Two-level orthogonal screening designs with 80, 96, and 112 runs, and up to 29 factors

Two-level orthogonal screening designs with 80, 96 and 112 runs, and up to 29 factors

Alan R. Vazquez[1], Eric D. Schoen[2], and Peter Goos[2]

[1]Department of Statistics, University of California, Los Angeles, U.S.A.
[2]Department of Biosystems, Faculty of Bioscience Engineering, KU Leuven, Belgium

June 2, 2021

Abstract

Due to recent advances in the development of laboratory equipment, large screening experiments can be conducted to study the joint impact of a few dozen factors. While much is known about orthogonal designs involving 64 and 128 runs, there is a lack of literature on screening designs with intermediate run sizes. In this article, we construct screening designs with 80, 96 and 112 runs which allow the main effects to be estimated independently from the two-factor interactions and limit the aliasing among the interactions. We motivate our work using a 14-factor tuberculosis inhibition experiment and compare our new designs with alternatives from the literature using simulations.

Keywords: Drug combination experiment, large run size, minimal aliasing, nonregular design, orthogonal array, strength three

1 Introduction

The first stages of product and process development involve screening experiments to identify, from a list of potentially influential factors, those that are indeed influential. Many screening experiments are conducted using two-level orthogonal designs. An attractive feature of these designs is that the two levels of every factor occur equally often and the factors' main effects are not aliased with each other. Two-level orthogonal designs are divided into two broad groups: regular and nonregular designs. In regular designs, all pairs of factors'

effects are either fully aliased or not aliased at all, while nonregular designs involve pairs of effects which are only partially aliased. Mee (2009) and Wu and Hamada (2009) provide comprehensive treatises on two-level orthogonal designs.

Due to recent advances in the miniaturization of laboratory equipment, orthogonal designs involving many factors and runs are being used to study complex biological and chemical processes, such as drug combination regimens. For instance, in order to develop a treatment that inhibits tuberculosis, Silva et al. (2016) studied the 14 drugs shown in Table 1. The overall goal of the study was to identify the concentrations of the drugs that maximize the percentage of inhibition of *Mycobacterium tuberculosis*—the causative agent of tuberculosis—in infected human cells.

The first stage of the study involved a screening experiment to identify the influential drugs. These drugs would subsequently be studied in later stages to find their optimal concentrations. Table 1 shows the factor levels used in the screening experiment, which correspond to the absence or presence at a given concentration of the drugs. Previous knowledge about the drugs suggested that many of them would have an active individual effect on the response. It also suggested the presence of active drug-drug interactions. For instance, the drug Rifampicin is known to have both synergistic and antagonistic interactions with some of the other drugs in Table 1. Therefore, all 14 main effects (MEs) and all 91 two-factor interactions (TFIs) of the 14 drugs were considered at the screening stage. In this article, we explore cost-efficient two-level designs for the tuberculosis inhibition experiment.

Two-level orthogonal designs (either regular or nonregular) may be generally classified by their so-called strength (Hedayat et al., 1999). In technical terms, a two-level orthogonal design has a strength of t if every subset of t factor columns is an equally replicated full 2^t factorial design, where t is an integer larger than or equal to two. Consequently, a two-level orthogonal design of strength t has a run size that is a multiple of 2^t. Classifying two-level regular designs in terms of their strength is equivalent to classifying them according to their well-known resolution (Wu and Hamada, 2009, ch 5), which is a Roman numeral equal to the smallest order of factors' interactions that are fully aliased with the intercept. In fact, the strength of a regular design equals its resolution minus one. In what follows, we categorize regular and nonregular designs in terms of their resolution and strength, respectively.

To study the effects of the 14 drugs, Silva et al. (2016) used a two-level nonregular design of strength 4 with 128 runs. Similar to regular resolution-V designs, nonregular

Table 1: Factors considered in the tuberculosis inhibition study and their concentrations for the screening experiment. The concentrations of the drugs are shown in micrograms per milliliter (μg/ml).

		Levels	
Label	Drug	-1	$+1$
X_1	Amoxicillin/clavulanate	0	0.7000
X_2	Clofazimine	0	0.0780
X_3	Cyclosorine	0	1.5000
X_4	Ethambutol	0	0.0500
X_5	Isoniazid	0	0.0030
X_6	Linezolid	0	0.0130
X_7	Moxifloxacin	0	0.0310
X_8	PA-824	0	0.0038
X_9	*para*-aminosalicylic acid	0	0.0700
X_{10}	Prothionamide	0	0.0110
X_{11}	Pyrazinamide	0	4.0000
X_{12}	Rifampicin	0	0.0008
X_{13}	SQ-109	0	0.1000
X_{14}	TMC-207	0	0.0060

strength-4 designs do not involve any aliasing among the MEs and the TFIs, and allow the estimation of all MEs and all TFIs with full precision. Therefore, they are excellent for screening MEs and TFIs simultaneously. For 12 to 15 factors, two-level nonregular strength-4 designs with 128 runs can be obtained using dedicated combinatorial methods available in Hedayat et al. (1999). These nonregular designs require only half as many runs as regular resolution-V designs for the same numbers of factors.

Silva et al. (2016) conducted the 14-factor 128-run nonregular strength-4 design in a completely randomized fashion. The number of drugs present in each run ranged from zero to 10, where the run without drugs served as a control. The total time required to prepare the 128 combinations of the 14 drugs was around seven hours, using the automatic liquid handling workstation called Microlab STAR Line. The statistical analysis of the results from the 128-run design resulted in eight significant MEs and six significant TFIs. Three of

the significant TFIs satisfied strong effect heredity (Wu and Hamada, 2009, ch 4), meaning that the MEs of the factors involved in these interactions were also significant. These three TFIs involved the drug Rifampicin, which is in line with the existing knowledge about the drugs. The other three significant TFIs, however, did not satisfy effect heredity. The fact that only 14 of the 105 potential effects of the drugs turned out to be significant suggests that, in hindsight, these effects might also have been detected using a more economical experimental design, with fewer runs than 128.

As it is undesirable that MEs are aliased with TFIs in screening experiments, we only consider alternative designs in which there is no such aliasing. Moreover, since the motivating experiment demands screening as many as 105 potential effects, we focus on designs with at least 64 runs so as to limit the chances of missing the active effects. One possible 64-run design to study 14 factors is a 64-run regular design of resolution IV; see, e.g., Wu and Hamada (2009, ch 5). In resolution-IV designs, the MEs are not aliased with the TFIs and can be estimated with full precision. However, these designs involve pairs of TFIs which are fully aliased. A major limitation of this aliasing structure is that additional experimental effort may be needed to disentangle or acquire insight into fully aliased interactions.

Two-level nonregular strength-3 designs are attractive alternatives to regular resolution-IV designs. This is because strength-3 designs do not involve any aliasing between the MEs and the TFIs, and, in contrast with resolution-IV designs, any two TFIs may be only partially aliased in these designs. The partial aliasing between two TFIs permits at least some information to be retrieved about each of the interactions. Another advantage of nonregular strength-3 designs over resolution-IV designs is their flexible run sizes, since strength-3 designs are available for run sizes which are multiples of eight. This is unlike regular resolution-IV designs which exist only for run sizes that are powers of two. For 14 factors, nonregular strength-3 designs have the potential to provide cost-efficient alternatives that fill the large gap between a 64-run regular resolution-IV design and the 128-run nonregular strength-4 design used by Silva et al. (2016). However, not much is known about nonregular strength-3 designs with run sizes between 64 and 128.

In this article, we construct nonregular strength-3 designs with 80, 96 and 112 runs to fill the gap between the run sizes 64 and 128. In Section 2, we review the existing strength-3 designs with run sizes ranging from 64 to 128. In Section 3, we introduce criteria to assess the quality of nonregular strength-3 designs. In Section 4, we outline an effective construction procedure to generate our nonregular strength-3 designs with 80, 96 and 112 runs. This construction procedure concatenates two smaller equally-sized strength-

3 designs. In Section 5, we present a collection of strength-3 designs with 80, 96 and 112 runs and 9 to 29 factors we obtained using that construction. The collection includes alternative design options for the tuberculosis inhibition experiment. We show that the designs we found outperform or are competitive with the benchmark strength-3 designs available in the literature in terms of the aliasing among the TFIs.

In Section 6, we revisit the tuberculosis inhibition experiment and propose alternative strength-3 design options. Using simulations, we compare these orthogonal design options with the 128-run design actually used, with each other and with nonorthogonal designs that could have been considered as well. We end the article in Section 7 with a discussion and suggestions for future research. A practical conclusion of our research is that, for situations involving many active MEs and up to a moderate number of large active TFIs, our 80-, 96 and 112-run strength-3 designs have the potential to provide the same or almost the same statistical power to identify the active effects as 128-run strength-4 designs. Our collection of strength-3 designs is available in the online supplementary materials accompanying this article.

2 Literature review

Similar to two-level regular resolution-IV designs, two-level nonregular strength-3 designs with N runs can accommodate up to $N/2$ factors. Complete catalogs of two-level strength-3 designs with 32, 40 and 48 runs are available from Schoen et al. (2010). For 32 runs, these catalogs include both regular resolution-IV and nonregular strength-3 designs with up to 16 factors. For 40 and 48 runs, the catalogs include nonregular strength-3 designs with up to 20 and 24 factors, respectively. For run sizes larger than 48, it is computationally infeasible to enumerate all strength-3 designs (Bulutoglu and Margot, 2008; Schoen et al., 2010). To overcome the lack of complete catalogs of large designs, several authors have presented partial collections of attractive strength-3 designs with run sizes from 64 to 128. Table 2 provides an overview of these collections and highlights the contributions of the present article in bold font. For completeness, the table also includes available collections of regular resolution-IV designs with 64 and 128 runs.

Chen et al. (1993) enumerated all regular resolution-IV designs with 64 runs and up to 32 factors. Using different enumeration approaches, Xu (2009) and Block and Mee (2005) identified the most attractive resolution-IV designs with 128 runs and up to 64 runs. Using so-called quaternary linear codes, Xu and Wong (2007) obtained nonregular strength-3

Table 2: Literature review on two-level regular and nonregular designs of strength 3 with run sizes from 64 to 128. The notation '$\leq k$' means that designs are available for up to k factors.

