Journal of Statistical Planning and Inference 139 (2009) 629 – 641

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Response surface design evaluation and comparison

Christine M. Anderson-Cook[a,*], Connie M. Borror[b], Douglas C. Montgomery[c]

[a]*Statistical Sciences, Los Alamos National Laboratory, USA*
[b]*Division of Mathematical and Natural Sciences, Arizona State University West, USA*
[c]*Department of Industrial Engineering, Arizona State University, USA*

ARTICLE INFO

Available online 12 April 2008

Keywords:
Design optimality
Graphical methods
Variance dispersion graphs
Fraction of design space plots

ABSTRACT

Designing an experiment to fit a response surface model typically involves selecting among several candidate designs. There are often many competing criteria that could be considered in selecting the design, and practitioners are typically forced to make trade-offs between these objectives when choosing the final design. Traditional alphabetic optimality criteria are often used in evaluating and comparing competing designs. These optimality criteria are single-number summaries for quality properties of the design such as the precision with which the model parameters are estimated or the uncertainty associated with prediction. Other important considerations include the robustness of the design to model misspecification and potential problems arising from spurious or missing data. Several qualitative and quantitative properties of good response surface designs are discussed, and some of their important trade-offs are considered. Graphical methods for evaluating design performance for several important response surface problems are discussed and we show how these techniques can be used to compare competing designs. These graphical methods are generally superior to the simplistic summaries of alphabetic optimality criteria. Several special cases are considered, including robust parameter designs, split-plot designs, mixture experiment designs, and designs for generalized linear models.

1. Introduction

Designed experiments allow the analyst to control the factors thought to be important in characterizing or explaining the response variable(s) of the experiment. Recent years have seen dramatic growth in the use of designed experiments, not just in the classical industrial, life sciences, and agricultural settings, but in many areas of business, such as marketing and financial services. This interest in the design of experiments has led to much new research on the subject. There are many types of experimental designs in the literature, and there are also many criteria on which experimental designs are based. There are an increasing number of computer software packages that give recommendations to the experimenter regarding the choice of a design based on some specific criteria and input from the experimenter. It is critical for an experimenter to understand the characteristics and features that should be taken into account when a design is chosen. That is the general topic of this paper.

Because there are many types of designs intended for specific areas of applications, we must limit the scope of our presentation. We focus on response surface designs. Myers et al. (2004) observe that the response surface framework has become the standard approach for much of the experimentation carried out in industrial research, development, manufacturing, and technology commercialization. Response surface methodology (RSM) is mostly concerned with approximating a complex unknown function

* Corresponding author. Tel.: +1 505 606 0347.
E-mail address: c-and-cook@land.gov (C.M. Anderson-Cook).

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.04.004

with a low-order polynomial, usually either a first-order model or a second-order model. Consequently, designs for fitting these models are of considerable interest.

The choice of a design for fitting a first-order model is relatively clear cut. If the region of experimentation is cuboidal or spherical, first-order orthogonal designs possess many desirable characteristics. However, choosing a response surface design to fit a second-order model is a much more complex and interesting problem, because while there are many "standard" designs to choose from, these designs have considerable flexibility as to their final specification. It is also possible (indeed, even necessary) to construct "nonstandard" designs to account for certain requirements of a specific experiment. Therefore, we concentrate primarily on designs for fitting the second-order response surface model. However, we will discuss some specific variations of this problem, such as mixture experiment designs and those for the robust parameter design problem.

When selecting a second-order response surface design, there are many design criteria and characteristics that could be considered. Myers and Montgomery (2002) suggest that a good design should:

1. Result in a good fit of the model to the data.
2. Provide sufficient information to allow a test for lack of fit.
3. Allow models of increasing order to be constructed sequentially.
4. Provide an estimate of "pure" experimental error.
5. Be insensitive (robust) to the presence of outliers in the data.
6. Be robust to errors in control of design levels.
7. Be cost-effective, that is, not require too many runs.
8. Allow for experiments to be done in blocks.
9. Provide a check on the homogeneous variance assumption.
10. Provide a good distribution of the variance of the predicted response throughout the design region.

Box and Draper (1975) and Anderson-Cook (2005) are other useful references on the desirable characteristics of a design.

Not all of the characteristics in the above list are required, nor are they necessarily equally important, in every RSM application. However, most of them must be given serious consideration when one designs an experiment. In the response surface framework, first-order models are used primarily for factor screening and for fitting a model to assist the experimenter in locating the region of the optimum through a method such as steepest ascent. In these applications, characteristic 1 mentioned above is likely to be very important, because both factor screening and steepest ascent depend critically on obtaining a model that is a good fit to the data and whose parameters are well-estimated. Second-order models are used primarily in response optimization, so a good-fitting model is important in that situation as well, but a model that performs well in response prediction is critical. Thus, prediction variance is an important concern. Criterion 10, which speaks to the distribution of prediction variance and its stability throughout the design region, is a key concern. In this paper, we pay special attention to this issue of the characterization of prediction variance for a design, how it can be evaluated graphically, and how it can be used for evaluating and comparing competing designs.

The list of 10 characteristics is a reminder that designing an experiment is not necessarily an easy task. Indeed, several of the 10 characteristics may be important and yet the experimenter may not be fully aware of the magnitude of their importance. Some items on the list present potential conflicts with each other. As a result, there are trade-offs that almost always exist when one chooses an appropriate design. Some of the methods that we will discuss are excellent ways to highlight some of these potential conflicts and can help the experimenter make more informed decisions in design selection.

2. Numerical and graphical assessment methods

In this section we consider a number of different measures that can be used to compare the estimation and prediction capability of potential designs.

2.1. Optimality criteria

Optimal design methods use a single criterion in order to construct designs for RSM; this is especially relevant when fitting second-order models. Kiefer (1959) detailed the theory behind optimum designs. His work was motivated in part by Wald (1943) who suggested the use of *D*-optimality for design comparison and Elfving (1952) with work in linear regression. The work of Kiefer and Wolfowitz (1960) provided the framework for practical implementation of *D*-optimality for design construction.

Several optimality criteria have been developed to address estimation or prediction through the use of variance characteristics. *D*- and *A*-optimality criteria provide a measure of the variance of the model coefficients through the moment matrix, $\mathbf{M} = \mathbf{X}'\mathbf{X}/N$ where $\mathbf{X}$ is the model matrix and N is the number of runs in the design. A *D*-optimal design is one that maximizes the determinant of $\mathbf{M}$, equivalently minimizing the volume of the confidence region on the model coefficients. An *A*-optimal design is one that maximizes the trace of the moment matrix, $\mathbf{M}$, and is directly related to minimizing the individual variances of the model coefficients. The *A*-criterion is not invariant, and as a result a design that is *A*-efficient for a particular setting of design variables

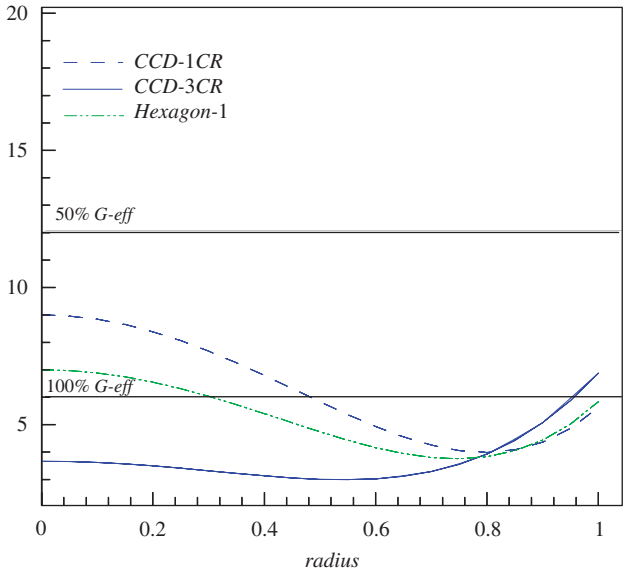

Fig. 1. Variance dispersion graph for three designs on a circular region with two factors.

may no longer be *A*-efficient if the design variable coding changes. Optimality criteria related to prediction variance include *G*- and *I*-optimality. The scaled prediction variance is given by

$$\text{SPV} = N\mathbf{x}'^{(m)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^m$$

where N is the number of observations in the design and $\mathbf{x}^m$ is a vector of design factor values expanded to the form of the model. A *G*-optimal design is one that minimizes the maximum SPV over the experimental design region. Designs that minimize the average SPV over the design region are known as *I*-optimal. *I*-optimality is also known as *IV*-, *Q*-, and *V*-optimality.

