Robustness of Classical and Optimal Designs to Missing Observations

Byran J. Smucker∗1, Willis Jensen2, Zichen Wu1, and Bo Wang1

1Department of Statistics, Miami University, Oxford, OH
2W.L. Gore & Associates, Flagstaff, AZ

Abstract

Missing observations are not uncommon in real-world experiments. Consequently, the robustness of an experimental design to one or more missing runs is an important characteristic of the design. Results of an evaluation of the robustness of classical and optimal designs to missing observations are presented, and optimal designs fare relatively well in terms of robustness compared to classical designs. Additionally, a modified version of an existing robustness criterion is used to construct designs that are robust to missing observations.

Keywords: Central Composite Designs; Factorial Designs; Missing Data; D-efficiency; I-efficiency

1 Introduction

When designing experiments, there is a tension between robustness and optimality. Optimality suggests a single, best design while robustness implies quality under a broad range of conditions. Robustness is crucial in the real world of experimentation, as noted generally by Wendelberger (2010). There is also a growing acknowledgement that multiple criteria should be used when constructing optimal designs (e.g. Lu et al. 2011; Gilmour and Trinca 2012). In the face of many different kinds of variability and uncertainty, it is important for designs that are optimal for estimation or prediction to be robust in other ways. Since there is generally not a single design that is optimal in all aspects, an acceptable compromise is to consider appropriate tradeoffs between key characteristics. Though Box and Draper (1975) gave a list of fourteen attributes that generally relate to a design’s robustness and practicality, one aspect not specifically mentioned is robustness to missing values. Box and Draper do mention outliers or “wild observations”, and to the extent that these observations can be omitted from the analysis in a principled way they would fall under the purview of our work. However, we discuss in particular the situation where no values are obtained for a particular run. The primary goal of our work is to consider some common classical and

*corresponding author, 311 Upham Hall, 100 Bishop Circle, Oxford, OH 45056; smuckerb@miamioh.edu; 513-529-0955 (tel); 513-529-0989 (fax)
Several robust-to-missing-observations criteria have been proposed in the literature. Box and Draper [1975] pointed out the connection between the diagonal values of the hat matrix and a design's robustness and applied their related criterion to central composite designs. Siddiqi [2011] applied a similar criterion to the smaller response surface designs of Draper and Lin [1990]. Herzberg and Andrews [1976] considered the probability that a design will not estimate the desired model, and Andrews and Herzberg [1979] suggested maximizing the expected value of the determinant of the information matrix under possible missing observations. Akhtar and Prescott [1986] developed a criterion that minimizes the maximum loss due to missing observations and applied it to the evaluation and generation of central composite designs, and Ahmad and Gilmour [2010] used this measure to study the robustness of so-called subset designs (Gilmour 2006). Herzberg et al. [1987] proposed equi-information designs, which retain equal information when up to two design points are missing. Imhof et al. [2002] presented results based on a different maximin criterion assuming continuous designs.

Missing from the literature is (1) a systematic comparison of classical and optimal designs in terms of their robustness to missing observations; and (2) a general algorithmic approach to generating designs that are robust to missing observations. We note that Hackl [1995] did generate robust designs for small response surface experiments, though they have not been formally published. Furthermore, a referee pointed us to a recent paper (da Silva et al. 2016) that incorporates a leverage-based missing-robustness criterion used within a larger compromise criterion that balances pure error and lack-of-fit degrees of freedom (see Gilmour and Trinca 2012).

The rest of the paper unfolds as follows. We first motivate our work by explaining why missing runs occur in practical situations. We then outline several evaluative measures of missing-robustness. Next we demonstrate the impact of missing runs on some standard designs, including fractional factorial and D-optimal designs for screening experiments, and central composite and I-optimal designs for prediction/optimization. Included in these comparisons are some new designs constructed using a version of a robustness criterion from the literature. We conclude with a discussion and some thoughts on future research.

2 Motivation

A common solution in the statistical literature for dealing with missing observations is imputation. Imputation methods fill in the missing values based on the other data points, to allow for model fitting. Though imputation methods are common for large observational datasets, we believe these methods are problematic in an experimental design situation because of the relatively small number of runs. Thus, we do not consider this a viable solution to many experimental, missing-observation problems.

