Supplementary sections to

Projections of Definitive Screening Designs by Dropping Columns: Selection and Evaluation

Alan R. Vazquez[1, 2], Peter Goos[1, 2], and Eric D. Schoen[1, 3]

[1]KU Leuven, Belgium
[2]University of Antwerp, Belgium
[3]TNO, Zeist, Netherlands

November 8, 2018

This document includes the following sections:

A Details on the simulation protocol

B Design properties and derivations

C Table with detailed results

A Details on the simulation protocol

In this section, we first introduce the model used for the simulation study in Section 2 of the main text. Next, we outline our simulation protocol and discuss the two-step approach of Jones and Nachtsheim (2017), which we used to analyze the data simulated from projected definitive screening designs (pDSDs). The pDSDs under study in the simulations are constructed by dropping four columns from the 21-run 10-factor standard definitive screening design (sDSD) in Table 1 of the main text. More specifically, pDSD_a is constructed by dropping the last four columns from this sDSD, while pDSD_b is constructed by dropping the columns C_6, C_8, C_9 and C_{10}.

A.1 Model

The model assumed for the motivating example in Section 2 is

$$\begin{aligned} Y_i = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_1 X_3 + \\ \beta_7 X_1 X_4 + \beta_8 X_2 X_3 + \beta_9 X_2 X_4 + \beta_{10} X_3 X_4 + \beta_{11} X_1^2 + \epsilon_i, \end{aligned} \tag{S1}$$

where Y_i denotes the i-th response, X_1, X_2, X_3 and X_4 denote the four active factors, and $\epsilon_i \sim N(0, 1)$. The two inactive factors are denoted by X_5 and X_6. This model has four linear effects (LEs), six two-factor interactions (TFIs) and one quadratic effect (QE).

A.2 Simulation protocol

For each pDSD option, each of our 1,000 simulations consisted of the following steps:

1. We assigned the active factors to four columns of the design. The inactive factors are arbitrarily assigned to the remaining two columns. We conducted the simulation study for each of the $\binom{6}{4} = 15$ possible assignments of the active factors to the columns. Table S1 shows one possible factor assignment.

2. We obtained the coefficients, β_j, for the active effects by randomly sampling (with replacement) 11 values from the set $\{-2, +2\}$. The SNRs are therefore 2 for all active effects.

3. Let $\mathbf{X}$ be the matrix consisting of the columns corresponding to the active effects in model (S1) and $\boldsymbol{\beta}$ be the vector of coefficients for the active effects. We generated a response vector as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with $\epsilon_i \sim N(0, 1)$.

4. The set of effects declared active was determined using the approach of Jones and Nachtsheim (2017), described in Section A.3.

After the 1,000 simulations, we calculated the powers for the active effects and type-I error rates for the inactive effects. We calculated the power of an active effect as the fraction of the simulations for which it was declared active. We calculated the type-I error rate of an inactive effect as the fraction of the simulations for which it was declared active. An R implementation of our simulation protocol is included in the supplementary files.

A.3 Analysis strategy

The set of effects declared active was determined using the two-step approach of Jones and Nachtsheim (2017). In the first step of these authors' method, standard significance tests are used to identify the active LEs. To this end, the k columns that are dropped from the $(m+k)$-factor sDSD, are used to provide an unbiased estimate of the error variance based on k degrees of freedom. In the second step, all subsets regression is used to identify the active TFIs and QEs among those factors identified in the first step. More specifically, the best subset of second-order effects is determined using the residual sum of squares. The largest subset size to be explored is $(m+k)/2$ but the search can be stopped prematurely using significance tests which compare subsets of different sizes.

For our simulations, we use the settings recommended by Jones and Nachtsheim (2017), with one exception. In the second step, we use seven, the number of active second-order effects, as the maximum subset size to be considered in all subsets regression, instead of $(m+k)/2 = 5$. Note also that, for pDSD$_a$, we use the columns C_7-C_{10} from the 10-factor sDSD in Table 1 of the main text to compute the unbiased estimate of the error variance in the first step. For pDSD$_b$, we use the columns C_6, C_8, C_9 and C_{10}.

A.4 Simulation results

We addressed the detection of the active LEs and second-order effects separately. One of the results was that both pDSDs had powers for the active LEs equal to one for all factor assignments. The type-I error rates for the inactive LEs were very close to zero for both designs and all factor assignments; the maximum type-I error rate for these effects was 0.07. So, both pDSDs were comparable when considering the power and FDR for the LEs.

Table S2 shows the powers and type-I error rates for the second-order effects for both pDSDs. The first column of the table shows the assignment of the active factors to the columns in the designs. The second and fourth column show the ranges of the powers for the six active TFIs and the active QE for pDSD$_a$ and pDSD$_b$, respectively. The third and and fifth column show the maximum type-I error rates for the inactive second-order effects for pDSD$_a$ and pDSD$_b$, respectively.

Table S1: Example of an assignment of the active factors to the columns in pDSD$_a$ and pDSD$_b$. The active factors X_1, X_2, X_3 and X_4, are assigned to the last columns of the designs. The inactive factors X_5 and X_6 are arbitrarily assigned to the first two columns. pDSD$_a$: design constructed by dropping the last columns of the 10-factor 21-run sDSD in Table 1 of the main text; pDSD$_b$: design constructed by dropping the columns C_6, C_8, C_9 and C_{10}.

