Supplementary sections for

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)

February 25, 2016

A Criteria for regular fractional factorial designs

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

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

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 pairs 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_{IV}^{k-f} 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.

B Non-orthogonal designs

In this section, we briefly describe the details for construction of the Bayesian D-optimal, MEPI, and PEC designs used in the main article.

B.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".

B.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 1 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 1: 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

B.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 2 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 2: 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

C Computing Q_B

Tsai and Gilmour (2010, Section 5.1) select two-level FFDs according to the Q_B criterion (Tsai et al., 2007), which approximates the average estimation efficiency of main effects and two-factor interactions over the class of all sub-models of the full two-factor-interaction model. Each sub-model's efficiency is weighted with the prior probability that the model turns out to be the best model for the data at hand. All sub-models satisfy the marginality requirement that a main effect may be omitted only if that factor does not appear in any interaction. The original derivation includes 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. In this section, 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. 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:

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

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 $\xi_{10}k = 0.734(7) = 5.14$. The main paper shows the ranking of 20-run 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 $\pi_1 < 1$.

D Study of strength-3 designs with 11 factors and 40 runs

Section 2.9 contrasted four OA$(40, 11, 3)$. Here we provide additional insights by comparing all strength 3 designs of this size. There are 260 OA$(40, 11, 3)$, all with generalized resolution 4.4; B_4 ranges from 18.96 to 22.8 (see Table 8 in the main paper). There are 48 minimum G-aberration designs, which have a cfv = [$F_4(24, 8) = (18, 312)$; $F_6(16) = 142$; $F_(24, 8) = (4, 161)$; $F_{10}(16) = 4$)]. 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 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.

The main paper's Table 8 highlights three of the 48 minimum G-aberration designs; Design 40.11.1a is the PEC optimal design among all the strength 3 arrays, while Design 40.11.1b has the best galp and Design 40.11.1c is best with respect to EC$_{15}$ among the strength 3 arrays. Figure 1 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 1 reveals a strong negative correlation between PIC_6 and IC_{15} for designs with the

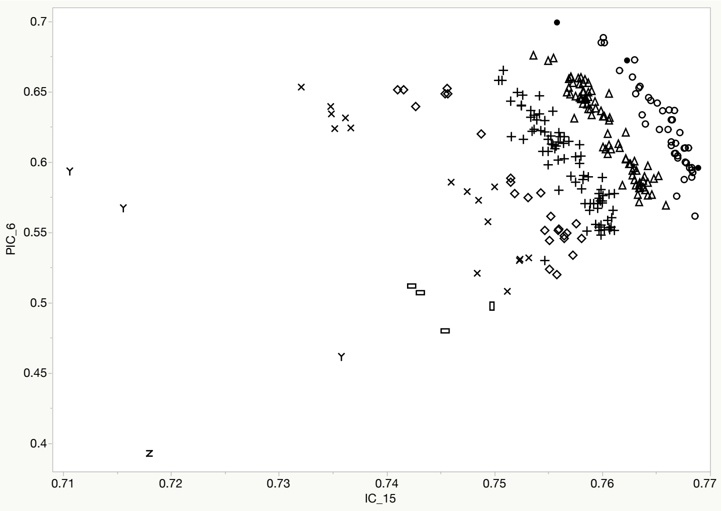

Figure 1: Comparison of 260 OA(40, 11, 3): average D-efficiency for two-factor interaction models across all 462 6-factor projections (PIC$_6$) vs. average D-efficiency for models with 11 main effects and 15 interactions (IC$_{15}$). Different symbols indicate different values of B_4, with empty and filled circles marking the 48 minimum G-aberration designs and z marking the maximum G-aberration design.

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.

E Selection of orthogonal designs

In this section, we select orthogonal designs for seven factors and 12–32 runs and 11 factors in 12–48 runs for screening purposes, based on complete sets of nonisomorphic designs. In the main paper, the selected designs are studied further with simulated models and a screening strategy to detect active effects.