Obviously, if possible, the first priority is to redo the runs to fill in the missing observations. However, if an experimental run fails during an experiment, it is sometimes difficult or impossible to redo. Consider, as an example, a manufacturing process experiment that uses actual production equipment. Laboratory or pilot plant experiments may be useful in the early phases of experimentation to suggest acceptable factor ranges and a rough idea of optimal designs in terms of their robustness to missing observations. Secondarily, we adapt an existing criterion to construct some new designs that are missing-robust.
the optimal factor region. Ultimately a final experiment using the production equipment
will be necessary to determine the best operating conditions for that equipment, because
the results from the laboratory or pilot plant are not typically directly generalizable due to
varying conditions in the respective facilities. In order to maximize manufacturing through-
put and minimize production costs, it is not unusual to have production equipment running
constantly. It becomes very expensive—though necessary—to pause production to execute
a series of experimental runs on that equipment. As soon as the experimental runs are
completed, production resumes and the experimental runs are tested and analyzed. If a
run was inadvertently missed, there will be strong resistance to shutting down production
again for a single run. In this environment, testing errors made on the experimental runs
could render one or more of the runs unusable. These missing values are generally not tied
to one of the experimental factors of interest and can thus be treated as missing at random.

Missing runs can also occur in new product development. With proper planning, an
adequate amount of custom-made raw material can be obtained from a supplier for use
in an experiment. In some situations, however, the material is limited because of cost or
scarcity. In this case, a failed run would require additional material which is likely to be
costly and time-inefficient to obtain. A probable outcome is that the run is simply lost to
the experiment.

Complex processes can also cause missing runs. In today’s global and complex manufac-
turing environment, it is not unusual to have processes whose steps span multiple, possibly
far-flung, locations. In these situations, it may be difficult or impossible to do a full ex-
periment across all process steps involving all factors so experiments will be performed on
localized portions of the process. However, in order to obtain the final measurements on
the experimental runs it may be necessary to continue processing them for multiple steps
beyond those within the scope of the experimental factors of interest. Inadvertent events
in downstream processing may result in a loss of some of the experimental runs and it is
difficult to redo the runs without a significant time lag or change in processing conditions.

As mentioned previously, an assumption throughout is that any missing observations
are missing at random. There are certainly many cases for which the assumption does
not hold—for instance, runs near the edge of the experimental region may have a higher
probability of yielding unusable or no information. This can occur with a poor selection of
factor levels and can be prevented with more thorough exploration of factors ranges prior
to formal experimentation. We concentrate on those cases where the missing observations
are independent of the factor levels used.

Industrial experience suggests that the number of missing observations is generally low
(< 20%). If the percentage of missing observations is large, it is clear that little information
may be obtained from the experimental results and it is likely that more fundamental issues
need to be addressed before attempting to execute the experiment. Here, we focus on
situations in which an observation is missing with a low probability. This is important to
note because the evaluations we perform implicitly or explicitly make this low-probability
assumption.
3 Evaluating the Impact of Missing Observations

In preparation for the evaluation of designs in terms of their robustness to missing observations, we establish the linear model setting within which we are working, and present and motivate several ways to summarize a design’s missing-robustness.

Throughout, we assume that the typical linear regression model will be fit:

\[ y = X\beta + \epsilon, \]  

(1)

with \( \epsilon \sim N(0, \sigma^2 I) \), where \( y \) and \( \epsilon \) are \( n \times 1 \) vectors, \( X \) is the \( n \times p \) expanded design matrix, and \( \beta \) is a \( p \times 1 \) vectors of unknown parameters. The least squares estimates of \( \beta \) are \( \hat{\beta} = (X^TX)^{-1}X^Ty \) which implies that \( \text{Cov}(\hat{\beta}) = \sigma^2(X^TX)^{-1} \). We consider two particular model forms. First, for screening applications in which the goal is to find the subset of active factors, the model includes just the main effects. Referring to the linear model (1), let \( f^T(x) \) be a row of \( X \) so that

\[ f^T(x)\beta = \beta_0 + \sum_{i=1}^{k} \beta_ix_i, \]  

(2)

where \( x = \{x_1, x_2, \ldots, x_k\} \) and \( k \) is the number of factors. For response surface experiments, the usual model is

\[ f^T(x)\beta = \beta_0 + \sum_{i=1}^{k} \beta_ix_i + \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \beta_{ij}x_ix_j + \sum_{i=1}^{k} \beta_{ii}x_i^2. \]  

(3)

In what follows, we will emphasize two sorts of criteria. First, we will describe the evaluation of design efficiency when one, two, or three runs are missing. For screening designs we focus on D-efficiency, and for response surface designs we examine both I- and D-efficiencies. Secondly, we consider two summative missing-robust criteria. The first relates to leverage values (Box and Draper 1975) and the second is an adaptation of the criterion of Andrews and Herzberg (1979), which is intuitively appealing because it balances D-optimality with robustness against missing values.