D-, *G*-, and *I*-optimality criteria are the primary criteria used in practice when constructing and evaluating response surface designs. *D*- and *I*-efficient designs can easily be constructed using commercially available software. There is no commercially available software that currently constructs designs using the *G*-criterion, but historically exchange algorithm approaches were used to find near optimal designs for a given set of candidate points. Several authors have implemented branch and bound methods, simulated annealing and genetic algorithms (GAs) for constructing optimal designs for various regions of interest (see Borkowski, 2003; Haines, 1987; Hamada et al., 2001; Heredia-Langner et al., 2003, 2004; Welch, 1982; Zhou, 2001).

For a first-order model, standard 2^k factorial and fractional factorial 2^{k-p} designs of resolution III or higher have been shown to be *D*-, *G*-, and *I*-optimal. For second-order models, this property for standard designs no longer holds true. Second-order designs such as the central composite design (CCD) and Box-Behnken design (BBD) have high *D*- and *G*-efficiencies, but are not *D*- or *G*-optimal.

2.2. Graphical displays

Single-number criteria such as *D*- and *G*-efficiency do not completely reflect the prediction variance characteristics of the design in question. For example, two designs can have the same *G*-efficiency, yet be quite different in terms of prediction variance in the design region. Alternatives to single-number summaries include graphical displays of the prediction variance across the design regions. Two such plots include the variance dispersion graph (VDG) and the fraction of design space (FDS) plot. VDGs were developed by Giovannitti-Jensen and Myers (1989) and plot the minimum, average, and maximum SPVs against distances from the overall center of the design space. VDGs are useful for examining and comparing the prediction variance characteristics of competing experimental designs. A VDG for three different two-factor designs located on a unit circle is given in Fig. 1. These designs include (1) a CCD with one center run (9 points), (2) a CCD with three center runs (11 points), and (3) a hexagon design with one center run (7 points). The model being fit in this situation is a full second-order model.

All three of these designs are rotatable, so the lines for the minimum, average, and maximum prediction variance coincide. A horizontal reference line at SPV $= p$ (where p is the number of model parameters) is often displayed on the VDG representing 100% G-efficiency. The FDS plot was developed by Zahran et al. (2003). On the FDS plot, the prediction variance is plotted against the FDS that has prediction variance at or below the given value. Fig. 2 displays the FDS plot for the same designs of Fig. 1.

Unlike the VDG, the FDS takes into account the proportion of the volume of the design space for a given radius. That is, the prediction variances on the VDG are given the same weight regardless of distance from the design center. The FDS plot consists

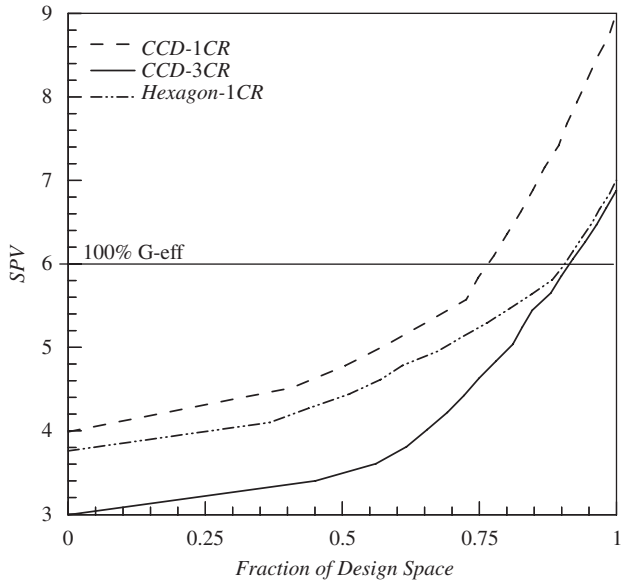

Fig. 2. Fraction of design space plot for three designs on a circular region with two factors.

of a single line for a given design and thus can easily display the prediction variance characteristics of several designs at once. Several modifications of both VDG and FDS plots have been made to compare designs involving mixture components, noise variables, split-plot structures, and the slope of the response model (Jang and Park, 1993).

3. Evaluation and comparison of designs

Important applications of VDG and FDS plots include design evaluation and comparison. The plots in Figs. 1 and 2 are good illustrations of these applications. For example, Fig. 1 illustrates that a single center run is insufficient to provide a good distribution of prediction variance throughout the design space for a rotatable CCD. The design with three center runs has much smaller prediction variance through most of the design space and is only inferior to the one center run design at the boundary of the region. The FDS plot in Fig. 2 clearly shows the dominance of the CCD with three center runs. VDGs have been used extensively to study the aspects of standard designs such as the allocation of center runs and most of the contemporary recommendations about selecting the number of center runs are derived from these studies (for example, see Myers et al. (1992) and Myers and Montgomery (2002)). Myers et al. (1992) dealt primarily with spherical second-order designs. An important finding in that paper was that the hybrid designs of Roquemore (1976) are excellent candidates for second-order designs when the number of runs is small. Park et al. (2005) focused on second-order designs on cubes and provided general guidance regarding the choice of a design that provides a good distribution of prediction variance. See also Lucas (2007) and Anderson-Cook et al. (2007). They showed that the designs developed by Hoke (1974) are excellent choices when the number of runs is small, and that computer-generated designs using the *I* criterion and the *G* criterion are also very good small designs (the *I* designs were created using JMP and the *G* designs were created using a GA).

Figs. 1 and 2 also provide an interesting comparison of the CCD with three center runs and the hexagon design. Notice from Fig. 2 that both designs have approximately the same *G*-efficiency (about 86%); however, they have very different SPV behavior throughout the design space. The model that results from the hexagon design would perform very poorly relative to the model for the CCD throughout a significant portion of the design space, as shown by the higher curve for the hexagonal design for most fractions of the design space. Predictions at the design center can often be more important in second-order models than predictions at the boundary of the region because the experimenter usually has a reasonably good idea about where the most important conditions are located. VDGs and FDS plots can be used in combination to provide precise information about which areas of the design space are likely to have poor prediction performance and the relative proportions at various SPV values. This example nicely illustrates the potential problems associated with relying on single-number efficiencies when evaluating second-order designs.

Another important consideration is whether to incorporate the cost of the experiment into the comparison of designs. For completely randomized designs, the use of the SPV in graphical summaries such as the VDG and the FDS plots is relatively standard. In this case the prediction variance has been scaled (multiplied) by the number of design points based on the assumption that the cost of the experiment is proportional to the number of observations collected. Using SPV, we are finding the best design after penalizing for their different costs. Sometimes, it is not as important to consider the quality of the design as a function of cost,

and in these cases the unscaled prediction variance

$$\mathrm{PV} = \mathbf{x}'^{(m)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^m$$

may be more appropriate. With this approach, the relative ranking of the designs considered in Figs. 1 and 2 would change since the SPV penalizes the CCD with three center runs more severely for its larger size than the other designs, and with the PV plots, it would become even more superior in performance. Within the design of experiments' research community, there has been considerable debate about the appropriateness of combining quality of prediction and cost into a single metric, or whether these attributes of the design should be kept separate. We feel that there are clearly situations when both approaches have merit, and the choice of which approach to use should depend on the priorities of the set-up and constraints of the experiment.

4. Special methods

In this section, we consider a number of special considerations for assessing designs. Although many of the qualitative and quantitative aspects of a good design are universal to various experiments, there are a number of specialized cases where methods have been developed to assess some additional trade-offs. In this section, we discuss some of the tools with which we have been involved.

4.1. Model robustness

Design assessment typically begins with the specification of a model that is proposed for the ensuing analysis once the data have been collected. Frequently, there may be uncertainty associated with what model may be appropriate, or after the data are collected, the model may be subsequently changed. Although it is tempting to proceed confidently assessing our design with a single model, this may naively lead us into subsequent difficulties. Most seriously, if our model does not have sufficient complexity to adequately approximate the true underlying function, and we have not allowed for some ability to check for lack of fit, we may never realize that we have fit an inappropriate model to our data. Alternately, if we focus on too large a model, and later reduce the model by eliminating unnecessary terms, our design might be considerably less optimal than intended for the final model. Since the intent of many experiments is to gain understanding about a largely unknown underlying relationship between factors and responses, being realistic about entertaining several candidate models when choosing a best design can be highly beneficial. The well-known bias-variance trade-off says that if we underfit a model, we need to be concerned about the fitted model making biased predictions of the response, while overfitting a model may lead to an inflation of the prediction variance. Hence, our goal is to find a model that fits just right.

Several different approaches to this problem have been considered. Heredia-Langner et al. (2004) use GA-based strategies for creating or augmenting a design using weights from the user-specified probability of several nested models. For example, for a particular design, we may feel that the most likely model is second order, but there is some chance that either a first-order or a third-order model might be appropriate. Based on the subjective weights, the algorithm determines a best design, which balances a chosen estimation or prediction criterion across the different candidate models. The designs generated typically are robust to small variations in the subjective weights and exhibit good characteristics with the ability to estimate the parameters and predict the response for each of the models, while maintaining the ability to test for lack of fit.