E.1 Designs with 7 factors

Table 8 of the main paper shows 22 designs that were selected for further study. Nine of these were selected based on our sets of orthogonal nonisomorphic designs; the other 13 designs were constructed by optimal design methods. Here we consider the selection of orthogonal designs.

Table 3: Five seven-factor designs in 16 runs

ID	$F_3(16,8)$		$F_4(16,8)$		Q_B	p_4	PIC_4	PIC_5
16.7.1	0	0	7	0	0.105	28/35	0.80	0.00
16.7.2	0	4	3	8	0.113	32/35	0.86	0.15
16.7.3	0	6	1	12	0.116	34/35	0.88	0.22
16.7.4	0	8	0	12	0.120	35/35	0.89	0.19
16.7.5	0	8	0	14	0.128	35/35	0.89	0.22

The tabulated Q_B values are based on strong effect heredity, a prior probability of an active main effect of 0.5 and a prior probability of 0.8 for an active interaction given that the corresponding main effects are also active.

E.1.1 12 runs

There is a single seven-factor design in 12 runs. Its generalized resolution is 3.667, which is attractive. Each main effect and each two-factor interaction is partially aliased with 15 other effects with a correlation of $\pm 1/3$, so that each entry of the galp equals 2.667. The correlations are not problematic when considering projections into three or four factors ($PIC_3 = 0.95; PIC_4 = 0.81$). However, the correlations result in the presence of 18 MDS of size 2. This suggests that model discrimination with this design may not be easy. For this reason, the design is not selected for power comparisons with other designs.

E.1.2 16 runs

The seven-factor designs of 16 runs designs were considered earlier by Deng and Tang (2002). We ordered the 55 designs of 16 runs according to the G-aberration. Table 3 shows characteristics of the top-5 designs.

Design 16.7.1 is a strength-3 design. This should make the design particularly suitable to detect main effects. However, the design is not the best in p_4, because of the defining words of length 4. The same is the case for the second-best and third-best designs in terms of G-aberration. The designs ranked fourth and fifth in terms of G-aberration, have no full words of length 3 or 4 and permit estimation of all interaction models in subsets of four factors with an average D-efficiency of 0.89. In addition, despite their higher Q_B value, they permit estimation of all models with two interactions. Therefore, they are better able to discriminate among models with one or two interactions than the three designs with smaller G-aberration.

In the simulation study, we include 16.7.1 because of its strength and both 16.7.4 and 16.7.5 because of their greater potential to detect two-factor interactions.

Table 4: Three seven-factor designs in 20 runs

ID	$F_3(12,4)$		$F_4(12,4)$		Q_B	p_5	PIC_5
20.7.1a	0	35	2	33	0.066	19/21	0.76
20.7.1b	0	35	2	33	0.066	19/21	0.76
20.7.18	0	35	5	30	0.078	18/21	0.67

Table 5: Three seven-factor designs in 24 runs

ID	$F_3(8)$	$F_4(8)$	Q_B	p_5	p_6	PIC_5	PIC_6
24.7.1	0	35	0.039	21/21	0/7	0.87	0
24.7.2	4	21	0.034	21/21	6/7	0.90	0.63
24.7.3	4	23	0.037	21/21	3/7	0.89	0.21

E.1.3 20 runs

There are two minimum G-aberration designs with seven factors and 20 runs which we designate 20.7.1a and 20.7.1b, respectively. Table 4 shows that the two designs have identical p_5 and PIC_5 in addition to identical F_3 and F_4 factors. Design 20.7.1a is slightly better in EC_6, because only 3 out of the 54,264 models with all the main effects and six two-factor interactions are not estimable, as opposed to 5 of such non-estimable models for 20.7.1b. On the other hand, the galp of design 20.7.1b, $(f_{1.92} = 12; f_{1.6} = 16)$ is slightly better than design 20.7.1a's galp $(f_{2.24} = 1; f_{1.92} = 10; f_{1.6} = 17)$. The first entry of the latter galp is due to the AD interaction.