Runs	Reference	Factors	Technique
64	Chen et al. (1993)	≤ 32	complete enumeration of regular strength-3 designs
	Xu and Wong (2007)	≤ 32	designs from selected quaternary linear codes
	Cheng et al. (2008)	17	partial fold over of a 32-run strength-3 design
	Mee (2009, ch 7)	32	fold over of 32-run Paley Hadamard matrix
	Vazquez et al. (2019)	≤ 17	concatenation of two 32-run strength-3 designs
	Vazquez et al. (2019)	≤ 32	projections of the folded-over 32-run Paley Hadamard matrix
72	Box and Hunter (1961), Mee (2009, ch 7)	36	fold over of 36-run Plackett-Burman design
80	Box and Hunter (1961), Mee (2009, ch 7)	40	fold over of 40-run Plackett-Burman design
	Cheng et al. (2008)	21	partial fold over of a 40-run strength-3 design
	present work	$\mathbf{\leq 21}$	**concatenation of two 40-run strength-3 designs**
88	Box and Hunter (1961), Mee (2009, ch 7)	44	fold over of 44-run Plackett-Burman design
96	Box and Hunter (1961), Mee (2009, ch 7)	48	fold over of 48-run Plackett-Burman design
	Cheng et al. (2008)	25	partial fold over of a 48-run strength-3 design
	Vazquez and Xu (2019)	≤ 16	concatenation of three 32-run regular strength-3 designs
	present work	$\mathbf{\leq 25}$	**concatenation of two 48-run strength-3 designs**
104	Box and Hunter (1961)	52	fold over of 52-run Plackett-Burman design
112	Box and Hunter (1961)	56	fold over of 56-run Plackett-Burman design
	present work	$\mathbf{\leq 29}$	**concatenation of two 56-run strength-3 designs**
120	Box and Hunter (1961)	60	fold over of 60-run Plackett-Burman design
128	Block and Mee (2005), Xu (2009)	≤ 64	partial enumeration of regular strength-3 designs
	Xu and Wong (2007)	≤ 64	designs from selected quaternary linear codes
	Box and Hunter (1961)	64	fold over of 64-run Plackett-Burman design
	Vazquez et al. (2019)	≤ 33	concatenation of two 64-run strength-3 designs
	Vazquez and Xu (2019)	≤ 16	concatenation of four 32-run regular strength-3 designs

designs with 64 and 128 runs for up to 32 and 64 factors, respectively. Vazquez et al. (2019) provide an alternative collection of 64- and 128-run nonregular strength-3 designs for up to 17 and 33 factors, respectively. Their designs were constructed by concatenating two equally-sized strength-3 designs with 32 and 64 runs. Based on a similar idea, Vazquez and Xu (2019) constructed nonregular strength-3 designs with 96 and 128 runs for up to 16 factors by concatenating multiple copies of a regular resolution-IV design with 32 runs.

Regular resolution-IV designs can be constructed by folding over regular resolution-III designs. Similarly, nonregular strength-3 designs can be constructed by folding over nonregular strength-2 designs. Box and Hunter (1961) suggested to construct large nonregular strength-3 designs by folding over well-known nonregular strength-2 designs such

as Plackett-Burman designs (Plackett and Burman, 1946). More specifically, consider a $(N-1)$-factor N-run Plackett-Burman design with coded levels -1 and $+1$. A nonregular strength-3 design with $2N$ runs and N factors is obtained by augmenting this design with an extra column of $+1$s and folding over the resulting design. Plackett and Burman (1946) report designs with run sizes N a multiple of four and at most 100. Using their designs with 36 to 64 runs, we obtain a collection of nonregular strength-3 designs with 72-128 runs and 36-64 factors. The 64-run 32-factor strength-3 design generated from the 32-run Plackett-Burman design is regular, and so it is included in the enumeration of 64-run strength-3 regular designs of Chen et al. (1993).

Mee (2009, ch 7) provides another collection of strength-3 nonregular designs with 64-, 72-, 80-, 88- and 96-run strength-3 designs with 32, 36, 40, 44 and 48 factors, respectively. These strength-3 designs are constructed by folding over Hadamard matrices of orders 32, 36, 40, 44 and 48. In technical terms, a Hamard matrix of order N is an $N \times N$ orthogonal matrix with all its elements ± 1. In the literature of combinatorics, one of the most popular methods to construct Hadamard matrices is that of Paley (1933). Hadamard matrices constructed using this method are commonly refereed to as Paley Hadamard matrices. The 64-run 32-factor nonregular strength-3 design in Mee (2009, ch 7) is constructed by folding over the Paley Hadamard matrix of order 32. For 72, 80, 88 and 96 runs, the strength-3 designs in Mee (2009, ch 7) provide similar aliasing among TFIs as those derived by folding over Plackett-Burman desings with 36 to 48 runs. To save space, Table 2 shows the strength-3 designs obtained from Placket-Burman designs only.

Vazquez et al. (2019) provide alternative 64-run strength-3 designs with up to 32 factors. Their designs are obtained from attractive projections of the folded-over 32-run Paley Hadamard matrix onto subsets of factors. Cheng et al. (2008) provide one 64-run 17-factor strength-3 design, one 80-run 21-factor strength-3 design and one 96-run 25-factor strength-3 design. These nonregular designs were constructed by folding over one or more columns of a specific strength-3 design.

Table 2 shows that nonregular strength-3 designs with 72, 80, 88, 96, 104, 112 and 120 runs are scarce in the current literature on two-level orthogonal designs. In this article, we present 80-run designs with up to 21 factors, 96-run designs with up to 25 factors, and 112-run designs with up to 29 factors, and thereby fill an important gap in the literature.

3 Evaluation of strength-3 designs

Two-level nonregular strength-3 designs are commonly evaluated in terms of the extent to which the TFIs are aliased. It is thereby assumed that three-factor and higher-order interactions are negligible. In this section, we review the most commonly used criteria to assess the severity of the aliasing among the TFIs in strength-3 designs. First, we show how to measure and visualize this aliasing using the color map of absolute correlations. Next, we introduce the generalized resolution, the F_4 vector and the B_4 value (Deng and Tang, 1999; Tang and Deng, 1999), which summarize the most important correlations in the color maps of strength-3 designs. As our final criteria to measure the quality of our strength-3 designs, we use the largest number of estimable TFIs (Cheng et al., 2008) and, for strength-3 designs that can estimate all these effects, the D-efficiency criterion (Goos and Jones, 2011, ch 2), which we discuss at the end of this section.

3.1 Aliasing among two-factor interactions

To assess the severity of the aliasing among the TFIs, we start from the interaction model matrix. For a given m-factor strength-3 design in which the two levels of every factor are coded as -1 and $+1$, the interaction model matrix involves columns corresponding to the intercept, the m MEs and the $m(m-1)/2$ TFIs. We measure the extent to which two effects are aliased using the absolute correlation between the corresponding effects' columns of this matrix. An absolute correlation close to 1 implies that the two effects are strongly aliased, while an absolute correlation of 0 means that the effects are not aliased at all. Ideally, all absolute correlations between pairs of columns of the two-factor interaction matrix are 0, because this means there is no aliasing among all MEs and all TFIs.

As an example, we consider the nonregular strength-3 design with 32 runs and 10 factors in Table 3, obtained from Schoen and Mee (2012). The 10 factors are labeled $X_1, X_2, \ldots, X_{10}$, and their levels are coded as -1 and $+1$. It is easy to see that, for each factor column in Table 3, there are 16 -1s and 16 $+1$s. Therefore, the two levels of every factor occur equally often in the design.

A popular way to visualize the absolute correlations between columns corresponding to the MEs and the TFIs is a color map. Figure 1 shows the color map of the absolute correlations for the ME and TFI columns of the model matrix of the 10-factor 32-run strength-3 design in Table 3. In the color map, the largest absolute correlations are visualized by the darkest cells while zero correlations are indicated in white. The color map shows that

Table 3: A two-level nonregular strength-3 design with 32 runs and 10 factors from Schoen and Mee (2012).

Runs	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
2	−1	−1	−1	−1	−1	−1	1	1	1	1
3	−1	−1	−1	−1	1	1	−1	−1	1	1
4	−1	−1	−1	−1	1	1	1	1	−1	−1
5	−1	−1	1	1	−1	−1	−1	−1	−1	−1
6	−1	−1	1	1	−1	1	−1	1	1	1
7	−1	−1	1	1	1	−1	1	1	−1	1
8	−1	−1	1	1	1	1	1	−1	1	−1
9	−1	1	−1	1	−1	−1	1	−1	1	1
10	−1	1	−1	1	−1	1	1	1	−1	−1
11	−1	1	−1	1	1	−1	−1	1	1	−1
12	−1	1	−1	1	1	1	−1	−1	−1	1
13	−1	1	1	−1	−1	1	−1	1	1	−1
14	−1	1	1	−1	−1	1	1	−1	−1	1
15	−1	1	1	−1	1	−1	−1	1	−1	1
16	−1	1	1	−1	1	−1	1	−1	1	−1
17	1	−1	−1	1	−1	1	−1	1	−1	1
18	1	−1	−1	1	−1	1	1	−1	1	−1
19	1	−1	−1	1	1	−1	−1	1	1	−1
20	1	−1	−1	1	1	−1	1	−1	−1	1
21	1	−1	1	−1	−1	−1	1	1	1	−1
22	1	−1	1	−1	−1	1	1	−1	−1	1
23	1	−1	1	−1	1	−1	−1	−1	1	1
24	1	−1	1	−1	1	1	−1	1	−1	−1
25	1	1	−1	−1	−1	−1	−1	1	−1	1
26	1	1	−1	−1	−1	1	−1	−1	1	−1
27	1	1	−1	−1	1	−1	1	−1	−1	−1
28	1	1	−1	−1	1	1	1	1	1	1
29	1	1	1	1	−1	−1	−1	−1	1	1
30	1	1	1	1	−1	−1	1	1	−1	−1
31	1	1	1	1	1	1	−1	−1	−1	−1
32	1	1	1	1	1	1	1	1	1	1

the 32-run 10-factor design in Table 3 is indeed an orthogonal strength-3 design. This is because the off-diagonal cells corresponding to the absolute correlations between two ME columns, and between any ME and any TFI column, are white in the color map. So, there is no aliasing among the MEs and between the MEs and the TFIs.

In two-level orthogonal designs, pairs of TFI columns sharing a common factor are never correlated. The only pairs of TFI columns that can be correlated are those involving four different factors. For the 10-factor design in Table 3, there are 630 pairs of that type. Figure 1 shows that the largest absolute correlations for two TFI columns involving four factors equal 1. These are visualized by the six darkest off-diagonal cells, corresponding to three pairs of TFIs which are fully aliased. These pairs are X_1X_2 and X_3X_4, X_1X_3 and X_2X_4, and X_1X_4 and X_2X_3, each of which involves the factors X_1, X_2, X_3 and X_4. In general, TFI pairs involving the same four factors have the same correlation. For every set of four different factors, there are three associated pairs of TFIs.

Figure 1 shows 372 gray off-diagonal cells. They correspond to 186 pairs of TFI columns with an absolute correlation of 0.5. These 186 pairs can be divided into $186/3 = 62$ groups of three TFI pairs involving the same four factors. The fact that the 10-factor 32-run design has absolute correlations of 0.5 implies that certain effects are partially aliased and that the design is nonregular. The figure also shows that the remaining 441 pairs of TFIs involving four factors are not aliased at all as the corresponding cells in the color map are white.

For a two-level strength-3 design with N runs, Deng and Tang (1999) showed that the possible values for the absolute correlations between pairs of TFI columns involving four factors necessarily equal $1 - 16q/N$, where q is an integer ranging from zero to the largest integer smaller than or equal to $N/16$. For the 32-run strength-3 design in Table 3, the possible values for the absolute correlations between pairs of TFI columns involving four factors are therefore $1 - 0 \times 16/N = 1$, $1 - 1 \times 16/N = 0.5$ and $1 - 2 \times 16/N = 0$, since q is at most two when $N = 32$. This is why the off-diagonal cells in the color map in Figure 1 have only three different colors.

