates (FDR) as well as power in a simulation study, addressing detection of main effects and two-factor interactions separately.

In this section, we report simulations for comparing selected 7-factor designs with 16–32 runs and 11-factor designs with 20–48 runs. For each run size, there is at least one OA, along with a PEC-optimal, a MEPI-optimal and a Bayesian D-optimal design. For the case of 7 factors and 32 runs, the latter three types of design are replaced by a D-optimal design for the full interaction model (7 main effects and 21 two-factor interactions).

All designs for 7 and 11 factors used in the simulations are characterized in Tables 8 and 9, respectively. Each orthogonal design ID is of the form $n.k.i$, where i is the G-aberration rank of the design. It can be seen that the included OAs are primarily minimum G-aberration designs. Design 16.7.1 is a regular design, while all other OAs are nonregular. Some of the designs generated with optimal design software turned out to be orthogonal arrays; for these cases and for all OAs, we include the generalized resolution in the table. A detailed account of the selected

Table 8: Selected seven-factor designs in 16–32 runs

n	ID	ρ	B_1	B_2	B_3	B_4	Q_B	x: p_x	PIC_x
16	16.7.1	4.0	0	0	0	7	0.105	4: 0.8	0.80
	16.7.4	3.5	0	0	2	3	0.120	4: 1	0.89
	16.7.5	3.5	0	0	2	3.5	0.128	4: 1	0.89
	MEPI		0.047	0.188	0.672	5.125	0.128	4: 0.829	0.76
	BAYESD	3.0	0	0	2.5	2.5	0.131	4: 0.886	0.80
	PEC		0.094	0.219	2.125	2.813	0.159	4: 1	0.84
20	20.7.1	3.8	0	0	1.4	2.04	0.066	5: 0.905	0.76
	MEPI		0.04	0.16	0.48	3.16	0.070	5: 1	0.78
	BAYESD		0	0.04	1.68	1.64	0.073	5: 1	0.81
	20.7.18	3.8	0	0	1.4	3	0.078	5: 0.857	0.67
	PEC		0.1	0.18	1	2	0.082	5: 1	0.78
24	BAYESD	3.67	0	0	0.667	1.667	0.033	5: 1	0.90
	MEPI		0	0.028	0.472	2.167	0.035	5: 1	0.89
	24.7.1	4.67	0	0	0	3.889	0.039	5: 1	0.87
	PEC		0.028	0.083	0.806	1.333	0.042	5: 1	0.88
28	MEPI		0	0.143	0.163	1.122	0.021	6: 1	0.86
	28.7.1	3.86	0	0	0.714	0.878	0.023	6: 1	0.87
	BAYESD		0.02	0.082	0.571	0.796	0.026	6: 1	0.86
	PEC		0.026	0.112	0.566	0.735	0.027	6: 1	0.86
32	32.7.2	4.5	0	0	0	1.5	0.011	6: 0.857	0.78
	D-Opt		0.043	0.109	0.293	0.406	0.018	6: 1	0.92
	32.7.x	3.75	0	0	0.812	0.375	0.018	6: 1	0.90

designs is available in the supplementary materials.

For $k = 7$, we computed Q_B using $(\pi_1, \pi_2, \pi_3) = (0.5,\ 0.8,\ 0)$ as discussed previously in Section 3.3. For $k = 11$ we used the prior $(0.5,\ 0.4,\ 0)$, so that we expect 5.5 main effects and 5.5 interactions. For seven cases, the minimum G-aberration design had lower Q_B than the non-orthogonal designs; the exceptions were $(n = 24, k = 7)$, where Bayes-D and MEPI designs were better, plus $(n = 28, k = 7)$ and $(n = 20, k = 11)$, where the MEPI design was best.

The method chosen for variable selection can certainly have an impact on power. To make comparisons in this regard, we perform simulations using both forward selection and the Dantzig selector. A brief description of each analysis method is given in Appendix C.

4.1 Simulation protocol

For each design under comparison, our simulation is carried out as follows. In each of 1,000 iterations:

1. From the columns of the design matrix, m columns are randomly assigned as the active main effects. The simulation study is conducted for each of m from 2 to 5.

Table 9: Selected 11-factor designs in 20–48 runs

n	ID	ρ	B_1	B_2	B_3	B_4	Q_B	x: p_x	PIC_x
20	MEPI		0.08	0.4	1.8	38.8	0.173	5: 1	0.76
	20.11.1	3.4	0	0	8.2	22.8	0.191	5: 0.848	0.67
	BAYESD		0	0	9.48	18.96	0.199	5: 0.719	0.57
	PEC		0.06	0.66	9.12	17.68	0.229	5: 0.887	0.63
24	24.11.1	4.67	0	0	0	36.67	0.092	5: 1	0.87
	MEPI	4.67	0	0	0	36.67	0.092	5: 1	0.87
	BAYESD		0.097	0.208	6.83	13.94	0.139	5: 0.998	0.80
	PEC		0.125	0.847	6.028	15.33	0.161	5: 1	0.76
32	32.11.1	4.0	0	0	0	25.5	0.048	6: 0.221	0.18
	MEPI		0	0.25	0	24.5	0.053	6: 0.264	0.21
	32.11.x	3.75	0	0	4.25	15.75	0.069	6: 0.983	0.76
	BAYESD		0.07	0.18	4.469	9.16	0.070	6: 1	0.79
	PEC		0.035	0.305	3.957	11.38	0.070	6: 1	0.79
40	40.11.1a	4.4	0	0	0	18.96	0.028	6: 0.831	0.70
	40.11.1b	4.4	0	0	0	18.96	0.028	6: 0.799	0.67
	40.11.1c	4.4	0	0	0	18.96	0.028	6: 0.701	0.60
	MEPI		0	0.24	0	18.16	0.033	6: 0.851	0.69
	40.11.260	4.4	0	0	0	22.8	0.034	6: 0.468	0.39
	PEC		0.03	0.210	2.7	8.18	0.039	6: 1	0.86
	BAYESD		0.053	0.295	2.953	5.8	0.041	6: 1	0.86
48	48.11.1	4.67	0	0	0	9.11	0.011	6: 0.935	0.88
	48.11.x	4.67	0	0	0	10.89	0.014	6: 0.974	0.90
	MEPI		0.009	0.073	0.745	8.54	0.017	6: 1	0.92
	BAYESD		0.059	0.156	2	4.14	0.024	6: 1	0.91
	PEC		0.045	0.149	2.160	6.12	0.026	6: 1	0.89

2. Based on the selection of active main effects, g two-factor interaction columns are randomly assigned as active under the assumption of weak effect heredity. The simulation study is conducted for each of g from 1 to 7.

