The Annals of Applied Statistics
2016, Vol. 10, No. 2, 864–883
DOI: 10.1214/16-AOAS916

LEVEL-SCREENING DESIGNS FOR FACTORS WITH MANY LEVELS

BY PHILIP J. BROWN AND MARTIN S. RIDOUT

University of Kent

We consider designs for f factors each at m levels, where f is small but m is large. Main effect designs with mf experimental points are presented. For two factors, two types of designs are investigated, termed *sawtooth* and *dumbbell* designs, based on a graphical representation. For three factors, cyclic sawtooth designs are considered. The paper seeks optimal and near optimal designs which involve factors with many levels but few observations. It also investigates issues of robustness when as much as one third of the data is structurally missing. An important area of application is in screening for drug discovery and we compare our designs with others using a published data set with two factors each with fifty levels, where the dumbbell design outperforms others and is an example of an inherently unbalanced design dominating more balanced designs.

1. Introduction. In the context of factorial experimentation, the term *screening design* usually refers to designs in which there are many factors and the aim is to identify the most important of these using designs with a small number of levels of each factor; see, for example, Box, Hunter and Hunter (2005), Chapters 6, 7. In contrast, this paper is concerned with experiments with a small number of factors, but where each factor has a large number of unordered levels. Our defining example, from Pickett et al. (2011), has two factors each with 50 levels, comprising different chemical modifications made at two sites on a target molecule. We do not assume any knowledge of these levels, either physical or chemical. The aim is to identify modifications that result in a high level of pharmaceutical activity. This aim presents a different type of screening problem that has wide applicability in drug development and other areas and can be supplemented by more informed knowledge. We refer to such designs as *level-screening* designs. They should provide fertile ground in widely used *combinatorial chemistry* [Thayer (1996)], a laboratory technique in which many molecular combinations, exploiting chemical series and synthesis routes, are tested for biological activity.

One common feature of both problems is that it may be feasible to include only a small fraction of all possible combinations of factor levels in the experiment. Another common feature is a focus on main effects at this initial experimentation stage, with interactions assumed to be negligible. Conventional interaction models,

Received December 2014; revised February 2016.

Key words and phrases. Screening designs, lead optimization in drug discovery, main effects, microarray loop designs, connectivity, identifiability, prediction and contrast variance.

with a parameter for each combination of factor levels, are in any case unwieldy when factors have many levels and alternative approaches, for example, grouping factor levels and exploring interactions at the group level [Franck, Nielsen and Osborne (2013)], seem more promising. In the example of Section 4, the main effects of the factors appear to be large in relation to any interactions.

In this paper we consider designs for f factors, each at m levels, where $f = 2$ or $f = 3$, but m is large, assuming an additive main-effects model. A saturated design that allows estimation of all main effect parameters, but has no residual degrees of freedom, requires $(m-1)f+1$ experimental units. We consider slightly larger designs, with mf units, since these allow the possibility for the levels of each factor to be equally replicated. Section 2 presents two basic designs for two factors, the *sawtooth*, which has equal replication, and the *dumbbell*, which does not. We compare these designs by various criteria, including the average variance of the estimated difference between two factor levels and the average variance of the estimated expected response for all combinations of factor levels. Section 3 extends these designs to three factors and focuses particularly on cyclic sawtooth designs. Section 4 explores the application of these designs to the data of Pickett et al., and compares them with designs that select combinations of factor levels in a randomized way [Pickett et al. (2011), Borrotti et al. (2014)]. There are a large number of missing values in this pharmaceutical application and so an important issue is the robustness of the designs to missing values. Our results indicate that for two factors the dumbbell design offers the best combination of efficiency and robustness. Finally, in Section 5 we discuss extensions and areas for future work.

2. Designs for two factors. We consider factors A and B, each at m levels, and assume that the response Y_{ijk} to replicate k of level i of factor A in combination with level j of factor B follows the additive model

$$Y_{ijk} = \alpha_i + \beta_j + \varepsilon_{ijk} \qquad (1 \le i, j \le m, 1 \le k \le n_{ij}), \tag{1}$$

where the ε_{ijk} are independent $\mathrm{N}(0, \sigma^2)$ random variables and where n_{ij} is the number of replicates of the treatment combination $A_i B_j$; usually, $n_{ij} = 1$ and the subscript k can be omitted. The model is over-parametrized and for estimation we impose the constraint $\beta_1 = 0$. Here we have followed the parametrization of Tjur (1991), whose results we use later. We could also have chosen the more conventional parametrization

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk} \qquad (1 \le i, j \le m, 1 \le k \le n_{ij}),$$

with constraints $\alpha_1 = \beta_1 = 0$. However, the optimality results that we develop are invariant to the chosen parametrization. In either parametrization, the number of parameters to be estimated is $1 + 2(m-1)$.

We consider designs with $2m$ experimental units in which all parameters are estimable. Such designs only provide a single degree of freedom for estimating σ^2.

However, the objective of this initial screening stage is to identify the "better" levels of each factor to take through to further stages of experimentation, where in our application "better" means the factor levels with the largest estimated effects, interpreted fairly liberally, and in reality further responses would be investigated, for example, safety and ADME (absorption, distribution, metabolism, excretion) [Balani et al. (2005)]. The assays for such extra attributes were not done for the current application. At this stage we are not interested in making formal statistical inferences, for which a better estimate of σ^2 would be required.

One criterion for comparing designs is to consider the average variance of the estimated differences $\hat{\alpha}_{i_1} - \hat{\alpha}_{i_2}$ across all pairs of levels of factor A. Of course, one might equally consider the corresponding average variance for factor B, but the designs that we consider are symmetrical in A and B, so there is no loss of generality in considering factor A. We denote this average variance by V_A. A design that minimizes V_A within a particular class of designs is said to be A-optimal within that class [Atkinson and Donev (1992), Section 10.1].

Another criterion, which is arguably more important for the application in Section 4, is the average variance over all combinations (i, j) of the estimated expected responses

$$\hat{\mu}_{ij} = \hat{\alpha}_i + \hat{\beta}_j.$$

For brevity, and because most factor combinations $A_i B_j$ are not included in the experiment, we refer to the $\hat{\mu}_{ij}$ as *predictions*, and denote their average variance by V_P. A design that minimizes V_P within a particular class of designs is said to be P-optimal within that class; for continuous predictors, minimizing the average (integrated) prediction variance over a range of the predictor space is sometimes known as V-optimality [Atkinson and Donev (1992), Section 10.7] or I-optimality [Dette and O'Brien (1999)].

A further criterion that is widely used in experimental design is the generalized variance of a set of parameter estimates, which is proportional to the volume of a confidence ellipsoid for the parameters. Here, the relevant parameters are the estimated treatment effects for factor A or, equivalently, for factor B. A design that minimizes the generalized variance within a class of designs is said to be D-optimal within that class [Atkinson and Donev (1992), Section 10.1]. Although we include some results on this criterion, we regard it as less relevant for applications involving qualitative factors than the earlier criteria.

