# Two-Level Designs to Estimate All Main Effects and Two-Factor Interactions

Pieter T. Eendebak
Department of Engineering Management, University of Antwerp, Belgium
(pieter.eendebak@gmail.com)

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

November 25, 2015

**Abstract**

We study the design of two-level experiments with $N$ runs and $n$ factors large enough to estimate the interaction model, which contains all the main effects and all the two-factor interactions. Yet, an effect hierarchy assumption suggests that main effect estimation should be given more prominence than the estimation of two-factor interactions. Orthogonal arrays (OAs) favor main effect estimation. However, complete enumeration becomes infeasible for cases relevant for practitioners. We develop a partial enumeration procedure for these cases and we establish upper bounds on the D-efficiency for the interaction model based on arrays that have not been generated by the partial enumeration. We also propose an optimal design procedure that favors main effect estimation. Designs created with this procedure have smaller D-efficiencies for the interaction model than D-optimal designs, but standard errors for the main effects in this model are improved. Generated OAs for 7–10 factors and 32–72 runs are smaller or have a higher D-efficiency than the smallest OAs from the literature. Designs obtained with the new optimal design procedure or strength-3 OAs (which have main effects that are not correlated with two-factor interactions) are recommended if main effects unbiased by possible two-factor interactions are of primary interest. D-optimal designs are recommended if interactions are of primary interest.

*Keywords: coordinate exchange; D-efficiency; $D_s$-efficiency; optimal design; orthogonal array; partial enumeration*

# 1 INTRODUCTION

Experimenters using two-level factorial experiments usually think of the data as being generated from an additive model with main effects, two-factor interactions and higher-order interactions. To structure the analysis, they assume that main effects are more important than two-factor interactions, while two-factor interactions are more important than higher-order interactions. The assumption, called effect hierarchy, was coined first by Wu and Hamada (2000).

Empirical evidence in support of the effect hierarchy assumption was given by Li et al. (2006). These authors considered 46 two-level experiments with 3–7 factors. They found that about 40% of the main effects were active, as opposed to 11% of the two-factor interactions and 6.8 % of the three-factor interactions. In addition, the median main effect size was four times larger than the median size of the two-factor interactions and eight times larger than the median size of the three-factor interactions. While three-factor interactions evidently cannot be ruled out, including only main effects and two-factor interactions in a model for responses from two-level experiments seems a reasonable first approach.

In this paper, we consider the design of two-level experiments large enough to estimate a model with all the main effects and all the two-factor interactions. Yet the effect hierarchy assumption suggests that there are not so many two-factor interactions active, and that the size of active two-factor interactions is considerably smaller than the size of the main effects. Under these conditions, it makes sense to maximize precision of main effects that are unbiased by possible two-factor interactions. An example of this type of experiment was carried out recently at TNO, Eindhoven, the Netherlands. The experiment was concerned with the making of 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 concentrations of seven colorants. The main interest was in the size of the factorial effects. Only a few of the colorants are expected to be active for any given wave length. Further, optical laws suggest that main effects are much more likely than interaction effects. The experimental budget permitted construction of as many as 40 phantoms. Clearly, this number should be sufficient to construct a model with an intercept, all seven main effects and all 21 two-factor interactions. In the rest of this paper, we call such a model the interaction model.

The purpose of this paper is to develop procedures for generating designs that can fit the interaction model, while giving the main effect estimators more precision than the estimators of the two-factor interactions. Design alternatives that might be considered for the phantom experiment include orthogonal arrays (OAs) and D-optimal designs. We contribute to the development of both types of design. In the rest of this section, we provide more details on OAs

and D-optimal designs and we outline the further organization of the paper.

## 1.1 Orthogonal arrays

Generally, an OA of strength $t$, $N$ runs and $n$ factors at $s$ levels is an $N \times n$ array of $s$ symbols such that for every $t$ columns every $s^t$ $t$-tuple occurs equally often (Rao, 1947; Hedayat et al., 1999). Such an array is denoted $\mathrm{OA}(N, n, s, t)$. Our present interest is in arrays with $s = 2$, and we omit the reference to the number of levels of the factors in the notation for an OA.

An attractive feature of OAs is that a model with only main effects can be estimated with the maximum possible precision. Therefore, OAs seem ideal candidate designs if the effect hierarchy assumption applies. However, the extent to which the maximum precision for main effects is retained in the interaction model depends on the strength of the OA.

OAs of strength 4 are D-optimal for the interaction model, because all subsets of four factors form an equally replicated full factorial design. For this reason, all main effect contrast vectors and all two-factor interaction contrast vectors are orthogonal to each other, and both the main effect estimators and two-factor interaction estimators have a maximum precision.

A disadvantage of these arrays is their run size. For the seven factor phantom design, an OA of strength 4 requires 64 runs, which is a substantial larger than both the experimental budget of 40 runs and the number of parameters in the interaction model, which equals 29. At the same time, the effect hierarchy assumption suggests that it is not important that all effects in the interaction model have maximum precision. It is therefore natural to study OAs of strength $t < 4$ capable of fitting the interaction model with smaller run sizes than a strength-4 array.

OAs of strength 3 retain mutual independence of main effects and independence of main effects with interactions. Therefore, main effects in an interaction model are estimated with maximum precision. The estimators of two-factor interactions are correlated. Therefore, at least some of these interactions are not estimable with maximum precision in a full interaction model. This need not be a problem if the effect hierarchy assumption holds, however. It is therefore of interest to study strength-3 OAs with maximum D-efficiencies for the interaction model (to be defined formally later in the paper).

OAs of strength 2 have orthogonal main effect contrast vectors, but these are correlated with the contrast vectors of two-factor interactions. Therefore, the main effect estimators have maximum precision only in a first-order model. At the same time, the D-efficiencies for the interaction model can be higher than in strength-3 arrays, because the combinatorial restrictions are less severe.

A naive way to find out OAs of strength 2 or 3 with the best possible D-efficiency is to enumerate all $\mathrm{OA}(N, n, t)$, and to check subsequently their D-efficiencies. The problem one is faced with in carrying out such a procedure is the large number of different designs. For example, we were able to establish a set of 530,469,996 $\mathrm{OA}(32, 7, 2)$. Any $\mathrm{OA}(32, 7, 2)$ not in the set can

be obtained from an array in the set by a sequence of column permutations, row permutations or level switches in a column. The arrays in the set cannot be obtained from each other by such a sequence of permutations. There are five OAs with the best D-efficiency for the interaction model; its value is 0.8432.

The enumeration of the set of $\mathrm{OA}(32, 7, 2)$ took about 7 days on a PC with an Intel Core i7 870 CPU at 2.93GHz. It is computationally infeasible to enumerate all $\mathrm{OA}(N, n, 2)$ for $N \geq 36$ and $n \geq 6$. Similarly, it is not feasible to enumerate all $\mathrm{OA}(N, n, 3)$ for $N \geq 64$ and $n \geq 8$. The first contribution of this paper is the introduction of a partial enumeration procedure for cases with $t \leq 3$ where a complete enumeration is not feasible and to introduce a simple method to establish upper bounds on the D-efficiency of arrays that have not been generated by the partial enumeration. Our partial enumeration of $\mathrm{OA}(32, 7, 2)$ took just 5 hours of computing time, produced all five D-optimal arrays and resulted in an upper bound of 0.8799 for D-efficiencies in arrays that were not generated.

## 1.2 Optimal designs

In the D-optimal approach, a design of $N$ runs and $n$ factors is sought that maximizes the D-efficiency of the interaction model (Atkinson et al., 2007). Because no combinatorial restrictions are imposed, the D-efficiency of a D-optimal design for this model will generally be higher than the D-efficiency of the best OAs under this model. However, such a D-optimal design does not support effect hierarchy. For example, we generated a D-optimal design for the phantom experiment with average standard errors for main effects and two-factor interactions of 0.1644 and 0.1652, respectively. The second contribution of this paper is the development of an optimal design procedure that favors the main effect estimation. For the phantom example, we created a design with average standard errors for main effects and two-factor interactions of 0.1600 and 0.1905, respectively. In this case, there is only a small improvement in main effect precision and a more substantial loss of precision in two-factor interactions. For many other cases, however, the improvement in main effect precision is much more substantial.