3. The coefficients, β, for the active effects are obtained via two scenarios:

   - Main effects and two-factor interactions are of the same size (Equal): Coefficients for the m active main effects and g active two-factor interactions are obtained by randomly sampling (with replacement) $m + g$ values from $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5\}$. A sign (+ or -) is randomly applied to each coefficient.

   - Two-factor interactions are smaller than main effects (Smaller): Coefficients for the active main effects are obtained by randomly sampling (with replacement) m values from $\{2, 2.5, 3, 3.5\}$ while coefficients for the active two-factor interactions are obtained by randomly sampling (with replacement) g values from $\{0.5, 1, 1.5, 2\}$. A sign (+ or -) is randomly applied to each coefficient.

4. Letting X be the matrix consisting of the columns corresponding to the active effects, the response vector is generated as $y = X\beta + \epsilon$ with $\epsilon_i \sim N(0, 1)$.

5. The set of significant effects is decided by performing one of the analysis methods.

At the end of 1,000 iterations, the average proportion of correctly identified main effects (power for main effects) and the average proportion of correctly identified two-factor interactions (power for two-factor interactions) are recorded. We also calculate the average proportion of main effects declared active which were, in fact, inactive (FDR for main effects), and the average proportion of two-factor interactions declared active which were inactive (FDR for two-factor interactions).

4.2 Simulation Results

4.2.1 $k = 7, n = 20$

Section 3 examined various criteria for nonregular designs using the case of 7 factors in 20 runs. Based on our simulation protocol, we display the results of our simulations for the alternative 7-factor, 20-run designs in Figure 4. The figure consists of two rows and four columns of panels. These differ according to the number of active main effects (first row: 2 main effects; second row: 4 main effects), size of two-factor interactions (Equal or Smaller), and analysis method (forward selection or Dantzig selector). The plotted lines show the power/false discovery rate estimates for 1 up to 7 active two-factor interactions. Solid and dashed lines are used to help distinguish design types.

Overall, the Dantzig selector outperforms forward selection for both main effect and two-factor interaction power (at the expense of slightly larger FDRs). However, we do see some benefit of forward selection for two-factor interaction detection in the case of two-factor interactions being smaller than main effects. This is especially the case when the true underlying model is parsimonious.

From Figure 4, it is evident that the 7-factor, 20-run design alternatives do not show dramatic differences with respect to power and FDR. However, some useful observations can be made. These are given as follows:

1. Design 20.7.1 and the MEPI-optimal design both appear to be good candidates for main effect detection. In particular, the MEPI-optimal design enjoys the lowest FDRs for main effects.

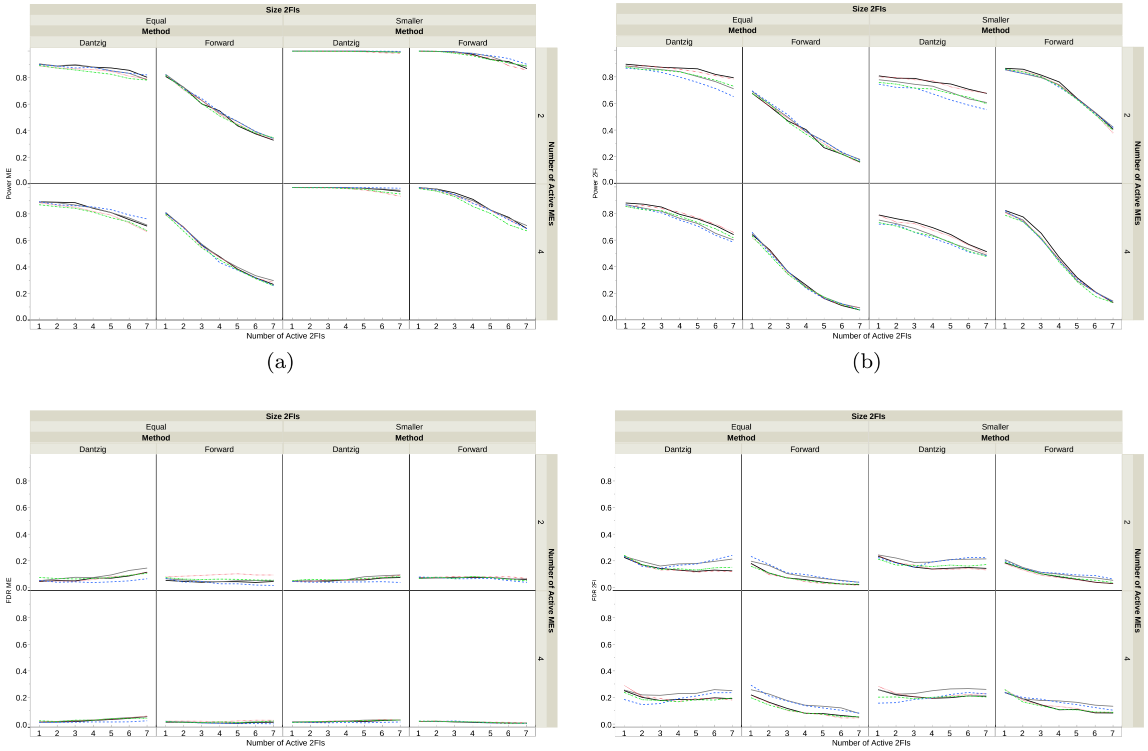
2. The Bayesian D-optimal design exhibits the largest (albeit only slightly) FDRs for main effects.

3. With respect to power for two-factor interactions, both Design 20.7.1 and the Bayesian D-optimal design are recommended. The more favorable performance for these two designs is most evident with the Dantzig selector.

4. Design 20.7.18 and the MEPI-optimal design have inferior FDRs for two-factor interactions.

5. Given its favorable performance overall, Design 20.7.1 could be safely recommended in the case of $n = 20, k = 7$.

That the min G-aberration and MEPI designs would perform well for main effect detection is consistent with Q_B. The uneven performance of the MEPI design for main effects and two-factor interactions was anticipated via galp (see Figure 1).