Ozol-Godfrey et al. (2005) look at assessing existing designs with FDS plots for a variety of reduced models nested inside the primary model. This approach allows comparison of prediction quality in the design space for a variety of model simplifications. Since which terms of the model may be removed during the analysis phase are dependent on the data collected and are unknown during the design phase, robustness to the final model is desirable. The total number of reduced models possible expands exponentially with the number of initial terms in the model. Hence it is helpful to look at a small number of strategic reductions that help characterize what quality of prediction can be expected across the simplified models. Two natural types of model simplifications can be considered: reducing the degree of the polynomial across all of the factors (second order, first order with interactions and first order) or reducing the number of active factors. The final simplified model is typically some combination of both of these types of reductions, so examining plots of both types of reductions can be informative.

Fig. 3 shows an FDS plot comparing two common designs (the central composite and Hoke D6 designs) for a cuboidal design space involving five factors, considering a reduction in the number of active factors. This plot illustrates a key result for this type of robustness comparison, which is directly related to the bias-variance trade-off. Reducing the number of terms in the model can only reduce the prediction variance at a given location, and hence for a given model the curves for reduced models are necessarily below the curve for the largest model. This is true for both types of model reduction mentioned above and is helpful for worst-case design assessment, since the FDS plot of the SPV for the largest model considered gives an upper bound for all nested models.

Finally, focusing on the mean squared error criterion, which balances prediction variance with bias directly, provides another approach for examining design robustness to model misspecification. Welch (1983), Vining and Myers (1991), and Piepel et al. (1993b) highlight the important effects that bias can have on model prediction and suggest methods for quantifying and visualizing the effect of model misspecification when the model is not rich enough to model the true underlying surface. Adaptations of VDGs allow for separation of variance and bias contributions to the mean squared error of prediction for any combination of

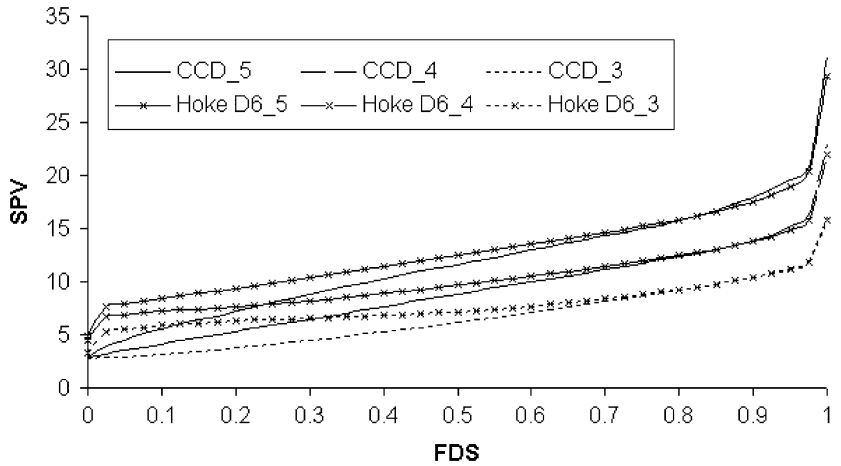

Fig. 3. Fraction of design space plot for comparing the robustness of a five-factor CCD to a Hoke D6 design for model reduction based on several factors not being active.

specified and true model. Anderson-Cook et al. (2008) show a number of graphical methods for comparing the effect on expected squared bias and mean squared error for different sizes of missing terms on a misspecified model.

Although the problem of assessing a design for model robustness may initially seem difficult because of the many potential options for the true model, gaining an understanding of how well different designs compete for a variety of possible models can greatly enhance our ability to select well. Being realistic early about our true level of uncertainty and what potential models might be used in the analysis phase can reduce the likelihood of a failed experiment incapable of producing the analysis needed, missing key features in the response, or poor prediction.

4.2. Mixture experiments

Mixture experiments are a special type of response surface problem where the factors are the components of a mixture and the response depends on the relative proportions of the components and not on their total amount. Because the component proportions must sum to one, the levels of the components are dependent and standard designs in a cuboidal or spherical region are not possible or appropriate. A mixture experiment with q components, where x_i is the proportion of the ith component, satisfies the constraints

$$0 \leqslant x_i \leqslant 1 \quad \forall i = 1, 2, \ldots, q \quad \text{and} \quad \sum_{i=1}^{q} x_i = 1.$$

The forms of mixture experiment models differ from the general polynomials used in the RSM because of the second constraint. The second-order canonical mixture model, known as the second-order Scheffé (1958) model, is

$$\eta = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j}^{q} \sum \beta_{ij} x_i x_j,$$

where β_i is the expected response at the pure component i and β_{ij} is a measure of the quadratic blending behavior of components i and j. The intercept and pure quadratic terms of the standard second-order polynomial do not appear because their inclusion would over-parameterize the model. For many applications, there are additional constraints on the component proportions, which will produce useful values of the response. This leads to design space shapes which can be quite irregular. In these cases, algorithmic approaches for finding good designs are common and have been integrated into the software packages Design-Expert and JMP.

Piepel and Anderson (1992) and Piepel et al. (1993a) extended the concept of VDGs to mixture experiments and other designs on irregular regions by calculating the minimum, average, and maximum prediction variance on shrinkage regions, which take the original design region shape and shrink it proportionately to the centroid of the region. Goldfarb et al. (2004a) show how the flexible structure of the FDS plots allows for the same interpretation of the FDS curve regardless of the shape of the region.

Mixture designs are also uniquely affected by measurement error, as a mis-measurement of one ingredient will not only change the amount of this component, but it will also affect the total amount of the mixture, and thereby alter the other components' proportions. Hamada et al. (2005) and Ozol-Godfrey and Anderson-Cook (2007) consider how prediction of the response optimum and prediction variance, respectively, are affected. Although not always the case, measurement errors typically yield a general worsening of the prediction variance for most of the design space, but with the minimum prediction variance only being slightly affected.

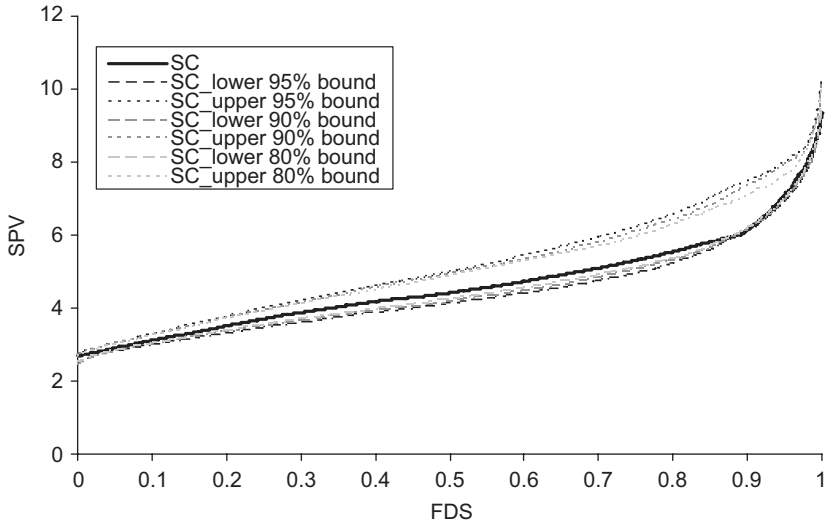

Fig. 4. FDS plot of the three-factor simplex-centroid design with 80%, 90%, and 95% bounds on the observed prediction variance values.

Fig. 4 shows an FDS plot with the range of changes for the SPV observed from simulation with absolute measurement error from a Uniform distribution added to all factors for a three-component simplex-centroid (SC) design. For this design, the distribution of SPV values shown in the FDS curves is quite skewed toward larger values, with a general worsening of prediction performance. Different designs can be affected quite differently by measurement error, and so if it may be a problem, this type of robustness can be worthwhile to explore.

Mixture-process experiments naturally have two groups of factors that we can expect to have quite different characteristics and impacts on the design performance. The mixture components are interdependent and constrained to sum to one. Frequently there are additional restrictions on acceptable proportions of the ingredients which may lead to oddly shaped design regions. The other set of factors, the process variables, can be varied independently and more traditional design region shapes for this subset are typical. In addition, how the prediction variance changes throughout the design region is likely to be quite different within the subspaces defined by each group of factors. Goldfarb et al. (2004a, b) show how FDS plots and surface or contour plots of 3-dimensional VDGs can be adjusted to provide not only a global summary of the prediction variance values, but also how these differ between the mixture and process variable subspaces. Again, the two complementary graphical techniques allow for a rich visual understanding of the changes in prediction variance throughout the design space. For prediction-based or optimality-based criteria, approximations to the average or worst prediction precision are available for the entire design space as well as for particular sub-regions using FDS plots.