3.1 Evaluation of D-Efficiencies when \( m \) Runs are Missing

D-optimal designs take their cue from the variance of the least-squares estimators, and include design points such that \( |(X^TX)^{-1}| \) is minimized. In order to assess the impact of missing runs, Herzberg and Andrews (1976) and Andrews and Herzberg (1979) used a criterion based upon the loss of D-efficiency. We adopt a similar approach but consider the performance of a design when any set of \( m \) runs are missing. Because there are \( \binom{n}{m} \) combinations of \( m \) missing runs, we calculate \( D_F(i, m) \) as the efficiency of a design, with
respect to the $n$-run optimal design, when the $i^{th}$ combination of $m$ runs is missing:

$$D_F(i, m) = \left( \frac{|X_{n-m,i}^T X_{n-m,i}|}{|X^*_{n}^T X^*_{n}|} \right)^{1/p}, \quad i = 1, 2, \ldots, \binom{n}{m},$$

where $X_{n-m,i}$ is the $i^{th} (n-m) \times p$ design submatrix and $(X^*_{n})$ is the corresponding D-optimal design with $n$ runs. This metric gives a sense, in an absolute way, of how much information is being lost when $m$ runs are missing. We are interested in the distributions of $D_F$ for various designs and values of $m$. Those designs with relatively large medians, small variances, and small reductions in efficiency as $m$ increases are desirable as robust to missing observations.

### 3.2 Evaluation of I-Efficiencies when $m$ Runs are Missing

Though there are many optimal design criteria, I-optimal designs are appropriate for response surface models because they seek to minimize the average prediction variance across the design space. The variance of the fitted value at $x = (x_1, x_2, \ldots, x_k)$ is $V(\hat{Y}|x) = \sigma^2 f^T(x)(X^T X)^{-1} f(x)$. The I-criterion for a design is

$$I = \int_R f^T(x)(X^T X)^{-1} f(x) \, dx,$$

where $R$ represents the design region. This criterion can be computed as $I = \text{trace}[(X^T X)^{-1} A]$ where $A = \int_R f(x) f^T(x) \, dx$ is the so-called moments matrix which does not depend upon the design. To facilitate straightforward comparisons between the designs, we limit ourselves to cuboidal design regions. For this design region and the full quadratic model, the I-criterion can be computed in closed-form (see Goos and Jones 2011). This design region also means that the central composite designs that we study are face-centered. The I-efficiency of a design with respect to another design with I-criterion value $I^*$ can be defined, then, as $I_{\text{eff}} = I^*/I$.

Analogous to the D-efficiency criterion, we define $I_F(i, m)$ as the I-efficiency of a design with respect to the $n$-run I-optimal design, when the $i^{th}$ combination of $m$ runs is missing:

$$I_F(i, m) = \frac{I^*_n}{I_{n-m,i}}, \quad i = 1, 2, \ldots, \binom{n}{m},$$

where $I_{n-m,i}$ is the $i^{th} n-m$-run design evaluated according to the I-criterion and $I^*_n$ is the $n$-run I-optimal design. As before, we are interested in the distribution of these I-efficiencies for various designs and values of $m$.

### 3.3 Leverage

Based on the regression model (1) and the least squares estimates $\hat{\beta} = (X^T X)^{-1} X^T y$, we have $\hat{y} = X(X^T X)^{-1} X^T y = H y$, where $H = \{h_{ij}\}$ is the so-called hat matrix and its diagonals the leverage values.

To justify the use of leverage values in evaluating the missing-robustness of a design,
we can make an argument similar to Box and Draper (1975). Since \( \hat{y}_i = h_{i1}y_1 + h_{i2}y_2 + \ldots + h_{in}y_n \), if the \( u \)th observation is missing, the change in the \( i \)th predicted value is simply \( \delta_i = -h_{iu}y_u \). If we wish to aggregate the effect of the missing observation \( u \) over all observations, we might define

\[
\sum_{i=1}^{n} \delta_i^2 = \sum_{i=1}^{n} h_{iu}^2 = y_u^2 h_{uu},
\]

where the last equality follows from the fact that \( H \) is idempotent. Since \( \sum_{i=1}^{n} h_{ii} = p \), we cannot minimize the sum of the leverage values. Instead, as Box and Draper suggest, we might desire the leverage values to be as uniform as possible. This can be accomplished by minimizing the leverage criterion \( h = \sum_{i=1}^{n} h_{ii} \). We compare designs in Sections 4 and 5 using this criterion.