In general, when evaluating a strength-3 design based on the color map, we focus only on the off-diagonal cells corresponding to pairs of TFIs involving four factors, because the other off-diagonal cells—corresponding to pairs of MEs, pairs of a ME and a TFI, and pairs of TFIs involving three factors—are always white in the color map. Good strength-3 designs possess few colored off-diagonal cells corresponding to pairs of TFIs involving four factors, because this means that few pairs of TFIs are aliased. Moreover, the lighter the

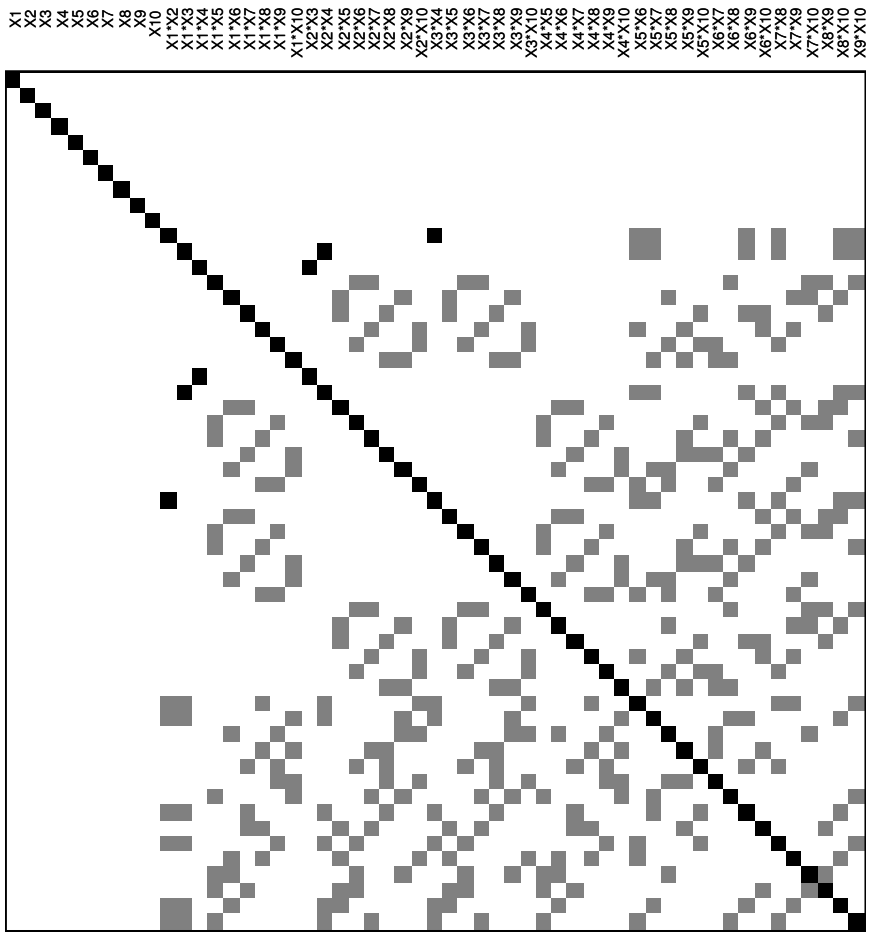


Figure 1: Color map showing the absolute correlations between the ME and TFI columns for the 10-factor 32-run strength-3 design in Table 3. The white, gray and dark cells correspond to absolute correlations equal to 0, 0.5 and 1, respectively.

colors of these cells, the better, because any existing aliasing is mild then. The generalized resolution, the F_4 vector and the B_4 value provide alternative summaries of the information contained in the color map about the aliasing between pairs of TFIs involving four factors.

3.2 Generalized resolution

The maximum absolute correlation between pairs of TFI columns for a given strength-3 design determines the largest extent to which two TFIs are aliased. It also determines the generalized resolution of the design. More specifically, for a strength-3 design, the generalized resolution is $5 - \rho_{\max}$, where $\rho_{\max}$ is the maximum absolute correlation between pairs of TFI columns. For instance, for the 10-factor 32-run strength-3 design in Table 3,

the maximum absolute correlation between pairs of TFI columns, $\rho_{\max}$, is 1, as witnessed by the darkest off-diagonal cells in the color map. So, the design has a generalized resolution equal to 4.

Ideally, the generalized resolution of a strength-3 design is large, since small absolute correlations between pairs of TFI columns are desirable. In other words, the lighter is the darkest off-diagonal cell corresponding to pairs of TFI columns in its color map, the larger the generalized resolution of a strength-3 design.

The generalized resolution criterion can be used to evaluate and compare two-level regular and nonregular designs. For two-level regular designs, the generalized resolution equals the resolution (Deng and Tang, 1999). This means that a two-level regular resolution-IV design has a generalized resolution of 4.

3.3 The $\mathbf{F_4}$ vector

The F_4 vector summarizes all the correlations between pairs of TFI columns involving four factors and shown in the color map in Figure 1. For strength-3 designs, the entries of the F_4 vector are the numbers of pairs of TFI columns involving four factors with an absolute correlation value of 1, $1 - 16/N$, $1 - 32/N$, and so on, divided by three. The division by three is due to the fact that the F_4 vector counts each set of three TFI pairs involving the same four factors only once, because they have the same correlation. For example, the F_4 vector of the 10-factor 32-run strength-3 design in Table 3 is $F_4(1, 0.5, 0) = (1, 62, 147)$. This means that there are $1 \times 3 = 3$, $62 \times 3 = 186$, and $147 \times 3 = 441$ pairs of TFI columns with absolute correlations of 1, 0.5 and 0, respectively, in the design. These pairs are visualized by the 3 dark, 186 gray and 441 white off-diagonal cells in the color map in Figure 1.

When comparing two strength-3 designs in terms of their F_4 vectors, we prefer the design whose value for the first entry in which the F_4 vectors differ is smaller, because this means that it involves less severe aliasing among the TFIs between the two designs (Deng and Tang, 1999). In other words, we prefer strength-3 designs that sequentially minimize the F_4 vector from left to right, because they have color maps involving the smallest number of dark off-diagonal cells corresponding to pairs of TFI columns, and, subject to this, the smallest number of gray off-diagonal cells, and so on.

The F_4 vector is a more refined criterion than the generalized resolution because it allows to detect differences in quality between designs of equal generalized resolution. Note

that sequentially minimizing the F_4 vector maximizes the generalized resolution.

3.4 The B_4 value

The B_4 value is a numerical summary of all the correlations between pairs of TFI columns involving four factors and thus of all entries in the F_4 vector. More specifically, the B_4 value of a strength-3 design equals the sum of the squared correlations between all pairs of TFI columns involving four factors divided by three. The divisor of three eliminates the redundancy due to TFI pairs involving the same four factors. The B_4 value of a strength-3 design can be calculated from its F_4 vector (in which the division by three has already taken place). For example, the 10-factor 32-run strength-3 design in Table 3 with $F_4(1, 0.5, 0) = (1, 62, 147)$ has a B_4 value equal to $1 \times (1)^2 + 62 \times (0.5)^2 + 147 \times (0)^2 = 16.5$. The B_4 value of 16.5 is a compact summary of all the dark and gray non-diagonal cells in the color map in Figure 1.

In general, a small B_4 value implies that the off-diagonal cells corresponding to pairs of TFI columns in the color map mainly show light colors. When comparing two strength-3 designs in terms of their B_4 values, we then prefer the design with the smaller B_4 value as it involves less overall aliasing among the TFIs (Tang and Deng, 1999).

As with the F_4 vector, the B_4 value is also a more refined criterion than the generalized resolution because it allows to detect differences in quality between designs of equal generalized resolution. The F_4 vector and the B_4 value can be used to evaluate two-level regular resolution-IV designs. For these designs, the first entry of the F_4 vector equals the B_4 value, while the rest of the entries are zero.

3.5 Degrees of freedom for estimating two-factor interactions

Similar to regular resolution-IV designs, nonregular strength-3 designs permit the estimation of the intercept and all the MEs, independently from the TFIs. The question is then, how many TFIs can we estimate with these designs? The rank of the matrix involving all TFI columns answers this question as it quantifies the number of degrees of freedom available for estimating TFIs. The more degrees of freedom, the more TFIs can be estimated.

For a two-level strength-3 design with N runs and m factors, the largest possible number of degrees of freedom for estimating TFIs is $N-m-1$. If a design does provide that specific number of degrees of freedom, it is referred to as a *second-order saturated* design (Cheng et al., 2008). Second-order saturated designs are attractive because all their experimental

runs provide information concerning the MEs and the TFIs.

The 10-factor 32-run strength-3 design in Table 3 provides 15 degrees of freedom for TFIs. Therefore, this design allows the estimation of the intercept, the 10 MEs and up to 15 TFIs. This design, however, is not second-order saturated because the number of degrees of freedom for TFIs is smaller than $N-m-1=21$. The fact that six degrees of freedom do not provide information concerning the TFIs means that they involve higher-order effects, which are generally assumed negligible.

3.6 D-efficiency to estimate the interaction model

If an m-factor strength-3 design permits the estimation of all $m(m-1)/2$ TFIs, we calculate its D-efficiency (Goos and Jones, 2011, ch 2) for estimating the interaction model. Let $\mathbf{X}$ denote the $N \times p$ interaction model matrix including a column of ones and the columns associated with the m MEs and the $m(m-1)/2$ TFIs, where $p = 1 + m + m(m-1)/2$. The D-efficiency for the interaction model then is $|\mathbf{X}^T\mathbf{X}|^{1/p}/N$, which ranges from zero to one. For a given run size N, a higher D-efficiency is preferred because this implies a higher precision for the estimates of the parameters in the interaction model. Strength-3 designs that do not permit the estimation of the interaction model have a D-efficiency of zero. This is the case for the 10-factor 32-run strength-3 design in Table 3.

4 Construction of large designs by concatenation

Given the availability of complete catalogs of strength-3 designs with up to 48 runs, one way to construct strength-3 designs with 80 and 96 runs is by concatenating two equally-sized strength-3 designs with 40 and 48 runs, respectively. Vazquez et al. (2019) introduced an effective column change/variable neighborhood search (CC/VNS) algorithm for this purpose. For any two equally-sized strength-3 designs, referred to as parent designs, the CC/VNS algorithm searches for the best concatenated design in terms of the F_4 vector or the B_4 value.

In this section, we briefly introduce the CC/VNS algorithm and outline the concatenation procedure. We also provide details on the 40- and 48-run parent designs we used to construct our strength-3 designs with 80 and 96 runs, respectively. Since, in this article, we also wish to explore strength-3 designs with 112 runs, we also need 56-run strength-3 parent designs. For lack of complete catalogs of 56-run strength-3 designs, we obtained

56-run parent designs by adding or removing columns from 56-run 27-factor strength-3 designs generated using the fold over technique. We discuss our 56-run parent designs at the end of this section.

4.1 The CC/VNS algorithm

The CC/VNS algorithm is a heuristic algorithm (Michalewicz and Fogel, 2004) to concatenate two strength-3 parent designs so as to sequentially minimize the F_4 vector or to minimize the B_4 value of the concatenated design. One of the parent designs is referred to as the upper parent design while the other as the lower parent design. The CC/VNS algorithm is composed of two interconnected algorithms: the column change (CC) algorithm and the variable neighborhood search (VNS) algorithm. Both algorithms perform column permutations and sign switches of the elements of one or more columns in the lower parent design, so as to improve the concatenated design.

4.1.1 Building block 1: CC algorithm

The CC algorithm performs systematic changes to each of the columns of the lower parent design, starting from the leftmost column and ending at the rightmost column. For each column in turn, the CC algorithm evaluates three changes: switching the signs of the elements of the column, swapping it with another column to its right in the lower parent design, and swapping it with the sign-switched version of that column. If one of these changes improves the concatenated design in terms of the F_4 vector or the B_4 value, then the improved concatenated design replaces the original and the algorithm continues its operations on the improved design. The CC algorithm repeats the whole process until no changes to the columns in the lower parent design improve the concatenated design.

4.1.2 Building block 2: VNS algorithm

When concatenating parent designs involving many factors, the CC algorithm may end up with a concatenated design which is suboptimal. The VNS algorithm attempts to overcome this weakness by systematically changing specific sets of columns in the lower parent design. There are four types of changes the VNS algorithm can make to the lower design:

1. Switch signs of the elements of any column in the lower design.

2. Swap any two columns in the lower design.

3. Switch signs of the elements of any two columns in the lower design.

4. Choose a subset of three columns in the lower design, move the first two columns one position to the right and move the third column to the first position.