4.3. Robust design problems

The overall goal of robust parameter design (RPD) is the selection of levels of the controllable variables (denoted $\mathbf{x}$) that will be robust or insensitive to changes in the levels of the noise variables (denoted $\mathbf{z}$). Taguchi (1986, 1987) proposed the use of orthogonal arrays. Orthogonal arrays consist of the crossing of two orthogonal designs: one design involving the controllable variables and the other design for the noise variables. Taguchi further recommended the signal-to-noise ratio (SNR) as a response which combines the mean and variance into a single performance measure. Statistical analysis would then be carried out on the SNR.

Taguchi's contributions to RPD were primarily in three areas: (1) quality philosophy and practice, (2) experimental design, and (3) data analysis. There have also been several criticisms of Taguchi's work, particularly with respect to the analysis methods and choice of experimental design (see Nair, 1992). The crossed array designs become prohibitively large as the number of variables increases. Furthermore, the designs do not allow for estimation of the control-by-control interactions. Since the late 1980s several experimental design and data analysis alternatives have been promoted. Welch et al. (1990) recommended the use of a single experimental design for both the control and noise variables called combined arrays. With the combined array, a single model containing both the control factors and the noise factors can be fitted to the response of interest:

$$\mathbf{y}(\mathbf{x},\mathbf{z}) = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} + \mathbf{z}'\boldsymbol{\delta} + \mathbf{x}'\boldsymbol{\Delta}\mathbf{z} + \boldsymbol{\epsilon} \tag{1}$$

where $\mathbf{x}$ and $\mathbf{z}$ are vectors representing the control and noise factors, respectively. In addition, $\boldsymbol{\beta}$ represents the coefficients for the control factors, $\mathbf{B}$ is a matrix of coefficients for the quadratic terms in the control factors, $\boldsymbol{\delta}$ is the vector of coefficients for the

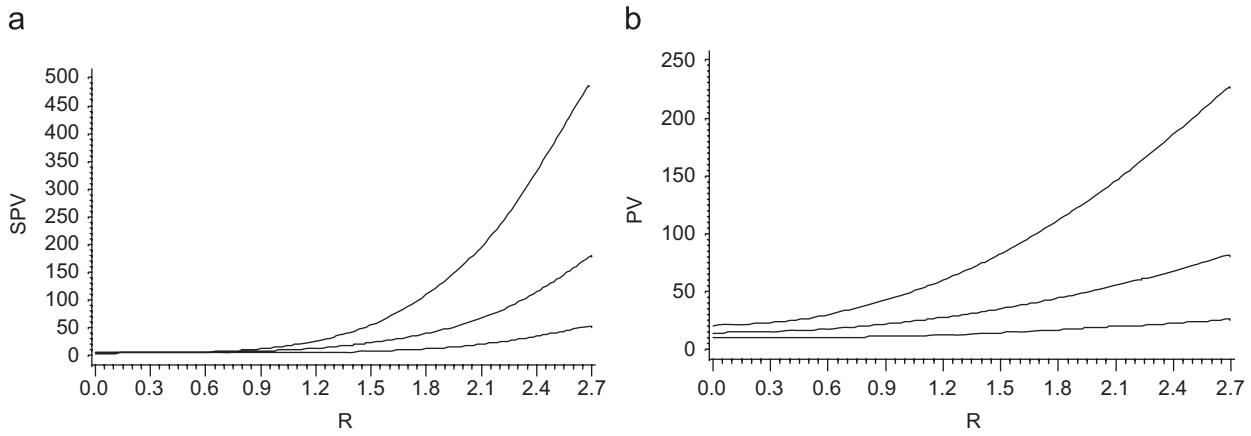

Fig. 5. Variance dispersion graphs for a small composite design with seven factors (four control and three noise): (a) SPV for the mean model and (b) prediction variance for the slope.

noise factors, and $\boldsymbol{\Delta}$ is the matrix of coefficients for the control-by-noise interactions. The errors, $\boldsymbol{\epsilon}$, are assumed to be NID$(0, \sigma^2)$. Myers et al. (1992) first introduced and discussed the use of a dual response approach to the robust design problem.

Evaluation and comparison of combined array designs for the RPD problem have generally involved optimality criteria such as *D*-, *G*-, and *I*-optimality, the prediction variance of the mean response, and prediction variance of the slope. From (1), the mean response can be modeled as

$$E_{\varepsilon,z}[\mathbf{y}(\mathbf{x},\mathbf{z})] = \beta_0 + \mathbf{x}'\boldsymbol{\beta} + \mathbf{x}'\mathbf{B}\mathbf{x} \tag{2}$$

The unconditional variance of y(x,z) in (1) can be shown to be

$$\mathrm{Var}_{\boldsymbol{\epsilon},\mathbf{z}}[\mathbf{y}(\mathbf{x},\mathbf{z})] = \mathbf{l}(\mathbf{x})\ \mathrm{Var}_z(\mathbf{z})\mathbf{l}'(\mathbf{x}) + \sigma^2 \tag{3}$$

where $\mathrm{Var}_z(\mathbf{z})$ is the variance–covariance matrix of $\mathbf{z}$ and $\mathbf{l}(\mathbf{x}) = \boldsymbol{\delta} + \mathbf{x}'\Delta$. Apart from the variance–covariance matrix for $\mathbf{z}$ and the error variance σ^2, the variance is a direct function of the terms $\boldsymbol{\delta} + \mathbf{x}'\Delta$. Notice that $\mathbf{l}(\mathbf{x})$ is a vector of partial derivatives of the model in (1) with respect to the noise variables, $\mathbf{z}$. As such, $\mathbf{l}(\mathbf{x})$ represents the *slopes* of the response in the direction of the noise variables, $\mathbf{z}$. Larger slopes result in a larger process variance. A large value of the slope indicates a possibly significant control-by-noise interaction and the *variance* of this slope would be important to examine, and in practice, minimize as much as possible. Atkinson (1970) first examined designs for estimating the slope of a response surface model at a fixed point, focusing on the expected mean square error. Other authors have examined the importance of the slope of a response surface including designs that are slope-rotatable (Hader and Park, 1978). See Murty and Studden (1972), Myers and Lahoda (1975), Mukerjee and Huda (1985) for designs and analyses involving the slope of a response surface.

The fitted response model for Eq. (1) is

$$\hat{\mathbf{y}}(\mathbf{x},\mathbf{z}) = \hat{\beta}_0 + \mathbf{x}'\hat{\boldsymbol{\beta}} + \mathbf{x}'\hat{\mathbf{B}}\mathbf{x} + \mathbf{z}'\hat{\boldsymbol{\delta}} + \mathbf{x}'\hat{\boldsymbol{\Delta}}\mathbf{z} \tag{4}$$

where $\hat{\beta}_0$, $\hat{\boldsymbol{\beta}}$, $\hat{\mathbf{B}}$, and $\hat{\boldsymbol{\Delta}}$ are found using least squares estimation. The prediction variance for (4), $\mathrm{Var}(\hat{\mathbf{y}}(\mathbf{x},\mathbf{z}))$, is an appropriate measure for evaluating designs with respect to the mean model in (2). This prediction variance will include the error associated with estimating the model parameters and the error transmitted to the response through the noise variables. We can evaluate the unscaled or scaled prediction variance for the mean. The SPV is obtained by dividing by the error variance, σ^2, and multiplying by the number of experimental observations, *N*.

In addition to the prediction variance for the mean model, it is of interest in RPD to model the variance for the estimated slopes. The variance of the slope, $\mathrm{Var}(\hat{\mathbf{l}}(\mathbf{x}))$, is a direct estimate of the precision of the variance model given in Eq. (3). The prediction variance of the mean model and the prediction variance of the slope can then be used to compare designs created for the RPD problem. Borror et al. (2002) evaluated response surface designs on spherical and cuboidal regions using VDGs displaying the SPVs of the mean model and the prediction variance of the slope. These SPVs can also be displayed graphically using FDS plots. Fig. 5 displays two VDGs for a 33-point small composite design (SCD) for four control factors and three noise factors. Fig. 5a displays the SPV for the mean model and Fig. 5b displays the prediction variance of the slope. The *x*-axis represents the Euclidean distance from the center of the design over the entire design region. The three lines on each figure represent the minimum, average, and maximum SPV for a single design only (in this case the SCD). When comparing several designs, there could be numerous lines on one VDG and the graph quickly becomes difficult to interpret. In contrast, FDS plots display only one line per design and are easy to interpret. Fig. 6 displays the FDS plots of the SPV of the mean model and prediction variance for the slope for six