(a) pDSD$_a$

X_5	X_6	X_1	X_2	X_3	X_4
C_1	C_2	C_3	C_4	C_5	C_6
0	1	1	1	1	1
1	0	−1	−1	−1	−1
1	−1	0	−1	1	1
1	−1	−1	0	1	1
1	−1	1	1	0	−1
1	−1	1	1	−1	0
1	1	−1	1	−1	1
1	1	−1	1	1	−1
1	1	1	−1	−1	1
1	1	1	−1	1	−1
−1	−1	−1	1	−1	1
−1	−1	−1	1	1	−1
−1	−1	1	−1	−1	1
−1	−1	1	−1	1	−1
−1	1	−1	−1	1	0
−1	1	−1	−1	0	1
−1	1	1	0	−1	−1
−1	1	0	1	−1	−1
−1	0	1	1	1	1
0	−1	−1	−1	−1	−1

(b) pDSD$_b$

X_5	X_6	X_1	X_2	X_3	X_4
C_1	C_2	C_3	C_4	C_5	C_7
0	1	1	1	1	1
1	0	−1	−1	−1	1
1	−1	0	−1	1	−1
1	−1	−1	0	1	1
1	−1	1	1	0	−1
1	−1	1	1	−1	1
1	1	−1	1	−1	0
1	1	−1	1	1	−1
1	1	1	−1	−1	−1
1	1	1	−1	1	1
−1	−1	−1	1	−1	−1
−1	−1	−1	1	1	1
−1	−1	1	−1	−1	1
−1	−1	1	−1	1	0
−1	1	−1	−1	1	−1
−1	1	−1	−1	0	1
−1	1	1	0	−1	−1
−1	1	0	1	−1	1
−1	0	1	1	1	−1
0	−1	−1	−1	−1	−1

Table S2 shows that pDSD_a has powers smaller than 0.42 and maximum type-I error rates larger than 0.58 for three factor assignments. For the other 12 assignments, pDSD_a has powers around 0.47-0.98 and maximum type-I error rates smaller than or equal to 0.38. In contrast, pDSD_b has powers smaller than 0.4 for only two of the 15 possible assignments of the active factors to the columns. For the rest of the assignments, this design had powers around 0.46-0.98 and maximum type-I error rates smaller than or equal to 0.37. The fact that pDSD_b had powers in the range 0.46-0.98 and maximum type-I error rates smaller than 0.37 for more factor assignments than pDSD_a, indicates that dropping the columns C_6, C_8, C_9 and C_{10} from the 10-factor sDSD is a better option than dropping the last columns.

B Design properties and derivations

In this section, we derive the properties of pDSDs constructed by dropping columns from sDSDs. For notational simplicity, let $n = m + k$ and assume that the pDSD with m factors and $N = 2n + 1$ runs is constructed by dropping k columns from an n-factor sDSD. Note that, if $k = 0$, then the design is the original sDSD; otherwise, it is a pDSD.

B.1 Models with main effects only

The D-efficiencies for models with m LEs and the standard errors for the LE estimates in the models do not depend on the set of k columns dropped from an n-factor sDSD, because (i) the LEs' contrast vectors in a sDSD are orthogonal to each other and to the column of ones in the model matrix (corresponding to the intercept) and (ii) the precision is the same for each LE estimate.

Consider the $N \times (m+1)$ matrix $\mathbf{X}_{\text{l}}$ including the intercept and all LE contrast vectors of an m-factor pDSD. It is easy to show that $\mathbf{X}_{\text{l}}^T\mathbf{X}_{\text{l}}$ is a diagonal matrix with determinant $|\mathbf{X}_{\text{l}}^T\mathbf{X}_{\text{l}}| = [2n+1][2n-2]^m$. For an m-factor sDSD, this determinant equals $(2m+1)(2m-2)^m$. Therefore, the relative D-efficiency of a N-run pDSD and a $(2m+1)$-run sDSD for a model with all the LEs is:

$$D_{\text{le}} = \left[\frac{2n+1}{2m+1}\left(\frac{2n-2}{2m-2}\right)^m\right]^{\frac{1}{m+1}} = \left(1+\frac{2k}{2m+1}\right)^{\frac{1}{m+1}}\left(1+\frac{k}{m-1}\right)^{\frac{m}{m+1}}. \tag{S2}$$

Table S2: Simulation results for each of the 15 possible assignments of the active factors to the columns in the pDSD options. pDSD_a: design constructed by dropping the last columns of the 10-factor 21-run sDSD in Table 1 of the main text; pDSD_b: design constructed by dropping columns C_6, C_8, C_9 and C_{10}.

	pDSD_a		pDSD_b	
Factor assignment	Range power	Maximum type-I error rate	Range power	Maximum type-I error rate
1, 2, 3, 4	0.27-0.42	0.59	0.28-0.40	0.60
1, 2, 3, 5	0.49-0.98	0.35	0.51-0.98	0.35
1, 2, 3, 6	0.50-0.97	0.38	0.51-0.97	0.35
1, 2, 4, 5	0.49-0.97	0.38	0.49-0.96	0.34
1, 2, 4, 6	0.49-0.98	0.35	0.48-0.96	0.36
1, 2, 5, 6	0.28-0.40	0.58	0.51-0.95	0.36
1, 3, 4, 5	0.52-0.98	0.34	0.49-0.95	0.35
1, 3, 4, 6	0.50-0.97	0.35	0.48-0.95	0.37
1, 3, 5, 6	0.50-0.98	0.32	0.46-0.97	0.37
1, 4, 5, 6	0.51-0.98	0.35	0.50-0.94	0.33
2, 3, 4, 5	0.51-0.98	0.35	0.48-0.97	0.35
2, 3, 4, 6	0.50-0.97	0.35	0.50-0.97	0.35
2, 3, 5, 6	0.50-0.97	0.33	0.27-0.40	0.58
2, 4, 5, 6	0.47-0.97	0.34	0.53-0.96	0.36
3, 4, 5, 6	0.27-0.41	0.60	0.48-0.95	0.35