We included 20.7.1a in the main paper's simulation study, and dropped the 'a' in the ID of the design. Further simulations, not reported here, show that there was no detectable difference in power between both minimum G-aberration designs.

Design 20.7.18 was included in the simulation study for two reasons. First, Table 4 suggests that this design will be inferior for screening. We wanted to check how appreciable the difference with the minimum G-aberration design is. Second, design 20.7.18 can be shown to contain half of the runs of the design actually used for the motivating example, which was the minimum G-aberration 40-run design with 7 factors. It is conceivable that this design is run in two parts, where the first part is 20.7.18. A decision on conducting the second part might depend on the success off the first part. The issue here is whether one should start with a minimum G-aberration 20-run design, which does not extend to the minimum G-aberration 40-run design, or whether one aims at ending with the minimum G-aberration 40-run design and, for this reason does not start with the best 20-run design.

E.1.4 24 runs

Deng and Tang (2002) presented top 3 designs according to G-aberration based on projections from 23-factor designs obtained from Hadamard matrices. These designs were also best, second best and third best according to G_2-aberration. Schoen et al. (2015) confirm by exhaustive search that these designs maintain this ranking among all possible seven-factor designs. Key properties are presented in Table 5.

The minimum G-aberration design has a strength of 3. It is a fold-over of the 12-run design discussed earlier. Each interaction is aliased with 10 other interactions that do not share either of the interacting factors. Therefore the GALP for each interaction is $1 + 10 * (1/3)^2 = 2.11$. For the second-best design, the worst GALP for an interaction is 1.89; it occurs for two interactions. Finally, for the third design the worst GALP for an interaction is 2.11. It occurs for a single interaction. It is interesting to note that the main effect of factor 6 in this design has a GALP of 1. So this is a clear main effect.

As shown by the p_6 value, the second-order model in any subset of six factors is not estimable for 20.7.1. In contrast, the second design only has a single of such a six-factor subset, while there are four such subsets for the third design. These features are parallelled in the Q_B vales for the respective designs. Based on Q_B we would recommend 24.7.2. However, the Bayesian D-optimal design that we constructed turned out to be an orthogonal array of strength 2 with an even smaller Q_B value, and we dropped 24.7.2 in favor of the Bayesian D-optimal design.

We noted before that it would be extremely rare if all interactions among six factors would be active. So, as $p_5 = 1$ for all three designs, we do not think that the bad p_6 of the first design should be a conclusive argument to discard this design. A boxplot of D-efficiencies of all models containing all the main effects and 1–4 two-factor interactions is shown in Figure 2. The strength-3 design clearly is better here. This is all the more true when looking at models with five interactions (not shown in the figure): the efficiencies for the first design are all above 0.9. Due to the presence of MDS of size 5, some efficiencies for the second and third design equal zero in models with 5 interactions. An additional advantage of the strength-3 design is the independent assessment of the main effects. We therefore include this design (along with the Bayesian D-optimal design of strength 2) in the simulation study.

E.1.5 28 runs

We searched through all possible OA(28,7,2) to find good ones in terms of G-aberration. Table 6 shows the top 6 designs. The minimum G-aberration design has $F_3(4) = 35$ and $F_4(12, 4) = (1, 34)$. The other five designs have the same F_3, but their $F_4(12, 4) = (2, 33)$. The rank of the two-factor interaction model matrix is 28 for all these cases, so the two-factor interaction model has only one linear dependency. The single MDS for the first design involves just 11 interactions, while the other designs' MDS are larger, some even 21. However, 11 is sufficiently long as to cause little or no confusion.