The VNS algorithm performs the four types of changes consecutively, starting with the first type and ending with the fourth type. More specifically, the most complex types of changes are explored only when the simpler ones did not allow the concatenated design to be improved.

4.1.3 The full CC/VNS algorithm

The inputs to the CC/VNS algorithm are the upper and the lower parent designs. The algorithm begins by creating an initial concatenated design. The initial design results from executing the CC algorithm on a concatenated design generated by random permutations and sign switches of the columns in the lower parent design. Next, the algorithm improves the initial concatenated design using the following steps:

I. Set the index i to 1.

II. Modify the current best concatenated design using the VNS algorithm and the i-th type of change.

IIIa. Minimize the B_4 value or sequentially minimize the F_4 vector of the resulting concatenated design from Step II using the CC algorithm. If a better concatenated design is found, then go back to Step I with the improved concatenated design as the new current best design. Otherwise, go back to Step II.

IIIb. If all possible modifications to the current best concatenated design using the i-th type of change in Step II have been exhausted, then increase the value of index i by 1 and, if $i \leq 4$, go back to Step II.

IV. If $i > 4$, terminate the algorithm.

The CC/VNS algorithm terminates in the event no better concatenated design has been found after exploring all four types of changes of the VNS algorithm. The output of the algorithm is the best concatenated design found for the two parent designs selected. Finally, the whole CC/VNS algorithm can be repeated several times so as to increase the likelihood of finding the optimal concatenated design.

4.2 Concatenation procedure

To construct the large strength-3 designs, we used the concatenation procedure of Vazquez et al. (2019). The procedure involves the following steps:

Step 1. Select either the F_4 vector or the B_4 value as the criterion of interest for the concatenated design.

Step 2. Select two strength-3 parent designs with m factors and $N/2$ runs, which perform well in terms of the criterion selected in Step 1.

Step 3. Repeatedly apply the CC/VNS algorithm to the parent designs from Step 2. The output of the algorithm is the best m-factor N-run concatenated design among all repetitions. Call that design C.

Step 4. Add the column $\mathbf{z} = [\mathbf{1}_{N/2}^T, -\mathbf{1}_{N/2}^T]^T$ to the concatenated design C, where $\mathbf{1}_{N/2}$ is the $N/2 \times 1$ vector of ones. The resulting design is an N-run strength-3 design with $k = m + 1$ factors, which minimizes the criterion selected in Step 1.

According to Vazquez et al. (2019), concatenating the best strength-3 parent designs in terms of the B_4 value and the F_4 vector generally leads to the best concatenated designs in terms of these criteria. For this reason, to construct good large designs in terms of the F_4 vector or the B_4 value, we use attractive strength-3 parent designs in terms of the F_4 vector or B_4 value, respectively, in Step 2.

Vazquez et al. (2019) showed that 10 and 40 repetitions of the CC/VNS algorithm are generally enough to obtain high-quality concatenated designs in terms of the F_4 vector and the B_4 value, respectively. Note that optimizing the F_4 vector automatically maximizes the generalized resolution of the concatenated design.

In Step 4, the column $\mathbf{z}$ defines an extra factor that can be added to the concatenated design C. This extra factor has several attractive properties. For instance, its two levels occur equally often and its ME is not aliased with the MEs and TFIs of the original factors. Moreover, its m TFIs are neither aliased with any of the MEs nor with any of the TFIs of the other factors (Vazquez et al., 2019). Therefore, adding the extra factor implies that one additional ME and m additional TFIs can be estimated independently and with maximum precision.

4.3 Parent designs

4.3.1 40-run parent designs

A complete catalog of two-level strength-3 designs with 40 runs and 8 to 20 factors is available in Schoen et al. (2010). From this catalog, we considered the designs with minimum B_4 value as parent designs for the 80-run concatenated designs that optimize the B_4 value. For the 80-run concatenated designs that optimize the F_4 vector, we picked the parent designs from the top three 40-run strength-3 designs in terms of the F_4 vector. When there are multiple best designs, we used the degrees of freedom for TFIs as tie breaker. For the series of 40-run strength-3 designs including more than three best designs in terms of the F_4 vector and the degrees of freedom for TFIs, we randomly selected three of these designs as parents.

4.3.2 48-run parent designs

A complete catalog of two-level strength-3 designs with 48 runs and 8 to 24 factors is also available in Schoen et al. (2010). However, the number of different 48-run strength-3 designs is too large to consider all designs with minimum B_4 value or F_4 vector as parent designs for the 96-run concatenated designs. Instead, we used the 48-run strength-3 designs with up to 24 factors recommended by Schoen and Mee (2012) both for B_4 and F_4 optimization. These designs are best in terms of the B_4 value and the F_4 vector.

4.3.3 56-run parent designs

Using two different techniques to enumerate orthogonal designs, Bulutoglu and Margot (2008) and Schoen et al. (2010) tried to enumerate two-level 56-run strength-3 designs with up to 28 factors. They concluded that enumerating all 56-run strength-3 designs with more than eight factors is computationally infeasible, because there are too many designs. For lack of a complete catalog of 56-run strength-3 designs, we obtained the parent designs for our 112-run concatenated designs by folding over selected two-level nonregular strength-2 designs with 28 runs and 27 factors, and adding an extra column or removing one or more of their columns.

Our procedure to obtain the 56-run strength-3 parent designs with 8 to 28 factors is as follows:

1. We started with the complete series of 7570 two-level nonregular strength-2 designs

with 28 runs and 27 factors, obtained from Schoen et al. (2017). This series includes the 28-run 27-factor Plackett-Burman design.

2. We then generated all 56-run nonregular strength-3 designs with 27 factors using the fold over technique.

3. Next, we selected the 27-factor 56-run designs that were best and second best in terms of the F_4 vector. There was one overall best design in terms of the F_4 vector with a generalized resolution as large as 4.57 and eight second best designs all with a generalized resolution of 4.29. Therefore, these nine designs did not involve fully aliased TFIs. We include these 56-run 27-factor strength-3 designs in the online supplementary materials.

4. We obtained 56-run strength-3 designs with fewer than 27 factors by dropping columns from the nine best 27-factor 56-run designs. More specifically, we evaluated all projections of the nine best 27-factor 56-run designs onto $8 \leq m \leq 26$ factors. We used the best designs in terms of the B_4 value and the F_4 vector, among all projections of all nine 27-factor designs, as parent designs for the 112-run concatenated designs that optimize the B_4 value and the F_4 vector.

5. Finally, our 56-run 27-factor strength-3 parent design was the best design in terms of the F_4 vector. We obtained the 56-run 28-factor strength-3 parent design by appending the column $[\mathbf{1}_{28}^T, -\mathbf{1}_{28}^T]^T$ to this 27-factor 56-run design.

A detailed account of the best projections of the folded-over 56-run designs is included in supplementary Section A. To the best of our knowledge, these 56-run strength-3 designs are new to the literature.

5 A collection of strength-3 designs with 80, 96 and 112 runs

For each combination of number of runs and number of factors, we considered all pairs of parent designs, including pairs of the same designs. We then used the concatenation procedure described in Section 4.2 to generate concatenated designs from each pair of parent designs. Tables 4, 5 and 6 show the best designs we found with 80, 96 and 112 runs, respectively, in terms of the F_4 vector or the B_4 value. In nine cases, we found that our

best F_4-optimized designs were equally good or better than our best B_4-optimized designs, or vice versa, in terms of both the F_4 vector and the B_4 value. For these cases, we only include the overall best design in terms of both the B_4 value and the F_4 vector in the tables. Supplementary Section B shows the detailed construction of our concatenated designs from their parent designs.

In Tables 4, 5 and 6, the designs are labeled as k.b, k.f or k.bf, where k is the number of factors in the concatenated design, 'b' indicates designs that are best in terms of the B_4 value, 'f' indicates designs that are best in terms of the the F_4 vector, and 'bf' indicates designs that are best in terms of both the B_4 value and the F_4 vector. The tables report the generalized resolution (GR), the F_4 vector, the B_4 value, and the number of degrees of freedom for estimating TFIs of the designs. For notational simplicity, the tables omit the last entry of the F_4 vector, corresponding to pairs of TFI which are not correlated.

We compare our designs to the few benchmark designs in Table 2. As additional benchmark designs, we consider strength-3 designs obtained by dropping the last columns from folded-over Plackett-Burman designs with 40, 48 and 56 runs. This is because of the following two reasons: (1) folded-over Placket-Burman designs are well-known among practitioners, and (2) it is the easiest way to obtain designs with a smaller number of factors. The folded-over Plackett-Burman designs include the corresponding column $\mathbf{z}$ in the first position. A detailed analysis of the benchmark strength-3 designs derived from folded-over Plackett-Burman designs is included in supplementary Section C.

5.1 80-run designs with up to 21 factors

Table 4 shows the properties of our best 80-run strength-3 designs with 9 to 15 factors. For 9 to 11 factors, the designs have a generalized resolution of 4.8, whereas, for 12 and 13 factors, they have a generalized resolution of 4.6. For 14 factors or more, the best designs in terms of the B_4 value and the F_4 vector have a generalized resolution of 4.4 and 4.6, respectively. Therefore, none of the 80-run designs in Table 4 involve fully aliased TFIs.

For each number of factors, the best 80-run designs in terms of the B_4 value and the F_4 vector provide the same degrees of freedom for TFIs. The 80-run designs with 9 to 11 factors permit the estimation of all the TFIs and so, they can estimate the full interaction model. For this model, designs 9.bf, 10.bf and 11.bf in Table 4 provide a D-efficiency of 0.943, 0.891 and 0.802, respectively.

For 21 factors, the designs in Table 4 are second-order saturated, since they employ all

Table 4: Strength-3 designs with 80 runs. $F_4(1, 0.8) = (0, 0)$ for all designs. GR: generalized resolution; df: degrees of freedom for TFIs; ^{e}: design permits the estimation of all TFIs; ^{s}: design is second-order saturated.

Design	GR	$F_4(0.6, 0.4, 0.2)$	B_4	df	Design	GR	$F_4(0.6, 0.4, 0.2)$	B_4	df
9.bf^e	4.8	(0, 0, 18)	0.72	36	17.b	4.4	(1, 78, 886)	48.28	54
10.bf^e	4.8	(0, 0, 44)	1.76	45	17.f	4.6	(0, 67, 989)	50.28	54
11.bf^e	4.8	(0, 0, 82)	3.28	55	18.b	4.4	(2, 101, 1178)	64.00	55
12.bf	4.6	(0, 4, 164)	7.20	49	18.f	4.6	(0, 95, 1250)	65.20	55
13.b	4.6	(0, 24, 186)	11.28	50	19.b	4.4	(1, 148, 1481)	83.28	56
13.f	4.6	(0, 6, 264)	11.52	50	19.f	4.6	(0, 130, 1614)	85.36	56
14.b	4.4	(1, 25, 330)	17.56	51	20.b	4.4	(1, 186, 1907)	106.40	57
14.f	4.6	(0, 16, 415)	19.16	51	20.f	4.6	(0, 167, 2064)	109.28	57
15.b	4.4	(6, 16, 507)	25.00	52	21.b^s	4.4	(4, 242, 2320)	132.96	58
15.f	4.6	(0, 28, 563)	27.00	52	21.f^s	4.6	(0, 216, 2557)	136.84	58
16.b	4.4	(1, 55, 658)	35.48	53					
16.f	4.6	(0, 43, 757)	37.16	53					

their 80 degrees of freedom for estimating the intercept, the 21 MEs and up to 58 TFIs.