The variances for the ordinary least squares (OLS) estimators of the intercept and all LEs in a pDSD are calculated from the matrix $\sigma^2(\mathbf{X}_{\text{l}}^T\mathbf{X}_{\text{l}}\,)^{-1}$, where σ^2 denotes the variance of the residual errors. It is easy to see that the variance for any LE equals $\sigma^2(2n-2)^{-1}$. For a sDSD, this variance equals $\sigma^2(2m-2)^{-1}$. Thus, the standard error for any LE estimate obtained from a N-run pDSD relative to the standard error produced by an $(2m+1)$-run sDSD equals

$$\text{SE}_{\text{le}} = \left(\frac{2m-2}{2n-2}\right)^{1/2} = \sqrt{\frac{m-1}{m+k-1}}. \tag{S3}$$

It is easy to see from Equations (S2) and (S3) that the relative D-efficiency increases and the relative standard error for a LE estimate decreases with the number k. In other words, the relative D-efficiency increases and the relative standard error for a LE estimate decreases with the run size of the sDSD used to construct the pDSD.

B.2 Models with linear and quadratic main effects

Just as the D-efficiency for a model with LEs only, the D-efficiency of a pDSD for a model with LEs and QEs is insensitive to the set of k columns dropped from an n-factor sDSD. Consider the $N \times (2m+1)$ model matrix $\mathbf{X}_{\text{lq}} = [\mathbf{1}_n, \mathbf{Q}, \mathbf{L}]$, where $\mathbf{1}_n$ is an $N \times 1$ column vector of ones (the intercept column), $\mathbf{Q}$ is the $N \times m$ matrix including the QE contrast vectors, and $\mathbf{L}$ is the $N \times m$ matrix including the LEs' contrast vectors of the pDSD. The information matrix is

$$\mathbf{X}_{\text{lq}}^T\mathbf{X}_{\text{lq}} = \begin{pmatrix} N & 2(n-1)\mathbf{1}_{1\times m} & \mathbf{0}_{1\times m} \\ 2(n-1)\mathbf{1}_{m\times 1} & 2(n-2)\mathbf{J}_{m\times m} + 2\mathbf{I}_{m\times m} & \mathbf{0}_{m\times m} \\ \mathbf{0}_{m\times 1} & \mathbf{0}_{m\times m} & 2(n-1)\mathbf{I}_{m\times m} \end{pmatrix}, \tag{S4}$$

where $\mathbf{I}_{m\times m}$ is the identity matrix of order m, $\mathbf{J}_{m\times m}$ is the $m \times m$ matrix with all its entries equal to 1 and $\mathbf{0}_{p\times q}$ denotes a $p \times q$ matrix of zeroes. Note that matrix (S4) is a block diagonal matrix and can be expressed as

$$\mathbf{X}_{\text{lq}}^T\mathbf{X}_{\text{lq}} = \begin{pmatrix} \mathbf{A} & \mathbf{0}_{(m+1)\times m} \\ \mathbf{0}_{m\times(m+1)} & \mathbf{B} \end{pmatrix},$$

where

$$\begin{aligned}\mathbf{A} &= \begin{pmatrix} N & 2(n-1)\mathbf{1}_{1\times m} \\ 2(n-1)\mathbf{1}_{m\times 1} & 2(n-2)\mathbf{J}_{m\times m}+2\mathbf{I}_{m\times m}\end{pmatrix}, \text{ and} \\ \mathbf{B} &= 2(n-1)\mathbf{I}_{m\times m}.\end{aligned}$$

Using Harville (2011, p.p. 187), we have that $|\mathbf{X}_{\text{lq}}^T\mathbf{X}_{\text{lq}}| = |\mathbf{A}||\mathbf{B}|$. We can calculate the determinant of $\mathbf{A}$ by taking $c_0 = N$, $c = 2(n-1)$, $a = 2(n-2)$ and $b = 2$, and applying lemma $2iii$ of Zhou and Xu (2017). The determinant of the information matrix $|\mathbf{X}_{\text{lq}}^T\mathbf{X}_{\text{lq}}|$ then is $2^{2m}(n-1)^m\left[(m-1)^2+k(m+2)\right]$. For an m-factor sDSD, this determinant equals $2^{2m}(m-1)^{m+2}$ for the information matrix of a model including the intercept, all QEs and all LEs. As a result, the D-efficiency of an N-run pDSD relative to that of a $(2m+1)$-run sDSD for a model including all LEs and all QEs, is:

$$\begin{aligned}D_{\text{le+qe}} &= \left\{\left(\frac{n-1}{m-1}\right)^m\left[\frac{(m-1)^2+k(m+2)}{(m-1)^2}\right]\right\}^{\frac{1}{2m+1}} \\ &= \left(1+\frac{k}{m-1}\right)^{\frac{m}{2m+1}}\left(1+\frac{k(m+2)}{(m-1)^2}\right)^{\frac{1}{2m+1}}\end{aligned}$$

Clearly, this relative efficiency also increases with k and thus with the run size of the pDSD constructed.