3.4 The Truncated Herzberg-Andrews Criterion

We will now present a final criterion that we will use to evaluate designs in terms of missing-robustness. However, here we are slightly more ambitious: we will also use this criterion to construct missing-robust designs.

3.4.1 The Full Criterion

Herzberg and Andrews (1976) introduced a quantity that reduces to the determinant of the reduced information matrix when any set of \( i \) (\( i = 1, 2, \ldots, n \)) observations are missing. Andrews and Herzberg (1979) went further, proposing a design construction criterion based on the expected value of a related quantity. That is, they constructed a criterion that serves as a measurement of the expected precision after taking into account all the possible ways observations could be missing. The larger this expectation, the greater the robustness.

Consider the value of \( |X^T D^2 X|^{\frac{1}{p}} \), where \( p \) is the number of model parameters and \( D^2 \) is a diagonal matrix with \( d^2_{ii} \) on the diagonal, such that \( d^2_{ii} \) is:

\[
d^2_{ii} = \begin{cases} 
0 & \text{with probability } \alpha(x), \\
1 & \text{with probability } 1 - \alpha(x).
\end{cases}
\]

The value of \( d^2_{ii} \) indicates whether observation \( i \) is missing. For instance, if \( d^2_{ii} = d^2_{jj} = 0 \), then \( |X^T D^2 X|^{\frac{1}{p}} = |X^T_{ij} X_{ij}|^{\frac{1}{p}} \) with \( X_{ij} \) the design matrix resulting when the \( i \)th and \( j \)th points are omitted. If all diagonal entries in \( D^2 \) are 1, then this quantity corresponds to a version of the \( D \)-criterion, \( |X^T X|^{\frac{1}{p}} \). Here \( \alpha(x) \) specifies the probability of losing the unit that received treatment \( x \) in the design, and we note that in general its value could vary depending on where the points are in the design space.

The criterion which we call HA, studied by Andrews and Herzberg (1979), is defined as the expectation of \( |X^T D^2 X|^{\frac{1}{p}} \). If we assume that missing a particular run is independent of missing any other run, then \( \alpha(x) = \alpha \) for every design point and an expression for the
HA criterion is:

\[
E\left(\left|X^TD^2X\right|^{\frac{1}{p}}\right) = (1 - \alpha)^n \left|X^TX\right|^{\frac{1}{p}} + \alpha(1 - \alpha)^{n-1} \sum_{i=1}^{n} \left|X_i^TX_i\right|^{\frac{1}{p}}
\]

\[
+ \alpha^2(1 - \alpha)^{n-2} \sum_{i\neq j, i<j} \left|X_{ij}X_{ij}\right|^{\frac{1}{p}} + \alpha^3(1 - \alpha)^{n-3} \sum_{i\neq j\neq k, i<j<k} \left|X_{ijk}X_{ijk}\right|^{\frac{1}{p}} + \ldots
\]

\[
= \sum_{\ell=0}^{n} \alpha^\ell (1 - \alpha)^{n-\ell} T_\ell,
\]

where conceptually \(T_\ell\) stands for the sum of all the possible \(\left|X^TX\right|^{1/p}\) when there are \(\ell\) observations missing. Though there could be a non-zero probability that more than \(n - p\) observations will go missing, this would result in a singular system and the element in the criterion would be taken as 0.

The HA criterion is a composite of design efficiency and robustness since it incorporates the \(D\)-criterion not only when all observations have been retained, but also when one or more are missing. The main challenge associated with this criterion is its computational difficulty. Andrews and Herzberg (1979) showed that \(\left|X^TD^2X\right|^{\frac{1}{p}}\) can be simplified to \(\left|X^TX\right|^{(1-\alpha)p}\) when we disregard the \((1/p)\)th-root and let \(\alpha(x) = \alpha\), which suggests that by this measure \(D\)-optimal designs may be relatively missing-robust. However, one can imagine the complexity of the criterion in general since to evaluate it involves the computation of up to \(\sum_{i=0}^{n-p} \binom{n}{i}\) determinants of \((p - i) \times (p - i)\) matrices.