For 21 factors, Cheng et al. (2008) report a strength-3, second-order saturated design with 80 runs. This design has a generalized resolution of 4.4, $F_4(1, 0.8, 0.6, 0.4, 0.2, 0) = (0, 0, 125, 0, 2288, 3572)$, a B_4 value of 136.52 and 46 degrees of freedom for estimating TFIs. Both the designs 21.b and 21.f in Table 4 outperform the second-order saturated design of Cheng et al. (2008) in terms of the F_4 vector. Design 21.b also outperforms this benchmark design in terms of the B_4 value.

All 80-run designs in Table 4 outperform the strength-3 designs obtained by dropping the last columns from the folded-over 40-run Plackett-Burman design in terms of the F_4 vector, the B_4 value and the degrees of freedom for estimating TFIs. For 9 to 13 factors, our 80-run designs outperform these benchmark designs in terms of the generalized resolution too. The same is the case for the best 80-run F_4-optimized designs with 14 to 21 factors. However, the best 80-run B_4-optimized designs with 14 to 21 factors have the same generalized resolution as strength-3 designs obtained by dropping columns from the folded-over 40-run Plackett-Burman design; see supplementary Table S6.

After a close inspection of the structure of the 80-run designs, we found that some of them have replicate runs. More specifically, designs 10.bf and 11.bf in Table 4 have four duplicate runs each, while design 9.bf has two duplicate runs. The reason behind the presence of replicate runs in these 80-run designs is that their parent designs have replicate runs too. Replicate runs provide a pure error estimate of the error variance, which can be used in significance tests for the MEs and the TFIs. None of our other 80-, 96- and 112-run designs have replicate runs.

5.2 96-run designs with up to 25 factors

Table 5 shows the properties of our best 96-run strength-3 designs with 9 to 25 factors. For 9 and 10 factors, the designs have a generalized resolution of 4.83. For 11 factors, the best designs in terms of the B_4 value and the F_4 vector have a generalized resolution of 4.6 and 4.83, respectively. For 12 factors or more, the 96-run designs have a generalized resolution of 4.67, except for designs 15.b and 17.b which have a generalized resolution of 4 and 4.5, respectively. Therefore, with the exception of design 15.b, the 96-run designs in Table 5 provide pairs of TFIs which are only partially aliased.

For 9, 10 and 16 to 25 factors, the best 96-run designs in terms of the B_4 value and the F_4 vector provide the same degrees of freedom for TFIs. For the other numbers of factors, the best designs in terms of the F_4 vector provide more degrees of freedom for TFIs than the best designs in terms of the B_4 value. Four of the six 96-run designs with 9 to 12 factors permit the estimation of all TFIs, since the number of degrees of freedom for these effects equals the number of TFIs. For the 9- and 10-factor 96-run designs, the D-efficiency for the interaction model is 0.988 and 0.959, respectively. The D-efficiencies of designs 11.f and 12.f are 0.875 and 0.799, respectively.

Designs 14.f and 15.f perform extremely well in terms of the degrees of freedom for TFIs. Both designs provide as many as 80 degrees of freedom for TFIs, the largest number in the table. Table 5 further shows that designs 15.f, 25.f and 25.b are second-order saturated and thus employ all their 96 degrees of freedom for estimating the intercept, the MEs and the TFIs.

Regarding benchmark strength-3 designs, Vazquez and Xu (2019) provide 96-run strength-3 designs with 9 to 16 factors, while Cheng et al. (2008) provide one 96-run strength-3 design with 25 factors. The 96-run designs in Table 5 outperform the designs of Vazquez and Xu (2019) in terms of both the F_4 vector and the B_4 value. The designs in Table 5 also provide

Table 5: Strength-3 designs with 96 runs. $F_4(1, 0.83, 0.67) = (0, 0, 0)$ for all designs except 15.b. For that design $F_4(1, 0.83, 0.67) = (1, 0, 0)$. GR: generalized resolution; df: degrees of freedom for TFIs; ^{e}: design permits the estimation of all TFIs; ^{s}: design is second-order saturated.

Design	GR	$F_4(0.5, 0.33, 0.17)$	B_4	df	Design	GR	$F_4(0.5, 0.33, 0.17)$	B_4	df
9.bf^e	4.83	(0, 0, 6)	0.17	36	18.b	4.67	(0, 181, 1118)	51.17	63
10.bf^e	4.83	(0, 0, 24)	0.67	45	18.f	4.67	(0, 144, 1312)	52.44	63
11.b	4.67	(0, 10, 16)	1.56	54	19.b	4.67	(0, 223, 1506)	66.61	64
11.f^e	4.83	(0, 0, 70)	1.94	55	19.f	4.67	(0, 197, 1676)	68.44	64
12.b	4.67	(0, 18, 32)	2.89	62	20.b	4.67	(0, 300, 1876)	85.44	65
12.f^e	4.67	(0, 1, 126)	3.61	66	20.f	4.67	(0, 260, 2092)	87.00	65
13.b	4.67	(0, 14, 144)	5.56	67	21.b	4.67	(0, 397, 2286)	107.61	66
13.f	4.67	(0, 2, 222)	6.39	77	21.f	4.67	(0, 333, 2624)	109.89	66
14.b	4.67	(0, 21, 216)	8.33	69	22.b	4.67	(0, 477, 2918)	134.06	67
14.f	4.67	(0, 8, 330)	10.06	80	22.f	4.67	(0, 424, 3224)	136.67	67
15.b	4.00	(0, 108, 0)	13.00	74	23.b	4.67	(0, 608, 3500)	164.78	68
15.f^s	4.67	(0, 31, 440)	15.67	80	23.f	4.67	(0, 531, 3906)	167.50	68
16.b	4.67	(0, 100, 614)	28.17	61	24.b	4.67	(0, 695, 4434)	200.39	69
16.f	4.67	(0, 71, 762)	29.06	61	24.f	4.67	(0, 583, 4996)	203.56	69
17.b	4.50	(3, 120, 875)	38.39	62	25.b^s	4.67	(0, 861, 5240)	241.22	70
17.f	4.67	(0, 102, 1012)	39.44	62	25.f^s	4.67	(0, 708, 5984)	244.89	70

more degrees of freedom for TFIs than the designs of Vazquez and Xu (2019), except for 9 and 10 factors. For these numbers of factors, the degrees of freedom for TFIs of our 96-run designs and the designs of Vazquez and Xu (2019) are the same.

The 25-factor 96-run design of Cheng et al. (2008) has a generalized resolution of 4, $F_4(1, 0.83, 0.67, 0.5, 0.33, 0.17, 0) = (30, 0, 0, 0, 1940, 0, 10680)$, a B_4 value of 245.55 and 70 degrees of freedom for TFIs. Since the strength-3 design of Cheng et al. (2008) provides 70 degrees of freedom for TFIs, it can estimate up to 96 effects: an intercept, 25 MEs and 70 TFIs. So, this design is second-order saturated. Designs 25.b and 25.f in Table 5, which are also second-order saturated, outperform the design of Cheng et al. (2008) in terms of the generalized resolution, the F_4 vector and the B_4 value.

The 96-run strength-3 designs in Table 5 have the same generalized resolution as the

strength-3 designs obtained by dropping the last columns from the folded-over 48-run Plackett-Burman design, except for designs 9.bf, 10.bf, 11.f and 15.b and 17.b. Designs 9.bf, 10.bf and 11.f have a larger generalized resolution than the benchmark designs, while designs 15.b and 17.b have a smaller generalized resolution than these designs. In terms of the F_4 vector, the B_4 value and the degrees of freedom for estimating TFIs, our 96-run designs outperform the strength-3 designs derived from the folded-over 48-run Plackett-Burman design, but there are some exceptions. These exceptions are our best 96-run B_4-optimized designs with 15, 17, 21, 23 and 25 factors, which have a worse F_4 vector than the benchmark designs. However, the latter designs provide a larger B_4 value and a smaller number of degrees of freedom for estimating TFIs than the designs constructed by concatenation.

5.3 112-run designs with up to 29 factors

Table 6 shows the properties of our best 112-run strength-3 designs with 9 to 29 factors. For 9 and 10 factors, the designs in Table 6 have a generalized resolution of 4.86. For 11, 12 and 14 to 20 factors, the best designs in terms of the B_4 value and the F_4 vector have a generalized resolution of 4.57 and 4.71, respectively. For 13 factors, the best design in terms of both the F_4 vector and the B_4 value has a generalized resolution of 4.71. The 112-run designs for 21 factors or more have a generalized resolution of 4.57.

For 11 and 21 to 25 factors, the 96-run designs in Table 5 have a larger generalized resolution than the corresponding 112-run designs in Table 6. A partial explanation for this result is in the generalized resolution of the parent designs. For the 96-run designs, the 48-run parent designs with 10 and 20 to 24 factors all have a generalized resolution of 4.67; see Table 5 in Schoen and Mee (2012). For the 112-run designs, the 56-run parent designs with these numbers of factors have a generalized resolution of 4.57 or less; see Table S2 in the supplementary sections. Therefore, in terms of generalized resolution, the parent designs for the 112-run designs are inferior to those used to construct the 96-run designs. Similarly, for 9 to 15 factors, the 96-run designs in Table 5 have a smaller B_4 value than the 112-run designs in Table 6. This result can also be explained by the parent designs used, since the B_4 values of the 48-run parent designs with 8 to 14 factors are generally smaller than those of the 56-run parent designs. This can be seen from the B_4 values of the 48-run strength-3 designs in Table 5 in Schoen and Mee (2012) and the 56-run strength-3 designs in supplementary Table S2.

For each number of factors, the best 112-run designs in terms of the B_4 value and the

Table 6: Strength-3 designs with 112 runs. $F_4(1, 0.86, 0.71, 0.57) = (0, 0, 0, 0)$ for all designs. GR: generalized resolution; df: degrees of freedom for TFIs; ^{e}: design permits the estimation of all TFIs; ^{s}: design is second-order saturated.