Now, consider the information matrix (S4) for a pDSD. Using Harville (2011, p.p. 89), the variances for the OLS estimators of the model including the intercept, all LEs and all QEs in a pDSD can be calculated from

$$\sigma^2\left(\mathbf{X}_{\text{lq}}^T\mathbf{X}_{\text{lq}}\right)^{-1} = \sigma^2\begin{pmatrix}\mathbf{A}^{-1} & \mathbf{0}_{(m+1)\times m} \\ \mathbf{0}_{m\times(m+1)} & \mathbf{B}^{-1}\end{pmatrix}.$$

Note that sub-matrix $\sigma^2\mathbf{A}^{-1}$ contains the variances for the OLS estimators of the QEs. Consider the following sub-matrices

$$\mathbf{W} = N, \mathbf{T} = 2(n-2)\mathbf{J}_{m\times m}+2\mathbf{I}_{m\times m}, \text{ and } \mathbf{V} = \mathbf{U}^T = 2(n-1)\mathbf{1}_{1\times m},$$

that partition matrix $\mathbf{A}$ into four parts. Using Harville (2011, p.p. 99) and Lemma $2i$ of Zhou and Xu (2017), it is straightforward to show that the variance of the OLS estimator

for any QE, denoted as $\hat{\beta}_{ii}$, in a model including the intercept, all LEs and all QEs based on a pDSD is

$$\mathrm{Var}(\hat{\beta}_{ii}) = \sigma^2 \frac{(m-1)^2 + m(k-1) + k + 4}{2\left[(m-1)^2 + k(m+2)\right]},$$

For an m-factor sDSD, $\mathrm{Var}(\hat{\beta}_{ii}) = \sigma^2(m^2-3m+5)/(2(m-1)^2)$. The standard error of any QE estimate obtained from a N-run pDSD relative to the one produced by a $(2m+1)$-run sDSD is therefore

$$\begin{aligned} \mathrm{SE}_{\mathrm{qe}} &= \left\{ \left[\frac{(m-1)^2}{(m-1)^2 + k(m+2)} \right] \left[\frac{m^2 - 3m + 5 + k(m+1)}{m^2 - 3m + 5} \right] \right\}^{1/2} \\ &= \frac{m-1}{\sqrt{(m-1)^2 + k(m+2)}} \sqrt{1 + \frac{k(m+1)}{m^2 - 3m + 5}}. \end{aligned}$$

For any given number of factors m, this relative standard error decreases with k and thus with the run size of the pDSD. The relative standard error approaches

$$(m-1)\sqrt{(m+1)/[(m^2-3m+5)(m+2)]}$$

for large values of k. The relative standard errors for the LE estimates are not affected by the inclusion of QEs in the model. Therefore, they can still be calculated using Equation (S3). This is because the LEs' contrast vectors are orthogonal to those of the QEs whenever a sDSD or a pDSD is used.

B.3 Correlations between specific second-order effects' contrast vectors

Regardless of which k columns are dropped from an $(m+k)$-factor sDSD, the correlation between the contrast vectors of any two QEs β_{ii} and β_{jj} equals

$$r_{ii,jj} = \frac{1}{3} - \frac{2}{N-3},$$

This correlation increases with the run size and approaches 1/3 for large values of N or k. So, the QEs' contrast vectors of m-factor pDSDs exhibit larger correlations when they are based on larger sDSDs and involve more runs.

The correlation between the contrast vector of any QE β_{ii} and the contrast vector of a TFI β_{ij} is always zero. The correlation between the contrast vector of a QE β_{ii} and the

contrast vector of a TFI β_{jk} is non-zero, and equals

$$r_{ii,jk} = \pm\sqrt{\frac{4N}{3(N-3)(N-5)}}.$$

The expressions for $r_{ii,jj}$ and $r_{ii,jk}$ are obtained by substituting $m+k$ for m in the corresponding expressions given in Jones and Nachtsheim (2011).

If $m+k$ is a multiple of 4, the three correlations $r_{ii,jk}$, $r_{jj,ik}$ and $r_{kk,ij}$ corresponding to any triplet of factors i, j and k are all positive, or one correlation is positive, while the other two are negative. If $m+k$ is not a multiple of 4, two of the three correlations are positive and one is negative, or all three correlations $r_{ii,jk}$, $r_{jj,ik}$ and $r_{kk,ij}$ are negative (Schoen et al., 2018). The correlations tend to zero as the run size N or k increases. So, when using an m-factor pDSD, the correlations between the contrast vectors for the QE of one factor and the interaction between two other factors are closer to zero than when using an m-factor sDSD. In other words, the aliasing is smaller.

Finally, the correlation between the contrast vector of a TFI β_{ij} and the contrast vector of another TFI β_{ik}, involving the same factor i, is non-zero too. To show this, consider a TFI contrast vector $\mathbf{x}_{ij}$ formed by the element-wise multiplications of the LE contrast vectors i and j and a TFI contrast vector $\mathbf{x}_{ik}$ formed by the element-wise multiplication of the LE contrast vectors i and k in the pDSD. Denote the elements in $\mathbf{x}_{ij}$ and $\mathbf{x}_{ik}$ as $x_{s,ij}$ and $x_{s,ik}$, respectively; $s = 1, \ldots, N$. Note that the average of these elements, denoted as $\bar{x}_{s,ij}$, equals zero and that their sum of squares equals $2(m+k) - 4 = 2n - 4$. The correlation between the contrast vectors corresponding to β_{ij} and β_{ik} then is