## 1.3 Organization

The rest of this paper is organized as follows. In Section 2, we return to the motivating example in more detail. We consider four different candidate designs and introduce design measures to characterize the designs. In Section 3, we introduce the enumeration algorithm for OAs and the optimal design algorithm for D-efficient designs that favor main effect estimation. In Section 4, we detail the numbers and best efficiencies of the generated OAs, give upper bounds for those that might have been obtained by enumeration of the complete set, and contrast these with efficiencies obtained by optimal design algorithms. Next, we study in detail the statistical

properties of the best designs for up to 72 runs and up to 10 factors and compare these with the best literature designs known to us. Finally, there is a brief discussion of the strengths and weaknesses of our approach in Section 6. Software to generate orthogonal arrays and optimal designs is provided in supplementary materials.

# 2 OPTIMALITY MEASURES AND CANDIDATE DESIGNS

In this section, we introduce four optimality measures for designs that fit the interaction model. We illustrate these measures with four candidate designs for the phantom experiment.

## 2.1 Optimality measures

The interaction model based on a two-level design $A$ can be stated formally as $y = \mathbf{X}\beta + e$, where $y$ is an $N \times 1$ vector of responses and $\mathbf{X}$ an $N \times p$ model matrix with an intercept, $n$ main effect contrast vectors and $n(n-1)/2$ two-factor interaction contrast vectors. Finally, $\beta$ is the $p \times 1$ vector of the factorial effects and $e$ is an $N \times 1$ vector of random errors with expectation zero and variance $\sigma^2$.

The parameters of the model can be estimated with the OLS estimator $b$ of $\beta$ with $b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$; its covariance matrix is $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$. Therefore, the capability of a design to return precise estimates of the factorial effects is maximized if $\mathbf{X}^T\mathbf{X}$ is maximized in some sense. One meaningful way is the maximization of $|\mathbf{X}^T\mathbf{X}|$, because maximizing this determinant minimizes the volume of a joint confidence region of the parameters under normal distribution of the random error $e$ (Atkinson et al., 2007).

For convenience, the determinant $|\mathbf{X}^T\mathbf{X}|$ is scaled by the number of parameters in the model and the run size. The scaled version of the determinant is designated $\mathrm{D}(A)$, with $\mathrm{D}(A) = |\mathbf{X}^T\mathbf{X}/N|^{1/p}$. We call $\mathrm{D}(A)$ the D-efficiency of $A$, thereby omitting the reference to the interaction model. It is well known that $0 \leq \mathrm{D}(A) \leq 1$. $\mathrm{D}(A) = 0$ if and only if the columns of $\mathbf{X}$ are linearly dependent, while orthogonal columns of $\mathbf{X}$ give a D-efficiency of 1.

To address the joint precision of the main effects in the interaction model, we slightly modify a criterion used by Schoen (2010) based on the concept of $\mathrm{D}_s$-optimality (Atkinson et al., 2007). For this purpose, we divide the parameter vector $\beta$ in a vector $\beta_1$ with main effect coefficients and a vector $\beta_{02}$ with the coefficients for intercept and the two-factor interactions. The model matrix $\mathbf{X}$ is split in an analogous way into $\mathbf{X_1}$ and $\mathbf{X_{02}}$ so that $y = \mathbf{X_1}\beta_1 + \mathbf{X_{02}}\beta_{02} + e$. A $\mathrm{D}_s$ optimal design maximizes

$$\mathcal{D}_s = |\mathbf{X}^T\mathbf{X}|/|\mathbf{X_{02}}^T\mathbf{X_{02}}|, \tag{1}$$

assuming that $\mathbf{X}_{02}$ is of maximum rank. It is easy to show that

$$|\mathbf{X}^T\mathbf{X}|/|\mathbf{X_{02}}^T\mathbf{X_{02}}| = |\mathbf{X}_1^T(I - \mathbf{X_{02}}(\mathbf{X_{02}}^T\mathbf{X_{02}})^{-1}\mathbf{X_{02}}^T)\mathbf{X_1}|. \quad (2)$$

The right hand side of (2) is the determinant of the residual sums of squares and products matrix after regressing the main effects collected in $\mathbf{X_1}$ on the intercept and two-factor interactions collected in $\mathbf{X_{02}}$. The scaled version of the determinant is designated $\mathrm{D}_s(A)$, with $\mathrm{D}_s(A) = \mathrm{D}_s^{1/n}$. In the rest of the paper, we call $\mathrm{D}_s(A)$ the $\mathrm{D}_s$-efficiency of $A$.

If $\mathbf{X_{02}}$ is indeed of maximum rank, $0 \leq \mathrm{D}_s(A) \leq 1$. $\mathrm{D}_s(A) = 0$ if the columns of $\mathbf{X}$ are linearly dependent, while $\mathrm{D}_s(A) = 1$ if the main effect columns of $\mathbf{X_1}$ are orthogonal to each other and also orthogonal to the intercept and two-factor interaction columns in $\mathbf{X_{02}}$.

It might seem unusual to maximize a determinant for main effect contrast vectors after accounting for two-factor interaction contrast vectors, as carried out in (2), because this reverses the roles of main effects and two-factor interactions. Indeed, interactions are defined as the part of the joint effect of factors left over when main effects are accounted for. We nevertheless think that the $\mathrm{D}_s$ criterion is useful as a design selection criterion in case main effects are of primary interest. If we fit the interaction model, the main effects are unbiased by possible two-factor interactions, and $\mathrm{D}_s$ captures the estimation efficiency of the unbiased main effects. Fitting a model where interactions involving one or more of the factors are dropped will improve the efficiency, but the main effect estimates risk bias from the omitted interactions.

Finally, our third and fourth optimality criteria are $\mathrm{A}_1$-efficiency and $\mathrm{A}_2$-efficiency, which are average variances of main effects and interactions, respectively, scaled by $\sigma^2/N$.

## 2.2 Candidate designs for the motivating example

To illustrate the optimality criteria outlined in the previous section, we introduce four candidate designs for the phantom case with 40 runs and 7 factors. (The designs are available in the supplementary materials to this paper.)

1. Using a complete set of OA$(40, 7, 3)$, we establish that the OA reported by Schoen and Mee (2012) is the only strength-3 OA of this size capable of fitting the interaction model. Its D-efficiency equals 0.8030.

2. Using a procedure that is discussed later in the paper, we generated 300 D-efficient OAs of strength 2 and include the most D-efficient array as a candidate design.

3. Using a coordinate exchange algorithm, we generated a D-optimal design.

4. Using another procedure discussed later in the paper, we generated designs with $\mathrm{D} + 2\mathrm{D}_s$ as optimality criterion. We include the best design according to this criterion as the fourth candidate; the design is designated compromise design.

An overview of the various efficiency measures for the candidate designs is given in Table 1.

OAs of strength 3 are $D_s$-optimal and $A_1$-optimal. This follows from equation (2), because $\mathbf{X_{02}}^T\mathbf{X_1} = \mathbf{0}$. The high $D_s$-efficiency and $A_1$-efficiency for the compromise design show that this design has near orthogonality of the main effects with respect to each other and to the interactions.

The D-efficiency of the strength-3 candidate is substantially smaller than the D-efficiency of the D-optimal design, while the $D_s$-efficiency of the D-optimal design is worse than the $D_s$-efficiency of the strength-3 design. The compromise design is indeed a compromise as it has improved D- and $A_2$-efficiencies when compared to the strength-3 design and improved $D_s$- and $A_1$-efficiencies when compared tot the D-optimal design.