Design	GR	$F_4(0.43, 0.29, 0.14)$	B_4	df	Design	GR	$F_4(0.43, 0.29, 0.14)$	B_4	df
9.bf^e	4.86	(0, 0, 18)	0.37	36	21.b	4.57	(21, 447, 2417)	89.67	74
10.bf^e	4.86	(0, 0, 60)	1.23	45	21.f	4.57	(1, 559, 2383)	94.45	74
11.b^e	4.57	(1, 0, 101)	2.25	55	22.b	4.57	(25, 578, 2930)	111.57	75
11.f^e	4.71	(0, 2, 112)	2.45	55	22.f	4.57	(3, 674, 3040)	117.61	75
12.b	4.57	(1, 10, 149)	4.04	65	23.b	4.57	(41, 716, 3506)	137.53	76
12.f	4.71	(0, 10, 174)	4.37	65	23.f	4.57	(8, 805, 3661)	141.89	76
13.bf	4.71	(0, 16, 250)	6.41	65	24.b	4.57	(58, 848, 4301)	167.65	77
14.b	4.57	(1, 49, 334)	11.00	67	24.f	4.57	(13, 1000, 4452)	174.88	77
14.f	4.71	(0, 43, 409)	11.86	67	25.b	4.57	(59, 1008, 5341)	202.12	78
15.b	4.57	(3, 69, 490)	16.18	68	25.f	4.57	(17, 1185, 5246)	206.92	78
15.f	4.71	(0, 72, 531)	16.71	68	26.b	4.57	(79, 1225, 6219)	241.43	79
16.b	4.57	(2, 114, 639)	22.71	69	26.f	4.57	(24, 1408, 6312)	248.16	79
16.f	4.71	(0, 112, 725)	23.94	69	27.b	4.57	(89, 1460, 7393)	286.41	80
17.b	4.57	(6, 155, 866)	31.43	70	27.f	4.57	(31, 1674, 7369)	292.74	80
17.f	4.71	(0, 161, 981)	33.16	70	28.b	4.57	(90, 1760, 8642)	336.57	81
18.b	4.57	(16, 197, 1138)	42.25	71	28.f	4.57	(40, 1987, 8658)	346.25	81
18.f	4.71	(0, 232, 1216)	43.76	71	29.b^s	4.57	(113, 2066, 10037)	394.25	82
19.b	4.57	(15, 261, 1519)	55.06	72	29.f^s	4.57	(52, 2296, 10175)	404.63	82
19.f	4.71	(0, 314, 1594)	58.16	72					
20.b	4.57	(25, 331, 1929)	70.98	73					
20.f	4.71	(0, 425, 1912)	73.71	73					

F_4 vector provide the same degrees of freedom for estimating TFIs. The 112-run designs with 9 to 11 factors permit the estimation of all TFIs. Designs 9.b, 10.bf, 11.b and 11.f in Table 6 provide a D-efficiency for the interaction model of 0.974, 0.924, 0.855 and 0.832, respectively. For 12 to 15 factors, the 112-run designs do not provide more degrees of freedom for TFIs than the 96-run designs in Table 5. Once again, this can be explained by the fact that the 48-run parent designs with 11 to 14 factors generally provide more degrees for freedom for TFIs than the 56-run parent designs; see Table 5 in Schoen and Mee (2012) and supplementary Table S2. Table 6 shows that the 29-factor 112-run designs are

second-order saturated, since they employ all their 112 degrees of freedom for estimating the intercept, the 29 MEs and up to 82 TFIs.

All 112-run strength-3 designs in Table 6 outperform the strength-3 designs obtained by dropping the last columns from the folded-over 56-run Plackett-Burman in terms of the F_4 vector, the B_4 value and the degrees of freedom for estimating TFIs. For 21 to 29 factors, our 112-run designs have the same generalized resolution as these benchmark designs. This is also the case for our best 112-run designs in terms of the B_4 value with 11 to 20 factors. The rest of the 112-run designs in Table 6 have a larger generalized resolution than strength-3 designs obtained by omitting columns from the folded-over 56-run Plackett-Burman; see supplementary Table S8.

6 Alternative designs for the tuberculosis inhibition experiment

The practical example that motivated this article is a tuberculosis (TB) inhibition experiment (Silva et al., 2016). In order to develop a treatment to increase the inhibition of TB, the absence and presence of the 14 drugs in Table 1 were studied. The goal of the experiment was to detect the active effects of the drugs, among their 14 MEs and their 91 TFIs, on the percentage of inhibition of *Mycobacterium tuberculosis* in infected human cells. The design actually used was a two-level strength-4 nonregular design with 14 factors and 128 runs, which provided full precision to estimate all the MEs and all the TFIs. The total time required to prepare the 128 drug combinations given by this design was around seven hours. This means that preparing a single combination of the 14 drugs requires, approximately, three minutes and 17 seconds.

The work we did in this article allowed us to find six alternative cost-efficient strength-3 designs with 80, 96 and 112 runs, for the TB inhibition experiment. We first introduce these alternatives together with several 64-run design options from the literature. Next, we discuss alternative nonorthogonal designs for the TB inhibition experiment. We end this section by comparing the orthogonal and nonorthogonal design options using a simulation study.

6.1 Strength-3 orthogonal design options

Table 7 shows several strength-3 design options with 64, 80, 96 and 112 runs for the TB inhibition experiment. For each design option, the table reports the maximum absolute correlation (which equals 5 minus the generalized resolution) and the sum of squared correlations between pairs of TFI columns (which equals three times the B_4 value), as well as the number of degrees of freedom available for estimating TFIs. The table also reports the number of pairs of TFI columns that possess the maximum absolute correlation. That number—which is shown in the fourth column of the table and corresponds to three times the first nonzero entry of the F_4 vector—should only be used to compare designs with the same maximum absolute correlation.

In Table 7, the 80-, 96- and 112-run designs are labeled N.b or N.f, where N is run size of the design, 'b' indicates designs that minimize the B_4 value and 'f' indicates designs that sequentially minimize the F_4 vector. The table also includes 64-run strength-3 designs from Xu and Wong (2007) and Vazquez et al. (2019). These designs are Pareto optimal among all the available 14-factor 64-run strength-3 designs when considering the F_4 vector, the B_4 value and the degrees of freedom for TFIs; see Vazquez et al. (2019) for details. Design 64.q in the table is obtained from the quaternary linear codes in Xu and Wong (2007). Designs 64.p and 64.f are obtained from Vazquez et al. (2019). Design 64.p is a specific projection of the folded-over 32-run Paley Hamadard matrix, while design 64.f is constructed by concatenating two 32-run strength-3 designs and sequentially minimizing the F_4 vector of the concatenated design.

Table 7 shows that design 64.p has a maximum absolute correlation between pairs of TFI columns involving 4 factors equal to 0.25, the smallest maximum absolute correlation among all design options. However, this design provides only 31 degrees of freedom for TFIs, the smallest number in Table 7. The second best design in terms of the maximum absolute correlation between pairs of TFI columns involving four factors is design 112.f with a maximum absolute correlation of 0.29, followed by the 96-run designs with maximum absolute correlations of 0.34. In terms of the sum of squared correlations between pairs of TFI columns, Table 7 shows that designs 96.b and 96.f are the best and second-best, respectively, among all the design options. More specifically, designs 96.b and 96.f have sum of squared correlation values of 24.99 and 30.18, respectively. The 96-run designs also provide the largest numbers of degrees of freedom for TFIs; design 96.b provides 69 degrees of freedom while design 96.f provides 80. Although 112 runs are, in principle, enough to estimate the intercept, 14 MEs and 91 TFIs, the 112-run designs do not allow

Table 7: Strength-3 design options for the 14-factor tuberculosis inhibition experiment. N.f: N-run concatenated design that minimizes the F_4 vector; N.b: N-run concatenated design that minimizes the B_4 value; $\rho_{\max}$: maximum absolute correlation between pairs of TFI columns; #Pairs($\rho_{\max}$): number of pairs of TFI columns with an absolute correlation of $\rho_{\max}$; SSC: sum of squared correlations between pairs of TFI columns; df: number of degrees of freedom for estimating TFIs. Design 64.q is from Xu and Wong (2007); design 64.p and 64.f are from Vazquez et al. (2019). The generalized resolution value is calculated as $5 - \rho_{\max}$, while the B_4 value is SSC/3.

Runs	Label	$\rho_{\max}$	#Pairs($\rho_{\max}$)	SSC	df
64	64.q	0.50	168	42.00	49
	64.p	0.25	1578	99.00	31
	64.f	0.50	72	72.75	43
80	80.b	0.60	3	52.68	51
	80.f	0.40	48	57.48	51
96	96.b	0.34	63	24.99	69
	96.f	0.34	24	30.18	80
112	112.b	0.43	3	33.00	67
	112.f	0.29	129	35.58	67

the estimation of all these effects simultaneously.

Overall, designs 96.b and 96.f are attractive alternatives for the TB inhibition experiment because of their good performance in terms of the maximum absolute correlation, sum of squared correlations and degrees of freedom for TFIs. Design 96.f has fewer pairs of TFI columns with the maximum absolute correlation, 0.34, and offers more degrees of freedom for TFIs than design 96.b. On the other hand, design 96.b has a smaller sum of squared correlations value than design 96.f. So, no design dominates the other in terms of all criteria. Given that preparing a single combination of the 14 drugs requires around three minutes and 17 seconds, the total time needed to prepare the drug combinations in the 96-run designs would have been, approximately, five hours and 15 minutes. This would have saved around one hour and 45 minutes of sample preparation using the Microlab STAR Line station, when compared to the 128-run strength-4 design used by Silva et al. (2016).

An attractive feature of the 14-factor 96-run designs (as well as of the 80- and 112-run designs) in Table 7 is that, by construction, they include a factor whose 13 TFIs with the other factors are neither aliased with any of the MEs nor with any of the other TFIs. Since previous knowledge suggested the presence of active TFIs involving the drug Rifampicin, a sensible advice would have been to assign this drug to that factor.

6.2 Nonorthogonal design options

Attractive alternatives for screening that may or may not belong to the class of two-level orthogonal designs are D-optimal designs (Goos and Jones, 2011, ch 2) and Bayesian D-optimal designs (DuMouchel and Jones, 1994). For a given number of runs, a D-optimal design maximizes the D-efficiency (defined in Section 3.5) for a specific model, for instance, the interaction model involving the 14 drugs. A higher D-efficiency implies a higher precision for the estimates of the parameters in the specified model. D-optimal designs are available for any number of runs larger than or equal to the number of model parameters. So, for the TB inhibition experiment, we can generate D-optimal designs with at least 106 runs, since the interaction model for the 14 drugs includes one intercept, 14 MEs and 91 TFIs.

Bayesian D-optimal designs are constructed by maximizing a Bayesian modification to the D-efficiency. When considering MEs and TFIs, these designs can be constructed to provide an efficient estimation of the intercept and all the MEs, while allowing for some detectability for TFIs. To construct a Bayesian D-optimal design we need to specify a tuning parameter called the prior variance. This tuning parameter defines a trade-off between the design's estimation efficiency for the intercept and all the MEs and its ability to detect TFIs. Larger values of the prior variance result in a Bayesian D-optimal design with a larger emphasis on the TFIs. For the TB inhibition experiment, Bayesian D-optimal designs are available with run sizes as small as 15, since these designs must only estimate the intercept and the 14 MEs of the drugs. Therefore, an attractive advantage of these designs when compared to D-optimal designs and strength-3 designs is their flexible run sizes.

Using the coordinate-exchange algorithm (Meyer and Nachtsheim, 1995), as implemented in the statistical software package JMP v14, we constructed D-optimal and Bayesian D-optimal designs with 14 factors and run sizes comparable to those of the strength-3 designs in Table 7. More specifically, we generated a D-optimal design with 112 runs and

Bayesian D-optimal designs with 64, 80 and 96 runs. We used 1,000 iterations for the coordinate-exchange algorithm and a prior variance equal to 1/16, the default parameter value in JMP v14 for generating Bayesian D-optimal designs (with the TFIs specified as "If Possible").

Table 8 shows the main properties of the D-optimal and Bayesian D-optimal designs for the TB inhibition experiment. In contrast with the strength-3 designs, the two levels of each factor in these designs do not occur equally often and the MEs are aliased with each other to some extent. So, the D-optimal and Bayesian D-optimal designs are not orthogonal. Table 8 shows that the D-optimal and Bayesian D-optimal designs also provide ME columns which are correlated with the TFI columns to some extent. This is not the case for the strength-3 designs in Table 7.

Regarding the TFIs, with one exception, the D-optimal and Bayesian D-optimal designs outperform the strength-3 designs in terms of the maximum absolute correlation, the sum of squared correlations and the degrees of freedom for TFIs. The exception is the 64-run Bayesian D-optimal design, which provides a larger sum of squared correlations than the strength-3 design 64.q in Table 7. Nevertheless, the 64-run Bayesian D-optimal design has the same number of degrees of freedom for TFIs and a smaller maximum absolute correlation between these effects than design 64.q. The reason behind the good performance of the D-optimal and Bayesian D-optimal designs for the TFIs is that, unlike the strength-3 designs, they allow some aliasing among the MEs and between the MEs and the TFIs.