(a)

(b)

(c)

(d)

Figure 4: Simulation Results for $n = 20, k = 7$. X-axes for the plots represent the number of active two-factor interactions. Y-axes: (a) Power for Main Effects, (b) Power for 2FIs, (c) FDR for Main Effects, (d) FDR for 2FIs. Black (solid): 20.7.1; grey (solid): 20.7.18; red (dotted): Bayesian D-optimal; blue (dashed): MEPI-optimal; green (dash-dot): PEC-optimal.

4.2.2 $k = 11, n = 40$

The results of the simulation for $k = 11, n = 40$ are displayed in Figure 5. Rather than displaying results for two and four active main effects as in the 7-factor case, the simulation results are now shown for three and five active main effects. This difference is so that the number of active main effects relative to the total number of factors remains similar to the 7-factor case. In all other aspects, the figures are organized in the same as way as those for the 7-factor case. Our comments on the displayed results are as follows:

1. As with the 7-factor case, the Dantzig selector provides more favorable performance overall. However, we once again see a benefit with forward selection for two-factor interaction detection in case the two-factor interactions are smaller than the main effects.

2. The orthogonal arrays and the MEPI-optimal design are clearly recommended for main effect detection. For both forward selection and the Dantzig selector, these design choices display the highest power and lowest FDRs for main effects.

3. While we do note favorable performance with regards to power for two-factor interactions for the orthogonal arrays and MEPI-optimal design *when performing forward selection*, the Bayesian D-optimal and PEC-optimal designs are the clear winners for this metric.

4. FDRs for two-factor interactions differ depending on the analysis method used. With forward selection, the Bayesian D- and PEC-optimal designs consistently show the smallest FDR for two-factor interactions. However, for the Dantzig selector, a tradeoff is evident. That is, the orthogonal arrays and MEPI-optimal design exhibit the smallest FDRs for models with fewer truly active effects while the Bayesian D-optimal and PEC-optimal designs are superior for models with a larger number of truly active effects.

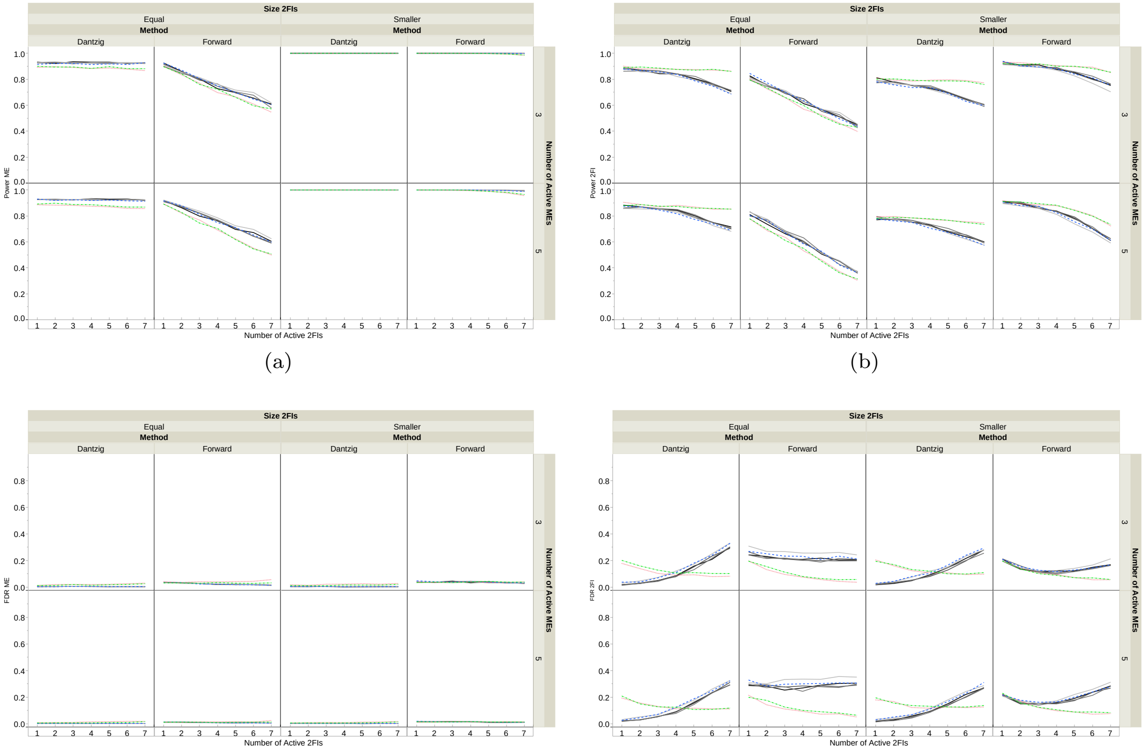
5. For power, little to no differences are observed among the 40-run orthogonal arrays. Differences among these four designs are more evident with FDR for two-factor interactions. Here, design 40.11.260 is the poorest performer.

4.2.3 Other run sizes

The results reported for $k = 7, n = 20$ and $k = 11, n = 40$ show patterns that are repeated for the other cases. Generally, the Dantzig selector outperforms forward selection except for the detection of two-factor interactions when these are smaller than the main effects. Both OAs and MEPI-optimal designs are generally powerful in detecting main effects, while they are also control the error rate better than the alternative designs. For detection of two-factor interactions, PEC-optimal and Bayesian D-optimal designs stand out. Several of the OAs (like 20.7.1) also perform well, while others (like 40.11.1a-1c) are inferior as regards detection of two-factor interactions. These findings emphasize the importance of simulation to select among competing designs.

4.3 Partial Replication and False Discovery Rates

In a separate set of simulations for forward selection only, we follow a protocol similar to that outlined above but now consider computation of power and false discovery rates when making use of a model independent error estimate. That is, at each iteration of the simulation study, we simulate a random error estimate via a chi-square distribution with 3 degrees of freedom and



(a)

(b)

(c)

(d)

Figure 5: Simulation Results for $n = 40, k = 11$. X-axes for the plots represent the number of active two-factor interactions. Y-axes: (a) Power for Main Effects, (b) Power for 2FIs, (c) FDR for Main Effects, (d) FDR for 2FIs. Black and grey (solid): orthogonal arrays; red (dotted): Bayesian D-optimal; blue (dashed): MEPI-optimal; green (dash-dot): PEC-optimal.

utilize this estimate in the forward selection procedure instead of the usual model dependent mean square error. This is analogous to having collected four additional observations at the center of the design; this replication yields a 3-df pure error estimate plus one df to assess curvature. We do not address using the latter df in the data analysis.