The generalized variance is calculated here as the product of the nonzero eigenvalues of the covariance matrix of the estimated treatment effects for factor A. Since this matrix has rank $m - 1$, the positive $(m - 1)$th root of the generalized variance, which we denote by V_D and which is sometimes termed the *standardized* generalized variance, is more useful for comparing different values of m.

2.1. *Sawtooth designs*. Consider the following design, in which the $2m$ design points are partitioned into two sets of m, such that each level of A and each level of B occurs once in each set:

	Set 1					**Set 2**				
Factor A	1	2	...	$m-1$	m	1	2	...	$m-1$	m
Factor B	1	2	...	$m-1$	m	2	3	...	m	1

The left panel of Figure 1 shows a graphical representation of this design for $m = 25$, in which a solid line segment joins level i of factor A to level j of factor B if and only if the combination $A_i B_j$ occurs in the design. Based on this graphical representation, we refer to this design as a *sawtooth design*.

An alternative representation of the sawtooth design is obtained by drawing a line segment to join two levels of factor A if and only if they occur together in the design with the same level of factor B, and similarly for the levels of factor B. It gives rise to the concentric circles shown with dashed lines in the left panel of Figure 1. Structurally, we can regard the design for factor A as an incomplete block design with blocks of size 2 defined by the levels of factor B, and similarly with the factors interchanged, so that A acts as a blocking factor for B. These are known as *loop* designs and have been studied, for example, in the design of microarray experiments [Kerr and Churchill (2001), Vinciotti et al. (2005)]. The sawtooth design therefore comprises a pair of interlinked loop designs.

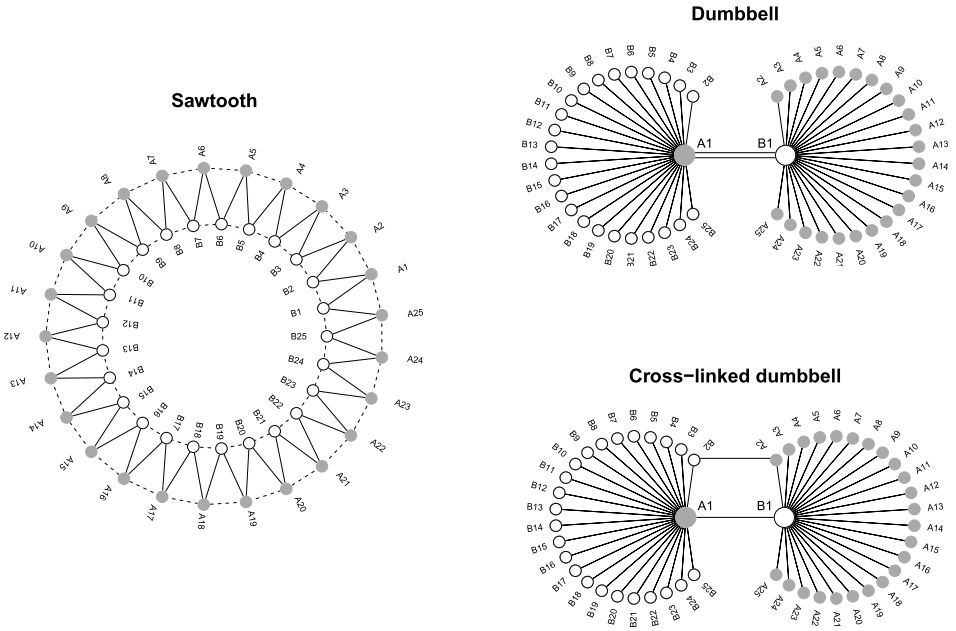

FIG. 1. *Sawtooth and dumbbell designs for factors with $m = 25$ levels.*

A general design in which each level of each factor occurs in each set has the following form:

	Set 1					Set 2				
Factor A	1	2	...	$m-1$	m	1	2	...	$m-1$	m
Factor B	1	2	...	$m-1$	m	$\pi(1)$	$\pi(2)$	...	$\pi(m-1)$	$\pi(m)$

where the levels of factor B are relabeled if necessary to give the factor combinations in Set 1 and where $\pi(\cdot)$ is a permutation of $\{1, \ldots, m\}$.

Recall that a *cycle* of length r is a circular sequence $x_1 \to x_2 \to \cdots \to x_r \to x_1$, where the elements $x_1, \ldots, x_r$ are distinct. It is well known that any permutation can be decomposed into disjoint cycles, and the following lemma, which is proved in Section 1 of the supplemental material [Brown and Ridout (2016)], shows that the cycle decomposition of the permutation $\pi(\cdot)$ is key to understanding the statistical properties of this design. Specifically, let c_π denote the number of cycles of $\pi(\cdot)$ and let X_π denote the corresponding model matrix for the additive model (1).

LEMMA 1. *The model matrix X_π is of rank $2m - c_\pi$, in particular, it is of rank $2m - 1$, implying that all parameters are estimable if and only if the permutation $\pi(\cdot)$ consists of a single cycle and is not the identity permutation. All such permutations are isomorphic to one another and the matrices $X_\pi' X_\pi$ have the same set of singular values. The product of the $(2m - 1)$ nonzero singular values is $4m^2$.*

That all single-cycle permutations are isomorphic to one another can be seen by relabeling the sawtooth design according to the permutation. For example, for $m = 6$ levels, an example of a single cycle permutation is $1 \to 4 \to 2 \to 3 \to 6 \to 5 \to 1$. We would relabel around the sawtooth to give the factor combinations A_1B_1, A_1B_4, A_4B_4, A_4B_2, A_2B_2, A_2B_3, ..., A_5B_1. Because the permutation is a cycle, the final line segment connects to the initial line segment, completing the circular sawtooth. Conversely, if the permutation has more than a single cycle, then the sawtooth cannot be relabeled in this way, and the design becomes disconnected, implying that it is no longer possible to estimate all the parameters in the model.

The results of this section indicate that, among designs in which the $2m$ points can be partitioned into two sets of size m such that each level of each factor occurs in each set, any design that allows estimation of all $2m - 1$ model parameters is equivalent to the sawtooth design after suitable relabeling.

2.2. *Dumbbell designs.* The left panel of Figure 1 shows that, although the sawtooth design is connected, with a path from every factor level to every other factor level, the path lengths can be large. We show below that, as a result, some contrasts and predictions have large variances. An alternative design, with short path lengths, which we term a *dumbbell design*, is shown in the upper right panel of Figure 1. In this design, each level of factor B is paired with a single level, say A_1, of factor A and similarly each level of factor A is paired with one level,

say B_1, of factor B. We refer to the treatment combination A_1B_1 which links the two parts of the design as the *anchor point* of the design. By suitable relabeling, any treatment combination can be chosen as the anchor point of the dumbbell design. We note that the dumbbell design is a particular instance of a *one-factor-at-a-time* (OFAT) design [Daniel (1973)]. Notwithstanding suboptimal behavior in general, this two factor situation is special and certain optimality results are given in the next section.