$$r_{ij,ik} = \frac{\sum \left[x_{s,ij} - \bar{x}_{ij}\right]\left[x_{s,ik} - \bar{x}_{ik}\right]}{\sqrt{\sum \left[x_{s,ij} - \bar{x}_{ij}\right]^2 \sum \left[x_{s,ik} - \bar{x}_{ik}\right]^2}} = \frac{\sum x_{s,ij}x_{s,ik}}{\sqrt{\sum \left(x_{s,ij}\right)^2 \sum \left(x_{s,ik}\right)^2}} = \frac{\sum x_{s,ij}x_{s,ik}}{2n-4},$$

where all sums run from $s = 1, \ldots, N$. Note that $\sum x_{s,ij}x_{s,ik} = \sum x_{s,ii}x_{s,jk}$ in the above expression, where $x_{s,ii}$ is the s-th element of the QE contrast vector $\mathbf{x}_{ii}$. It is easy to show that $\sum x_{s,ii}x_{s,jk} = \pm 2$. If we substitute $\sum x_{s,ij}x_{s,ik} = \pm 2$ and $2n+1$ by N in the expression above and simplify, we obtain the following expression for the correlation:

$$r_{ij,ik} = \pm\frac{2}{N-5}.$$

The correlations $r_{ij,ik}$, $r_{ji,jk}$ and $r_{ki,kj}$ exhibit the same patterns as the correlations $r_{ii,jk}$, $r_{jj,ik}$ and $r_{kk,ij}$. They also decrease with the run size. So, they are closer to zero for an

m-factor pDSD than for an m-factor sDSD, which means that a pDSD reduces the aliasing between two interaction effects β_{ij} and β_{ik}.

B.4 Power calculations

In Sections 3.1.1 and 3.1.2 of the main text, powers for t tests of the hypotheses L_1, L_{me}, Q_{me}, I_{me}, Q_3 and I_3 are discussed. The statistical power provided by the pDSD to test these hypotheses is computed as Power $= 1 - \text{Prob}(-\text{t}_{\nu,\alpha/2} < T_{\nu,\lambda} < \text{t}_{\nu,\alpha/2})$, where $T_{\nu,\lambda}$ is a random variable following a non-central t-distribution with ν degrees of freedom and non-centrality parameter λ, and $-\text{t}_{\nu,\alpha/2}$ and $\text{t}_{\nu,\alpha/2}$ are the critical values based on a central t-distribution with ν degrees of freedom for a significance level equal to α. For all the hypothesis tests $\nu = N - p$, where p is the number of parameters included in the model. The non-centrality parameter of the t distribution is given by

$$\lambda = \frac{\beta_i/\sigma}{\sqrt{\text{Var}(\hat{\beta}_i)}},$$

where the OLS estimate $\hat{\beta}_i$ is computed for a given model. Showing that the power calculations are independent of the set of k columns to drop from the $(m+k)$-factor sDSD is equivalent to showing that the value of $\text{Var}(\hat{\beta}_i)$, and thus of λ, is the same for any subset. We show below that λ only depends on the values of m and k.

B.4.1 Power for L_1.

Since the all LE contrast vectors are orthogonal to the intercept, it is easy to see that the variance for the OLS estimator of any LE is $\sigma^2(2(m+k)-2)^{-1}$. Since only the intercept and a LE are included in the model, $\nu = N - 2 = 2(m+k) - 1$, and the expression for the power calculations follows.

B.4.2 Power for L_{me}.

Using the model matrix $\mathbf{X}_{\text{l}}$ and the calculations in Section B.1, it is easy to show that the variance for the OLS estimator is the same as for L_1. Since the number of parameters already in the model is one plus the number of factors, the expression for the power calculations follows.

B.4.3 Power for $Q_{\mathbf{me}}$.

Consider the $N \times (m+2)$ matrix, $\mathbf{X}_{\mathrm{q}m}$, including the intercept column, one QE contrast vector and all LE contrast vectors of the pDSD. For whatever QE contrast vector is chosen, the information matrix is

$$\mathbf{X}_{\mathrm{q}m}^T \mathbf{X}_{\mathrm{q}m} = \begin{pmatrix} N & 2(n-1) & \mathbf{0}_{1\times m} \\ 2(n-1) & 2(n-1) & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & \mathbf{0}_{m\times 1} & 2(n-1)\mathbf{I}_{m\times m} \end{pmatrix},$$

where $n = m + k$. Note that this matrix is a block diagonal matrix with blocks

$$\mathbf{A} = \begin{pmatrix} N & 2(n-1) \\ 2(n-1) & 2(n-1) \end{pmatrix}, \text{ and } \mathbf{B} = 2(n-1)\mathbf{I}_{m\times m},$$

Using Harville (2011, p.p. 89), we can easily calculate the inverse of the information matrix:

$$\left(\mathbf{X}_{\mathrm{q}m}^T \mathbf{X}_{\mathrm{q}m}\right)^{-1} = \begin{pmatrix} \mathbf{A}^{-1} & \mathbf{0}_{2\times m} \\ \mathbf{0}_{m\times 2} & \mathbf{B}^{-1} \end{pmatrix} = \begin{pmatrix} 1/3 & -1/3 & \mathbf{0}_{1\times m} \\ -1/3 & (2n+1)/(6n-6) & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & \mathbf{0}_{m\times 1} & (2n-2)^{-1}\mathbf{I}_{m\times m} \end{pmatrix}.$$

Therefore, the variance for the OLS estimator of any QE is $\sigma^2(2n+1)/(6n-6)$. Given that the number of parameters already included in the model is $m+1$, the expression for the power follows.

B.4.4 Power for $I_{\mathbf{me}}$.