The strength-2 candidate design cannot be recommended, because the D-optimal design is better in all efficiency measures considered here.

To illustrate the connection between the various optimality criteria and the precision of individual main effects and interactions, we present boxplots of the standard errors of the coefficients in Figure 1, assuming an error variance of 1. We consider two model classes. The first one is the single full interaction model in seven factors. The second class consists of the seven models where all interactions of one particular factor are dropped from the full interaction model. We call these models reduced models.

The upper panel of the figure shows the standard errors of the main effects. There are four pairs of boxes, one pair for each candidate design. Each broad box shows the seven standard errors for the full interaction model based on the respective designs. Each narrow box shows the 49 standard errors for all the main effects in the reduced models.

The minimum standard error is $1/\sqrt{40}$, equalling about 0.1581. All main effects of the strength-3 option have this minimum value. Three of the main effect standard errors of the compromise design also have this value. The four other standard errors all equal 0.1614, which is in between the main effects standard errors of the D-optimal design and the strength-3 design. The strength-2 design has the worst values of the main effect standard errors.

As expected, there is no change in main effect standard errors for the reduced models based on the strength-3 option. Remarkably, this is also the case for the compromise design. The reason is that the main effects in this design are orthogonal to the two-factor interactions. The

Table 1: Efficiencies of four candidate designs for the motivating example.

| Design | D-efficiency | $D_s$-efficiency | $A_1$-efficiency | $A_2$-efficiency |
|---|---|---|---|---|
| strength 3 | 0.8030 | 1 | 1 | 0.4483 |
| strength 2 | 0.9245 | 0.8495 | 0.8483 | 0.8483 |
| D-optimal | 0.9534 | 0.9343 | 0.9248 | 0.9157 |
| compromise | 0.8875 | 0.9884 | 0.9767 | 0.6860 |

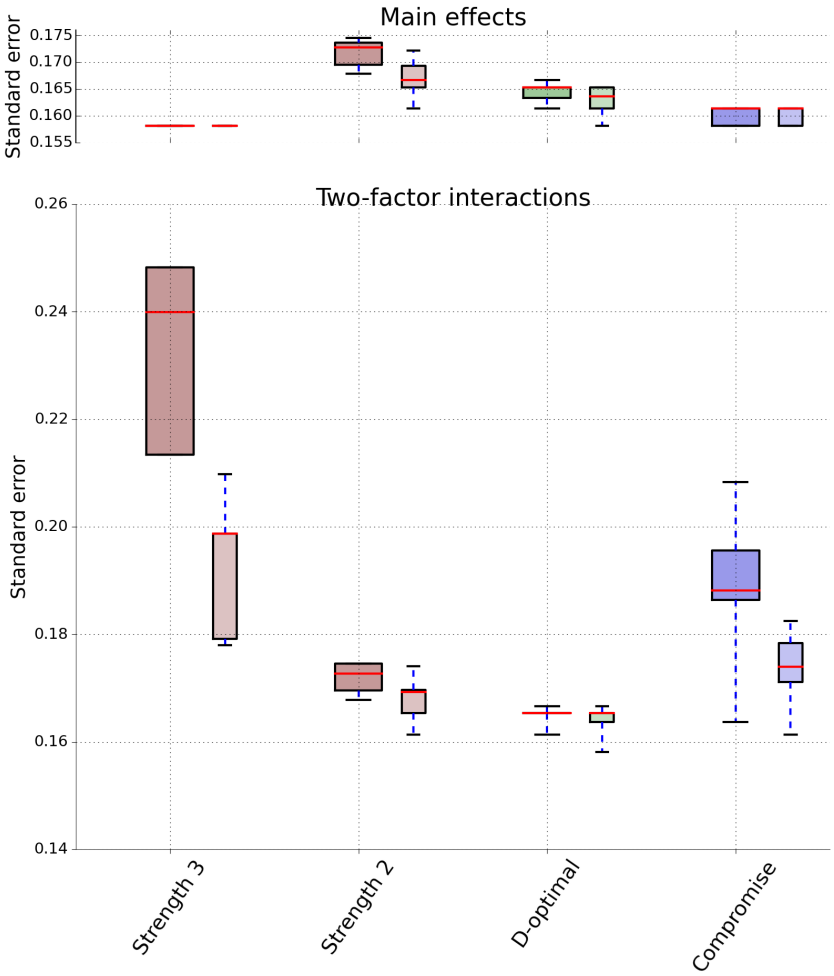


Figure 1: Standard errors in interaction models based on four designs of 40 runs and 7 factors. Broad boxes: full interaction model; narrow boxes: seven models lacking interactions of one of the factors.

standard errors for the strength-2 option are considerably improved, while those of the D-optimal option are also improved. The ranking of the four options regarding the main effect standard errors does not change, however.

The lower panel in the figure shows standard errors of the 21 interaction coefficients in the interaction model (broad boxplots) and of the 105 coefficients in the seven reduced models (narrow boxplots). The D-optimal design is clearly superior here, while the compromise design is intermediate between the D-optimal design and the strength-3 design. The standard errors in the reduced models of the strength-3 and compromise designs are considerably lower than those in the full interaction model. This shows the value of these designs in case only a subset of the interactions is active.

The design actually used for the phantom experiment was the strength-3 option. Statistical analysis of the results (not shown) revealed that, depending on the response variable, there were two or three substantial main effects sized 2.5 or more times the estimated standard deviation of

individual observations. The most substantial interactions (one or two per response) were roughly between 0.5 and 1 times this standard deviation. These findings show that effect hierarchy assumptions were in place here.

We would prefer the D-optimal design for cases when the focus is on the search for interactions among a limited number of factors known to be active. The strength-2 alternative has little to add, because it is outperformed by the three other options regarding the standard errors of the main effects and by the D-optimal design regarding those of the interactions. If the compromise design had been known in time, we might have recommended that design for the phantom experiment.

# 3 GENERATION OF DESIGNS

## 3.1 Orthogonal arrays

We want to generate OAs with good D-efficiency for the interaction model. Earlier work on D-efficient OAs (Tang and Zhou, 2013) is restricted to strength-2 OAs for the special case that the OA is embedded in a saturated OA with $N$ runs and $N-1$ factors, while there are only a few specified interactions of interest. In this paper, we consider the case that all interactions are of equal interest, while the OA need not be embedded in a saturated OA. To generate OAs, we slightly modified the algorithm of Schoen et al. (2010) (SEN). The complete source for the system is available on the world wide web (Eendebak, 2015) as well as in the supplementary materials. Here, we review the key elements of the original algorithm. Its goal is to obtain a set of all OA$(N, n, t)$. For any specific set of parameters $t$, $N$ and $n$, there may be many arrays. These can be partitioned in isomorphism classes. All arrays within one isomorphism class can be obtained from each other by a sequence of row permutations, column permutations or level permutations. These arrays are mathematically and statistically equivalent. Therefore, it suffices to study only one instance of every isomorphism class.

The algorithm of SEN features lexicographic ordering of arrays. An array $Q$ is lexicographically smaller than an array $R$ if there exists a column index $k > 0$ such that $Q_k < R_k$, whereas $Q_i = R_i, i = 1, \ldots, k-1$. Here, $Q_k < R_k$ if there exists a row index $m > 0$ such that $Q_{mk} < R_{mk}$ , while $Q_{jk} = R_{jk}, j = 1, \ldots, m-1$. So, reading column-wise, the first element for which $Q$ and $R$ differ has a smaller value in $Q$.

For any set of parameters of an OA, the algorithm produces a minimum complete set of arrays. This is a set with one unique representative array for each isomorphism class called the lexicographically minimal (LM) array.

**Definition 1.** An array is lexicographically minimal (LM) in its isomorphism class if no row, column, or level permutation results in a lexicographically smaller array.