### 3.4.2 The Truncated Criterion

We wish to construct designs using the HA criterion, as well as evaluate them, so it is particularly important that the criterion is computationally feasible. To ameliorate the computational difficulty, we propose two simplifications. First, we truncate the criterion so that we only include the first four terms of the original. Secondly, we take advantage of some previously developed algebraic simplifications that reduce the computational complexity of the truncated quantity. The truncated HA criterion is as follows:

\[
\text{THA} = E\left(\left|X^TD^2X\right|^{\frac{1}{p}}\right) = (1 - \alpha)^n \left|X^TX\right|^{\frac{1}{p}} + \alpha(1 - \alpha)^{n-1} \sum_{i=1}^{n} \left|X_i^TX_i\right|^{\frac{1}{p}}
\]

\[
+ \alpha^2(1 - \alpha)^{n-2} \sum_{i\neq j, i<j} \left|X_{ij}X_{ij}\right|^{\frac{1}{p}}
\]

\[
+ \alpha^3(1 - \alpha)^{n-3} \sum_{i\neq j\neq k, i<j<k} \left|X_{ijk}X_{ijk}\right|^{\frac{1}{p}}.
\]

Under the assumption that the number of missing observations is binomially distributed,
the mean number of unusable observations out of $n$ trials is $n\alpha$. In industrial applications, both $n$ and $\alpha$ are typically small. In fact, based on experience we believe that the interval $(0,0.2]$ is a reasonable range for $\alpha$. This implies that the distribution for the number of runs missing is right-skewed, suggesting that focusing on the first four terms will not compromise the full criterion heavily in many realistic situations. However, we acknowledge that this does limit somewhat the applicability of this criterion and the designs constructed from it.

The truncated criterion can be simplified in a way that will reduce computational demands. More specifically, we can rewrite the criterion in terms of $|X^TX|$ and elements of the hat matrix, $H = \{h_{ij}\}$. These simplifications were first introduced by Andrews and Herzberg (1979), but we reproduce them here for clarity and completeness.

Consider the second term of the truncated criterion, $\alpha(1-\alpha)^{n-1} \sum_{i=1}^{n} |X_i^TX_i|^{\frac{1}{p}}$. This corresponds to the case where exactly one observation is missing. We focus on the determinant part of the term, $|X_i^TX_i|$, in which we assume that the $i$th observation, represented by the $1 \times p$ expanded design point $x_i$, is missing. Because $X^TX = \sum_{j=1}^{n} f(x_j)f^T(x_j)$, we have that $X_i^TX_i = X^TX - f(x_i)f^T(x_i)$. So, using standard determinant results, we have that $|X^TX - f(x_i)f^T(x_i)| = |X^TX|R_i$ where $R_i = 1 - h_{ii}$. Thus, the summation in the second term in (5) can be simplified to $|X^TX|^{\frac{1}{p}} \sum_{i=0}^{n} R_i^{\frac{1}{p}}$. Similarly, when two observations are missing, $|X^TX|$ becomes $|X_{ij}^TX_{ij}| = |X^TX|R_{ij}$ where $R_{ij} = (1-h_{ii})(1-h_{jj}) - h_{ij}^2$ and we can also find that $|X_{ijk}^TX_{ijk}| = |X^TX|R_{ijk}$ with

$$R_{ijk} = (1-h_{ii})(1-h_{jj})(1-h_{kk}) - h_{ij}^2(1-h_{kk}) - h_{ik}^2(1-h_{jj}) - h_{jk}^2(1-h_{ii}) - 2h_{ij}h_{ik}h_{jk}.$$  

Putting these together, we can compute the truncated criterion as

$$\text{THA} = \left|X^TX\right|^{1/p} \left( (1-\alpha)^n + \alpha(1-\alpha)^{n-1} \sum_{i=0}^{n} R_i^{1/p} + \alpha^2(1-\alpha)^{n-2} \sum_{i\neq j}^{n} R_{ij}^{1/p} \right) + \alpha^3(1-\alpha)^{n-3} \sum_{i\neq j \neq k}^{n} R_{ijk}^{1/p} \right).$$  

### 3.4.3 Constructing THA-Optimal Designs

In this section, we discuss how the coordinate exchange algorithm (Meyer and Nachtsheim 1995) can be adapted to use the THA criterion. The basic coordinate exchange algorithm starts with a randomly generated initial design and considers exchanges of the first element in the first row, comparing the corresponding criterion values and choosing the exchange which results in the largest increase in the criterion value. Then, it proceeds to consider exchanges with the second element of the first row and finds its best level. After all the factors in one row of the design have been examined, we jump to the next row and repeat the process until all the factors in the different rows are updated. We continue this process
until the algorithm converges to a locally optimal design.