Consider the $N \times (m+2)$ matrix, $\mathbf{X}_{\mathrm{i}m}$, including the intercept column, a single TFI contrast vector and all LE contrast vectors for the pDSD. Regardless of the TFI chosen, the information matrix is

$$\mathbf{X}_{\mathrm{i}m}^T \mathbf{X}_{\mathrm{i}m} = \begin{pmatrix} 2n+1 & 0 & \mathbf{0}_{1\times m} \\ 0 & 2n-4 & \mathbf{0}_{1\times m} \\ \mathbf{0}_{m\times 1} & \mathbf{0}_{m\times 1} & 2(n-1)\mathbf{I}_{m\times m} \end{pmatrix},$$

which is a diagonal matrix. Then the variance for the OLS estimate of the TFI is $\sigma^2(2n-4)^{-1}$. Given that the number of parameters already included in the model is $m+1$, the expression for the power follows.

B.4.5 Power for Q_2 and I_2.

Consider the $N \times 6$ matrix, $\mathbf{X}_{\text{q2}}$, including the intercept column, the two QE contrast vectors, the two LE contrast vectors and the TFI contrast vector of any two-factor projection of the pDSD. It is easy to see that the information matrix for any two-factor projection is

$$\mathbf{X}_{\text{q2}}^T\mathbf{X}_{\text{q2}} = \begin{pmatrix} N & 2n-2 & 2n-2 & 0 & 0 & 0 \\ 2n-2 & 2n-2 & 2n-4 & 0 & 0 & 0 \\ 2n-2 & 2n-4 & 2n-2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2n-2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2n-2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2n-4 \end{pmatrix}.$$

Using Harville (2011, p.p. 89, 99), it is easy to show that the inverse of this information matrix equals

$$\left(\mathbf{X}_{\text{q2}}^T\mathbf{X}_{\text{q2}}\right)^{-1} = \begin{pmatrix} \frac{2n-3}{4n-7} & \frac{1-n}{4n-7} & \frac{1-n}{4n-7} & 0 & 0 & 0 \\ \frac{1-n}{4n-7} & \frac{3(n-1)}{2(4n-7)} & \frac{-(n-4)}{2(4n-7)} & 0 & 0 & 0 \\ \frac{1-n}{4n-7} & \frac{-(n-4)}{2(4n-7)} & \frac{3(n-1)}{2(4n-7)} & 0 & 0 & 0 \\ 0 & 0 & 0 & (2n-2)^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 & (2n-2)^{-1} & 0 \\ 0 & 0 & 0 & 0 & 0 & (2n-4)^{-1} \end{pmatrix}.$$

So the variances for the OLS estimates of the TFI and of any QE equal $\sigma^2(2n-4)^{-1}$ and $\sigma^2(3n-3)/(8n-14)$, respectively. Since the number of parameters already included in the model is $N-6$, the power expressions for I_2 and Q_2 follow.

B.4.6 Power for Q_3 and I_3.

Schoen et al. (2018) considered pDSDs in three factors. They made a distinction between DSDs with $n \equiv 0 \pmod 4$ and those with $n \equiv 2 \pmod 4$, where $n = (N-1)/2 = m + k$ and they showed that for a given N all three-factor pDSDs are isomorphic or statistically equivalent.

Consider now the second order model matrix of a three-factor projection of a DSD, $\mathbf{X}_{\text{q3}}$. This is an $N \times 10$ model matrix including the intercept column, three QE contrast vectors,

three LE contrast vectors and three TFI contrast vectors. When $(N-1) \equiv 0 \pmod 8$, the information matrix for one of the isomorphic DSDs is

$$\mathbf{X}^T_{\mathrm{q}3_0}\mathbf{X}_{\mathrm{q}3_0} = \begin{pmatrix} 2n+1 & 2n-2 & 2n-2 & 2n-2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2n-2 & 2n-2 & 2n-4 & 2n-4 & 0 & 0 & 0 & 0 & 0 & -2 \\ 2n-2 & 2n-4 & 2n-2 & 2n-4 & 0 & 0 & 0 & 0 & 2 & 0 \\ 2n-2 & 2n-4 & 2n-4 & 2n-2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2n-4 & -2 & 2 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 & -2 & 2n-4 & -2 \\ 0 & -2 & 0 & 0 & 0 & 0 & 0 & 2 & -2 & 2n-4 \end{pmatrix}.$$

Alternatively, when $(N-1) \equiv 4 \pmod 8$, the information matrix for one of the isomorphic DSDs is

$$\mathbf{X}^T_{\mathrm{q}3_2}\mathbf{X}_{\mathrm{q}3_2} = \begin{pmatrix} 2n+1 & 2n-2 & 2n-2 & 2n-2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 2n-2 & 2n-2 & 2n-4 & 2n-4 & 0 & 0 & 0 & 0 & 0 & -2 \\ 2n-2 & 2n-4 & 2n-2 & 2n-4 & 0 & 0 & 0 & 0 & -2 & 0 \\ 2n-2 & 2n-4 & 2n-4 & 2n-2 & 0 & 0 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2n-2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 0 & 0 & 0 & 2n-4 & -2 & -2 \\ 0 & 0 & -2 & 0 & 0 & 0 & 0 & -2 & 2n-4 & -2 \\ 0 & -2 & 0 & 0 & 0 & 0 & 0 & -2 & -2 & 2n-4 \end{pmatrix}.$$