The dumbbell design illustrated in the upper right panel of Figure 1 has two replicates of the anchor point, indicated by the double edge joining A_1 to B_1. This replication ensures that the design has $2m$ points, but a single replicate only is needed to allow all parameters in the model to be estimated. An alternative $2m$-point design, which we refer to as a *cross-linked* dumbbell design, shown in the lower right panel of Figure 1 has a single replicate of A_1B_1, but has an additional factor combination A_jB_k $(j > 1, k > 1)$ linking the two sides of the dumbbell; in Figure 1, $j = k = 2$.

2.3. *Comparison of the sawtooth and dumbbell designs.* Since we are interested in the relative values only of variances from different designs, and since these are all proportional to σ^2, we assume without loss of generality that $\sigma^2 = 1$.

We first consider variances of estimated pairwise differences of factor levels. The sawtooth design for factor A is equivalent to a loop design, so an argument of Bailey (2007), Section 4.1, applies. Bailey notes that $\hat{\alpha}_i - \hat{\alpha}_{i'}$ can be estimated by a contrast involving all the observations that link A_i to $A_{i'}$ in the sawtooth diagram. The contrast coefficients are alternately $+1$ and -1, so the variance of the contrast is just the sum of the variances of the observations involved which are independent by virtue of independence of compounds. Each individual observation has variance σ^2. With $k = |i - i'|$, there are $2k$ independent steps in one direction, and $2(m - k)$ in the other direction, giving two uncorrelated estimators, with variances $2k\sigma^2$ and $2(m - k)\sigma^2$. The least squares estimator $\hat{\alpha}_i - \hat{\alpha}_{i'}$ is a weighted linear combination of these two estimators, with weights chosen to minimize the variance of the resulting estimator. This calculation gives

$$\operatorname{var}(\hat{\alpha}_i - \hat{\alpha}_{i'}) = \frac{2k(m-k)}{m},$$

where we have set $\sigma^2 = 1$ as indicated above. It follows that the minimum variance is $2(m - 1)/m$, the maximum variance is $2t(m - t)/m$, where t is the integer part of $m/2$, and the average variance is $V_A^{\text{saw}} = (m + 1)/3$.

This result is a simple example of the approach of Tjur, which exploits an analogy with electrical networks to calculate variances of contrasts of the form $\hat{\alpha}_i - \hat{\alpha}_{i'}$ and variances of predictions of the form $\hat{\alpha}_i + \hat{\beta}_j$ for any 2-factor design, based on representing the design as a 2-color graph as in Figure 1. The general approach is more complicated when the graph has circuits, as for the cross-linked dumbbell design. In Section 2 of the Supplemental Material [Brown and Ridout (2016)], we use this approach to prove the following theorem.

TABLE 1
Expressions for V_A, V_P and V_D for the 2-factor sawtooth, dumbbell and cross-linked dumbbell designs

Design	V_A	V_P	V_D
Sawtooth	$\frac{m+1}{3}$	$\frac{2m^2+1}{6m}$	$(\frac{2^{m-1}}{m^2})^{1/(m-1)}$
Dumbbell	$\frac{2m-1}{m}$	$\frac{5m^2-6m+2}{2m^2}$	$(\frac{m+1}{2m})^{1/(m-1)}$
Cross-linked dumbbell	$\frac{2m-3}{m-1}$	$\frac{11m^2-20m+8}{4m^2}$	$(\frac{1}{2})^{1/(m-1)}$

THEOREM 1. *Among all designs with $2m$ points, the dumbbell design is P-optimal for $m \geq 8$ and the cross-linked dumbbell design is A-optimal for $m \geq 4$.*

Table 1 gives exact expressions for V_A, V_P and V_D for the three designs shown in Figure 1. By comparing variance to that of a (hypothetical) orthogonal design we can examine *efficiencies*. As $m \to \infty$, the A-efficiency and the D-efficiency of the sawtooth design both approach zero, whereas the A-efficiency and the D-efficiency of the dumbbell design both approach one. Conversely, as $m \to \infty$, the P-efficiency of the sawtooth design again approaches zero, but the limiting P-efficiency of the cross-linked dumbbell design is $10/11$. Therefore, our preference on the basis of efficiency is for the dumbbell design, since it has very similar A- and D-efficiency to the cross-linked dumbbell, but is somewhat more P-efficient.

For the sawtooth design, the pattern of the variances of pairwise differences $\text{var}(\hat{\alpha}_i - \hat{\alpha}_j)$ is very similar to the pattern of variances of predictions $\text{var}(\hat{\alpha}_i + \hat{\beta}_j)$. This pattern is illustrated in Figure 2 for the sawtooth design with $m = 25$.

A further advantage of the dumbbell design over the sawtooth is that estimates of differences between factor levels of the form $\hat{\alpha}_i - \hat{\alpha}_{i'}$ are less highly correlated in the dumbbell design, as are the estimates of expected response $\hat{\alpha}_i + \hat{\beta}_j$; details are provided in Section 3 of the Supplemental Material [Brown and Ridout (2016)].

Throughout this paper we assume that the number of levels of factors are the same. This equality is not necessary but just simplifies formulae. For example, the dumbbell with an unequal number of levels would just be a lopsided dumbbell; the sawtooth for $m_A > m_B$ levels of A and B would require $m_A + m_B$ observations where some B levels would be recycled.

2.4. *Missing values*. The effect of occasional missing values is rather different for the three designs.

If one of the two replicates of the anchor point of the dumbbell design is lost, then $V_A = 2$ and $V_P = (3m^2 - 4m + 2)/m^2$. If both replicates are lost, then the design becomes disconnected and not all parameters can be estimated. Thus, the design is problematic unless the anchor point can be chosen to have small probability of being lost; we return to this point in Section 4.