The coordinate exchange procedure is very simple, yet quite powerful because it can be easily adapted to many criteria of interest, with the computational difficulty commensurate to the computational complexity of the criterion. In our implementation, we adapted Matlab’s cordexch to the THA criterion. Clearly, the truncated HA criterion is significantly more computationally intensive than the D-criterion. However, based on (6), we only need to update $|X^T X|$ and the hat matrix. Furthermore, both the computation of $|X^T X|$ and the hat matrix can be done by $QR$ factorization. Thus we are able to combine the coordinate exchange algorithm and (6) to produce designs that are robust against missing observations. For all THA-optimal designs, we used $\alpha = 0.1$ and 100 algorithm tries. The value of $\alpha$ is application-specific and THA designs could be created for any reasonable value. For our purposes, we used the midpoint of the range of values that we believe are most likely.

4 Impact of Missing Observations on Screening Designs

Given that the primary objective of our work is to compare classical and optimal designs in terms of their robustness to missing observations, in this section we consider the impact of missing observations on designs commonly used for factor screening. We focus primarily on fractional factorial designs (FFDs), along with D-optimal designs that have become more commonly used in recent years. We also include in our comparisons the THA-optimal designs of Section 3.4. The model of interest is (2), which includes only main effects.

Two-level factorial designs are widely used in screening experiments to investigate main effects and interaction effects. When the number of factors is more than about 4 or 5, the number of runs required in a factorial design is often prohibitively large, so fractional factorial designs are frequently used instead. Resolution III fractions—those that cannot estimate any interactions free of the main effects—are often used when main effects are of primary interest (see Wu and Hamada 2009). Alternatively, D-optimal designs can also effectively estimate the first-order model (see Goos and Jones 2011). For each design, we calculate the D-efficiencies from (4), for each set of 1, 2 and 3 missing runs, as well as the leverage criterion of Section 3.3 (lower is better) and the THA criterion of Section 3.4.2 (higher is better).

Consider the situation in which $k = 5$ and the experimental goal is to determine which of the main effects are significant, using $n = 16$ or fewer runs. Two classical designs are the resolution III $2^{5-2}$ fraction with $I = ABCD = BCE = ADE$, and the resolution V $2^{5-1}$ fraction with $I = ABCDE$ which have 8 and 16 runs, respectively. We assess these five-factor fractional factorial designs, along with D-optimal and THA-optimal designs with $n = \{8, 10, 12, 16\}$. Results are shown in Figure 1 and Table 1.

From the results, we observe little difference among the designs with the same number of runs. For $n = 8$, the D-optimal and THA-optimal designs that we used are regular fractions, with defining relations $I = ABCDE = ABD = CDE$ and $I = ABCDE = BCD = ADE$, respectively. Thus, it is unsurprising that the three designs appear equivalent in terms of missing-robustness. For $n = 16$, the D-optimal and THA-optimal designs are orthogonal, but not regular fractions. All three 16-run designs are D-optimal, and from Table 1 it is clear that they are equivalent in terms of the THA and leverage criteria as well. This illustrates that there can be multiple D-optimal designs and in fact neither the
Figure 1: D-efficiencies for possible main effects designs, for $k = 5$ and $n = \{8, 10, 12, 16\}$, according to the number of missing runs.

Table 1: THA and leverage criterion values for screening designs.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$n$</th>
<th>Design</th>
<th>THA Value</th>
<th>Leverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>8</td>
<td>FFD</td>
<td>6.30</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D-optimal</td>
<td>6.30</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>6.30</td>
<td>4.5</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>D-optimal</td>
<td>8.15</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.19</td>
<td>3.63</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>D-optimal</td>
<td>10.25</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>10.25</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>FFD</td>
<td>13.33</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D-optimal</td>
<td>13.33</td>
<td>2.25</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>13.33</td>
<td>2.25</td>
</tr>
</tbody>
</table>

THA nor leverage criteria will necessarily discriminate between them. Interestingly, there appears to be a slight difference between the regular fraction and the other two designs in terms of D-efficiencies when two or three runs are missing (Figure 1). That is, the fractional factorial has less variation in terms of D-efficiency than the other two. This may be because the FFD is a regular fraction while the others are not. There is little or no difference between the D-optimal and THA-optimal designs when $n = 10$ and $n = 12$.