One finding from these simulations is the superior power when screening is based on a model independent error estimate. This pleasing outcome led us to emphasize a model independent error estimate in earlier drafts of this article. However, our initial investigations did not include a study of false discovery rates. Figure 6 displays this result for all 11-factor designs in Table 9; each boxplot shows simulation results across all scenarios. Clearly, FDRs can be quite high (especially with smaller sample sizes) when utilizing a model independent error estimate. We later discovered that this result was previously recognized by Westfall et al. (1998) in the context of forward selection. These authors noted that while power can be dramatically increased when the error variance is known, experimentwise error rates can also be largely inflated.

The reader will likely notice that the Dantzig selector was not considered in this separate set of simulations. More investigation is warranted to appropriately and satisfactorily incorporate a

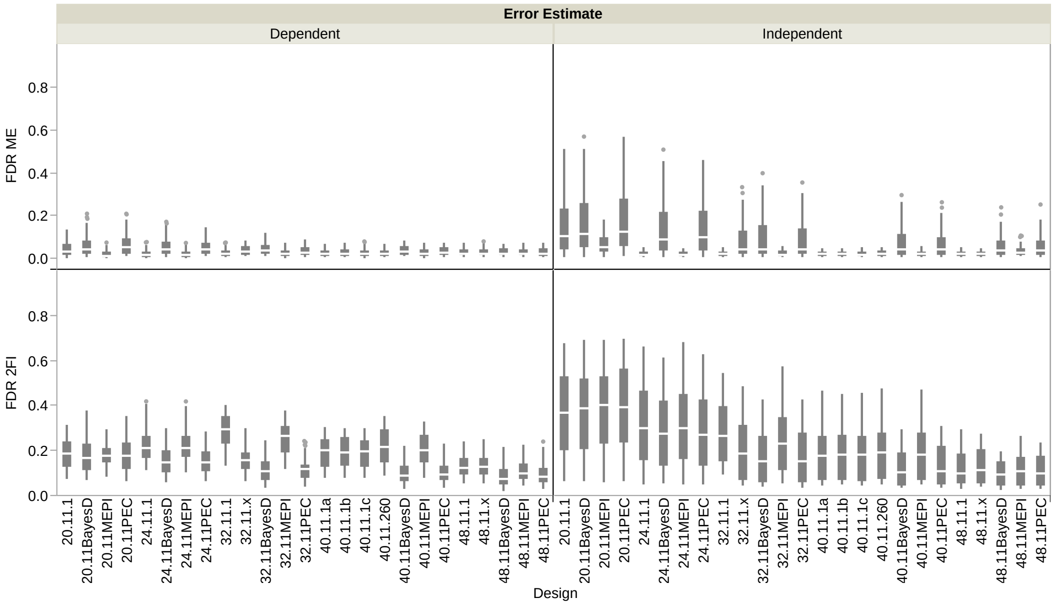


Figure 6: FDR Comparison: Dependent versus Independent Error Estimate for $k = 11$.

model independent estimate in the execution of the Dantzig selector.

5 PRACTICAL EXPERIMENT REVISITED

In the Introduction, we mentioned a practical screening application involving the making of phantoms to calibrate medical devices. We now return to this application to discuss suitable design options.

The design actually used for this experiment was the OA(40,7,3) listed by Schoen and Mee (2012) for intensive screening of 7 factors in 40 runs. This type of screening is recommended where many of the main effects could be active, while there could also be interactions. The design was large enough to estimate a model with all the main effects and all the two-factor interactions. The main response variables were the reflections of a phantom measured at eight wavelengths. Most of the models turned out to contain three substantial main effects sized two or more times the estimated standard deviation of the observations, two interactions of about one time the standard deviation and smaller interactions of half this size.

The 40-run design can be split in three different ways into two OA(20,7,2). It is conceivable to start with the first half of the 40-run design and decide after the first 20 runs whether or not to follow-up experimentation based on the second half. Each of these 20-run OAs turns out to be isomorphic to design 20.7.18 studied in this paper.

We can consult the simulation results to check how helpful the designs studied in this paper

Table 10: Alternative designs for the phantom experiment: power (left of dash) and FDR (right of dash) for detection of main effects (ME) and two factor interactions (2FI) through forward selection or the Dantzig selector.

		Forward		Dantzig	
n	Design	ME	2FI	ME	2FI
16	16.7.1	0.99 / 0.03	0.36 / 0.49	1.00 / 0.04	0.27 / 0.66
16	16.7.4	0.95 / 0.04	0.62 / 0.24	0.98 / 0.05	0.57 / 0.38
16	16.7.5	0.96 / 0.04	0.62 / 0.21	0.98 / 0.05	0.57 / 0.39
16	MEPI	0.97 / 0.03	0.59 / 0.27	1.00 / 0.03	0.55 / 0.32
16	Bayes-D	0.96 / 0.06	0.55 / 0.21	0.98 / 0.08	0.52 / 0.31
16	PEC	0.93 / 0.03	0.59 / 0.20	0.98 / 0.03	0.56 / 0.30
20	20.7.1	1.00 / 0.03	0.81 / 0.18	1.00 / 0.03	0.77 / 0.21
20	20.7.18	0.99 / 0.04	0.79 / 0.17	1.00 / 0.03	0.75 / 0.22
20	MEPI	1.00 / 0.04	0.79 / 0.18	1.00 / 0.02	0.71 / 0.17
20	Bayes-D	0.99 / 0.03	0.80 / 0.16	1.00 / 0.03	0.77 / 0.20
20	PEC	0.99 / 0.04	0.80 / 0.16	1.00 / 0.03	0.72 / 0.20
24	24.7.1	1.00 / 0.04	0.85 / 0.15	1.00 / 0.01	0.80 / 0.11
24	MEPI	1.00 / 0.04	0.87 / 0.16	1.00 / 0.01	0.78 / 0.16
24	Bayes-D	1.00 / 0.03	0.85 / 0.16	1.00 / 0.03	0.82 / 0.17
24	PEC	1.00 / 0.03	0.85 / 0.15	1.00 / 0.03	0.80 / 0.15
28	28.7.1	1.00 / 0.03	0.91 / 0.14	1.00 / 0.03	0.86 / 0.18
28	MEPI	1.00 / 0.05	0.90 / 0.15	1.00 / 0.01	0.86 / 0.16
28	Bayes-D	1.00 / 0.03	0.89 / 0.14	1.00 / 0.03	0.85 / 0.17
28	PEC	1.00 / 0.03	0.89 / 0.15	1.00 / 0.02	0.82 / 0.10
32	32.7.2	1.00 / 0.04	0.91 / 0.15	1.00 / 0.00	0.85 / 0.09
32	32.7x	1.00 / 0.04	0.92 / 0.14	1.00 / 0.02	0.87 / 0.11
32	D-optimal	1.00 / 0.04	0.93 / 0.15	1.00 / 0.01	0.88 / 0.09