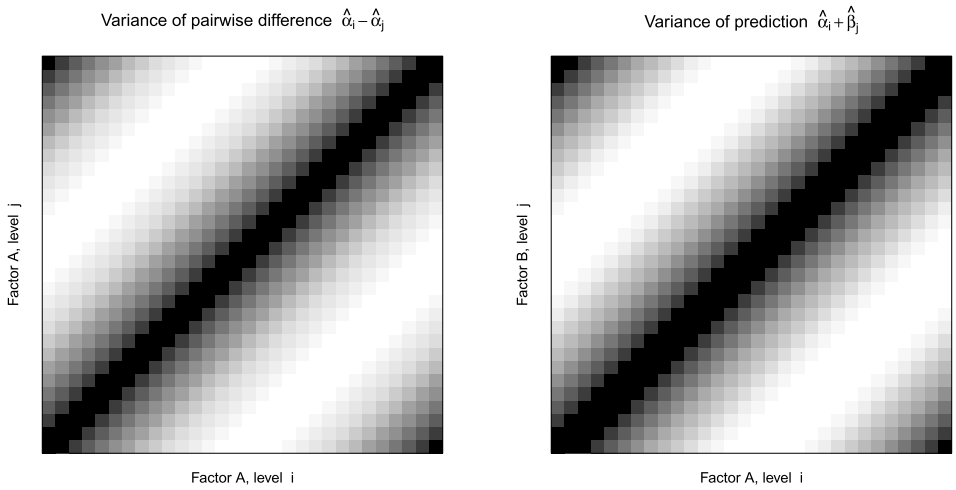

FIG. 2. *Comparison of variances of pairwise differences and variances of estimated expected responses for the sawtooth design with* $m = 25$: *darker shading corresponds to smaller variances. The scales of the two graphs differ slightly*: *LH panel*, 0 (*black*) *to* 15 (*white*); *RH panel*, 0.983 (*black*) *to* 14.983 (*white*).

If any other design point is lost from the dumbbell design, say A_1B_j, then the design provides no information about the parameter β_j, and therefore no predictions can be made that involve this parameter. On the other hand, the values of V_A and V_P for differences and predicted values that *can* be estimated are increased only modestly.

The cross-linked dumbbell provides some robustness against loss of the anchor point. If the single replicate of the anchor point is lost, but the cross-link is retained, then we find

$$V_A = \frac{2(m^2 + m - 4)}{m(m-1)} \quad \text{and} \quad V_P = \frac{5m^2 - 12m + 9}{m^2}.$$

In contrast, for the sawtooth design, if a single point is lost, then all parameters can still be estimated, but the average variances of differences and predictions increase considerably. Moreover, if a second point is lost, then the design becomes disconnected and not all parameters can be estimated.

In Section 4 we examine the practical performance of these designs in the presence of a considerable number of missing values.

3. Designs for three factors. Although we specifically consider designs for three factors with $3m$ points, the ideas in this section could be extended to designs for $f > 3$ factors with mf points.

3.1. *The* 3*-factor sawtooth designs*. The 2-factor sawtooth design illustrated in Figure 1 consists of two sets of m design points. Within each set, the levels

of the two factors are developed cyclically from the generators (1, 1) (Set 1) and (1, 2) (Set 2).

With three factors, there are many more possibilities and full-rank designs need not have a cyclic structure. However, we have focused on designs that extend the cyclic structure to three sets of three factors so that a design is specified by its three generators. For example, with $m = 5$, the generators (1, 1, 1), (1, 2, 4) and (1, 4, 3) produce the design

	Set 1					**Set 2**					**Set 3**				
Factor A	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Factor B	1	2	3	4	5	2	3	4	5	1	4	5	1	2	3
Factor C	1	2	3	4	5	4	5	1	2	3	3	4	5	1	2

Designs constructed in this way have been termed *cyclic superimposed designs* by Hall and Williams (1973). However, their motivation was different. They were interested in superimposing an additional set of treatments on an existing design, assuming no interaction between the original treatments and the new treatments. They focused on two sets of treatments where the initial design was a cyclic incomplete block design. Our focus here is on designs with three factors, arranged in three sets, where each level of each factor occurs once in each set.

The factor levels can always be reordered so that the generator for Set 1 is (1, 1, 1), as above. We have investigated the performance of generators for the other two sets for $m \leq 12$ numerically. Typically, designs with the same characteristics can arise from several different sets of generators, but we found that designs with optimal properties always had at least one representative with generators of the form

$$\text{Set 1:} \quad (1, 1, 1), \qquad \text{Set 2:} \quad (1, 2, k+1), \qquad \text{Set 3:} \quad (1, k+1, k),$$

where $1 \leq k \leq m$, and where $k + 1$ is replaced by 1 when $k = m$. We therefore consider generators of this form for larger values of m. Finding an optimal design requires a linear search over the possible values of k, which can be done efficiently even for moderately large values of m using methods described in Section 4 of the Supplemental Material [Brown and Ridout (2016)]; for $m = 100$, for example, evaluating designs for all possible values of k takes just a few seconds of computing time. Moreover, the search time can be further reduced by a factor of two by noting that designs with $k = j$ and $k = m + 1 - j$ have equivalent statistical properties so that we need consider only $1 \leq k \leq r$, where $r = \text{int}\{(m + 1)/2\}$. This equivalence arises because the design generated by $k = j$ for factors A, B and C is identical to the design generated by $k = m + 1 - j$ except that factors A and C are interchanged.

To compare different choices of k using different criteria, we work with efficiencies, as defined below, rather than variances, since the variances for different

criteria are not directly comparable. In the cyclic design, each factor level is replicated 3 times. To calculate the *efficiency*, we therefore calculate the ratio of the variance for an (hypothetical) orthogonal design with three replicates of each factor level to the variance of the cyclic design. This approach leads to the following formulas for A-, P- and D- efficiency:

$$E_A = \frac{2}{3} V_A^{-1},$$

$$E_P = \frac{3m-2}{3m} V_P^{-1},$$

$$E_D = \frac{1}{3} V_D^{-1/(m-1)},$$

where V_A, V_P and V_D are analogous to the corresponding quantities for 2-factor designs defined in Section 2.

Figure 3 shows these efficiencies for different values of k for $m = 7$ and $m = 20$. The efficiencies E_A and E_P are numerically very similar and lead to the same optimal choice of k in each case. More generally, we found numerically that, for all m in the range $3 \leq m \leq 100$, the same value(s) of k maximize(s) E_A and E_P.

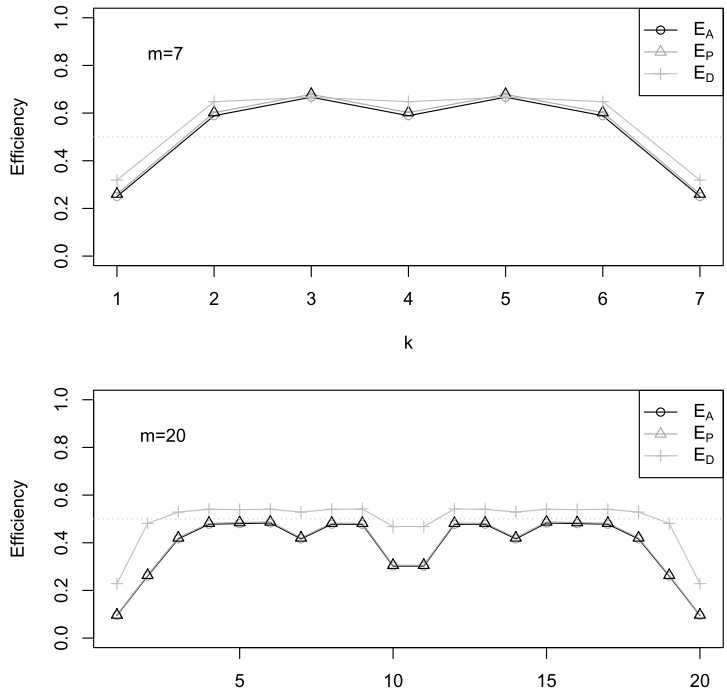

FIG. 3. *Efficiencies of cyclic designs with different choices of k for* $m = 7, 20$.