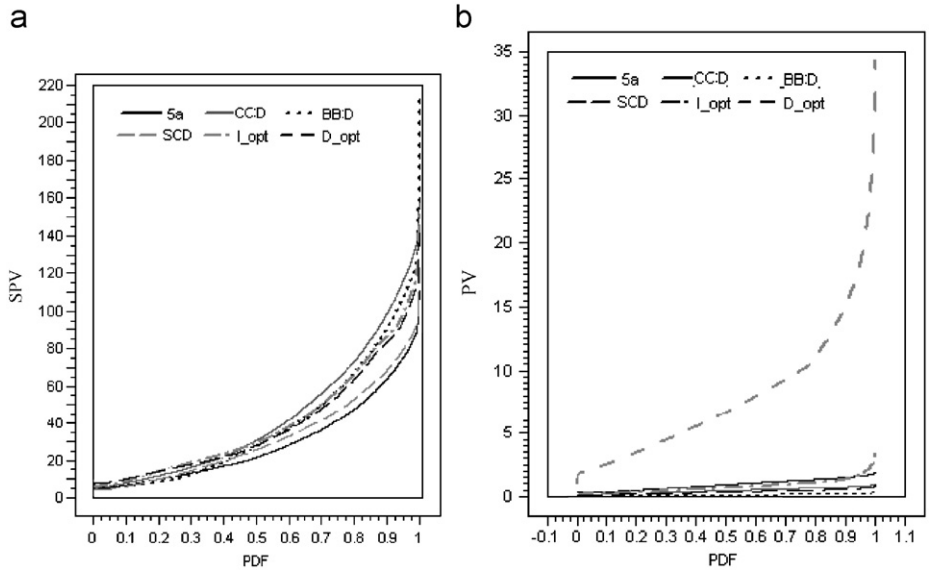

Fig. 6. FDS plots: (a) SPV for the mean and (b) prediction variance for the slope for six designs with five variables (three control and two noise).

competing designs. The number of factors investigated in this case was five (three control and two noise). The designs displayed in Fig. 6 include a 22-point mixed resolution design (M_5a), a 38-point standard CCD (M_CCD), a 21-point SCD (M_SCD), a 32-point I-optimal design (I-Opt), a 32-point D-optimal design (D-Opt), and a 40-point BBD (M_BB). From Fig. 6, it can be seen that all six designs are competitive with respect to the SPV of the mean model (Fig. 6a) over 95% of the design region. Near the design boundary, the SPV of the mean model for the Box-Behnken design increases significantly. However, with respect to the prediction variance of the slope, the SCD is clearly a poor design. Larger values of the prediction variance of the slope indicate that more variability is transmitted from the noise variables to the predicted response.

Miró-Quesada and del Castillo (2004) presented two new approaches for the dual response problem in robust parameter design by considering the prediction characteristics of the models. The characteristics emphasized include the expected MSE of the variance model and the unbiased estimator for a combined variance. The first approach presented by the authors involves a criterion used to scale the noise variables at the design stage in a dual response model so that the experimenter can avoid negative values of the estimated process variance. The second approach involves optimizing a criterion that involves all sources of variability, including noise factors, estimated prediction variance, and prediction error. Although Miró-Quesada and del Castillo (2004) do not evaluate designs using graphical approaches in their paper, it may be beneficial to construct FDS plots and VDGs for these methods.

Robust parameter design problems can often occur in a framework where mixture components are also present. Hamada et al. (2005) present mixture experiments with errors in the mixture components. Steiner and Hamada (1997) and Goldfarb et al. (2003) analyzed a mixture-process variable RPD problem where some of the process variables are noise variables. Because these are response surface problems, the predictive capabilities of the models obtained from designed experiments are of major importance. Chung et al. (2008) evaluate the predictive capabilities of a model obtained from designs involving noise variables. In their work, the prediction variance for the mean and the variance of the slope are considered when both quantitative and qualitative variables are present. Designs are constructed such that the prediction variance for the mean model and the prediction variance of the slope are simultaneously optimized. These designs can be easily evaluated using optimality criteria and FDS plots.

The FDS plot in Fig. 7 compares two competing designs for a mixture-process problem involving three mixture components, one continuous noise variable and one categorical noise variable. One design was created using the D-optimality criterion and a standard statistical software package. The other design was created using a GA with the prediction variance of the mean response and the variance of the slope as the objective functions combined into a single desirability function. The D-optimal and GA designs are both 24-point designs. The FDS displays the SPV for both designs. Based on this plot, the GA design is clearly better with respect to SPV over the design region.

4.4. Split-plot designs

Split-plot designs (SPDs) have natural groupings of the factors into those with hard or costly-to-change levels and others with levels that are relatively easy to change. When hard-to-change factors exist, it is typically more cost-effective to reduce or minimize the number of times the levels of these factors are changed. The separate randomizations for the whole plot

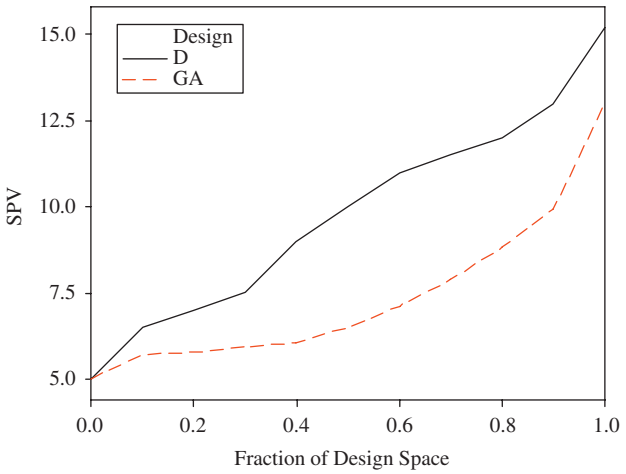

Fig. 7. FDS plot for comparing two designs involving mixture components, process variables, a continuous noise variable, and one categorical noise variable.

(hard-to-change) and sub-plot (easy-to-change) factors of an SPD lead to a compound symmetric error structure involving two error terms, which must be accounted for not only when doing inference, but also when determining an optimal design. For an SPD with a whole plots, the following linear mixed model can be written to explain the variation in the $N \times 1$ response vector, $\mathbf{y}$,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\partial} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect model parameters including the intercept; $\mathbf{X}$ is the $N \times p$ fixed effects model matrix; $\mathbf{Z}$ is an $N \times a$ incidence matrix of ones and zeroes where the ijth entry is 1 if the ith observation ($i = 1, \ldots, N$) belongs to the jth whole plot ($j = 1, \ldots, a$); $\boldsymbol{\partial}$ is an $a \times 1$ vector of random effects where the elements are assumed *i.i.d* $\mathrm{N}(0, \sigma_\delta^2)$ with σ_δ^2 denoting the variability among whole plots; and $\boldsymbol{\epsilon}$ is the $N \times 1$ vector of *i.i.d* $\mathrm{N}(0, \sigma_\varepsilon^2)$ residual errors for σ_ε^2 denoting the variation among subplot units. It is assumed that $\boldsymbol{\partial}$ and $\boldsymbol{\epsilon}$ are independent. The covariance matrix of the responses is

$$\mathrm{Var}(\mathbf{y}) = \sum = \sigma_\delta^2 \mathbf{Z}\mathbf{Z}' + \sigma_\varepsilon^2 \mathbf{I}_N = \sigma_\varepsilon^2 [d\mathbf{Z}\mathbf{Z}' + \mathbf{I}_N]$$

where $\mathbf{I}_N$ is an $N \times N$ identity matrix and $d = \sigma_\delta^2/\sigma_\varepsilon^2$ represents the variance component ratio. A variance ratio of 2 would imply that the variation among whole plots is twice as large as the variation among subplot units within a whole plot. Because this variance ratio affects the estimation and prediction, the "goodness" of a design is also affected. For example, the D-, G-, and V-optimal designs change depending on particular values of the variance ratio. Liang et al. (2006b) show how different variance ratio values and different prediction optimality criteria lead to different settings of the factorial points in a split-plot CCD. Since both σ_δ^2 and σ_ε^2 are typically unknown at the point of designing an experiment, choosing a good design in this setting requires some additional considerations. Determining whether good estimation of the model parameters (say, for model selection) or predicting well in the design space is of primary importance can help determine which optimality criteria are most relevant for design comparison. In addition, examining the robustness of the designs to changes in the variance ratio through graphical summaries can be beneficial to better understand the choices. Liang et al. (2006a, b) introduce variations of the VDG and FDS plots which can illustrate the impact of change.