could have been in detecting the substantial main effects and the large interactions detected in the 40-run design. The OA 20.7.18 is only 18th best in terms of G-aberration. It is natural to compare the design with 20.7.1, which is best in terms of G-aberration, and to the PEC-optimal, MEPI-optimal and Bayesian D-optimal alternatives of the same run size. It further pays off to consider whether alternatives of other run sizes would improve detection of active effects or would be a more economical, yet equally powerful, alternative to the 20-run design.

In Table 10, we tabulate power and FDR results for the case of three active main effects and two active interactions of smaller size than the main effects. The pairs of figures separated by a slash are power and FDR, given for detection of main effects (ME) and two-factor interactions (2FI), both for the forward selection procedure and the detection using the Dantzig selector.

A striking feature of the results is the very high power for the main-effect detections over all designs along with low FDR. There are no substantial differences between the designs or between the method to detect the main effects. So the main effects could as well have been detected with the 16-run design. Note, however, that a follow-up design with only three factors would not be appropriate, because the different responses have different sets of active main effects.

The detection of two-factor interactions clearly depends on the run size. The 16-run designs are not recommended, because of large FDR values and poor power. The OA 16.7.1 shows especially poor results due to the availability of only 7 degrees of freedom to detect interactions.

The 20-32-run designs all show an improved FDR for detecting two-factor interactions when compared with the 16-run designs. The forward selection has FDR values ranging from 0.14-0.18, while the Dantzig selector seems to be improving with increasing run size. The FDR for this method is still worse than for the forward selection in the 20-run designs, about equal for 24 runs or 28 runs and better for 32 runs.

As to the power for detecting two-factor interactions, the results for the 20-run designs under forward selection would seem acceptable to many practitioners. As expected, the power increases with the run size, although the increase when we go from 28 runs to 32 runs is not substantial. Interestingly, the Dantzig selector has an inferior power to detect two-factor interactions when compared to the forward selection procedure. As the results on main effect detection are equally good for both procedures, we recommend the forward selection procedure for the present cases.

Finally, we believe that all five 20-run designs are suitable options to detect three main effects and two interactions sized smaller than the main effects. Among designs of the same run size, there are only small differences in forward selection's power to detect two-factor interactions (the 16-run cases being an exception). For the 20-run cases, OA 20.7.1 is best in power, for the 24-run designs, the MEPI design is best, for the 28-run designs, the OA is best, and for the 32-run design, the D-optimal design is best. Note that design 32.7x and the 32-run D-optimal design can estimate the full interaction model so that other detection methods might be more appropriate.

6 DISCUSSION

In this paper, we demonstrated the selection of a two-level design to screen main effects and interactions. We exemplified the selection process by studying seven-factor designs in 16–32 runs and 11-factor designs in 20–48 runs. The selection process involved, first, a selection of candidate designs for each of the run sizes considered and, second, a comparison of the power of the intended screening procedure based on the selected designs both within the same run size and across run sizes.

Our selection procedure included orthogonal arrays as well as MEPI-optimal, PEC-optimal and Bayesian D-optimal designs. The latter three types of design were generated using specific parameters for the prior variance of the regression coefficients (Bayesian D-optimal designs), the desired number of two-factor interactions and related parameters (MEPI-optimal designs), and the desired dimension of the projections (PEC-optimal designs); refer to Appendix A for details.

For the orthogonal arrays, we considered the complete series of nonisomorphic OAs for each run size and number of factors. An OA's inclusion in the set of candidate designs was primarily based on G-aberration. However, even if a minimum G-aberration OA has no full aliasing words of length 3, we do not like it to have full aliasing words of length 4. For this reason, we included a few near minimum G-aberration designs in our selection of 16-run seven-factor OAs.

Criteria that apply to nonorthogonal as well as to orthogonal designs include the gwlp, Q_B, EC, PEC, IC, PIC and galp. We found these criteria useful to articulate the different model fitting capabilities of a design. We showed that IC and PIC are more informative that EC and PEC, respectively. Figure 1 illustrates for OAs with $n = 40$ and $k = 11$ that PIC may show greater differences among designs than IC, but the choice between these criteria should more crucially be based on the types of interaction models one expects to estimate.

Measures of model discrimination were discussed only briefly in Section 3.7. Of existing measures we prefer summaries of EPD for ease of computation and interpretability. Other measures focus on worst case and best case scenarios. It seems that measuring a design's ability to discriminate between models needs further development; some criterion based on the density of models in n-dimensional space might prove useful. Miller (2005) contains some ideas that may show a way forward.

The second step in the selection process of a suitable design involved estimating power via simulation for the set of designs selected in the first step. Power simulations need simulated datasets under a variety of known models that include a random error as well as a screening procedure. We implemented a scenario where both the main effects and the two-factor interactions had coefficients size 0.5–3.5 times the size of the random error. In another scenario, the main effect coefficients were 2–3.5 times the random error, while the interaction coefficients were 0.5–2 times the random error. Both scenarios realistically cover situations where an effect might be missed because of its size. The ranges of 2–5 active main effects and 1–7 active two factor interactions cover a wide range of situations that likely can occur in practice.