We calculated the inverses of the information matrices using Mathematica. For both information matrices, the output of Mathematica (not included here) showed that the variance of the OLS estimates for any QE is the same. The output also showed that the variances of the OLS estimates of the three TFIs are equal. The variances in case $(N-1) \equiv 0 \pmod 8$ differ from the corresponding variances when $(N-1) \equiv 4 \pmod 8$. For $n \equiv 0 \pmod 4$, the variances for a QE estimate, $\hat{\beta}_{ii}$, and a TFI estimate, $\hat{\beta}_{ij}$, are:

$$\mathrm{Var}(\hat{\beta}_{ii}) = \frac{4n^3 - 21n^2 + 24n + 2}{10n^3 - 66n^2 + 102n + 8}$$
$$\mathrm{Var}(\hat{\beta}_{ij}) = \frac{5n^2 - 19n + 14}{10n^3 - 66n^2 + 102n + 8}.$$

For $n \equiv 2 \pmod 4$, the variances for a QE estimate, $\hat{\beta}_{ii}$, and a TFI estimate, $\hat{\beta}_{ij}$, are:

$$\text{Var}(\hat{\beta}_{ii}) = \frac{4n^3 - 29n^2 + 54n - 26}{10n^3 - 86n^2 + 218n - 172}$$
$$\text{Var}(\hat{\beta}_{ij}) = \frac{5n^2 - 29n + 36}{10n^3 - 86n^2 + 218n - 172}.$$

Since the number of parameters already included in the model is $N - 10$, the power expressions for I_3 and Q_3 follow for the two cases.

C Table with detailed results

Table S3 shows the best and worst sets of columns to drop from sDSDs for $m + k \in \{6, 8, \ldots, 20, 24\}$ and $k \leq 8$. The table includes (i) the average absolute correlations, (ii) the maximum absolute correlations, and (iii) the sum of all squared correlations between contrast vectors for pairs of TFI contrast vectors. Each set of columns is labeled as *i.criteria*, where *i* can be "b" for best option or "w" for worst option. The *criteria* can include the letters "a", "m", or "s" corresponding to the average absolute correlation, maximum absolute correlation, and sum of squared correlations criteria, respectively. For instance, a set of columns labeled "b.ams" thus is best in terms of all three criteria, while a design labeled "w.a" is worst in terms of the average absolute correlation criterion.

When dropping one, two and three columns from sDSDs with 6-14 and 18-24 factors, we did not find differences in the resulting pDSDs. This was also the case when dropping four columns from the 6-factor sDSD and when dropping one column from the 16-factor sDSD. For this reason, all these cases are not shown in Table S3.

Table S3 shows that the best sets of columns are optimal in terms of all criteria for all combinations of numbers of factors in the sDSD and numbers of dropped columns, except when dropping five, six and eight columns from the 16-factor sDSD, and when dropping seven columns from the 20-factor sDSD. For these cases, there is one set that is best in terms of the average absolute correlation, and another one that is best in terms of both the maximum absolute correlation and the sum of squared correlations.

Finally, Table S3 shows that, for eight combinations of numbers of factors in the sDSD and numbers of dropped columns, dropping the last columns is the best option in terms

of at least one criterion. However, for 14 combinations, dropping the last columns is the worst option in terms of at least two criteria.

Table S3: Sets of k columns to drop from an n-factor sDSD. b: best option; w: worst option; a: average correlation; m: maximum correlation; s: sum of squared correlations. Last columns are shown in boldface.

m Factors in design	n	k	Option	Subset of columns to drop	Average correlation	Maximum correlation	Sum of Squared Correlations
4	8	4	b.ams	**5 6 7 8**	0.133333	0.167	0.3333
			w.ams	4 5 7 8	0.266667	0.667	1.6667
6	10	4	b.ams	6 8 9 10	0.207143	0.750	6.7500
			w.ams	**7 8 9 10**	0.221429	0.750	8.2500
5		5	b.ams	4 6 7 9 10	0.166667	0.250	1.4063
			w.ams	**6 7 8 9 10**	0.200000	0.750	2.9063
4		6	b.ams	4 6 7 8 9 10	0.150000	0.250	0.3750
			w.ams	**5 6 7 8 9 10**	0.250000	0.750	1.8750
8	12	4	b.ams	7 8 10 12	0.190476	0.400	23.7600
			w.ams	**9 10 11 12**	0.193651	0.400	24.2400
7		5	b.ams	7 8 10 11 12	0.181429	0.400	12.0900
			w.ams	**8 9 10 11 12**	0.192857	0.400	13.0500
6		6	b.ams	5 7 8 10 11 12	0.160000	0.400	4.9200
			w.ams	**7 8 9 10 11 12**	0.182857	0.400	5.8800
5		7	b.ams	5 7 8 9 10 11 12	0.146667	0.400	1.7400
			w.ams	**6 7 8 9 10 11 12**	0.200000	0.400	2.7000
4		8	b.ams	4 5 6 7 8 10 11 12	0.080000	0.100	0.1200
			w.ams	**5 6 7 8 9 10 11 12**	0.160000	0.400	0.6000
10	14	4	b.ams	**11 12 13 14**	0.193939	0.500	58.0000
			w.ams	9 12 13 14	0.194949	0.500	58.6667
9		5	b.ams	**10 11 12 13 14**	0.185714	0.500	34.2500
			w.ams	9 11 12 13 14	0.188889	0.500	35.5833