In Figure 3, we show D-efficiencies for second order models in 5 or 6 factors, based on the

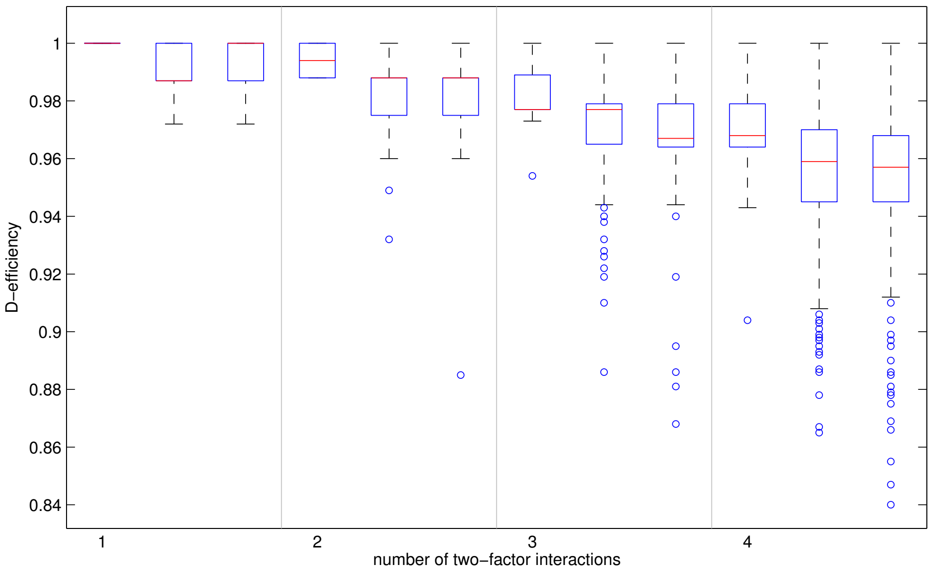

Figure 2: 24-run 7-factor designs: D-efficiencies of models with 1–4 interactions

Table 6: Six seven-factor designs in 28 runs

ID	$F_3(4)$	$F_4(12,4)$		$F_5(16,8)$		Q_B	p_6	PIC_6
28.7.1	35	1	34	2	11	0.023	7/7	0.87
28.7.2	35	2	33	0	15	0.024	7/7	0.85
28.7.3	35	2	33	0	17	0.024	7/7	0.85
28.7.4	35	2	33	1	11	0.024	7/7	0.85
28.7.5	35	2	33	1	13	0.024	7/7	0.85
28.7.6	35	2	33	1	15	0.024	7/7	0.85

top 6 designs. The minimum G-aberration design performs better, because it has only one correlation of 3/7, as opposed to 2 for the other designs. As the minimum G-aberration also has slightly better information capacities for up to 4 interactions (plot not shown), we include this design in the simulation study.

E.1.6 32 runs

Table 7: Three seven-factor designs in 32 runs

ID	$F_3(8)$	$F_4(16,8)$		Q_B	p_6	PIC_6
32.7.1	0	4	0	0.008	6/7	0.81
32.7.2	0	6	0	0.011	6/7	0.78
32.7.x	2	3	6	0.011	7/7	0.90