We considered two screening procedures. One was based on forward selection under weak effect heredity with a protection against overfitting, while the other employed the Dantzig selector. The Dantzig selector outperformed forward selection except for the detection of two-factor interactions in case these are smaller than the main effects. Li et al. (2006) found that this case occurs most often in practice.

The simulation results suggest that orthogonal arrays selected based on G-aberration are powerful for detecting main effects, while MEPI-optimal designs appeared to perform equally well. For detecting two-factor interactions, PEC- and Bayesian D-optimal designs can be recommended. However, some orthogonal arrays do just as well. As there is no clear winner here, it is important to conduct simulation studies among several candidate designs to decide on a final design. The software provided with this article should be helpful for this purpose.

One of the initial big surprises in our simulations was the superior power when the screening is based on forward selection with a model independent error estimate. However, the FDR for this procedure is seriously inflated so that the procedure cannot be recommended. We would welcome further research on procedures that reliably integrate an independent error estimate in the detection of main effects and two-factor interactions.

A Non-orthogonal designs

Here, we briefly describe the details for construction of the Bayesian D-optimal, MEPI, and PEC designs used in this article.
In decimal form, the three designs for k=7, n=20 are:
Bayes-D: 0, 7, 13, 19, 26, 36, 40, 55, 59, 60, 75, 78, 80, 85, 92, 97, 98, 111, 118, 121
MEPI: 0, 14, 23, 25, 34, 43, 45, 49, 52, 58, 67, 79, 82, 85, 92, 101, 102, 104, 115, 127
PEC: 0, 12, 15, 19, 24, 40, 50, 53, 62, 71, 73, 74, 82, 84, 93, 96, 110, 113, 123, 124
To convert these to binary columns, use the MATLAB function dec2bin.

A.1 Bayesian D-optimal designs

Jones et al. (2008) utilized Bayesian D-optimality to construct supersaturated designs. Specifically, their criterion selects a design which maximizes

$$\phi_D = |\mathbf{X}'\mathbf{X} + \mathbf{K}/\tau^2|^{1/p}$$

where $\mathbf{X}$ is the $n \times p$ model matrix, τ^2 is the prior variance of the regression coefficients, and

$$\mathbf{K} = \begin{pmatrix} 0 & \mathbf{0_{1\times p}} \\ \mathbf{0_{p\times 1}} & \mathbf{I_{p\times p}} \end{pmatrix}.$$

The Bayesian D-optimal designs in this work were constructed using JMP (version 11) software's Custom Design platform with a default value of $\tau^2 = 1$. The intercept term is specified to be “Necessary” with all other terms “If Possible”.

A.2 MEPI-optimal designs

The main effect plus interaction (MEPI) model space can be used to design experiments when two-factor interactions are suspected but unspecified. However, the size of this model space can become quite large and thus, a serious hindrance of its use in constructing model robust designs. Let $\mathcal{F}_g$ represent the MEPI model space for up to g two-factor interactions (g can be considered an upper bound for the number of two-factor interactions expected). A MEPI-optimal design is that which maximizes the estimation capacity over $\mathcal{F}_g$. Maximizing the information capacity can be used as a secondary criterion.

MEPI-optimal designs in this paper were constructed using the approximate model space approach of Smucker and Drew (2015). These authors show that the designs constructed via this approach sacrifice little in terms of robustness and can be constructed in a shorter amount of time. An outline of their algorithm is as follows:

1. Select a small sample of s_1 models from the full model space, $\mathcal{F}_g$. This is the approximate model space, denoted by $\mathcal{S}_1$; it is chosen to be close to balanced (i.e., pairs of two-factor interactions appear together in models as equally often as possible). Smucker and Drew (2015) elect to create $\mathcal{S}_1$ based on balanced incomplete block designs. They explore s_1 =16, 32, 64, 128, and 256.

2. Construct a number of designs that are robust for the models in $\mathcal{S}_1$. Designs were constructed via a coordinate exchange algorithm.

3. Evaluate these designs with respect to the models in $\mathcal{F}_g$. If $\mathcal{F}_g$ is too large for a quick evaluation of all models, take a sample of s_2 models and evaluate the designs with respect to this set.

Table 11 lists the specifications for the MEPI-optimal designs utilized in this work. The supplementary materials of Smucker and Drew (2015) contain MATLAB programs to implement their design construction approach.

Table 11: MEPI-optimal Design Parameters

n	k	g	s_1	s_2
16	7	3	32	1330
20	7	6	64	2000
24	7	6	32	2000
28	7	11	16	2000
20	11	2	64	1485
24	11	6	64	2000
32	11	6	64	2000
40	11	6	16	2000
48	11	6	128	2000

A.3 PEC-optimal designs

Assume that h of the main effects are active, along with the associated $\binom{h}{2}$ two-factor interactions. Thus, there are $\binom{k}{h}$ possible models for any choice of h. Denoting the projective model space as $\mathcal{P}_h$, a PEC-optimal design seeks to maximize the projection estimation capacity sequence over

$\mathcal{P}_h$ for $1 \leq h \leq \ell$. For the projective model space, the designs we utilize are constructed via the coordinate exchange algorithm of Smucker et al. (2012). As a secondary criterion, we maximize the minimum D-efficiency. Table 12 lists the choices of ℓ for the PEC-optimal designs in this paper. The supplemental materials of Smucker et al. (2012) contain the MATLAB programs that we utilized for design construction.