In Figure 1, we see that as the number of missing runs increases, the D-efficiency values tend to decrease while their variability increases. When $n = 8$, two missing runs often result in an inestimable model and three missing runs guarantees an inestimable model. Thus, the impact of a few missing runs is exaggerated when the original design is close to saturated.

We performed additional investigations for $k = 10$, but do not report them since similar conclusions were obtained. Overall, compared to the Truncated HA missing-robust criterion, the classical and optimal designs are not only robust, but nearly equivalent. Furthermore, the impact of missing runs on screening designs appears small as long as the number of runs is sufficient.
5 Impact of Missing Observations on Response Surface Designs

Now we move beyond screening designs to study the impact of missing values on designs used to fit the quadratic model (3). As before, we wish to compare classical and optimal designs, so we study the most common response surface design, the central composite design (CCD, see Myers et al. 2009), along with I-optimal designs (see Goos and Jones 2011) and the THA-optimal designs of Section 3.4. We consider experiments with both $k = 3$ and $k = 5$, for a variety of sample sizes, and assess the impact of 1, 2, and 3 missing observations on both I- and D-efficiencies. We also compare the designs in terms of the leverage and THA criteria.

5.1 Three-Factor Response Surface Experiments

For a CCD with three factors, there are typically $n_f = 2^3 = 8$ runs for the factorial portion of the design, 6 runs for the face-centered axial points, and $n_c$ center points. Here, we take $n_c = 2, 4, 6$ and consequently study the three-factor, face-centered CCD with 16, 18, and 20 runs, respectively. For these three run sizes, as well as $n = 12$, we constructed I-optimal designs using JMP® (JMP 2015) software and THA-optimal designs using the methods described in Section 3.4.

In Figure 2, we see that the efficiency losses due to missing observations are especially acute when $n = 12$. There appears to be little difference between the efficiency loss for the central composite and I-optimal designs, though the optimal designs are generally higher in both I- and D-efficiencies. Because the THA criterion is closely related to the D-criterion, the THA-optimal designs have a higher baseline D-efficiency than the other designs and in fact appear to be D-optimal. However, the THA-optimal designs have a lower baseline I-efficiency, though they seem to absorb less efficiency loss than the other designs as the number of missing observations increases.

The first part of Table 2 reveals that when $k = 3$ the CCDs and I-optimal designs have similar THA criterion values, while the THA-optimal designs are predictably superior. The THA-optimal designs also tend to have more uniform leverage values, while the I-optimal designs are a bit better on this measure than the CCDs.

5.2 Five-Factor Response Surface Experiments

For $k = 5$, the standard CCD has either $n_f = 2^{5-1} = 16$ or $n_f = 2^5 = 32$ factorial points, 10 axial points, along with $n_c$ center runs. We considered four CCDs with $n = \{28, 32, 36, 44\}$, $n_f = \{16, 16, 16, 32\}$, and $n_c = \{2, 6, 10, 2\}$, respectively. We also constructed I-optimal and THA-optimal designs with $k = 5$ and $n = \{24, 28, 32, 36, 44\}$. (For $n = 36$, the number of center runs in the CCD is excessive; better classical-type designs could be chosen, but we retained this one to maintain consistency in the types of designs we compared.)

Compared to $k = 3$, Figure 3 shows larger efficiency differences between the CCDs and the I-optimal designs, though this may be partially explained by the CCDs with 6 and 10 center runs. There is no obvious difference between the designs in terms of efficiency loss. The THA-optimal designs are again essentially D-optimal, and are typically inferior in terms of the I-criterion and superior in terms of the D-criterion. Interestingly, when design
resources are constrained ($n = 24$ and $n = 28$), the THA-optimal designs are competitive with the I-optimal and central composite designs, in terms of I-efficiency, as the number of missing observations increase.

Note that several of the THA-optimal designs have a D-efficiency greater than 100%. This is because the “D-optimal” designs generated to serve as D-efficiency baselines were algorithmically generated using JMP® software (JMP 2015) and thus do not guarantee a globally D-optimal design. This suggests that the D-criterion alone may be insufficient to evaluate a design; here we have THA-optimal designs which are clearly superior in terms of robustness (second part of Table 2) but also as good or better in terms of the D-criterion than the designs generated explicitly to produce D-optimal designs. These results also highlight the similarity between the THA-criterion and the D-criterion.
Table 2: THA and leverage criterion values for response surface designs.