TABLE 2
Optimal choice(s) of k in the range $1 \le k \le \text{int}\{(m+1)/2\}$, to maximize E_A and E_P, for selected values of m

m	5	10	15	20	25	30	35	40	45	50
Optimal k	2, 3	3, 4	4	6	10, 11	9	6	12	7, 13	7, 8
m	55	60	65	70	75	80	85	90	95	100
Optimal k	21	14, 23	15	16	14, 17, 23	15	16	25	40	28, 37

The efficiency E_D gives numerically different values, but these follow a similar, though less marked, pattern to the other two efficiencies as k varies. For $m = 7$, all three criteria give $k = 3$ (or the equivalent, $k = 5$) as the optimal choice. For $m = 20$, the optimal choice based on E_A and E_P is $k = 6$ (or $k = 15$), whereas for E_D the optimal choice is $k = 9$ (or $k = 12$); however, the value of E_D for $k = 6$ is 99.83% of the optimal value.

For a given value of m, we recommend choosing k to optimize E_A and E_P because we consider these to be more important criteria than E_D for the type of application that we are interested in, and because E_D fluctuates less markedly with k than E_A and E_P; for $3 \le m \le 100$, the resulting value of E_D is never less than 99.40% of the optimal value, with this minimum occurring for $m = 12$.

The optimal choice of k for selected values of m is shown in Table 2.

An efficient algorithm for computing V_A and V_P for 3-factor sawtooth designs is given in Section 4 of the Supplemental Material [Brown and Ridout (2016)]. An alternative visual representation of the effect of varying k for fixed m is based on this algorithm, which generates two sets of m points in $3m$-dimensional space, such that V_A is the mean squared distance between points within a group and V_P is the mean squared distance between points in different groups. Figure 4 shows a 2-dimensional representation of these two groups of points for $m = 50$ using classical multidimensional scaling (principal coordinate analysis), based on the matrix of *squared* distances between points. The points appear as a skew-symmetric "butterfly," with the wings representing the two groups. Thus, insofar as the relative distances are preserved, what is desirable is small wings and small distances between the wings; the empirical results indicate that choices of k that are good under one criterion will also be good under the other. Figure 4 indicates that, in addition to the optimal choice of $k = 7$ or $k = 8$ (circled), other good choices of k are $k = 15, 19, 22, 6, 9, 12, 14, 23$.

3.2. *OFAT designs for* 3 *factors*. A natural way to extend the dumbbell design is to have an anchor point $A_1B_1C_1$ replicated three times. The remaining design points are of the form $A_1B_1C_j$, $A_1B_jC_1$ and $A_jB_1C_1$ for $j = 2, \ldots, m$. However, this design gives larger values of V_A and V_P (although a smaller value of V_D) than an alternative OFAT design that fixes the factor combinations B_1C_1, A_2B_2 and

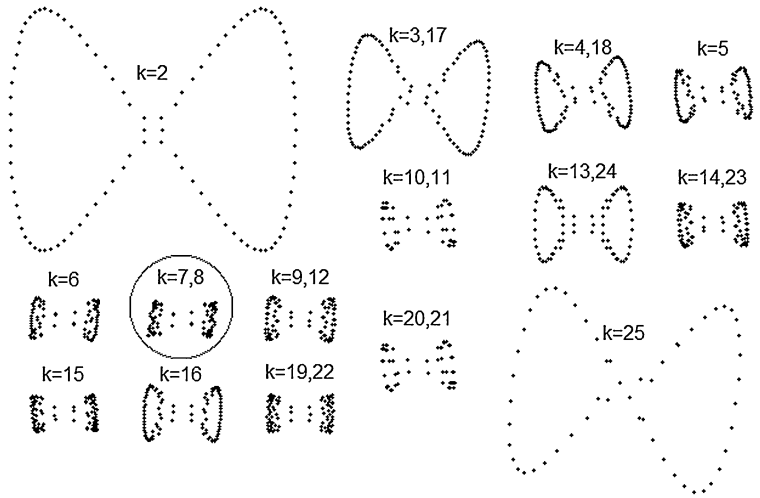

FIG. 4. *A butterfly plot for $m = 50$. The mean distance between points within each wing is representative of V_A and the mean distance between a point on one wing and a point on the other wing is representative of V_P. The circle indicates the optimal choice of k.*

A_1C_2 while varying the factors A, B and C in turn across all their levels. For this latter design, the following explicit expressions may be obtained:

$$V_A^{\mathrm{OFAT}} = \frac{2(3m^2 - 5m + 1)}{3m(m-1)},$$

$$V_P^{\mathrm{OFAT}} = \frac{11m^2 - 18m + 3}{3m^2},$$

$$V_D^{\mathrm{OFAT}} = \left(\frac{5m+2}{12m}\right)^{1/(m-1)}.$$

It follows that as $m \to \infty$, $E_A^{\mathrm{OFAT}} \to 1/3$, $E_D^{\mathrm{OFAT}} \to 1/3$ and $E_P^{\mathrm{OFAT}} \to 3/11$.

3.3. *Comparison of the 3-factor sawtooth and dumbbell designs.* Figure 5 shows efficiencies of the 3-factor OFAT design and the optimal 3-factor cyclic sawtooth design for $m \leq 150$. For the cyclic design, the values of E_A and E_P are similar and less than E_D, whereas for the OFAT design the values of E_A and E_D are similar and greater than E_P. Over this range of m values, the sawtooth design is considerably more efficient than the OFAT design in terms of E_P and E_D, but, in terms of E_A, the OFAT design becomes more efficient for $m > 120$.