Table 12: PEC-optimal Design Parameters

n	k	ℓ
16	7	4
20	7	5
24	7	5
28	7	6
20	11	4
24	11	5
32	11	6
40	11	6
48	11	6

B Computing Q_B

Tsai and Gilmour (2010, Section 5.1) take the full two-factor interaction model as the maximal model. Imposing marginality, they derive equation (5), where ξ_{ij} denotes the sum of prior probabilities for all models in which at least the main effects of i factors and j two-factor interactions of these i factors are included. In a related paper, Tsai and Gilmour (2007) focus on the second-order response surface model; there they impose the constraint that any model requiring more than $n-1$ degrees of freedom must have prior probability of 0. In the derivations to follow, we remove this constraint, so that the prior does not depend on the sample size. We assume the prior of Bingham and Chipman (2007) with (π_1, π_2, π_3) as defined in Section 3.3. When $\pi_3 = 0$, the sum ξ_{ij} equals the prior probability that a particular set of i main effects and j interactions are active, which is simply $\pi_1^i \pi_2^j$. However, when $\pi_3 > 0$ and marginality is imposed, the ξ_{ij} sums equal:

$$\begin{aligned}
\xi_{10} = &\quad \pi_1 + (1-\pi_1)(1 - C_1^{k-1}) \\
\xi_{20} = &\quad \pi_1^2 + 2\{\pi_1(1-\pi_1)[1-(1-\pi_3)C_1^{k-2}]\} + (1-\pi_1)^2\{1 - 2C_1^{k-2} + C_2^{k-2}\} \\
\xi_{21} = &\quad \pi_1^2\pi_2 + 2\pi_1(1-\pi_1)\pi_3 \\
\xi_{31} = &\quad \pi_1\xi_{21} + \pi_1^2(1-\pi_1)\pi_2[1-(1-\pi_3)^2 C_1^{k-3}] + 2\{\pi_1(1-\pi_1)^2\pi_3(1-(1-\pi_3)C_1^{k-3})\} \\
\xi_{32} = &\quad \pi_1^3\pi_2^2 + \pi_1^2(1-\pi_1)\pi_3^2 + 2\pi_1^2(1-\pi_1)\pi_3\pi_2 + \pi_1(1-\pi_1)^2\pi_3^2 \\
\xi_{42} = &\quad \xi_{21}^2,
\end{aligned} \tag{9}$$

where $C_r = [1 - \pi_1 + \pi_1(1-\pi_3)^r]$. If $\pi_3 = 0$, note how each ξ_{ij} simplifies to $\pi_1^i \pi_2^j$.

For $k = 7$, the priors (0.5, 0.8, 0) and (0.5, 0.4, 0.2) both have 3.5 main effects and 4.2 interactions expected to be active. However, the expected number of main effects to be included is $\xi_{10}k = 0.734(7) = 5.14$. Table 3 shows the ranking of designs under the strong heredity prior with $\pi_3 = 0$. Under the weak heredity prior given above, $Q_B = (3.134B_1 + 1.822B_2 + 0.917B_3 + 0.24B_4)/n$. This criterion elevates the MEPI design to the top, due to its small B_3 value. Under weak effect heredity, we expect to include 1 or 2 inactive main effects, due to those factors appearance in active interactions. Their inclusion increases the importance of low B_3.

The flexibility and simplicity of Q_B makes it an attractive criterion. A prior with $\pi_1 = 1$ corresponds to the MEPI family of models, with $g = \pi_2 k(k-1)/2$ active interactions expected. By contrast, the PEC family corresponds to $\pi_2 = 1$ and $\pi_1 < 1$.

C Analysis methods

C.1 Forward Selection

Despite documented shortcomings (e.g., high Type I error rates; see Westfall et al. (1998)) forward selection remains popular and commonly used in practice for variable selection. This is especially the case when $p > n$. Forward selection begins with the null model and adds the most significant term at each step based on an F-test. Here, we perform forward selection restricted by weak effect heredity. That is, an interaction term is not eligible to enter the model unless at least one of its parent main effects is selected for inclusion. To help avoid model overfitting, we control the experiment-wise error rate (EER) via Bonferroni adjusted p-values. Forward selection terminates when the adjusted p-value first exceeds the specified EER. In this article, we use EER=0.5. For additional justification, see Mee (2013).

C.2 Dantzig selector

The Dantzig selector (Candes and Tao, 2007) is a shrinkage method in which the estimator $\hat{\beta}$ is the solution to

$$\min_{\hat{\beta}\in\mathbb{R}} \left\|\hat{\beta}\right\|_1 \text{ subject to } \left\|\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\beta})\right\|_\infty \leq \delta,$$

where δ is a tuning constant. The Dantzig selector can be recast as a linear program and solved in a straightforward manner using linear programming algorithms available in many software packages. Our computations were performed using the "lpSolve" package in R.

To perform automated variable selection, we choose the value of the tuning parameter δ via the modified AIC (AIC$_c$):

$$\text{AIC}_c = n\log\left(\frac{RSS}{n}\right) + \frac{2n\tilde{p}}{n-\tilde{p}-1},$$

where RSS is the residual sum of squares and $\tilde{p}$ is the number of terms in the model under consideration. In this article, we perform the two-stage GaussDantzig selector. That is, active effects are first identified using the Dantzig selector and ordinary least-square estimates are obtained by regressing the response on the identified set of factors. The active effects selected by the Dantzig selector are those whose coefficient estimates exceed some threshhold γ. We use $\gamma = 0.5$ as this is the smallest effect size considered for active effects in the simulation study.

SUPPLEMENTARY MATERIAL

Design selection.pdf: Account of the selection of the designs in Tables 8 and 9 from complete series of nonisomorphic designs.

Programs.zip: Matlab programs for power simulations.

Designs.zip: Text files with the selected design.

ACKNOWLEDGEMENTS

We thank Byran Smucker for providing all the MEPI designs and Hongquan Xu for the result about sum(gwlp). The research of the second author was supported by a grant from the Research Foundation - Flanders (FWO).

References

Androulakis, E., Angelopoulos, P., and Koukouvinos, C. (2014). Model discrimination criteria on model-robust designs. *Communications in Statistics - Simulation and Computation*, 43:1575–1582.

Bingham, D. R. and Chipman, H. A. (2007). Incorporating prior information in optimal design for model selection. *Technometrics*, 49:155–163.

Bulutoglu, D. A. and Margot, F. (2008). Classification of orthogonal arrays by integer programming. *Journal of Statistical Planning and Inference*, 138:654–666.

Butler, N. A. (2003). Minimum aberration construction results for nonregular two-level fractional factorial designs. *Biometrika*, 90:891–898.

Candes, E. O. and Tao, T. (2007). The Dantzig Selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351.

Chen, J., Sun, D. X., and Wu, C. F. J. (1993). A catalogue of two-level and three-level fractional factorial designs with small runs. *International Statistical Review*, 61:131–145.

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

Cheng, C. S., Deng, L. Y., and Tang, B. (2002). Generalized minimum aberration and design efficiency for nonregular fractional factorial designs. *Statistica Sinica*, 12:991–1000.