<table>
<thead>
<tr>
<th>$k$</th>
<th>$n$</th>
<th>Design</th>
<th>THA Value</th>
<th>Leverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>12</td>
<td>I-optimal</td>
<td>3.50</td>
<td>8.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>3.90</td>
<td>8.43</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>5.46</td>
<td>6.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>5.88</td>
<td>6.49</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>I-optimal</td>
<td>5.46</td>
<td>6.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>5.88</td>
<td>6.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>5.33</td>
<td>6.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>6.56</td>
<td>5.67</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>I-optimal</td>
<td>5.96</td>
<td>6.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>6.56</td>
<td>5.67</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>6.31</td>
<td>6.56</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>7.15</td>
<td>5.07</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>I-optimal</td>
<td>6.16</td>
<td>18.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>7.09</td>
<td>18.57</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>6.54</td>
<td>17.76</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.16</td>
<td>16.03</td>
</tr>
<tr>
<td>5</td>
<td>28</td>
<td>I-optimal</td>
<td>7.14</td>
<td>17.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.16</td>
<td>16.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>5.87</td>
<td>17.50</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.38</td>
<td>13.95</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>I-optimal</td>
<td>7.08</td>
<td>15.94</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.38</td>
<td>13.95</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>5.11</td>
<td>17.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.25</td>
<td>12.30</td>
</tr>
<tr>
<td>5</td>
<td>36</td>
<td>I-optimal</td>
<td>7.02</td>
<td>14.34</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>8.25</td>
<td>12.30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CCD</td>
<td>6.24</td>
<td>10.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td>THA-optimal</td>
<td>7.05</td>
<td>10.10</td>
</tr>
</tbody>
</table>
Figure 3: I- and D-efficiencies for possible designs for the second-order model, for $k = 5$ and $n = 24, 28, 32, 36,$ and $44$, according to the number of missing runs.
The second part of Table 2, with \( k = 5 \), shows that the I-optimal designs are typically more robust than the CCDs in terms of the THA and leverage criteria. However, this is slightly misleading because the experiments for which the disparity is most pronounced (\( n = 32 \) and \( n = 36 \)) are cases in which the CCD has an abnormally large number of center points (6 and 10, respectively). When \( n = 44 \) and the full \( 2^5 \) factorial is used, the CCD actually outperforms the I-optimal design. The THA-optimal designs are clearly preferred on the THA and leverage criteria.

Overall, the CCDs and I-optimal designs do not appear to differ substantially in terms of their efficiency loss as the number of missing observations increases. However, since the I-optimal designs are more I-efficient, this advantage is retained even when design runs are missing. The THA-optimal designs are more robust and D-efficient, but less I-efficient. Their lack of I-efficiency reduces their utility as response surface designs except when resources are severely constrained.

### 6 Discussion and Conclusions

We have presented some empirical results for the effect of missing observations on various classical and optimal designs, as well as a new type of missing-robust design, in both screening and response surface settings. In general, we found missing observations have little impact on screening designs as long as the number of runs in the initial design is not too small. The D-optimal and fractional factorial designs were also robust when compared to the new THA-optimal missing-robust screening designs. A practitioner who wants to have a more robust design to protect against the possible loss of information from missing observations would do well to add a few additional runs to the appropriate design.

Secondly, we did not find evidence that optimal response surface designs are less robust to missing observations than classical designs. In terms of efficiency, they are superior though that is expected based on their construction. More interestingly, they performed well in terms of average efficiency when a few observations were lost. The classical CCDs had a similar amount of efficiency loss, but were less efficient overall. Since the new THA-optimal missing-robust designs are based on the D-criterion, they suffered significantly in terms of I-efficiencies. However, they appeared to have less efficiency loss in the face of missing observations. Notably, when the number of observations was small relative to the model being estimated, these missing-robust designs sometimes outperformed their I-optimal counterparts in terms of I-efficiency, when two or three runs were missing. These THA-optimal design also tend to be D-optimal, or nearly so, which suggests that D-optimal designs have good missing-robustness properties.

These results should inspire some confidence that optimal designs compare favorably to their classical counterparts in terms of robustness to missing observations. We also suggest that for severely resource-constrained experiments in which missing observations are a nonnegligible possibility, either extra runs should be added or an explicitly missing-robust design should be used. Finally, we note that there are no missing-robustness criterion in the literature that are based on I-optimality; such a criterion might produce designs that are attractive alternatives to an I-optimal design when the possibility of missing observations cannot be ignored.
A Appendix: Supplementary Material

The designs used in this study are included as material supplementary to this article.

References