4. Application to lead optimization. We utilize data generated and analyzed by Pickett et al. that are available at http://pubs.acs.org. The data are intended to be typical of data arising in the process of *lead optimization* in drug development,

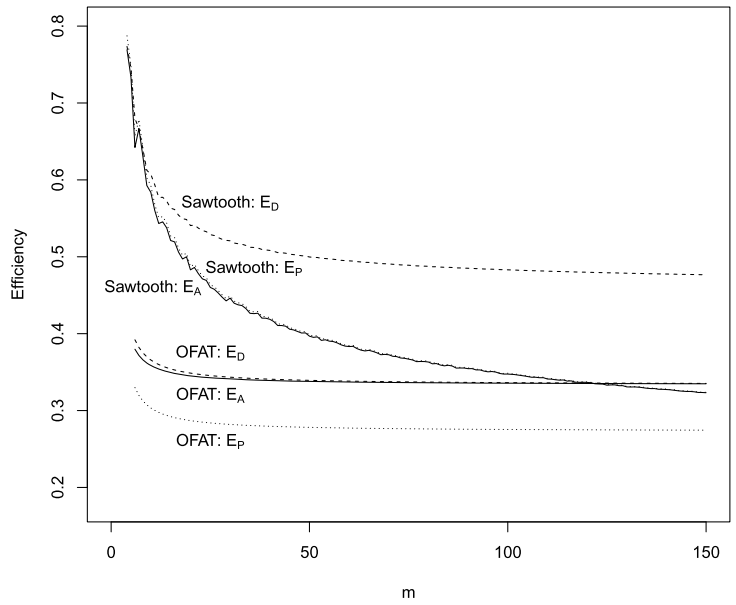

FIG. 5. *Efficiency measures for the optimal* 3*-factor cyclic sawtooth design and the* 3*-factor OFAT design*

where a promising compound (the *lead*) is improved by chemical modifications. In this application the compound can be modified at two sites, A and B, and 50 possible modifications (chemical reagents) were considered at each site. Thus, A and B are the factors and the different reagents are the levels of the factors. The basic compound was an inhibitor and the aim was to synthesize all 50×50 possible modifications and measure their inhibitory strength (pIC_{50}) by means of an assay. However, the "complete" data matrix has 796 missing values (32%), mostly because the modified compound could not be synthesized (23%) or because it was found to be inactive (7%), but also because occasionally the assay failed (1%) or was not undertaken (1%).

The paper of Pickett et al. (2011) also goes on to apply sequential design through genetic algorithms to achieve good designs with at most 140 observations. A similar brief directs Borrotti et al. (2014) to develop algorithms for sequential design combining both genetic algorithm and statistical modeling ideas. We have instead concentrated on statistical design and inference ideas for fixed sample size design, which may be augmented at a later stage by further samples to explore the design space and modeling alternatives in the light of information gained.

In the analysis below, we exclude the five levels of each factor that have the largest number of missing values (A10, A13, A20, A23, A36; B26, B30, B36, B45, B50) to leave a 45×45 array; most of the values omitted (397/426) were where the modified compound had not been synthesized and the overall incidence

of missing values is reduced to 18%. The motivation for excluding these factor levels was to give a fairer comparison of different designs, as explained below.

Fitting an additive model by adding the factors in either order yields the following analyses of variance:

Term	d.f.	MS	Term	d.f.	MS
A	44	2.92	B	44	31.29
$+B$	44	31.71	$+A$	44	3.34
Residual	1566	0.13	Residual	1566	0.13

Evidently, any interaction effects are small in relation to the magnitude of the main effects. Moreover, while the factors are not orthogonal due to the minimal design and missing values, it is clear that varying the chemical structure at position B has a much greater effect on the strength of inhibition than varying the structure at position A. Indeed, fitting factor B on its own explained about 80% of the variation, whereas factor A on its own explained only 7.5%.

We consider designs with $2m = 90$ points, except for the dumbbell design where there are no replicated values for the anchor point in the data set and the design therefore has only 89 points. While Section 2 has given the theoretical efficiencies of the sawtooth and dumbbell designs, these results assume that there are no missing values. Given that 18% of the data are missing, no design will be able to estimate all of the main effect parameters and different criteria are required for comparing designs. Moreover, the performance of a design depends on the particular alignment of missing values. We therefore generated 1000 designs of each type for comparison and analyzed the available data for each design as outlined below.

For the dumbbell design, we assume that background pharmaceutical knowledge will enable us to select an anchor point that does not yield a missing response. Alternatively, this inability to synthesize could be checked by testing a single compound prior to the main experiment. If successful, then no extra compounds would be needed. If unsuccessful, then further compounds would need to be tested. However, with a proportion of $q = 0.18$ compounds unable to be made, the expected waiting time is only $q/(1-q) = 0.22$ extra compounds. Therefore, for the simulation we selected the anchor point at random from all treatment combinations with a nonmissing response. We note also that the dumbbell here has potentially 1 less observation than the sawtooth and other designs since there is no replication of the anchor point. The remainder of the design is determined by the anchor point. For the sawtooth, we used the design on the left of Figure 1 (extended to $m = 50$), after first applying a random permutation to the levels of A and B as outlined in Section 2.1.

In addition, we considered two further designs. The first simply chooses $2m = 90$ treatment combinations at random, without replacement. The second is the initial design of Borrotti et al. who explored evolutionary designs based on the same

data set. Their initial $2m$ observation fixed design generates treatment combinations by pairing each level of factor A with a randomly chosen level of factor B, and each level of factor B with a randomly chosen level of factor A. It is unclear whether the random choice of levels of the second factor is made with or without replacement, but we have assumed that this choice is done without replacement, which gives slightly better performance.

Neither of these latter designs would be desirable in the absence of missing values, since the probability that the resulting design can estimate all main effect parameters is small. For the design of Borrotti et al., it follows from Lemma 1 that this probability is $1/m$, since this proportion is the fraction of all permutations of m objects that consist of just a single cycle. For the completely random design, the probability would be much smaller. However, in the presence of missing values, where no design can estimate all of the main effect parameters, these designs may be more competitive.

Designs were compared using the following criteria:

C1 The number of nonmissing values in the design (maximum 90).

C2 The proportion of the 45^2 ($= 2025$) treatment combinations for which predictions can be made.

C3 The standardized root mean squared prediction error, calculated as RMSPE$/\hat{\sigma}$, where $\hat{\sigma} = 0.3571$ is the square root of the mean square error from fitting the additive main effect model to the full set of 45^2 data points. The RMSPE is calculated for all treatment combinations for which prediction is possible, but excluding the combinations that are present in the design.

C4 The average true response of the 10 treatment combinations that are *predicted* to have the highest response. This measure is used by Pickett et al.

C5 The Spearman rank correlation between the parameter estimates $\hat{\alpha}_i$ from fitting the main effect of factor A to the data from the design and the corresponding parameter estimates from fitting the main effect of factor A to the full data set. A high value indicates reasonable agreement between the data from the experimental design and the full data set in terms of the rank ordering of the A parameters.

C6 Similar to C5, but based on factor B.