Cheng, C. S., Mee, R. W., and Yee, O. (2008). Second order saturated orthogonal arrays of strength three. *Statistica Sinica*, 18:105–119.

Cheng, C. S., Steinberg, D. M., and Sun, D. X. (1999). Minimum aberration and model robustness for two-level factorial designs. *Journal of the Royal Statistical Society Series B*, 61:85–94.

Deng, L. Y. and Tang, B. (1999). Generalized resolution and minimum aberration criteria for Plackett-Burman and other nonregular factorial designs. *Statistica Sinica*, 9:1071–1082.

Deng, L. Y. and Tang, B. (2002). Design selection and classification for Hadamard matrices using generalized minimum aberration criteria. *Technometrics*, 44:173–184.

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*, 56:1–16.

DuMouchel, W. and Jones, B. (1994). A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, 36:37–47.

Fries, A. and Hunter, W. G. (1980). Minimum aberration 2^{k-p} designs. *Technometrics*, 22:601–608.

Hamada, M. and Wu, C. F. J. (1992). Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137.

Ingram, D. and Tang, B. (2005). Minimum G aberration design construction and design tables for 24 runs. *Journal of Quality Technology*, 37:101–114.

Jones, B., Li, W., Nachtsheim, C., and Ye, K. Q. (2007). Model discrimination - another perspective on model-robust designs. *Journal of Statistical Planning and Inference*, 137:1576–1583.

Jones, B. A., Lin, D.K. J., and Nachtsheim, C. J. (2008). Bayesian D-Optimal Supersaturated Designs. *Journal of Statistical Planning and Inference*, 138:86–92.

Li, W. and Nachtsheim, C. J. (2000). Model-robust factorial designs. *Technometrics*, 42:345–352.

Li, X., Sudarsanam, N., and Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, 11:32–45.

Liao, C.-T. and Chai, F.-S. (2009). Design and analysis of two-level factorial experiments with partial replication. *Technometrics*, 51:66–74.

Lin, C. D., Miller, A., and Sitter, R. R. (2008). Folded-over non-orthogonal designs. *Journal of Statistical Planning and Inference*, 138:3107–3124.

Lin, D. K. J. and Draper, N. R. (1993). Generating alias relationships for two-level Plackett and Burman designs. *Computational Statistics and Data Analysis*, 15:147–157.

Loeppky, J. L. (2004). Ranking nonregular designs. PhD dissertation, Simon Fraser University, Dept. of Statistics and Actuarial Sciences, Burnaby BC, Canada.

Loeppky, J. L., Sitter, R. R., and Tang, B. (2007). Nonregular designs with desirable projection properties. *Technometrics*, 49:454–467.

Mee, R. W. (2013). Tips for analyzing nonregular fractional factorial experiments. *Journal of Quality Technology*, 45:330–349.

Mee, R., Liao, C. T., and Chai, F. S. (2009). Letter to the Editor [with response]. *Technometrics*, 51:475–478.

Miller, A. (2005). The analysis of unrelicated factorial experiments using all possible comparisons. *Technometrics*, 47:51–63.

Miller, A. and Sitter, R. R. (2001). Using the folded-over 12-run Plackett-Burman design to consider interactions. *Technometrics*, 43:44–55.

Miller, A. and Sitter, R. R. (2005). Using folded-over nonorthogonal designs. *Technometrics*, 47:502–513.

Montgomery, D. C. (2012). *Design and analysis of experiments.* 8th edition, Wiley, New York.

Mukerjee, R. and Wu, C. F. J. (2006). *A modern theory of factorial design.* Springer, New York, NY, USA.

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. *Journal of the Royal Statistical Society Supplement*, 9:128–139.

Schoen, E. D., Eendebak, P. T., and Nguyen, M. V. M. (2010). Complete enumeration of pure-level and mixed-level orthogonal arrays. *Journal of Combinatorial Designs*, 18:123–140.

Schoen, E. D. and Mee, R. W. (2012). Two-level designs of strength 3 and up to 48 runs. *Journal of the Royal Statistical Society Series C*, 61:163–174.

Smucker, B. J., del Castillo, E., and Rosenberger, J. L. (2012). Model-robust two-level designs using coordinate exchange algorithms and a maximin criterion. *Technometrics*, 54:367–375.

Smucker, B. J. and Drew, N. M. (2015). Approximate model spaces for model-robust experiment design. *Technometrics*, 57:54–63.

Sun, D. X. (1993). Estimation capacity and related topics in experimental design. PhD dissertation, University of Waterloo, Department of Statistics and Actuarial Science, Waterloo ON, Canada.

Sun, D. X., Li, W., and Ye, K. Q. (2008). An algorithm for sequentially constructing nonisomorphic orthogonal designs and its applications. *Statistics and Application*, 6:144–158.

Tang, B. and Deng, L. Y. (1999). Minimum G_2-aberration for nonregular fractional factorial designs. *Annals of Statistics*, 27:1914–1926.

Tsai, P.-W. and Gilmour, S. G. (2010). A general criterion for factorial designs under model uncertainty. *Technometrics*, 52:231–242.

Tsai, P.-W., Gilmour, S. G., and Mead, R. (2000). Projective three-level main-effects designs robust to model uncertainty. *Biometrika*, 87:467–475.

Tsai, P.-W., Gilmour, S. G., and Mead, R. (2007). Three-level main-effects designs exploiting prior information about model uncertainty. *Journal of Statistical Planning and Inference*, 137:619–627.

Wang, J. C. and Wu, C. J. F. (1995). A hidden projection property of Plackett-Burman and related designs. *Statistica Sinica*, 5:235–250.

Westfall, P. H., Young, S. S., and Lin, D.K. J. (1998). Forward selection error control in the analysis of supersaturated designs. *Statistica Sinica*, 8:101–117.

Xu, H. (2005). Some nonregular designs from the Nordstrom-Robinson code and their statistical properties. *Biometrika*, 92:385–397.

Xu, H. and Deng, L.-Y. (2005). Moment aberration projection for nonregular fractional factorial designs. *Technometrics*, 47:121–131.

Xu, H. (2015). Nonregular factorial and supersaturated designs, in *Handbook of Design and Analysis of Experiments*, Eds. Dean, A., Morris, M., Stufken, J., and Bingham, D. Chapman and Hall/CRC.