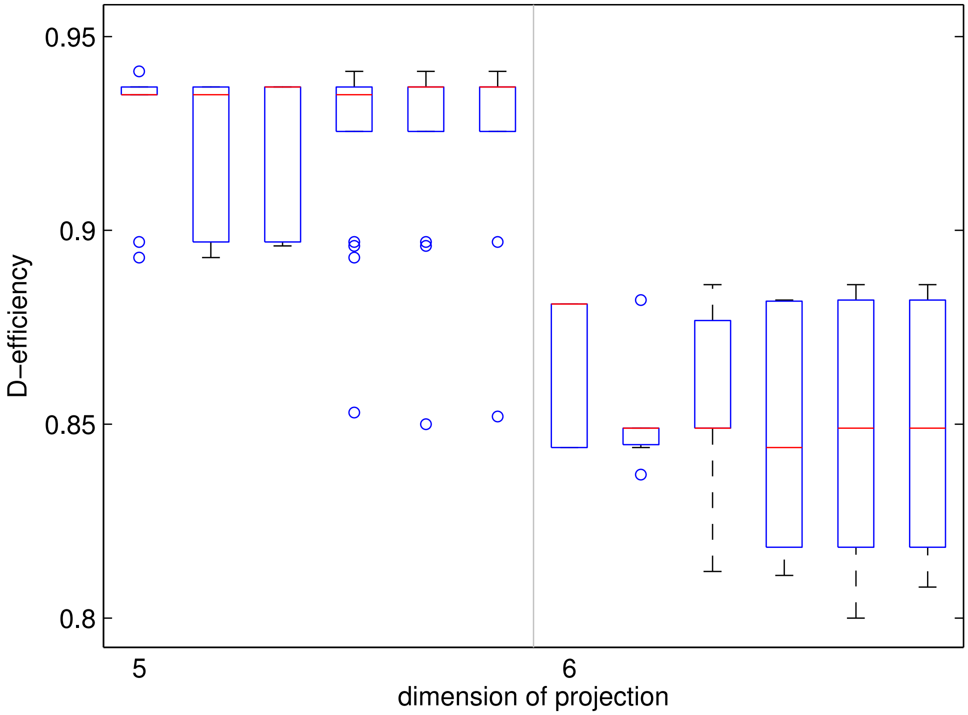

Figure 3: 28-run 7-factor designs: D-efficiencies of second order models involving 5 or 6 factors

There are only 17 7-factor 32-run designs of strength 3. Schoen and Mee (2012) checked these designs and presented the minimum G-aberration design. Here, we add the second best design and a third design designated x7, which maximizes the D-efficiency of the full interaction model among al orthogonal arrays in 32 runs ($D = 0.84$; Eendebak and Schoen, 2016). Table 7 shows key features of the designs.

The strength-3 arrays do not have a full rank for the second order model matrix in all seven factors. In addition, neither of these designs permits estimation of the second-order model in all 6-factor subsets. Design 32.7.1 has one MDS of size 5 involving factors 1-6. Therefore, it is this projection whose full model is not estimable. Design 32.7.2 has 4 MDS of size 6. All these involve factors 1-6. So, again, it is the projection into this set of 6 factors whose full model is not estimable. Because of the absence of MDS of size 5, we include 32.7.2 rather than 32.7.1 in the simulation study. Design 32.7.x is included as well, in view of its ability to estimate the full interaction model in 7 factors.

E.2 Designs with 11 factors

Table 9 of the main paper shows 25 designs that were selected for further study. Ten of these were selected based on our sets of orthogonal nonisomorphic designs; the other 15 designs were constructed by optimal design methods. Here we consider the selection of orthogonal designs. The tabulated Q_B values are based on strong effect heredity, a prior probability of an active main effect of 0.5 and a prior probability of 0.4 for an active interaction given that the corresponding main effects are also active.

E.2.1 12, 16, and 20 runs

Table 8: Five 11-factor designs in 20 runs

ID	$F_3(12,4)$		$F_4(12,4)$		Q_B	p_5	PIC_5
20.11.1	5	160	30	300	0.191	392/462	0.67
20.11.2a	6	159	26	304	0.192	383/462	0.66
20.11.2b	6	159	26	304	0.192	380/462	0.65
20.11.2c	6	159	26	304	0.192	368/462	0.63
20.11.2d	6	159	26	304	0.192	371/462	0.64

For 11 factors, the smallest strength 2 designs have 12, 16, and 20 runs. A 12-run design is a saturated main effect design. If one desires to explore possible interactions, such a design is surely inadequate. A 16-run design provides 4 degrees of freedom for two-factor interactions when all main effects are included. Such a design thus resembles the 12-run design for seven factors. For 20-run designs, there are 8 degrees of freedom for interactions, which is similar to a 16-run design for seven factors.