To generate a minimum complete set of $\mathrm{OA}(N, n, t)$, the algorithm starts with a single array with $t$ columns in lexicographically minimal form, which is called the root array. This array is a representative for the single isomorphism class in $\mathrm{OA}(N, t, t)$. Two further steps turn a minimum complete set of arrays with $t \leq k \leq n-1$ columns into a minimum complete set with $k+1$ columns:

1. *Extension*: for each array in the minimum complete set with $k$ factors, a set of extensions with the required strength is generated that is guaranteed to contain all LM arrays that can be reached from the original array.

2. *LM check*: for each generated array, a test is performed to check whether the array is in LM form or not. The arrays not in LM form are rejected.

SEN show that a repeated application of the two steps results in a minimal complete set of $\mathrm{OA}(N, n, t)$. The arrays generated by these authors include all $\mathrm{OA}(N, n, 2)$ with $N \leq 28$ and $n \leq 6$ and all $\mathrm{OA}(N, n, 3)$ with $N \leq 48$. So, strength-2 alternatives to the well known $\mathrm{OA}(32, 6, 5)$ with run sizes up to 28 can be found by searching through the list of designs that they generated. Similarly, Schoen and Mee (2012) found strength-3 alternatives to $\mathrm{OA}(64, 7, 4)$, $\mathrm{OA}(64, 8, 4)$ and $\mathrm{OA}(128, 9, 4)$ with run sizes up to 48 by searching through the list of strength-3 designs.

To address strength-2 cases with $N \geq 32$ and strength-3 cases with $N \geq 56$, we restrict the extension of arrays in the minimal complete sets. First, we partition each set in arrays that permit fitting the interaction model and those that do not. We only extend the arrays of the first set. The minimal complete set with extended arrays is guaranteed to contain the D-optimal array.

If there are many arrays that permit estimation of the interaction model we applied a further restriction. We order the arrays according to their D-efficiencies and we extend only the best designs with an additional column. There is no guarantee that the set thus generated contains the D-optimal array. However, it is possible to establish upper bounds for the best possible D-efficiency of arrays that might have been generated based on the best efficiencies of the arrays that were not extended. These bounds are based on two theorems. The first one predicts what might happen if an array is extended with one extra column. The result is as follows:

**Theorem 1.** *Let $A$ be an orthogonal array with $N$ rows and $k$ columns that can fit the interaction model. Let $P = [A\, E]$ be an array that results from extending $A$ with a single column $E$. Let $p_k = 1 + k + k(k-1)/2$. Then $\mathrm{D}(P) \leq \mathrm{D}(A)^{p_k/p_{k+1}}$.*

The purpose of Theorem 2 is to sharpen the bounds established by Theorem 1 by taking previous extensions into account.

**Theorem 2.** *Let $A$ be an orthogonal array in LM form. Define $\mathcal{D}(A) = |\mathbf{X}^T\mathbf{X}/N|$. Let $P_i = [A\, E_i]$ be an LM orthogonal array that results from extending $A$ with a single column $E_i$. Let $Q$ be an array that results from extending $P_i$ with $q$ columns and let $L_i = \mathcal{D}(P_i)/\mathcal{D}(A)$. Then*

$$\mathcal{D}(Q) \leq L^q \mathcal{D}(P_i) \tag{3}$$

*with $L = \max_j L_j$. The maximum is over all possible LM extensions $P_j$ derived from $A$.*

Supplementary Section A includes the proofs of the theorems. An illustration of the way that the bounds work out is given in Supplementary Section B.

Cheng et al. (2002) established an approximate relation between the average D-efficiency of a model containing all the main effects and $g$ two-factor interactions, and the first two elements of the generalized word length pattern (GWLP; Tang and Deng, 1999). The GWLP of an OA is a vector $(A_3, A_4, \ldots, A_n)$, where $A_i$ is the sum of squared correlations between $i$-factor interaction contrast vectors and the intercept. In case $g = 0.5n(n-1)$, there is just one D-efficiency to consider; the relation for this case is $1/\mathrm{D} \propto A_3 + A_4$. In the supplementary Section C, we confirm that the best D-efficiencies indeed are found for designs with small $A_3 + A_4$. It is important to note that the relation between D and $A_3$ on its own is much weaker. So a minimum $G_2$-aberration design, which minimizes the elements of the GWLP from left to right, does not necessarily have the best D-efficiency.

## 3.2 Optimal designs

We implemented a coordinate exchange algorithm in Python and Matlab. The algorithm is slightly more complicated than the original algorithm of Meyer and Nachtsheim (1995). It optimizes $O = \alpha_1 \mathrm{D} + \alpha_2 \mathrm{D}_s$, where D and $\mathrm{D}_s$ are defined in Section 2.1. A specification of the algorithm is given in the supplementary Section D. The implementations are available in further supplementary materials.

For all cases where we generated D-efficient orthogonal arrays, we also generated D-optimal designs for the interaction model using the Python implementation with $\alpha_2 = 0$ and 5,000 initial tries. To generate compromise designs, we need to set the parameters $\alpha_1$ and $\alpha_2$ in our exchange algorithm. We set $\alpha_1 = 1$ and made seven-factor designs in 40 runs for $\alpha_2$ ranging from 0 up to 6 in steps of 0.2. We repeated the process with eight-factor designs in 80 runs. We found that the $\mathrm{D}_s$-efficiencies of the designs became stable for $\alpha_2 \geq 2$, while, for the 80 run designs, the D-efficiencies slightly decreased from that value onward. For these reasons, we used $\alpha_1 = 1$ and $\alpha_2 = 2$ to construct the compromise designs. See supplementary Section D for more details.

Our compromise designs are intended for situations where main effects are more likely to be important than two-factor interaction effects. They permit efficient estimation of the full interaction model, because the D-efficiency is included in the goal function to be optimized. At

the same time, they favor estimation of main effects independently from two-factor interaction by the inclusion of $D_s$ in the goal function. Therefore, they can be considered as model-robust designs. The most important difference between our approach and earlier approaches to model robust designs is that the run sizes we consider permit estimation of the full interaction model. Therefore, no special attention is needed to account for nonestimable models, such as in Du-Mouchel and Jones (1994), Li and Nachtsheim (2000) and Smucker et al. (2012), or aliasing between primary and potential model terms such as in Jones and Nachtsheim (2011).

# 4 GENERATED DESIGNS

The strength-2 OAs we generated have 6–8 factors and 32–44 runs. An OA to estimate the interaction model in 9 factors requires at least 48 runs. It was infeasible to do even a partial enumeration of strength-2 arrays with this run size or larger run sizes because of the very large numbers of nonisomorphic arrays (even for the extension of an array with a single column).

As regards OAs with a strength $t \geq 3$, there is a single $\text{OA}(32, 6, 5)$, which naturally permits estimation of the interaction model with maximum D-efficiency. For $n \geq 7$, strength-3 arrays for the interaction model only exist for run sizes $N \geq 40$. We generated D-efficient OAs of strength 3 in up to 10 factors requiring up to 72 runs.

Unlike OAs, D-optimal designs and compromise designs are not restricted to run sizes that equal multiples of 4 (strength-2 OAs) or 8 (strength-3 OAs). However, for direct comparisons with OAs, we generated optimal and compromise designs for run sizes $28 \leq N \leq 72$ equalling a multiple of 4.

## 4.1 Strength-2 arrays and alternative designs

Table 2 shows selected results for the strength-2 OAs and alternative designs. For the alternatives to OAs, we used our optimal design software with 5,000 initial tries and we kept the best design either according to D-efficiency or to the compound criterion $D + 2D_s$. The OA series with five factors were fully generated (results not shown). OA series with $n > 5$ factors were partially generated by extending only a small fraction of the designs with $n - 1$ factors based on their D-efficiency. To get an appreciation of the arrays that were missed in a partial generation, we studied for $N = 32$ fully generated as well as partially generated series of OAs; more details are given in the supplementary Section E. That section also shows comprehensive results on generated OAs with other run sizes and numbers of factors.