Table 8: Nonorthogonal design options for the 14-factor tuberculosis inhibition experiment. N.bd: Bayesian D-optimal design with N runs; 112.d: D-optimal design with 112 runs for the interaction model; $\rho_{\max}$: maximum absolute correlation; SSC: sum of squared correlations; df: number of degrees of freedom for estimating TFIs. The D-efficiency for the interaction model of the 112-run D-optimal design is 0.747.

		Correlation among ME columns		Correlation between ME and TFI columns		Correlation among TFI columns		
Runs	Label	$\rho_{\max}$	SSC	$\rho_{\max}$	SSC	$\rho_{\max}$	SSC	df
64	64.bd	0.13	0.25	0.25	6.07	0.44	49.08	49
80	80.bd	0.10	0.12	0.25	5.82	0.30	30.16	65
96	96.bd	0.08	0.12	0.21	4.90	0.29	21.38	81
112	112.d	0.07	0.07	0.23	5.62	0.30	15.63	91

6.3 Comparing design options using simulations

Inspired by the TB inhibition experiment, we performed a simulation study involving eight active MEs and one to 10 active TFIs. We considered all the strength-3 designs in Table 7, all the nonorthogonal designs in Table 8, and the two-level nonregular strength-4 design with 128 runs and 14 factors used by Silva et al. (2016). To identify the active effects, we used the Dantzig selector (Candes and Tao, 2007; Phoa et al., 2009), which searches for the active effects by solving a linear programming problem. Using simulations, Marley and Woods (2010), Draguljić et al. (2014) and Mee et al. (2017) demonstrated the excellent performance of the Dantzig selector to correctly identify active MEs and TFIs, when compared to traditional model selection strategies such as forward selection. Details regarding the Dantzig selector and the selection of the required tuning parameters for an automatic model selection are given in supplementary Section D.

6.3.1 Simulation protocol

To explain how we performed our simulations, we need to introduce some notation. Let $\mathbf{X}$ be the $N \times [m + m(m-1)/2]$ interaction model matrix for 14 factors (excluding the intercept column) and g be the number of active TFIs. For each design and each g from one to 10, each of our 1,000 simulations consisted of the following steps:

1. We randomly selected eight ME columns of $\mathbf{X}$ and associated these with the eight active MEs. Next, we randomly selected g TFI columns of $\mathbf{X}$ subject to the constraint that they involved at least one factor with an active ME. In other words, we assumed that the TFIs satisfy weak effect heredity (Wu and Hamada, 2009, ch 4). The selected TFI columns were associated with the g active TFIs.

2. We generated the coefficients corresponding to the active effects according to two cases labeled 'minSNR1' and 'minSNR2', which consisted of adding 1 or 2, respectively, to an exponentially distributed random number. The coefficients corresponding to the inactive effects were drawn from $N(0, 0.25^2)$. A '+' or '−' sign was randomly assigned to each sampled value.

3. We generated an $N \times 1$ response vector $\mathbf{y}$ using the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where the $[m + m(m-1)/2] \times 1$ vector $\boldsymbol{\beta}$ contains the simulated coefficients for the active and inactive effects, and the $N \times 1$ vector of residuals $\boldsymbol{\epsilon}$ has elements ϵ_i drawn from $N(0, 1)$.

4. The set of effects declared active is determined using the Dantzig selector.

The numbers of active MEs and TFIs we considered cover the active effects in the TB inhibition experiment as well as other situations that may occur in practice. In the case labeled minSNR1, most signal-to-noise ratios (SNRs) for the active effects were close to 1 in absolute value, while a few were larger. In the case labeled minSNR2, most SNRs for the active effects were close to 2 in absolute value. Following the simulation protocols of Marley and Woods (2010) and Draguljić et al. (2014), we did not set the coefficients for the inactive MEs and TFIs to zero. Instead, they followed a normal distribution with zero mean and standard deviation of 0.25. Therefore, there was a small probability that, for a given simulation, the absolute value of an inactive effect would exceed that of an active effect. In such cases, the coefficient for the inactive effect was re-generated. An R implementation of our simulation protocol is included in the online supplementary materials.

6.3.2 Results

We used three measures to compare the designs: power, false discovery rate (FDR) and false positive rate. The power is the proportion of active effects that are successfully detected. The FDR is the proportion of effects declared active that are actually inactive. The false positive rate (or type-I error rate) is the proportion of inactive effects that are incorrectly declared active. Obviously, the power should be maximized, while the FDR and the false positive rate should be minimized. We computed the average power, average FDR and average false positive rate for each design option using the 1,000 simulations for each combination of number of active TFIs (1 to 10) and case ('minSNR1' or 'minSNR2').

Performance for detecting main effects

We found that all designs had average false positive rates well below 0.05 for the MEs, for all numbers of active TFIs in both the minSNR1 and the minSNR2 cases. Similarly, all designs had average FDRs of virtually zero for the MEs in all simulation settings we considered. So, all designs had an excellent performance in terms of the false positive rate and the FDR for the MEs.

Regarding the power for the MEs, the strength-3 designs had average powers above 0.95 for all numbers of TFIs in both the minSNR1 and minSNR2 cases. In contrast, the D-optimal and Bayesian D-optimal designs had average powers for MEs between 0.85 and 0.92. The Bayesian D-optimal designs with 64 and 80 runs had average powers strictly smaller

than 0.9 in the minSNR1 case. In the event that effect hierarchy holds (Wu and Hamada, 2009, ch 4)—that is, MEs are more important than TFIs—average powers smaller than 0.9 for the MEs may be regarded as unsatisfactory. We therefore do not consider these designs in the next section. We do include these results in the online supplementary materials.

Performance for detecting two-factor interactions

Figure 2 shows the average power and average FDR for the TFIs as a function of the number of active TFIs, for each of the designs in Tables 7 and 8 except for the 64- and 80-run Bayesian D-optimal designs. Each subfigure consists of two panels which differ according to the minimum SNR in the simulations. The figure does not show the average false positive rates for the TFIs because these are below 0.05 for each design, except for design 64.p. For this design, the average false positive rate for the TFIs was between 0.053 and 0.064 for case minSNR2 and more than eight active TFIs. One of the conclusions from Figure 2 is that the 128-run strength-4 design used by Silva et al. (2016) has the largest average powers and the smallest average FDRs for the TFIs in both cases.

Figure 2a shows that, when most SNRs for the active effects are close to 1 (case minSNR1), the best strength-3 designs in terms of the power for the TFIs are the 96- and 112-run designs, as these designs have average powers between 0.82 and 0.87 for all numbers of active TFIs. Among these designs, design 96.f is slightly better than the others for most numbers of active TFIs. Figure 2b shows that the best strength-3 designs in terms of the FDR for the TFIs are also the 96- and 112-run designs, as their average FDRs are below 0.07 for all numbers of active TFIs. Design 96.f also stands out as having smaller average FDRs for all numbers of active TFIs than the other strength-3 designs.

Regarding the nonorthogonal design alternatives, Figure 2a shows that the 96-run Bayesian D-optimal design, 96.bd, has similar average FDRs but slightly larger average powers than the strength-3 design 96.f for all numbers of active TFIs in the minSNR1 case. The average power of design 96.bd is between 0.87 and 0.9. In the minSNR1 case, the 112-run D-optimal design 112.d is the best design with fewer than 128 runs, since it has the largest average powers and the smallest average FDRs.

When most SNRs for the active effects are close to 2 (case minSNR2), the best designs in terms of power for the TFIs are the 96-, 80- and 112-run strength-3 designs, together with the 96-run Bayesian D-optimal design and the 112-run D-optimal design. This is because these designs have average powers larger than 0.95 for all numbers of active TFIs. Figure 2a shows that, for four or more active TFIs, the 96- and 112-run designs perform

slightly better than the 80-run designs. Regarding the FDR for the TFIs, Figure 2b shows similar results for both cases: the best designs for the high SNR case in terms of the FDR are again the 96- and 112-run designs, with average FDRs close to 0.

6.3.3 Discussion of the simulation results

Overall, our simulation results showed that the strength-3 and nonorthogonal designs with 96 and 112 runs provided a good performance in terms of power, false positive rate and false discovery rate for both the MEs and the TFIs. These designs, together with the 80-run strength-3 designs, will very likely identify all the active effects with a signal-to-noise ratio as small as 2. Nevertheless, our results also demonstrated that, in the presence of smaller active TFIs, choosing a design smaller than the 128-run nonregular design used by Silva et al. (2016) entails a decrease in the power for these effects. In situations where conducting 128 runs is feasible and several active effects as small as the error's standard deviation are expected and considered practically relevant, we would feel uncomfortable recommending one of the 96- or 112-run designs. The smaller designs are appropriate for situations where conducting 128 runs is too expensive or even infeasible, or where the goal is to identify large active TFIs. An important added value of our simulation results is that they show the trade-offs between using economical designs and being able to identify the active effects, when using the Dantzig selector as the data analysis method.

Regarding the alternative nonorthogonal designs with 96 runs, neither the Bayesian D-optimal design nor the strength-3 design 96.f dominated each other when considering the power for MEs and TFIs. If the MEs and TFIs are equally important, then the Bayesian D-optimal design is attractive. In contrast, if the MEs are considered to be more important than TFIs, then design 96.f is more attractive because it had a larger power for detecting the MEs than the Bayesian D-optimal design, while providing an only slightly smaller power for the TFIs. We reached similar conclusions for the strength-3 and D-optimal designs with 112 runs.

Our conclusions about the performance of strength-3 and nonorthogonal designs for identifying active MEs and TFIs are in line with the literature. Using simulations involving the Dantzig selector and 11-factor designs with 32 to 48 runs, Mee et al. (2017) showed that strength-3 designs are generally powerful for detecting the active MEs. For the detection of active TFIs, they showed that Bayesian D-optimal designs have larger powers. Using relative standard errors, Eendebak and Schoen (2017) and Vazquez and Xu (2019) reached similar conclusions when comparing strength-3 and D-optimal designs.

7 General discussion

This article features two-level nonregular strength-3 designs with run sizes between 64 and 128. Using an existing construction procedure for concatenating equally-sized strength-3 designs, we obtained a previously unknown collection of strength-3 designs with 80, 96 and 112 runs, and up to 29 factors. These designs minimize the aliasing among the two-factor interactions and, with one exception, provide pairs of two-factor interactions which are only partially aliased. Moreover, they outperform or are competitive with the available benchmark strength-3 designs in terms of the aliasing among interactions. Our new collection of designs is available in the online supplementary materials.

In most cases, we provide two types of strength-3 designs. One type of design minimizes the overall aliasing among all two-factor interactions, as measured by the B_4 value. The other type of design minimizes the most severe aliasing among these effects, as measured by the generalized resolution and the F_4 vector. Our preference for one of these two types of strength-3 designs depends on how we wish to cope with the aliasing of interactions. In practice, both types of designs are high-quality designs and the differences between them are not extremely important.

Our designs with 80, 96 and 112 runs include a column $\mathbf{z}$ which splits the experimental runs into two equally-sized strength-3 designs. Throughout the article, we evaluated this column as an additional treatment factor. Alternatively, we could use the column for a blocking factor to arrange the designs in two blocks with half of the runs. The blocking factor would then have the attractive property that its main effect is not confounded with any of the main effects or any of the two-factor interactions of the remaining treatment factors. Therefore, our large designs are suitable for experiments spanning two different days or requiring two different machines or operators.