Figure 6 shows boxplots of these different measures, summarizing the 1000 individual designs of each type. The average number of nonmissing values (C1) was broadly similar for all types of design, though the values in individual experiments were more variable for the dumbbell design. If the full data set is used, without omitting the factor levels with the most missing values, the dumbbell design usually has fewer missing values than the other designs, because the dumbbell design is chosen conditionally on the anchor point having a nonmissing response. Excluding the factor levels with the most missing values means that the designs are at least broadly comparable.

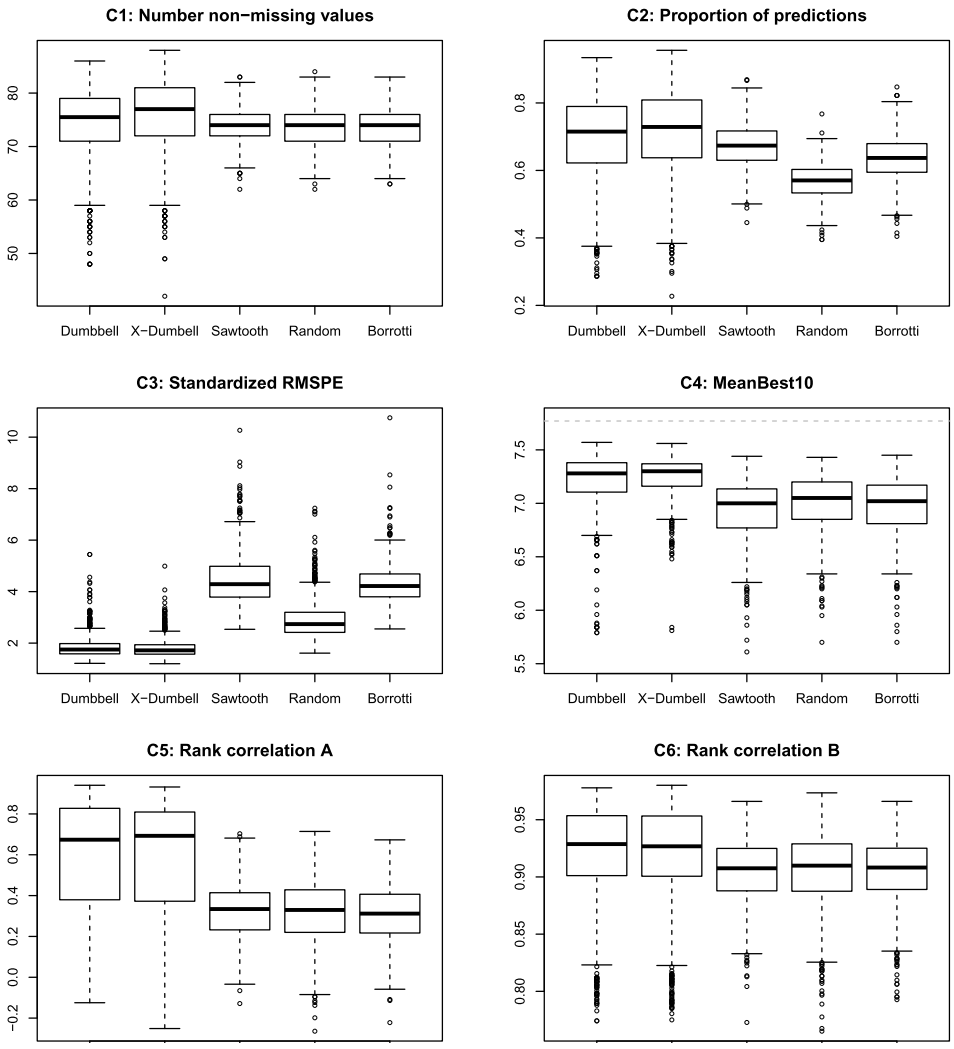

FIG. 6. *Boxplots for* 1000 *simulations and* $m = 45$ *levels of A and B*: *comparative criteria*, (C3) *root mean square prediction error*, (C4) *best ten predictors*, (C5, C6) *rank correlation of predictions*, (C1, C2) *direct effect of missingness*, *for five types of design when applied to the lead optimization activity data set*.

Although it has little advantage in terms of missing values, the structure of the dumbbell design allows it to predict responses for a higher proportion of the combinations of factor levels (C2) on average than the other designs, though it occasionally performs poorly. The purely random design, which has the least structure of the designs considered, performs poorly on this criterion. Despite making more predictions on average than other designs, the dumbbell gives the most accurate

predictions, as judged by the RMSPE (C3). The purely random design does better than the other two designs on this criterion, but it typically makes fewer predictions than the other designs.

On average, the mean response of the predicted top-10 treatment combinations (C4) is higher for the dumbbell design than for the other designs, which all perform similarly. The dumbbell design also has the best performance in terms of rank correlation (C5, C6). Rank correlation values are considerably higher for factor B than for factor A, reflecting the fact that the main effect parameter estimates differ more for factor B than for factor A in the full data set.

Overall, the dumbbell design appears to be the most satisfactory design, although occasionally it performs poorly due to missing data. This behavior is when an anchor point is chosen involving an A or B level (or both) for which there are very few compound combinations. The sawtooth design and the initial fixed design of Borrotti et al. perform similarly. This concordance is not surprising since the designs are similar in general structure. In the absence of missing data, the sawtooth design performs at least as well and usually better, but the presence of nearly 20% missing data removes this advantage. The random design is not recommended because typically it leads to fewer predictions than other designs.

These design comparisons remain pertinent to a wider inferential process in which further responses are compared (see Section 2).

5. Discussion. We have described approaches to design in situations where there are few factors but each is at many levels. These designs assume main-effects models. They are evidently good for the data in the application which shows little interaction. The fm observations or, more generally, $\sum_{i=1}^{f} m_i$ with m_i levels to the ith factor, are near minimal with just one spare degree of freedom when there are two factors. Various strategies are natural for augmenting such designs. One might surmise that interactions are much more likely between factor levels which have sizeable main effects. There might be information on quantitative variables, for example, solubility, molecular size and dipole moment, that underlie the factors, and they may enable grouping of levels through reagent similarity. All these considerations can be part of an augmentation process which is enabled and facilitated by the minimal designs we have described. Much is known about two-factor designs, but for three-factor and higher designs many questions are still open and even connectivity issues can be problematic [Wynn (2008)]. The appeal of cyclic designs [see John and Williams (1995)] is that they are connected and easy to analyze, and often near optimal, although not necessarily optimal (see below).

The sequential genetic algorithms of Pickett et al. (2011) and the hybrid versions such as Borrotti et al. (2014) take a more machine learning approach and the algorithms involve ideas with biological roots such as “crossover” and “mutation” of levels. In the genetic algorithm of Pickett et al. (2011) there are 10 equal sized generations, each of 14 observations, allowing a total experimental design of 140

observations. Borrotti et al. (2014), on the other hand, take a large initial sample of 100, as detailed in Section 4. This sample size potentially allows estimation of all main effects provided there are (i) no missing data and (ii) no internal cycles as discussed in Lemma 1. This initial sample acts as a base for subsequent data selection guided by random forest modeling, and further sets of 10 observations are generated with just four generations allowed by the total resource restriction of 140 observations.