The first two columns of Table 2 give the run size $N$ and the number of factors $n$. The third column shows the type of design, which is either an OA, a D-optimal design or a compromise design. Then, we show the D-efficiency of the designs. For OAs we subsequently provide an upper bound on D-efficiencies for arrays that might have been obtained if the series were generated

Table 2: Strength-2 arrays and alternative designs

| $N$ | $n$ | Type | D | $B$ | $D_s$ | $A_1$ | $A_2$ |
|---|---|---|---|---|---|---|---|
| 32 | 6 | OA | 1 | 1 | 1 | 1 | 1 |
| | | D-optimal | 1 | | 1 | 1 | 1 |
| | | compromise | 1 | | 1 | 1 | 1 |
| 32 | 7 | OA | 0.8432 | 0.8432 | 0.8131 | 0.8004 | 0.6150 |
| | | D-optimal | 0.8868 | | 0.8325 | 0.8094 | 0.7600 |
| | | compromise | 0.8033 | | 0.9406 | 0.9188 | 0.4796 |
| 36 | 6 | OA | 0.9374 | 0.9374 | 0.8713 | 0.8696 | 0.8696 |
| | | D-optimal | 0.9773 | | 0.9659 | 0.9612 | 0.9589 |
| | | compromise | 0.9743 | | 0.9884 | 0.9778 | 0.9476 |
| 36 | 7 | OA | 0.9022 | 0.9389 | 0.8000 | 0.8000 | 0.8000 |
| | | D-optimal | 0.9369 | | 0.9506 | 0.9476 | 0.8545 |
| | | compromise | 0.8716 | | 0.9836 | 0.9699 | 0.6420 |
| 40 | 7 | OA | 0.9245 | 0.9414 | 0.8495 | 0.8483 | 0.8483 |
| | | D-optimal | 0.9534 | | 0.9343 | 0.9248 | 0.9157 |
| | | compromise | 0.8875 | | 0.9884 | 0.9767 | 0.6860 |
| 40 | 8 | OA | 0.8019 | 0.9516 | 0.6411 | 0.6019 | 0.5373 |
| | | D-optimal | 0.8517 | | 0.6967 | 0.6788 | 0.7236 |
| | | compromise | 0.7463 | | 0.9734 | 0.9503 | 0.3575 |
| 44 | 7 | OA | 0.9449 | 0.9531 | 0.8926 | 0.8864 | 0.8864 |
| | | D-optimal | 0.9563 | | 0.9381 | 0.9380 | 0.8953 |
| | | compromise | 0.9113 | | 0.9895 | 0.9792 | 0.7737 |
| 44 | 8 | OA | 0.8524 | 0.9567 | 0.7789 | 0.7721 | 0.6593 |
| | | D-optimal | 0.8800 | | 0.8010 | 0.7927 | 0.7691 |
| | | compromise | 0.8034 | | 0.9796 | 0.9596 | 0.4806 |

fully. This bound, designated $B$, was obtained using Theorem 1 and Theorem 2 of Section 3.1. The final three columns in Table 2 show the $D_s$-, $A_1$- and $A_2$-efficiencies obtained.

The designs generated with the optimal design software were all nearly orthogonal; the smallest efficiency for the main-effects only model equals 0.9761.

The results for 32 runs and six factors show that the best designs obtained with any of the three methods all have a D-efficiency equalling 1. They correspond to the unique $\mathrm{OA}(32, 6, 5)$. The fact that the optimal design software returns a design that is known to be best in terms of D-efficiency and $D_s$-efficiency shows that the implementation is sufficiently powerful to find efficient designs.

The tabulated results on OAs with $N \geq 36$ are based on partially generated series. Although the generation is only partial, the series of $\mathrm{OA}(36, 6, 2)$ includes an array with the best possible D-efficiency of the interaction model. The seven-factor series have discrepancies of at most 0.0367 between the D-efficiency in the best array obtained and the upper bound.

For the best $\mathrm{OA}(40, 8, 2)$ and the best $\mathrm{OA}(44, 8, 2)$, the upper bound $B$ is higher by 0.1487 and 0.1043, respectively. These are substantial discrepancies. However, in both cases, the D-optimal design generated has a higher D-efficiency by 0.0498 and 0.0276, respectively. This

suggests that the upper bound for these two instances is particularly weak. Further, we show in supplementary section F.1 that the best OAs actually obtained are competitive with the best strength-2 arrays known from the literature. We therefore did not search for better OAs.

As expected, the D-optimal designs have a better D-efficiency than the OAs. The largest discrepancy is the one stated above for $\mathrm{OA}(40, 8, 2)$. In general, orthogonal arrays of strength 2 do not seem to be particularly favorable to estimate all the interactions or to estimate main effects independently from interactions, because all efficiencies for the D-optimal options are better than the efficiencies of the corresponding OAs. One notable seven-factor exception will be discussed in Section 5.1.

The compromise designs generally have the smallest D-efficiency of the three types of design. The most substantial discrepancy with respect to the D-optimal design again occurs for the case of 40 runs and 8 factors; the difference in efficiencies is 0.1054.

The differences in $\mathrm{D}_s$-efficiencies between the compromise designs and the D-optimal designs can be substantial. Particular examples are the designs for 32 runs and 7 factors, 40 runs and 8 factors and 44 runs and 8 factors. The $\mathrm{D}_s$-efficiencies along with $\mathrm{A}_1$- and $\mathrm{A}_2$-efficiencies of these designs show that the D-optimal designs should be preferred if the goal of the experiment is to estimate interactions, while the compromise designs should be preferred if one wishes to estimate main effects independently from interactions.

## 4.2 Strength-3 arrays and alternative designs

Table 3 shows results for the strength-3 OAs and alternative designs. The table has the same layout as the one for the strength-2 OAs plus alternatives. For these alternatives, we used our optimal design software with 5,000 initial tries and we kept the best design either according to D-efficiency or to the compound criterion $\mathrm{D} + 2\mathrm{D}_s$.

For the 40-run and 48-run OAs we used a complete enumeration. For the 56-run arrays, we completely generated the series with eight-factor arrays. Nine-factor arrays were only generated by extension of the 10,491 eight-factor arrays that support an interaction model. The extension resulted in only 28 nine-factor arrays that permit fitting an interaction model in this number of factors. Further extension resulted in a single ten-factor array. However, it is not possible to fit an interaction model based on this array. For the 64-run arrays, we completely generated the series with six factors. We extended all 326 arrays that support an interaction model. From seven factors onward, we retained at most a few thousands of arrays that support an interaction model. For the 72-run arrays, we again started with generating all six-factor arrays. We extended all 872 arrays that support an interaction model. Extension of these arrays resulted in more than two million seven-factor arrays. From seven factors onward, we retained only a part of the arrays that support an interaction model. We obtained five nine-factor arrays and we failed to obtain a ten-factor array. Details on the numbers of generated OAs and cutoffs used can be found in