Fig. 8 shows 3-D VDGs using the average SPV for the standard restricted split-plot CCD with the factorial locations adjusted to be I-optimal for one whole-plot factor and two sub-plot factors in a spherical region. This design takes the standard CCD and converts it to an SPD with the minimum number of required whole-plots see Kowalski (2002). Fig. 8 shows the VDGs for two variance ratio values, $d = 1$ and 10. The w- and x-dimensions show the radius from the center in the whole-plot and sub-plot directions, respectively. The I-optimal designs tend to have a large region with similar small values for the SPV near the center of the design space, at a cost of a larger value at the edge of the region. For small variance ratios (say, $d = 1$), SPV changes similarly as we move from the center of the design space toward the boundary of the design space in either the whole-plot or sub-plot directions. However, as the variance ratio gets larger, the majority of the changes in SPV occur as the location changes in the whole-plot (W) space. This can be seen by the flat surfaces when only the sub-plot or x-dimension is changed.

By examining the trade-offs in prediction quality throughout the region, we are better able to assess which of the candidate designs best suit the particular needs of our experiment. The single value comparison of the optimality criteria or their relative efficiency can mask some of the weaknesses of the design. The good news for practitioners is that for many situations, the best design can often be quite robust to misspecification of the variance ratio. Hence, an initial guess for this value that is reasonably close to the observed value will typically yield an acceptable result.

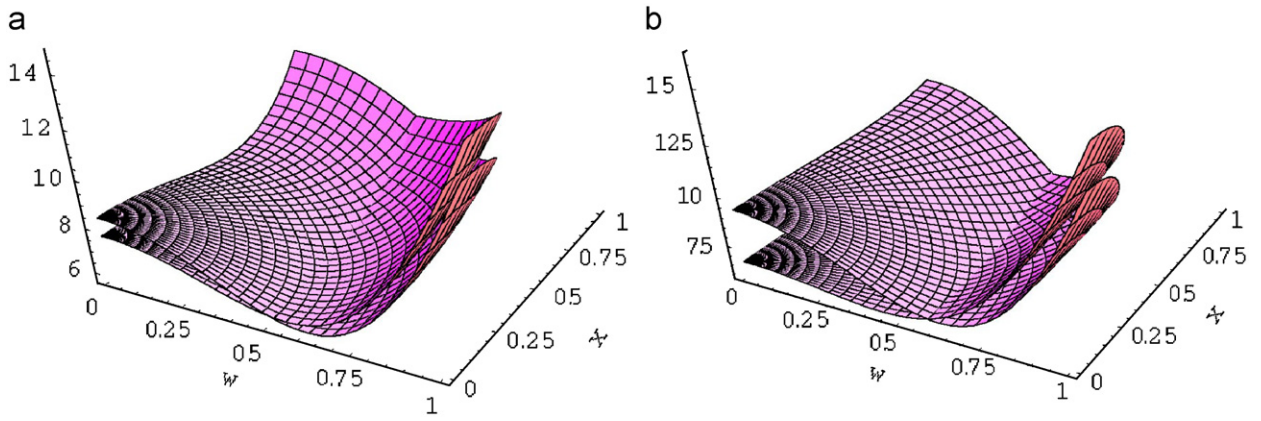

Fig. 8. 3-D VDGs of average SPV for the standard and *I*-optimal CCD for $d = 1$ (left) and 10 (right). The two surfaces cross each other and the *I*-optimal CCD has smaller values at the center.

An additional consideration for SPDs is to take into account the different costs associated with making changes at the whole-plot and sub-plot levels. For a completely randomized experiment, the total number of observations can be a reasonable surrogate for cost. Thus, many of the optimal design criteria use the total number of design points as the penalty to balance the decrease in prediction variance for larger designs. In some SPD cases, it may be prohibitively expensive to do the equipment set-up for each of the whole plots, and hence the only contributor to cost in these situations might be the total number of whole plots in the experiment. In other situations, a whole-plot might be slightly more expensive than a sub-plot, but the total cost of the experiment is more logically thought to be a weighted average of the number of whole plots and the number of sub-plots runs. Liang et al. (2007) develop methodology to compare designs based on a flexible cost structure, using $\#WP + rN$ (with #WP being the number of whole plots and N the total number of observations) to replace N. The user-specified parameter, r, summarizes the cost of the sub-plot observations relative to the cost of changing the whole-plot levels. Note that the cost incurred by measuring the response is considered a part of the sub-plot costs, because it will be applied to each observation. As discussed earlier, combining cost and quality of estimation into a single measure may not be appropriate for all situations, but this approach allows practitioners to make a fairer comparison between designs with different structures and assess the design on a per unit cost basis when this is an appropriate basis for comparison. Parker et al. (2008) give a detailed discussion of how to balance the many trade-offs among designs for the split-plot case.

4.5. Designs for generalized linear models

Similar to the split-plot experiment case considered above, designs involving generalized linear models for understanding the response are also dependent on some characteristics of the model parameters in order to assess the quality of the design. In this case, because the response variance is frequently a function of the response mean, the prediction variance, and hence the quality of the design, is dependent on the unknown model parameters. This complication of needing to know the model parameters before the experiment is even conducted might naturally encourage us to consider robustness to a variety of model parameter values.

If we are focused on the quality of prediction, Ozol-Godfrey et al. (2007) consider the GLM case with canonical links and develop two summaries to help compare designs. The prediction variance at location $\mathbf{x}_0$ in the design space for a generalized linear model is given by $\mathrm{Var}(\hat{\mu}(\mathbf{x}_0)) = [\mathrm{Var}(y(\mathbf{x}_0))]^2\mathbf{x}_0'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_0$, where V is the variance–covariance matrix of the design points. For standard linear models, the SPV is obtained by standardizing this quantity with a cost-factor, N, the number of observations, and dividing by the observation variance. If we do the equivalent standardization for the GLM case we obtain the SPV, $v(\mathbf{x}_0) = N\,\mathrm{Var}(y(\mathbf{x}_0))\mathbf{x}_0'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_0$. It can be shown for the canonical link that this SPV maintains the theoretical G-optimal maximum value in the design space of p, the number of model parameters. Alternately, if we are interested in prediction in the design regions, the penalized prediction variance, PPV, may be of more practical use and is defined as $\rho(x_0) = N\,\mathrm{Var}(\hat{\mu}(\mathbf{x}_0)) = N[\mathrm{Var}(y(\mathbf{x}_0))]^2\mathbf{x}_0'(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}_0$. Although in the standard linear model case these quantities, SPV and PPV, are scalar multiples of each other, they have very different characteristics and distributions in the design space for the generalized linear model setting. Hence, the trade-offs between theoretical and practical aspects of prediction emphases should be carefully considered. Regardless of whether SPV or PPV is more relevant to our assessment, examining the performance of a design across a range of response values is important to gain understanding of how well a design can be expected to perform once the experiment has been run. Ozol-Godfrey and Anderson-Cook (2007) consider several types of misspecification and examine adapted FDS plots to compare designs. Zahran and Myers (2003) show that if the ratio of the maximum to the minimum observation variances does not become too great (as influenced by the range of means observed), many standard designs continue to perform well for the first-order model case.

Although it is tempting to assume that good performance of classical designs continues to hold for generalized linear models, a more thorough investigation of design characteristics in this setting can help us understand the boundaries of when these designs will begin to falter. In addition, using the most appropriate metric for comparing designs needs special attention for this case as practical and theoretical considerations can lead to different conclusions.

5. Conclusions

As with the old adage "measure twice, cut once", considering the many trade-offs between potential designs is relatively simple to do and can pay big dividends with a substantially better result at the analysis stage. After the data have been collected, it may be impossible, impractical, or expensive to augment the design to resolve issues that might not have occurred if a better design were selected initially. Given the time and resources typically spent on running the experiment, investing in a careful selection of the right design makes sense.

In addition to optimal design criteria for both estimation of model parameters and prediction of future observations, a number of other aspects of the design should also be considered. Considering the cost of the experiment in a way that is appropriate to how the data are collected and analyzed as well as building in some protection for when things may go wrong are important aspects to be considered. Different types of experiments have different special features that should be incorporated into assessments and comparisons of experimental designs; this may require specialized tools or adaptation of these numerical and graphical assessment methods.

Although each experimental situation has different relative priorities and sometimes it is difficult to be precise about quantifying some aspects of these comparisons, thinking more broadly about what constitutes a good design is important. If we focus too narrowly on just a single objective, we may end up with a design that has poor performance in many other areas. However, if we balance several considerations, often we can achieve a design that is near-optimal in many areas.

References

Anderson-Cook, C.M., 2005. Statistics roundtable: how to choose the appropriate design. Quality Prog. October, 80–82.

Anderson-Cook, C.M., Borror, C.M., Montgomery, D.C., 2007. Response by the authors to letter to the editor. J. Quality Tech. 39, 91–92.

Anderson-Cook, C.M., Borror, C.M., Jones, B., 2008. Graphical tools for assessing the sensitivity of response surface designs to model misspecification. Technometrics, in press.

Atkinson, A.C., 1970. The design of experiments to estimate the slope of a response surface. Biometrika 57, 319–328.

Borkowski, J.J., 2003. Using a genetic algorithm to generate exact small response surface designs. J. Probab. Statist. Sci. 1, 65–88.