Deng and Tang (2002) identified the minimum G-aberration designs for 16 and 20 runs and these are the two designs we discuss further. The minimum G-aberration OA$(16, 11, 2)$ design has an attractive generalized resolution: $\rho = 3.5$. However 15 two-factor interactions are not estimable when a single interaction is included with the main effects. In fact, there are models with four main effects and one interaction that are not estimable. This is not much better than a regular resolution III 2^{11-7} fraction.

There are 1914 different OA(20,11,2). Table 8 shows salient features of the minimum G-aberration design and four designs that are second best according to G-aberration. All have a generalized resolution of 3.4, which is slightly worse than the minimum G-aberration 16-run design. However, based on 20.11.1 all two-factor interactions are estimable when a single interaction is included with the main effects, and just 1 model out of 1485 is not estimable when two interactions are included. The corresponding MDS involves all 11 factors, so it is not likely to be problematic. Design 20.7.11c has no MDS of size 2 and 7 MDS of size 3. The remaining three designs have 10 MDS of size 2.

The two-factor interaction model for 20.11.1 is estimable for all four-factor projections and for 392 out of 462 possible five-factor projections. For 20.11.2c, 368 out of 462 such models are estimable. We find this difference in favor of 20.11.1 more substantial than the difference in estimable models with 2 interactions in favor of 20.11.2c. Design 20.11.1 is thus selected in our simulation study.

E.2.2 24 runs

Table 9: Three 11-factor designs in 24 runs

ID	$F_3(8)$	$F_4(8)$	Q_B	p_5	PIC_5
24.11.1	0	330	0.092	462/462	0.87
24.11.2	18	270	0.100	462/462	0.86
24.11.3	28	246	0.107	458/462	0.85

Schoen et al. (2015) identified the top-3 designs according to both G-aberration and G_2-aberration. Key features are summarized in Table 9. The first design has a strength of 3. The other two designs have a strength of 2. The third design has four non-estimable projections into 5 factors, while all 5-factor projections are estimable in the other two designs. We prefer the strength 3 option, because it permits an independent assessment of main effects and two-factor interactions. Accordingly, this design is included in our simulation study.

E.2.3 32 runs

Table 10: Four 11-factor designs in 32 runs

ID	$F_3(8)$	$F_4(32, 24, 16, 8)$				Q_B	p_6	PIC_6
32.11.1	0	3	0	90	0	0.048	102/462	0.18
32.11.2	0	4	0	84	0	0.047	108/462	0.19
32.11.3	0	4	0	86	0	0.048	102/462	0.18
32.11.x	68	0	0	18	180	0.069	453/462	0.76

Table 10 shows key properties of four 32-run designs considered here for screening purposes. Schoen and Mee (2012) list the minimum G-aberration design (32.11.1) and the minimum G_2-aberration design (32.11.2). Here, we also consider design 32.11.3, which has the same G_2-aberration as the minimum G-aberration design and design 32.11.x, which is a strength 2 alternative constructed by concatenating a specific OA(20, 7 , 2) to the single OA(12, 7, 2).

Obviously, the strength-3 options are not particularly good in fitting interaction models, because all three designs considered here have defining words of length 4. This results in as many singular projections into 4 factors as there are words in the defining relation and three times as many nonestimable models with all the main effects and two two-factor interactions. In contrast, design 32.11.x permits estimation of second order models in all 5-factor subsets and 453 out of the 462 6-factor subsets. It is interesting to note that this feature is reflected in Q_B if we would use a prior probability of an active main effect of 0.5 and a prior probability of 1 for an active interaction given that the corresponding main effects are also active.

In the main paper, we study the screening process for design 32.11.1 and design 32.11.x by simulation. Designs 32.11.2 and 32.11.3 are not included in view of the extra defining word of length 4.