Table 3: Strength-3 arrays and alternative designs

| $N$ | $n$ | Type | D | $B$ | $D_s$ | $A_1$ | $A_2$ |
|---|---|---|---|---|---|---|---|
| 40 | 7 | OA | 0.8030 | 0.8030 | 1 | 1 | 0.4483 |
| | | D-optimal | 0.9534 | | 0.9343 | 0.9248 | 0.9157 |
| | | compromise | 0.8875 | | 0.9884 | 0.9767 | 0.6860 |
| 40 | 8 | OA | 0 | 0 | 1 | 0 | 0 |
| | | D-optimal | 0.8517 | | 0.6967 | 0.6788 | 0.7236 |
| | | compromise | 0.7463 | | 0.9734 | 0.9503 | 0.3575 |
| 48 | 7 | OA | 0.9585 | 0.9585 | 1 | 1 | 0.8750 |
| | | D-optimal | 0.9646 | | 0.9500 | 0.9459 | 0.9251 |
| | | compromise | 0.9585 | | 1 | 1 | 0.8750 |
| 48 | 8 | OA | 0.8365 | 0.8365 | 1 | 1 | 0.5973 |
| | | D-optimal | 0.9053 | | 0.8222 | 0.8099 | 0.8034 |
| | | compromise | 0.8450 | | 0.9859 | 0.9718 | 0.6043 |
| 48 | 9 | OA | 0.6753 | 0.6753 | 1 | 1 | 0.1439 |
| | | D-optimal | 0.7951 | | 0.5875 | 0.5574 | 0.5861 |
| | | compromise | 0.7250 | | 0.8759 | 0.8564 | 0.3440 |
| 56 | 7 | OA | 0.9192 | 0.0192 | 1 | 1 | 0.7826 |
| | | D-optimal | 0.9757 | | 0.9626 | 0.9609 | 0.9504 |
| | | compromise | 0.9585 | | 0.9912 | 0.9825 | 0.8903 |
| 56 | 8 | OA | 0.8642 | 0.8642 | 1 | 1 | 0.6478 |
| | | D-optimal | 0.9547 | | 0.9114 | 0.8992 | 0.9034 |
| | | compromise | 0.9040 | | 0.9903 | 0.9806 | 0.7396 |
| 56 | 9 | OA | 0.7610 | 0.7610 | 1 | 1 | 0.4522 |
| | | D-optimal | 0.8723 | | 0.7746 | 0.7600 | 0.7271 |
| | | compromise | 0.8067 | | 0.9256 | 0.9139 | 0.5363 |
| 64 | 8 | OA | 1 | 1 | 1 | 1 | 1 |
| | | D-optimal | 1 | | 1 | 1 | 1 |
| | | compromise | 0.9780 | | 1 | 1 | 0.9393 |
| 64 | 9 | OA | 0.9254 | 0.9626 | 1 | 1 | 0.8070 |
| | | D-optimal | 0.9190 | | 0.8097 | 0.8020 | 0.8410 |
| | | compromise | 0.8831 | | 0.9782 | 0.9681 | 0.7006 |
| 64 | 10 | OA | 0.8247 | 0.9692 | 1 | 1 | 0.5559 |
| | | D-optimal | 0.8371 | | 0.7074 | 0.6850 | 0.6494 |
| | | compromise | 0.7604 | | 0.8864 | 0.8734 | 0.4290 |
| 72 | 8 | OA | 0.9283 | 0.9439 | 1 | 1 | 0.8160 |
| | | D-optimal | 0.9824 | | 0.9759 | 0.9712 | 0.9668 |
| | | compromise | 0.9730 | | 0.9926 | 0.9855 | 0.9404 |
| 72 | 9 | OA | 0.8818 | 0.9391 | 1 | 1 | 0.6926 |
| | | D-optimal | 0.9473 | | 0.8844 | 0.8807 | 0.8931 |
| | | compromise | 0.9117 | | 0.9655 | 0.9599 | 0.7733 |
| 72 | 10 | OA | 0 | 0.9369 | 1 | | |
| | | D-optimal | 0.8935 | | 0.7752 | 0.7619 | 0.7708 |
| | | compromise | 0.8180 | | 0.9330 | 0.9239 | 0.5516 |

supplementary Section E.

The designs generated with the optimal design software were all nearly orthogonal; all D-efficiencies for the main effects only model were larger than 0.95.

The seven-factor OA in 48 runs is markedly better than the 40-run OA used in the motivating example. We prefer this OA to the D-optimal design of the same run size because the OA's $D_s$-efficiency is better, while its D-efficiency is only slightly less. The compromise design happens to be isomorphic to the OA.

The table shows two 64-run designs for 8 factors with a D-efficiency of 1. So these are strength-4 OAs. The compromise design is a strength-3 OA with a D-efficiency near 1, so that it has almost a strength of 4.

For the nine-factor OAs in 64 runs, the best D-efficiency found is 0.0372 lower than the upper bound, while the discrepancy is 0.1445 for the ten-factor arrays. However, the D-efficiencies obtained are very near those of the D-optimal designs; one is slightly worse and one is even slightly better. In addition, as shown in the Supplementary Section F for nine factor OAs and in Section 5 for ten-factor OAs, those we did obtain are competitive with the best literature arrays. For these reasons, we did not intensify the search of good OAs for nine or ten factors.

For the 72-run OAs, the discrepancy between the best D-efficiency and the upper bound for eight and nine factor arrays is 0.0156 and 0.0573, respectively. Our failure to find a 72-run ten-factor array did not prompt us to extend more than the few hundred seven-factor or eight-factor arrays with this run size, because even a few arrays extra lead to millions of new extensions. We believe that 72 is the maximum run size for which our methodology can yield useful OAs. Note that the compromise design of 10 factors has good $D_s$ and $A_1$-efficiencies, while the D-optimal 10-factor design has a good D-efficiency. However, $A_2$-efficiencies are not particularly good.

A further comparison among the designs in Table 3 leads to the following conclusions.

- The D-optimal designs have $A_1$-efficiencies that are nearly the same as their $A_2$-efficiencies.
- The D-optimal designs have substantially better D- and $A_2$-efficiencies than OAs of the same run size and number of factors. Exceptions are the 64-run cases of 9 and 10 factors and the OA$(48, 7, 3)$.
- D-optimal designs generally have substantially worse $D_s$ and $A_1$-efficiencies than the compromise designs and the OAs.

# 5 STUDY OF SPECIFIC CASES

We studied the standard errors in the interaction models based on the most efficient designs in 7–10 factors that we found. We compared our OAs with the smallest orthogonal arrays from the literature. Many of the new OAs are smaller, or have a higher D-efficiency, or have smaller standard errors for the coefficients, than the literature OAs. The OAs were contrasted with D-optimal designs and compromise designs. In general, we prefer designs for which the maximum standard error, either of the main effects or of the interactions, is minimized over the competing

designs of the same run size. All the designs studied are available in the supplementary materials for this paper. Section F includes a comprehensive table with efficiency measures and a discussion of the eight-factor and nine-factor cases. Seven-factor and ten-factor cases are presented here.

## 5.1 Seven factors

The two earliest literature OAs that we were able to find for the interaction model in seven factors have 48 runs, a strength of 2, a D-efficiency of 0.9222 and $D_s$-efficiencies of 0.8187 and 0.8823, respectively; see Mee (2009, p. 291) and Addelman (1961) for their construction. Schoen and Mee (2012) recommended seven-factor arrays of strength 3 in 40 and 48 runs capable of fitting the interaction model. These correspond to the strength-3 arrays characterized in Table 3. The 48-run OA has a D-efficiency of 0.9585 and a maximum $D_s$-efficiency, while the best seven-factor arrays of strength 2 and 40 or 44 runs from Table 2 have D-efficiencies of 0.9245 and 0.9449, and $D_s$-efficiencies of 0.8495 and 0.8926, respectively. So these alternatives have a better D-efficiency and either a greater strength or a smaller run size than the two earliest literature arrays.

In the remainder of this section, we restrict attention to 32-run and 36-run designs. While their D-efficiencies are smaller than those of the earlier literature designs, their run size is also smaller. The 32-run OAs are the smallest possible orthogonal arrays that support the interaction model for seven factors. The 36-run designs allow 7 residual degrees of freedom to conduct $t$ tests, assuming that higher-order effects are negligible.

Five OA$(32, 7, 2)$ have the globally best D-efficiency for OAs of this size, which equals 0.8432. The D-efficiency for the best 36-run OA we found is 0.9022. Assuming an error variance of 1, we calculated standard errors of the coefficients in the interaction model for the five best 32-run OAs, the 36 run OA, the D-optimal designs and the compromise designs of the same run sizes. Figure 2 shows the results. The upper panel in the figure is a dotplot of the standard error of the main effects, while the lower panel shows boxplots for the interactions. In the online version of the paper, the results for OAs are in brown-red, those for the D-optimal designs in green and those for the compromise designs in purple.