Borror, C.M., Montgomery, D.C., Myers, R.H., 2002. Evaluation of statistical designs for experiments involving noise variables. J. Quality Tech. 34, 54–70.

Box, G.E.P., Draper, N.R., 1975. Robust designs. Biometrika 62, 347–352.

Chung, P., Goldfarb, H., Montgomery, D., Borror, C., 2008. Optimal designs for mixture-process experiments involving continuous and categorical noise variables. Quality Tech. Quantitative Management, to appear.

Elfving, G., 1952. Optimum allocation in linear regression theory. Ann. Math. Statist. 23, 255–262.

Giovannitti-Jensen, A., Myers, R.H., 1989. Graphical assessment of the prediction capability of response surface designs. Technometrics 31, 159–171.

Goldfarb, H.B., Borror, C.M., Montgomery, D.C., 2003. Mixture-process variable experiments with noise variables. J. Quality Tech. 35, 393–405.

Goldfarb, H.B., Anderson-Cook, C.M., Borror, C.M., Montgomery, D.C., 2004a. Fraction of design space to assess the prediction capability of mixture and mixture-process designs. J. Quality Tech. 36, 169–179.

Goldfarb, H.B., Borror, C.M., Montgomery, D.C., Anderson-Cook, C.M., 2004b. Three-dimensional variance dispersion graphs for mixture-process experiments. J. Quality Tech. 36, 109–124.

Hader, R.J., Park, S.H., 1978. Slope-rotatable central composite designs. Technometrics 20, 413–418.

Haines, L.M., 1987. The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. Technometrics 29, 439–447.

Hamada, M., Martz, H.F., Reese, S., Wilson, A., 2001. Finding near optimal Bayesian experimental designs via genetic algorithms. Amer. Statist. 55, 175–181.

Hamada, M., Martz, H.F., Steiner, S.H., 2005. Accounting for mixture errors in analyzing mixture experiments. J. Quality Tech. 37, 139–148.

Heredia-Langner, A., Montgomery, D.C., Carlyle, W.M., Borror, C.M., 2003. Genetic algorithms for the construction of D-optimal designs. J. Quality Tech. 35, 28–36.

Heredia-Langner, A., Montgomery, D.C., Carlyle, W.M., Borror, C.M., 2004. Model-robust optimal designs: a genetic algorithm approach. J. Quality Tech. 36, 263–279.

Hoke, A.T., 1974. Economical second-order designs based on irregular fractions of the 3^n factorial. Technometrics 17, 375–384.

Jang, D.H., Park, S.H., 1993. A measure and a graphical method for evaluating slope rotatability in response surface designs. Comm. Statist. Theory Methods 22, 1849–1863.

Kiefer, J., 1959. Optimum experimental designs (with discussion). J. Roy. Statist. Soc. B 21, 229–249.

Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canad. J. Math. 12, 262–266.

Kowalski, S., 2002. 24 run split-plot experiments for robust parameter design. J. Quality Tech. 34, 399–410.

Liang, L., Anderson-Cook, C.M., Robinson, T.J., 2006a. Fraction of design space plots for split-plot designs. Quality Reliability Eng. Internat. 22, 275–289.

Liang, L., Anderson-Cook, C.M., Robinson, T.J., 2007. Cost penalized estimation and prediction evaluation for split-plot designs. Quality Reliability Eng. Internat. 23, 577–596.

Liang, L., Anderson-Cook, C.M., Robinson, T.J., Myers, R.H., 2006b. Three-dimensional variance dispersion graphs for split-plot designs. J. Comput. Graphical Statist. 15, 757–778.

Lucas, J., 2007. Letter to the Editor: comments on optimal designs for second-order polynomial models. J. Quality Tech. 39, 90–91.

Miró-Quesada, G., del Castillo, E., 2004. Two approaches for improving the dual response method in robust parameter design. J. Quality Tech. 36, 154–168.

Mukerjee, R., Huda, S., 1985. Minimax second- and third-order designs to estimate the slope of a response surface. Biometrika 72, 173–178.

Murty, V., Studden, W., 1972. Optimal designs for estimating the slope of a polynomial regression. J. Amer. Statist. Assoc. 67, 869–873.

Myers, R.H., Lahoda, S., 1975. A generalization of the response surface mean square error criterion with a specific application to the slope. Technometrics 17, 481–486.

Myers, R.H., Montgomery, D.C., 2002. Response Surface Methodology: Process and Product Optimization using Designed Experiments. second ed. Wiley, New York.

Myers, R.H., Khuri, A., Vining, G., 1992. Response surface alternatives to the Taguchi robust parameter design approach. Amer. Statist. 46, 131–139.
Myers, R.H., Montgomery, D.C., Vining, G.G., Borror, C.M., Kowalski, S.M., 2004. Response surface methodology: a retrospective and literature survey. J. Quality Tech. 36, 53–77.
Myers, R.H., Vining, G.G., Giovannitti-Jensen, A., Myers, S.L., 1992. Variance dispersion properties of second-order response surface designs. J. Quality Tech. 24, 1–11.
Nair, V., 1992. Taguchi's parameter design: a panel discussion. Technometrics 34, 127–161.
Ozol-Godfrey, A., Anderson-Cook, C.M., 2007. Fraction of design space plots for reexamining mixture design robustness to measurement errors. J. Statist. Appl. 1, 171–183.
Ozol-Godfrey, A., Anderson-Cook, C.M., Montgomery, D.C., 2005. Fraction of design space plots for reexamining model robustness. J. Quality Tech. 37, 223–235.
Ozol-Godfrey, A., Anderson-Cook, C.M., Robinson, T.J., 2007. Fraction of design space plots for generalized linear models. J. Statist. Plann. Inference 138, 203–219.
Park, Y.-J., Richardson, D.E., Ozol-Godfrey, A., Borror, C.M., Anderson-Cook, C.M., Montgomery, D.C., 2005. Prediction variance properties of second-order response surface designs for cuboidal regions. J. Quality Tech. 37, 253–266.
Parker, P.A., Anderson-Cook, C.M., Robinson, T.J., and Liang, L., 2008. Robust split-plot designs. Quality Reliability Eng. Internat., in press.
Piepel, G.F., Anderson, C.M., 1992. Variance dispersion graphs for designs on polyhedral regions. In: 1992 Proceedings of the Section on Physical and Engineering Sciences.American Statistical Association, Alexandria, Virginia, pp. 111–117.
Piepel, G.F., Anderson, C.M., Redgate, P.E., 1993a. Variance dispersion graphs for designs on polyhedral regions—revisited. In: 1993 Proceedings of the Section on Physical and Engineering Sciences.American Statistical Association, Alexandria, Virginia, pp. 102–107.
Piepel, G.F., Anderson, C.M., Redgate, P.E., 1993b. Response surface designs for irregularly-shaped regions (Parts 1–3). In: 1993 Proceedings of the Section on Physical and Engineering Sciences.American Statistical Association, Alexandria, Virginia, pp. 205–227.
Roquemore, R., 1976. Hybrid designs for quadratic response surfaces. Technometrics 18, 419–423.
Scheffé, H., 1958. Experiments with mixtures. J. Roy. Statist. Soc. Ser. B 20, 344–360.
Steiner, S.H., Hamada, M., 1997. Making mixtures robust to noise and mixing measurement errors. J. Quality Tech. 29, 441–450.
Taguchi, G., 1986. Introduction to Quality Engineering. UNIPUB/Kraus International, White Plains, NY.
Taguchi, G., 1987. System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost. UNIPUB/Kraus International, White Plains, NY.
Vining, G.G., Myers, R.H., 1991. A graphical approach for evaluating response surface designs in terms of the mean squared error of prediction. Technometrics 33, 315–326.
Wald, A., 1943. On the efficient design of statistical investigations. Ann. Math. Statist. 14, 134–140.
Welch, W.J., 1982. Branch and bound search for experimental designs based on D-optimality and other criteria. Technometrics 24, 41–48.
Welch, W.J., 1983. A mean squared error criterion for the design of experiments. Biometrika 70, 205–213.
Welch, W., Yu, T., Kang, S., Sacks, J., 1990. Computer experiments for quality control by parameter design. J. Quality Tech. 22, 15–22.
Zahran, A., Anderson-Cook, C.M., Myers, R.H., 2003. Fraction of design space to assess the prediction capability of response surface designs. J. Quality Tech. 35, 377–386.
Zahran, A., Myers, R.H., 2003. Use of standard factorial designs with generalized linear models. Virginia Tech Department of Statistics Technical Report 03-4.
Zhou, J., 2001. A robust criterion for experimental designs for serially correlated observations. Technometrics 43, 462–467.