Continued on next page

Table S3 (continued)

m Factors in design	n	k	Option	Subset of columns to drop	Average correlation	Maximum correlation	Sum of Squared Correlations
8		6	b.ams	**9 10 11 12 13 14**	0.177249	0.500	19.0000
			w.ams	7 9 10 11 12 14	0.185185	0.500	21.0000
7		7	b.ams	6 7 8 10 12 13 14	0.158333	0.500	8.3125
			w.ams	6 8 9 11 12 13 14	0.182143	0.500	11.6458
6		8	b.ams	6 7 8 10 11 12 13 14	0.147619	0.500	3.6667
			w.ams	6 7 8 9 11 12 13 14	0.176190	0.500	5.6667
14	16	2	b.ams	8 16	0.133333	0.857	231.8571
			w.ams	8 15	0.135845	0.857	231.8571
13		3	b.ams	**14 15 16**	0.131725	0.857	166.0102
			w.ams	8 14 15	0.134580	0.857	166.0102
12		4	b.ams	**13 14 15 16**	0.127473	0.857	115.0408
			w.ams	12 14 15 16	0.128671	0.857	117.2449
11		5	b.ms	8 13 14 15 16	0.124242	0.857	75.9949
			b.a	**12 13 14 15 16**	0.123088	0.857	78.4439
			w.a	7 8 12 13 14	0.133	0.857	77.464
			w.ms	12 13 14 15 16	0.123088	0.857	78.4439
10		6	b.ms	7 8 11 12 13 15	0.124675	0.857	47.3878
			w.ams	6 7 8 14 15 16	0.124675	0.857	53.2653
			b.a	8 11 13 14 15 16	0.118615	0.857	48.6122
9		7	b.ams	7 8 11 12 13 15 16	0.110204	0.857	27.7347
			w.ms	**10 11 12 13 14 15 16**	0.123810	0.857	39.0000
			w.a	6 7 8 9 10 11 13	0.140136	0.857	34.5918
8		8	b.ms	4 6 7 8 11 12 13 15	0.117914	0.857	14.5714
			w.ams	**9 10 11 12 13 14 15 16**	0.158730	0.857	35.1429
			b.a	7 8 11 12 13 14 15 16	0.104308	0.857	15.5510
14	18	4	b.ams	**15 16 17 18**	0.181593	0.375	201.1875
			w.ams	14 16 17 18	0.181777	0.375	201.5625

Continued on next page

Table S3 (continued)

m Factors in design	n	k	Option	Subset of columns to drop	Average correlation	Maximum correlation	Sum of Squared Correlations
13		5	b.ams	12 15 16 17 18	0.178072	0.375	143.3672
			w.ams	13 15 16 17 18	0.178821	0.375	144.4922
12		6	b.ams	9 12 13 14 15 18	0.172902	0.375	97.7813
			w.ams	9 13 15 16 17 18	0.176049	0.375	101.1563
11		7	b.ams	9 12 13 14 15 17 18	0.168813	0.375	65.4023
			w.ams	9 10 13 15 16 17 18	0.172348	0.375	68.0273
10		8	b.ams	9 10 12 13 14 15 17 18	0.162121	0.375	40.8750
			w.ams	8 10 11 12 14 15 16 18	0.168939	0.375	44.2500
16	20	4	b.ams	14 17 18 20	0.172923	0.444	322.2222
			w.ams	14 18 19 20	0.173109	0.444	322.8148
15		5	b.ams	14 16 17 19 20	0.170543	0.444	241.6944
			w.ams	12 16 17 19 20	0.171032	0.444	242.8796
14		6	b.ams	11 13 15 18 19 20	0.167359	0.444	176.5556
			w.ams	12 16 17 18 19 20	0.168661	0.444	178.6296
13		7	b.ms	11 13 15 17 18 19 20	0.164835	0.444	126.3519
			w.ams	12 15 16 17 18 19 20	0.166167	0.444	128.1296
			b.a	11 13 15 16 18 19 20	0.164391	0.444	126.3519
12		8	b.ams	10 12 13 14 17 18 19 20	0.159751	0.444	86.9259
			w.ams	9 12 13 14 15 16 17 20	0.164413	0.444	90.0370
20	24	4	b.ams	20 22 23 24	0.134792	0.364	693.3471
			w.ams	**21 22 23 24**	0.134852	0.364	693.7438
19		5	b.ams	17 20 22 23 24	0.133680	0.364	554.6343
			w.ams	**20 21 22 23 24**	0.133831	0.364	555.6260
18		6	b.ams	16 17 20 22 23 24	0.132283	0.364	437.0579
			w.ams	**19 20 21 22 23 24**	0.132845	0.364	440.0331
17		7	b.ams	15 17 19 20 22 23 24	0.130838	0.364	340.0165
			w.ams	**18 19 20 21 22 23 24**	0.131788	0.364	343.3884

Continued on next page

Table S3 (continued)

m Factors in design	n	k	Option	Subset of columns to drop	Average correlation	Maximum correlation	Sum of Squared Correlations
16		8	b.ams	15 16 17 19 20 22 23 24	0.128953	0.364	258.9421
			w.ms	13 18 19 20 21 22 23 24	0.130940	0.364	264.4959
			w.a	11 12 14 16 20 21 23 24	0.131092	0.364	263.7025

References

Harville, D. A. (2011). *Matrix Algebra From a Statistician's Perspective*. Springer.

Jones, B. and Nachtsheim, C. J. (2011). A class of three-level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology*, 43:1–15.

Jones, B. and Nachtsheim, C. J. (2017). Effective design-based model selection for definitive screening designs. *Technometrics*, 59:319–329.

Schoen, E. D., Eendebak, P. T., and Goos, P. (2018). A classification criterion for definitive screening designs. *The Annals of Statistics*. To appear.

Zhou, Y.-D. and Xu, H. (2017). Composite designs based on orthogonal arrays and definitive screening designs. *Journal of the American Statistical Association*, 112:1675–1683.