For the 32-run designs, the best design to estimate main effects independently from two-factor interactions is the compromise design, because it minimizes the maximum standard errors for the main effects as well as the $A_1$-efficiency. In case many interactions are likely to be substantial, we recommend the D-optimal design. This design minimizes the maximum standard error for interactions when compared with the alternative designs, while the standard errors also have a smaller range. There is no clear reason to recommend a 32-run OA, as judged by these standard errors or those of the main effects.

The most remarkable feature of Figure 2 is the complete uniformity of standard errors based on the 36-run design. Indeed, standard errors for all the main effects and all the two-factor interactions equal 0.1860. As this value also minimizes the maximum standard error of the

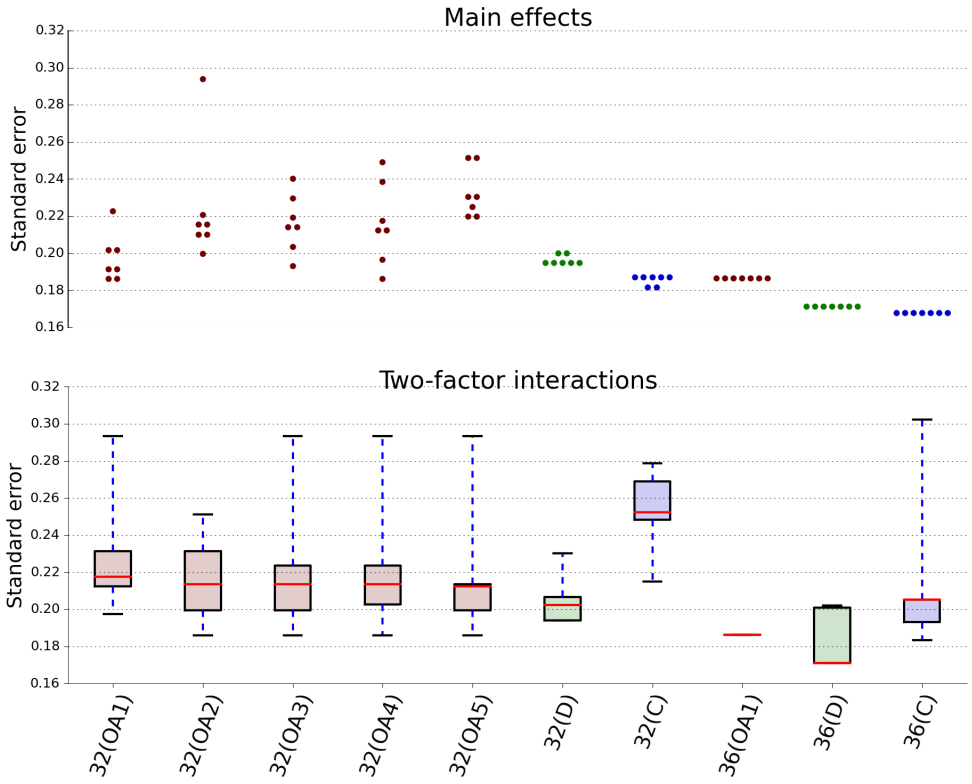


Figure 2: Standard errors for the coefficients of interaction models based on seven-factor designs of 32 or 36 runs. OA: orthogonal array of strength 2; D: D-optimal design; C: compromise design.

interactions, we would favor the OA in case the primary interest is in estimating interactions, even though the $A_2$-efficiency of the OA is 0.0545 less than for the D-optimal design (see Table 2). The compromise design has slightly smaller standard errors for the main effects than the D-optimal design, but the large standard errors of the interactions make us prefer the OA.

## 5.2 Ten factors

A ten-factor design that supports the interaction model must have at least 56 runs. Our work shows that there are no such strength-3 OAs. Our software returned D-optimal and compromise designs with D-efficiencies of 0.7492 and 0.6373 and $D_s$-efficiencies of 0.5275 and 0.8070, respectively. We think that these values are too low to recommend these designs. We obtained 60-run D-optimal and compromise designs with D-efficiencies of 0.7978 and 0.7100 and $D_s$-efficiencies of 0.6309 and 0.8627, respectively. However, it is usually prudent to include a few extra runs for model checking or estimation of random error. So it is natural to consider 64-run or 72-run designs.

The 64-run strength-3 OA given by Mee (2004), designated 64 (OA5), is the smallest and most D-efficient literature orthogonal design in 10 factors capable to fit the interaction model.

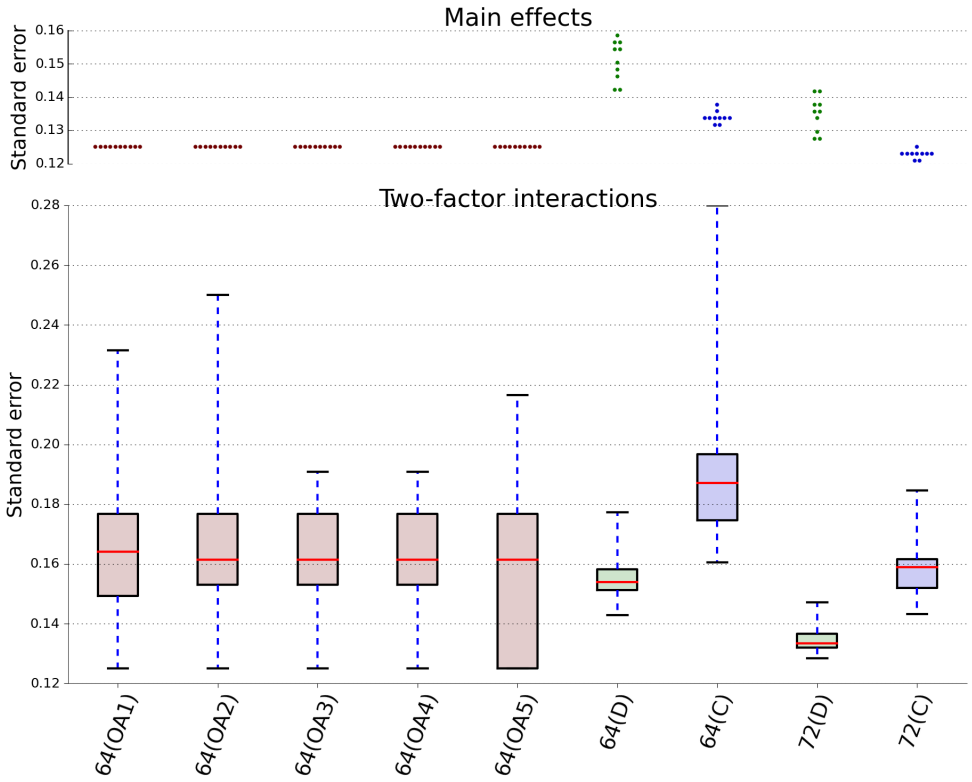


Figure 3: Standard errors in ten-factor designs of 64 and 72 runs. OA: orthogonal array of strength 3; D: D-optimal design; C: compromise design. D-efficiencies of OAs are 0.8247 (OA1) and 0.8238 (remaining OAs).

Its D-efficiency equals 0.8238. For reasons explained below, our enumeration did not include this design. However, we found three other designs with the same D-efficiency and one design with a slightly higher D-efficiency of 0.8247.

Figure 3 shows the standard errors of based on the five 64-run OAs, the D-optimal and compromise alternative designs, and 72 run D-optimal and compromise designs. A 72-run OA to fit the interaction model was not obtained; see Section 4.2.