The potential Achilles heel of the dumbbell design is the need for an anchor point. However, it is clear from Pickett et al. (2011) that it need not be a problem as there is sufficient pharmaceutical knowledge about the availability of some compounds. In fact, they seed their first generation with two compounds which were known to be active. When little is known, a pretest of the anchor point can ensure that the compound can be made.

Although the dumbbell is a one-factor-at-a-time (OFAT) design, with suboptimal properties in general, in this two-factor case we have proved it to be P-optimal and near A-optimal, with a slight modification giving strict A-optimality.

Our use of a particular class of generators for 3-factor cyclic designs was motivated by an exhaustive search of generators for factors with up to 12 levels. However, we emphasize that these designs are not necessarily optimal outside the class of cyclic designs. A concrete example of a slightly more efficient design for $m = 4$ is

Factor A	1	1	1	2	2	2	3	3	3	4	4	4
Factor B	1	3	4	1	2	4	2	3	4	1	2	3
Factor C	1	3	2	2	3	4	4	1	3	4	1	2

Like the cyclic design, it has equal replication of the levels of each factor, but, unlike the cyclic design, these cannot be arranged into complete "sets." The efficiency factors for this noncyclic design, with those of the optimal cyclic design in brackets, are $E_D = 0.777$ (0.774), $E_A = 0.776$ (0.769) and $E_P = 0.793$ (0.787). An interesting question for future research is whether noncyclic designs can give more substantial gains in efficiency for larger m.

As we have mentioned above, our statistical approach has focused on the first stage of experimentation and further observations may be added to tap into data missing by virtue of inability to synthesize compounds or possible interactions. A different supplementation in light of first phase analysis would take the k_A "significant" A effects and k_B significant B effects, ordering both sets of effects separately according to their size. It would then add a further supplementary set of $\max(k_A, k_B)$ additional observations, pairing the k_A levels of A with those of B in order, recycling if necessary, adding a sawtooth or even completely crossing the two sets if there are few of them. More formal decision theoretic methods or ones based on entropy [Meyer, Steinberg and Box (1996)] are also a possible rich area for further research, especially in the context of multiple responses and their competing demands. Our designs can offer an important ingredient to the process of lead optimization.

Acknowledgments. We are grateful to Dr. Darren Green and researchers at GlaxoSmithKline Research and Development, UK, for their advice and discussions on features of the data and to Professor Irene Poli and researchers at the European Centre for Living Technology, Universita Foscari, Venice, for discussions on Evolutionary Model Design. The first author is grateful for discussions with Professor Henry Wynn which led to clarification of connectivity issues. We also thank a referee and Associate Editor, whose comments led to considerable improvements to the paper.

SUPPLEMENTARY MATERIAL

Supplement: Level-screening, proofs, correlations and code (DOI: 10.1214/16-AOAS916SUPP; .pdf). The Supplement contains proofs of Lemma 1 and Theorem 1, results about correlations between estimators and details of efficient computations for the 3-factor sawtooth design.

REFERENCES

ATKINSON, A. C. and DONEV, A. N. (1992). *Optimum Experimental Designs*. Oxford Univ. Press, Oxford.

BAILEY, R. A. (2007). Designs for two-colour microarray experiments. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **56** 365–394. MR2409757

BALANI, S. K., MIWA, G. T., GAN, L. S. and LEE, F. W. (2005). Strategy of utilizing in vitro and in vivo ADME tools for lead optimisation and drug candidate selection. *Current Topics in Medicinal Chemistry* **5** 1033–1038.

BORROTTI, M., MARCH, D. D., SLANZI, D. and POLI, I. (2014). Designing lead optimisation of MMP-12 inhibitors. *Comput. Math. Methods Med.* **2014** 258627.

BOX, G. E. P., HUNTER, J. S. and HUNTER, W. G. (2005). *Statistics for Experimenters*: *Design*, *Innovation*, *and Discovery*, 2nd ed. Wiley-Interscience, Hoboken, NJ. MR2140250

BROWN, P. J. and RIDOUT, M. S. (2016). Supplement to "Level-screening designs for factors with many levels." DOI:10.1214/16-AOAS916SUPP.

DANIEL, C. (1973). One-at-a-time plans. *J. Amer. Statist. Assoc.* **68** 353–360.

DETTE, H. and O'BRIEN, T. E. (1999). Optimality criteria for regression models based on predicted variance. *Biometrika* **86** 93–106. MR1688074

FRANCK, C. T., NIELSEN, D. M. and OSBORNE, J. A. (2013). A method for detecting hidden additivity in two-factor unreplicated experiments. *Comput. Statist. Data Anal.* **67** 95–104. MR3079590

HALL, W. B. and WILLIAMS, E. R. (1973). Cyclic superimposed designs. *Biometrika* **60** 47–53. MR0359203

JOHN, J. A. and WILLIAMS, E. R. (1995). *Cyclic and Computer Generated Designs*, 2nd ed. *Monographs on Statistics and Applied Probability* **38**. Chapman & Hall, London. MR1382127

KERR, M. K. and CHURCHILL, G. A. (2001). Experimental design for gene expression microarrays. *Biostatistics* **2** 183–201.

MEYER, R. D., STEINBERG, D. M. and BOX, G. E. P. (1996). Follow-up designs to resolve confounding in multifactor experiments (with discussion). *Technometrics* **38** 303–313.

PICKETT, S. D., GREEN, D. V. S., HUNT, D. L., PARDOE, D. A. and HUGHES, I. (2011). Automated lead optimisation of MMP-12 inhibitors using a genetic algorithm. *ACS Medicinal Chemistry Letters* **2** 28–33.

THAYER, A. M. (1996). Combinatorial chemistry becoming a core technology at drug development companies. *Chem. Eng. News* **74** 57–64.

TJUR, T. (1991). Block designs and electrical networks. *Ann. Statist.* **19** 1010–1027. MR1105858

VINCIOTTI, V., KHANIN, R., D'ALEMONTE, D., DE JESUS, O., RASAIYAAH, J., SMITH, C. P., KELLAM, P. and WIT, E. (2005). An experimental evaluation of a loop versus a reference design for two channel microarrays. *Bioinformatics* **21** 492–501.

WYNN, H. P. (2008). Algebraic solutions to the connectivity problem for m-way layouts: Interaction-contrast aliasing. *J. Statist. Plann. Inference* **138** 259–271. MR2369631

SMSAS
UNIVERSITY OF KENT
CANTERBURY, KENT, CT2 7NF
UNITED KINGDOM
E-MAIL: pjb8@kent.ac.uk
msr@kent.ac.uk