Using a tuberculosis inhibition experiment and a simulation study, we investigated the usefulness of our 80-, 96- and 112-run designs with 14 factors. We discussed the strengths and limitations of the smaller strength-3 designs when compared to the 128-run nonregular strength-4 design actually used in the experiment. On the upside, our designs offer the same (or almost the same) chances to detect small and large active main effects as well as large active interactions as the 128-run design. On the downside, they are less effective to identify small active interactions. The tuberculosis inhibition experiment benefited from the fact that it was feasible to implement this excellent 14-factor 128-run nonregular strength-4 design, which allows to estimate all the effects of interest with full precision. The largest

number of factors for which 128-run nonregular strength-4 designs are available is 15. For 16 to 19 factors, both nonregular strength-4 designs and regular resolution-V designs require at least 256 runs, while, for 20 or more factors, they require at least 512 runs; see Mee (2004). Therefore, as the number of factors grows, our strength-3 designs become more and more attractive cost-efficient alternatives to strength-4 designs and resolution-V designs.

A byproduct of our research is that we found 56-run strength-3 designs with up to 28 factors, from projections of selected folded-over orthogonal designs. These 56-run designs fill the gap between 48 and 64 runs. When used as parent designs, some of these 56-run designs resulted in 112-run designs which are inferior to our 96-run designs in terms of the generalized resolution, F_4 vector, the B_4 value or the degrees of freedom for two-factor interactions. This is merely a consequence of the fact that a complete catalog of 56-run strength-3 designs is unavailable. Therefore, we could not identify the best possible parent designs for the 112-run designs. For 16 to 20 factors, however, our best 112-run designs in terms of the F_4 vector outperform the 96-run designs in terms of all the criteria we considered.

Finally, in this article we paid specific attention to nonregular strength-3 designs which can be constructed by concatenating two equally-sized strength-3 designs. In doing so, we ignored strength-3 designs with 72, 88, 104 and 120 runs, since they cannot be constructed by concatenating two strength-3 designs. As a matter of fact, the construction of these designs would require parent designs with run sizes of 36, 44, 52 and 60, which are not multiples of eight and for which strength-3 designs do not exist. So, an interesting topic for future research is to find attractive strength-3 designs with 72, 88, 104 and 120 runs, using a different construction method. One way to obtain these strength-3 designs is from the best projections of the folded-over Plackett-Burman designs with 36, 44, 52 and 60 runs. For instance, the strength-3 designs obtained by folding over the 36- and 44-run Plackett-Burman designs have generalized resolutions as large as 4.67 and 4.73, respectively. So, they provide attractive starting designs as they do not involve fully aliased pairs of two-factor interactions. We may extend the search for the best projections to the folded-over Plackett-Burman designs with 40, 48 and 56. Such study would reveal if the benchmark strength-3 designs with 80, 96 and 112 we considered in Section 5 were the best ones available. However, the identification of the best projections from all these folded-over Plackett-Burman designs is computationally very demanding. So, we leave it for future research.

Supplementary Materials

Supplementary sections.pdf Tables of 56-run parent designs; detailed construction of the concatenated designs with 80, 96 and 112 runs; details on the strength-3 designs derived from folded-over Plackett-Burman designs; details on the Dantzig selector.

Supplementary files.zip Collection of two-level strength-3 designs with 80, 96 and 112 runs and 9 to 29 factors, selected two-level strength-3 designs with 27 factors and 56 runs, and an R implementation of our simulation study.

Acknowledgements

The authors thank the three referees for their valuable remarks which helped to improve the presentation of the article. The authors also thank Bai-Yu Lee, Daniel L. Clemens and Marcus Horwitz for providing more details about the tuberculosis inhibition experiment.

Funding

The research that led to this article was financially supported by the Flemish Fund for Scientific Research FWO.

About the authors

Dr. Vazquez is an Assistant Adjunct Professor at the Department of Statistics at the University of California, Los Angeles. He is also a Junior Postdoctoral Fellow of the Flemish Fund for Scientific Research FWO. His email address is alanrvazquez@stat.ucla.edu.

Dr. Schoen is a Guest Professor at the Faculty of Bioscience Engineering of KU Leuven. His email address is eric.schoen@kuleuven.be.

Dr. Goos is a Full Professor at the Faculty of Bioscience Engineering of KU Leuven and at the Faculty of Business and Economics of the University of Antwerp. He is a Senior Member of the American Society for Quality. His email address is peter.goos@kuleuven.be.

References

Block, R. M. and Mee, R. W. (2005). Resolution IV designs with 128 runs. *Journal of Quality Technology*, 37:282–293.

Box, G. E. and Hunter, J. S. (1961). The 2^{k-p} fractional factorial designs. *Technometrics*, 3:311–351, 449–458.

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

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

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

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

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

Draguljić, D., Woods, D. C., Dean, A. M., Lewis, S. M., and Vine, A.-J. E. (2014). Screening strategies in the presence of interactions. *Technometrics*, 56:1–16.

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

Eendebak, P. T. and Schoen, E. D. (2017). Two-level designs to estimate all main effects and two-factor interactions. *Technometrics*, 59:69–79.

Goos, P. and Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach*. Wiley.

Hedayat, A., Sloane, N., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer.

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

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.

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

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

Michalewicz, Z. and Fogel, D. (2004). *How to Solve It: Modern Heuristics*. Springer.

Paley, R. E. A. C. (1933). On orthogonal matrices. *Journal of Mathematics and Physics*, 12:311–320.

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

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

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.

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

Silva, A., Lee, B.-Y., Clemens, D. L., Kee, T., Ding, X., Ho, C.-M., and Horwitz, M. A. (2016). Output-driven feedback system control platform optimizes combinatorial therapy of tuberculosis using a macrophage cell culture model. *Proceedings of the National Academy of Sciences*, 113:E2172–E2179.

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

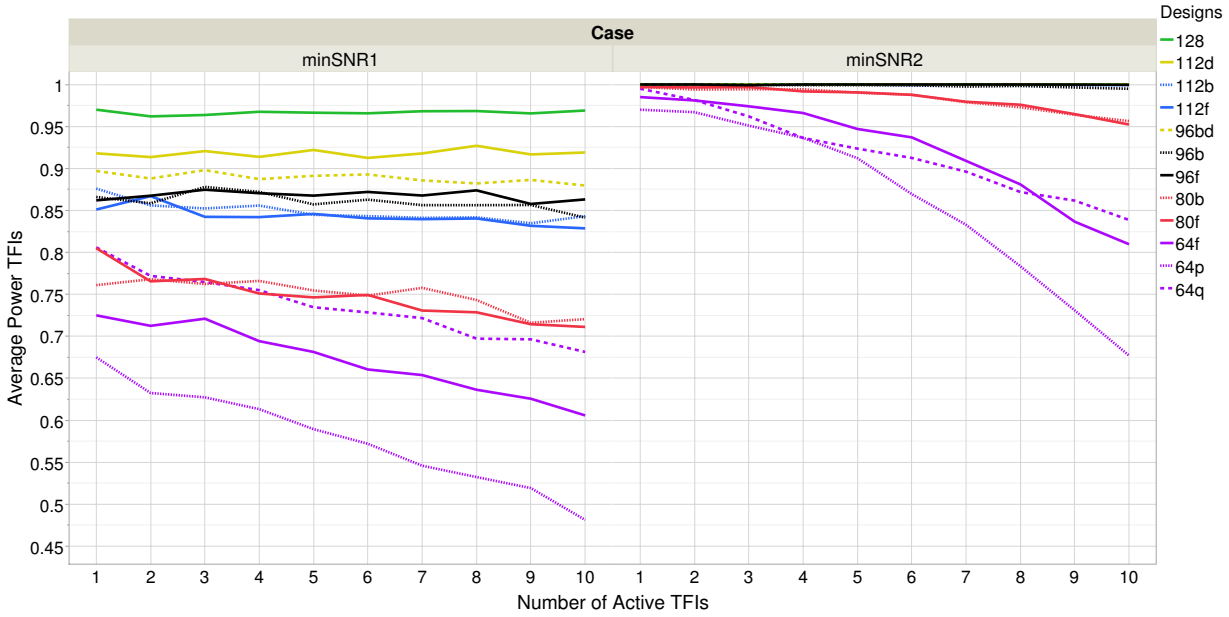
Vazquez, A. R., Goos, P., and Schoen, E. D. (2019). Constructing two-level designs by concatenation of strength-3 orthogonal arrays. *Technometrics*, 61:219–232.

Vazquez, A. R. and Xu, H. (2019). Construction of two-level nonregular designs of strength three with large run sizes. *Technometrics*, 61:341–353.

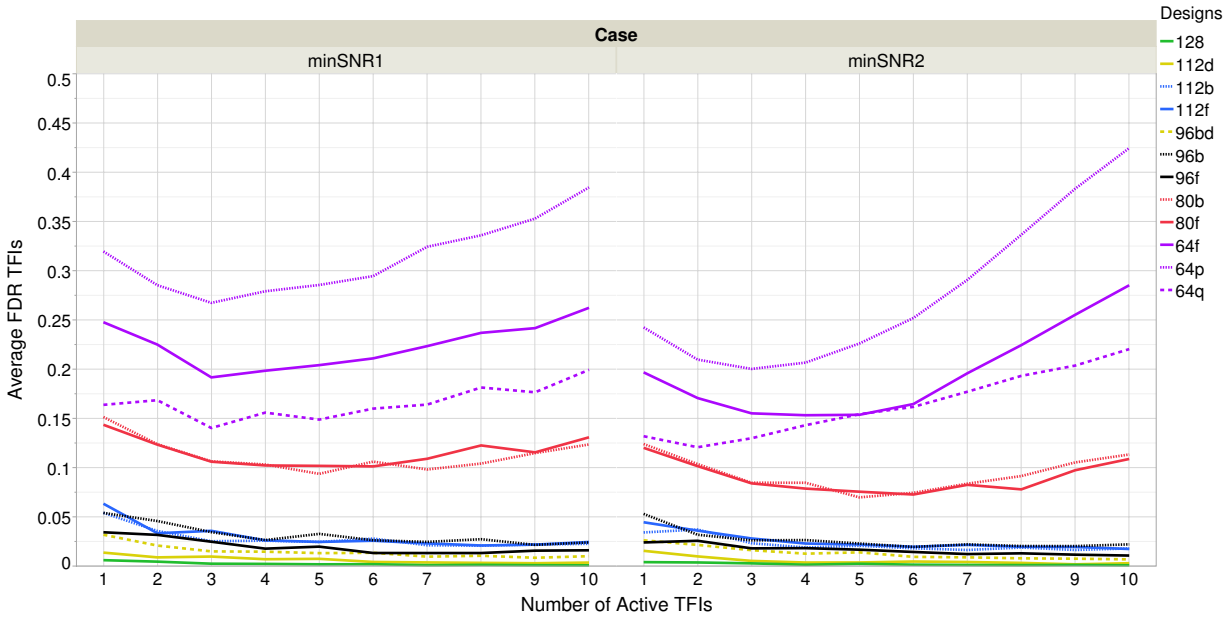
Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*. Wiley, 2nd edition.

Xu, H. (2009). Algorithmic construction of efficient fractional factorial designs with large run sizes. *Technometrics*, 51:262–277.

Xu, H. and Wong, A. (2007). Two-level nonregular designs from quaternary linear codes. *Statistica Sinica*, 17:1191–1213.



(a)



(b)

Figure 2: Average power and FDR for correctly identifying 1 to 10 active TFIs. Design '128' is the two-level nonregular strength-4 design used by Silva et al. (2016). The two-level strength-3 and nonorthogonal designs are labeled as in Tables 7 and 8, respectively. minSNR1: small SNRs for the active effects; minSNR2: large SNRs for the active effects. The online version of this figure is in color.