E.2.4 40 runs

In Section 3.9 of the main paper, we discussed all strength 3 designs in 40 runs to illustrate the interplay between various criteria. We refer to that section for particulars of recommended designs.

E.2.5 48 runs

Table 11: Three 11-factor designs in 48 runs

ID	$F_4(48,32,16)$			Q_B	p_6	PIC_6
48.11.1	0	0	82	0.011	432/462	0.88
48.11.x	0	0	98	0.014	450/462	0.90
48.11.y	1	0	77	0.012	434/462	0.88

Schoen and Mee (2012) recommend three designs based on G-aberration, G_2-aberration and rank of the second-order model matrix. Table 11 shows salient features of the designs. The first design is the minimum G-aberration design; it also minimizes G_2-aberration. The second design was recommended, because it minimizes G-aberration among the designs with a maximum rank of the second-order model matrix. The third design minimizes G_2-aberration among those designs.

In Figure 4, we show D-efficiencies of interaction models in subsets of 4–6 factors. The first design has 30 singular six-factor models, while the second one has 12 of these and the third design 28. Design 48.11.y features a defining word of length 4. Therefore, there are projections into 4 and more factors with a singular second-order model matrix. This defining word also causes model matrices with 2–5 interactions to be singular; singularity for the other two designs only starts at models with six interactions. For these reasons, the third design is not recommended for screening purposes. We include the minimum G-aberration design 48.11.1 and design 48.11.x, which has a better p_6, in the simulation study.

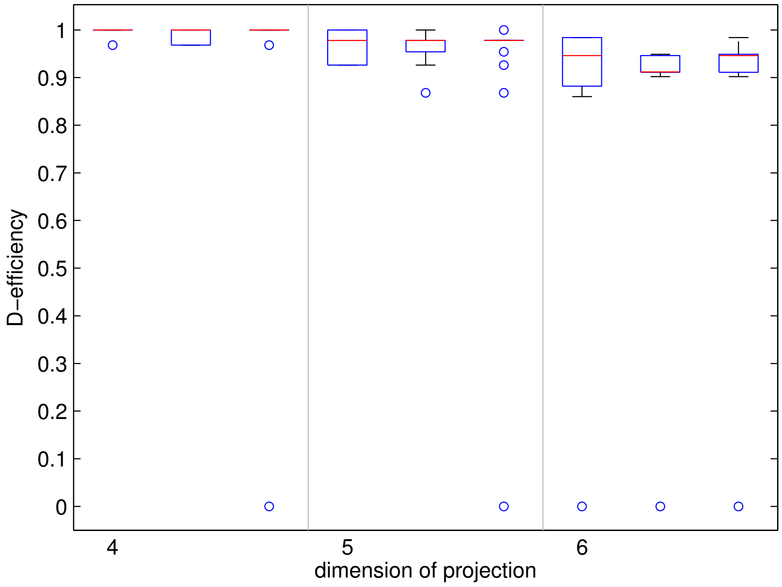

Figure 4: 48-run 11-factor designs: efficiencies of second-order models in subsets of the factors

F Analysis methods

F.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).

F.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.

References

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

Candes, E. O. and Tao, T. (2007). The Dantzig Selector: statistical estimation when p is much larger than n. *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., 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., 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. (2002). Design selection and classification for Hadamard matrices using generalized minimum aberration criteria. *Technometrics*, 44:173–184.

Eendebak, P. T. and Schoen, E. D. (2016). Two-level designs to estimate all main effects and two-factor interactions. *Technometrics*, http://dx.doi.org/10.1080/00401706.2016.1142903.

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

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

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

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

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.

Schoen, E. D., Vo-Than, N., and Goos, P. (2015). Two-level orthogonal designs in 24 and 28 runs. Working paper 2015-016, University of Antwerp, Faculty of Applied Economics.

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.

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. (2007). Three-level main-effects designs exploiting prior information about model uncertainty. *Journal of Statistical Planning and Inference*, 137:619–627.

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