The smallest standard errors for the main effects are reached for the compromise 72-run design, while the smallest standard errors for the interactions are reached for the 72-run D-optimal design. If budget allows, we prefer these designs to the 64-run alternatives.

For the 64-run cases, it is obvious that all OAs are equally good regarding the main effect standard errors, while the compromise design does better than the D-optimal design and worse than the OAs. As the compromise design has inferior standard errors for the interactions as well, it is not recommended. We prefer OA3 or OA4 if we want to estimate the main effects independently from the two-factor interactions, because the maximum standard error for the interactions is minimal in these OAs. OA3 and OA4 are nonisomorphic, but as it turns out, they

have the same frequency distribution of standard errors for the interactions.

The first seven columns of the lexicographically minimal version of design 64 (OA5) form a seven-factor design with a D-efficiency of 0.9310. In our enumeration, we did not extend 64-run seven-factor OAs with a D-efficiency of 0.9410 or less; see supplementary Section E. This explains why design 64 (OA5) was not included in our list of 10-factor designs. The finding illustrates that extension of two $k$-factor OAs $A_1$ and $A_2$, where $\mathrm{D}(A_1) < \mathrm{D}(A_2)$ can lead to extended designs $A_1^+$ and $A_2^+$ for which the order of D-efficiencies is reversed.

Finally, the set of standard errors for the interactions based on the D-optimal 64-run design is much more homogeneous and also has a smaller maximum than the corresponding sets for the OAs. As the D-efficiency of the D-optimal design is very similar to the best efficiency of a strength-3 OA, the standard errors of the main effects based on the D-optimal designs are considerably higher. We conclude that the optimal design is the preferred choice to assess two-factor interactions, while the OAs are preferred if we want to estimate the main effects independently from the two-factor interactions.

# 6 DISCUSSION

In this paper, we studied two-level experiments large enough to estimate a model with all the main effects and all the two-factor interactions. The assumption of effect hierarchy suggests that there are not so many two-factor interactions active, and that the size of active two-factor interactions is considerably smaller than the size of the main effects. Under these conditions, it makes sense to estimate main effects unbiased by possible two-factor interactions. We considered approaches based on orthogonal arrays and optimal designs.

Strength-3 OAs have maximum main effect precision irrespective of the number of interactions in the model. Therefore, the most D-efficient OA of strength 3 is an attractive option under effect hierarchy. If the priority of the experimenter is on detecting or estimating two-factor interactions, we would generally recommend D-optimal designs, because these have better precisions for these interactions.

Strength-2 OAs have maximum precision of the main effects only if no interactions are active. D-optimal designs for the interaction model were nearly orthogonal for the main-effect only model. In addition, these designs have a better precision for the interaction coefficients. For these reasons, we generally do not recommend a strength-2 OA to fit the interaction model, with one notable exception: we found an $\mathrm{OA}(36, 7, 2)$ for which all standard errors in the interaction model are equal. The D-optimal design we found has several standard errors for interactions that are higher than the standard error in the OA. Therefore, we recommend the OA for this case.

To attain a better $\mathrm{D}_s$-efficiency when using an optimal design approach, we implemented

a coordinate exchange procedure that optimizes $\alpha_1 \mathrm{D} + \alpha_2 \mathrm{D}_s$. The results presented here were obtained with the settings $\alpha_1 = 1$ and $\alpha_2 = 2$. The designs are a compromise between D-optimal designs and strength-3 OAs, both in terms of D-efficiency (D-optimal designs are generally better and strength-3 OAs worse) and $\mathrm{D}_s$-efficiency (D-optimal designs are substantially worse and strength-3 OAs are better).

As the $\mathrm{D}_s$-efficiency of the compromise designs is generally better than the $\mathrm{D}_s$-efficiency of strength-2 OAs, the compromise designs provide an attractive alternative to these OAs under effect hierarchy. Further, as the run size increases, it becomes increasingly difficult to obtain good strength-3 OAs, and compromise designs could be used instead. The compromise designs have the general advantage that they can be constructed for every run size that is compatible with fitting the interaction model. Interesting subjects for further research include the adaptation of this optimal design approach to fitting other models than the complete interaction model and optimization of this approach for larger cases than those studied here. Further, we might reverse the roles of main effects and two-factor interactions in our expression for $\mathrm{D}_s$. By replacing the original $\mathrm{D}_s$ with this modification in our criterion for compromise designs, we could prioritize interactions over main effects.

One problem in the generation of D-efficient OAs is the huge amount of different designs. The largest strength-2 case that we could handle completely is $\mathrm{OA}(32, 7, 2)$ with 530,469,996 different designs, while the largest strength-3 case is $\mathrm{OA}(48, 9, 3)$ with 166,081 different designs. This was the reason to develop a partial enumeration approach. We established upper bounds for the best possible D-efficiency of arrays that were not generated. Our approach resulted in smaller or more efficient alternatives to literature designs. We believe, however, that we reached the limits of its usefulness for the OAs with 10 factors and 64 or 72 runs.

Further interesting subjects for future research include exploration of multilevel designs either with our partial enumeration approach for orthogonal designs or with the compromise optimal design approach. For example, for four three-level factors (33 parameters in the interaction model), our methodology might yield a suitable 36-run design. For five factors (51 parameters in the interaction model), the nearest run size for an orthogonal design is 54. Sartono et al. (2012) showed that the four strength-3 designs of this size do not support the interaction model. Our methodology might give strength-2 designs or compromise designs of this run size that are capable of estimating this model. For six factors (73 parameters in the interaction model), there are strength-3 designs in 81 runs that support the interaction model (Sartono et al., 2012) so that our methods are less likely to be useful here.

## SUPPLEMENTARY MATERIAL

**Additional Results:** Proofs and applications of the theorems, relationship between D-efficiency and GWLP, details of the coordinate exchange algorithm, of the generation of OAs and of

specific eight-factor and nine-factor designs.

**Programs:** Software to generate OAs, D-optimal designs and compromise designs.

**Designs:** Designs for 7-10 factors studied in Section 5 and supplementary Section F.

# ACKNOWLEDGEMENTS

Comments by two referees and an associate editor led to a much broader scope and a more compelling presentation of results in the final version of the paper when compared with the initial submission. The research of the second author was supported by a grant from the Research Foundation - Flanders (FWO).

# References

Addelman, S. (1961). Irregular fractions of $2^n$ factorial experiments. *Technometrics*, 3:479–496.

Atkinson, A. C., Donev, A. N., and Tobias, R. D. (2007). *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

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

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.

Eendebak, P. T. (2015). The Orthogonal Array package. `http://www.pietereendebak.nl/oapackage/index.html`.

Hedayat, A., Sloane, N., and Stufken, J. (1999). *Orthogonal arrays : theory and applications*. Springer.

Jones, B. and Nachtsheim, C. J. (2011). Efficient designs with minimal aliasing. *Technometrics*, 53:62–71.

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.

Mee, R. W. (2004). Efficient two-level designs for estimating all main effects and two-factor interactions. *Journal of Quality Technology*, 36:400–412.

Mee, R. W. (2009). *A Comprehensive Guide to Factorial Two-Level Experimentation*. Springer-Verlag, New York.

Meyer, R. K. and Nachtsheim, C. J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, 37:60–69.

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

Sartono, B., Goos, P., and Schoen, E. D. (2012). Classification of three-level strength-3 arrays. *Journal of Statistical Planning and Inference*, 142:794–809.

Schoen, E. D. (2010). Optimum designs versus orthogonal arrays for main effects and two-factor interactions. *Journal of Quality Technology*, 42:197–208.

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.

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

Tang, B. and Zhou, J. (2013). D-optimal two-level orthogonal arrays for estimating main effects and some specified interactions. *Metrika*, 76:325–337.

Wu, C. F. J. and Hamada, M. (2000). *Experiments: Planning, Analysis and Parameter Design Optimization*. 1st edn, Wiley, New York.