Journal of Statistical Planning and Inference 139 (2009) 642 – 644

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison" by Christine Anderson-Cook, Connie Borror and Douglas Montgomery

Bradley Jones

Statistical Research and Development, SAS Institute Inc., USA

1. Discussion

I am honored and a little overwhelmed by the responsibility to discuss a paper having such a large scope. My strategy is to focus on a few issues and hope that the other discussants will fill in the gaps.

2. Make sure to match the design to the problem

To start, I would like to address the important question of what makes an RSM design good. The authors give their list of favorites. Naturally, this list has to do with statistical issues. It assumes that the really important questions have already been answered. What makes some experiments less useful than others is often a failure to seek answers to questions like these:

```
what is your response and how do you measure it?
is there more than one input stream to the process?
is day-to-day, setup to setup, or lot-to-lot variation a concern?
are there variables that are difficult to change from run to run?
are there many machines, operators or locations to consider?
have you removed factors due to budget considerations?
```

The crucial first step is to make sure that the proposed experimental plan takes account of the unique features of the real system under study. A design with perfect statistical properties, which relies on an over-simplification of the real system, is solving the wrong problem.

Simplification of the system description was, perhaps, a necessity in the days of hand calculation. Current computing power makes the generation and graphical comparison of many designs possible before making a final choice. This leads me to what I regard as a key point of the paper. That is, while computer-aided design construction requires a scalar function to optimize, graphs are an essential tool for comparing various design alternatives. Relying on single number efficiency comparisons is flying blind.

3. Use graphs to compare designs

Fraction of the design space (FDS) plots and variance dispersion graphs (VDG) are two graphical tools that the authors apply for design comparison to a wide variety of challenging design scenarios. Both of these plots are useful and I would like to recommend another graph called the prediction variance profile plot. Fig. 1 shows an example of this plot for the saturated I-optimal design in Table 1. Note that the *a priori* model for this design is the full quadratic model in three factors.

This graph is actually a vector of plots – one for each factor. Each plots the curve of the unscaled prediction variance for each factor conditioned on the values of the other factors. The displayed variance, 2.893333, is for the point [−1 1 1]. That is the point where each of the dashed vertical lines falls onto the x-axis. For example, the plot on the far left shows the conditional functional relationship between $X1$ and the prediction variance when $X2$ and $X3$ are both 1.

One powerful addition to this graph is to make it interactive. By dragging any of the vertical lines you can observe how the traces in the other plots change. It makes the prediction variance of any factor setting accessible.

The plotted point is one of the points that attain the maximum prediction variance in the cubic factor space. If one were to consider augmenting the design with one point, this would be the best one for improving the D-efficiency of the design.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.005

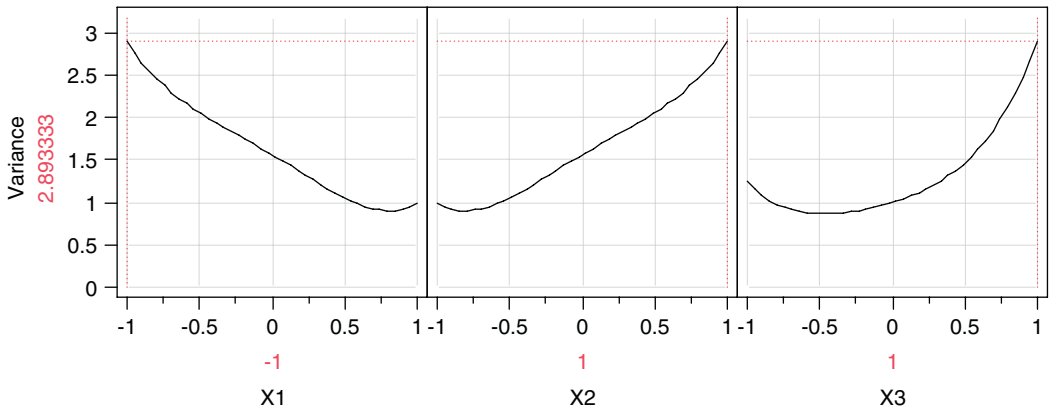

Fig. 1. Prediction variance profile plot of design in Table 1.

Table 1
Saturated I-optimal three-factor design

Run	*X*1	*X*2	*X*3
1	1	−1	1
2	1	−1	−1
3	0	1	−1
4	1	1	1
5	−1	−1	1
6	−1	0	−1
7	−1	1	0
8	0	−1	0
9	0	0	1
10	1	0	0

If the design were 100% G-efficient the maximum prediction variance would be the number of parameters divided by the sample size, which equals 1. For this design the maximum prediction variance is nearly three times larger, which indicates the statistical price that is being paid for the cost reduction of choosing a bare minimum for the number of runs.

By comparison, if you were able to afford six more runs you could run a face-centered CCD with 2 centerpoints. This design is also I-optimal for a cubic design space and a quadratic model. Its worst prediction variance is 0.79569, which is between 3 and 4 times smaller than the 10 run design. If additional runs are not costly, it is a good deal to spend 60% more for a greater than threefold reduction in variance.

4. Scaled or unscaled prediction variance—whatever buys you more runs

The use of scaled prediction variances allows one to overlay prediction variance traces for designs with different sample sizes on the same plot to assess their efficiency on a per run basis. If the budget for experimentation is negotiable, this kind of efficiency increase may help sell a larger sample size. On the other hand, the per run efficiency of the I-optimal 20 run design is worse than the face-centered CCD. Yet, I would always choose to run a design with more runs given the opportunity.

Another scaling to consider is the square root of the unscaled prediction variance. This quantity is in the units of the response; hence it is physically meaningful. It is also proportional to the length of prediction intervals.

5. Response surface designs for first-order models

The authors dispense with first-order RSMs in short order by recommending the use of orthogonal designs for cubic or spherical design regions. The two-level fractional factorial and Plackett-Burman designs are well known and exist for sample sizes that are powers of two or multiples of four, respectively. But, it is possible to create an orthogonal design on a sphere for the first-order model for any sample size. Here is a short algorithm.

(1) Start with an by $n-1$ matrix of random numbers and call it X.
(2) Center the columns by subtracting the column means.
(3) Add a column of ones to the matrix.

(4) Decompose X using singular value decomposition into $U * S * V'$.
(5) Remove the constant column from U.

The matrix, U, is an orthogonal design with n runs. Each row is a distance of 1 from the center of the design. You can scale U by multiplying by a constant, say the square root of n. These design alternatives are not generally known but could be useful. Note that these designs require the ability to set the factors precisely and they are not two-level designs.

6. Response surface designs for dozens of factors

Recently, I have seen several applications requiring designs for fitting full second-order models for two or three dozen factors. Traditionally, of course, screening these factors down to a smaller number would be recommended before doing an RSM design. Yet, the experimenters in each application were determined to proceed immediately to designing for a full second-order approximation model. Such a design for 30 factors would require at least 496 runs since there are 496 unknown coefficients in the model.

It seems doubtful that every one of these coefficients is non-negligible. More probably, assuming that the pareto principle holds, several dozen main effects, two-factor interactions and quadratic effects would be adequate to explain virtually all the nonrandom variation in the data. The problem is that *a priori* we do not know which several dozens of the 496 effects are the important ones. However, using a Bayesian D-optimal design as defined in DuMouchel and Jones (1994), an experimental plan requiring somewhere between 200 and 300 runs should be adequate to identify all the driving effects and make decent predictions. The prospect of saving hundreds of runs would seem an attractive alternative.

7. Conclusion

The key message of the paper is an important one and deserves one more repetition. Given the ability to generate several competitive designs for the solution of any particular design problem, it makes imminent sense to compare these designs graphically.

Reference

DuMouchel, W., Jones, B., 1994. A Simple Bayesian Modification of D-Optimal Designs to Reduce Dependence on an Assumed Model. Technometrics 36, 37–47.

Journal of Statistical Planning and Inference 139 (2009) 645 – 646

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion—"Response surface design evaluation and comparison" by Christine M. Anderson-Cook, Connie M. Borror, Douglas C. Montgomery

Peter A. Parker

National Aeronautics and Space Administration, Hampton, VA 23681, USA

I congratulate the authors on a thorough exposition of the evaluation and comparison of response surface designs that will serve as a useful reference to practitioners. In addition, they have highlighted the quintessential element in design comparison; namely, the need for trade-offs among competing criteria. In this opportunity to discuss the article, I focus on the statistician's responsibility to accurately translate the experimental objectives into statistical performance measures. After all, an accurate interpretation of the objectives is the necessary foundation to evaluate the suitability of a response surface design for a particular application. I also offer some suggestions to improve the communication between the statistician and the subject-matter expert (SME) to identify important design criteria.

The first step in a design planning exercise is to obtain an unambiguous definition of the experimental objectives. An excellent systematic framework to seek these sometimes elusive objectives is provided by Coleman and Montgomery (1993). In fact, most statisticians would likely agree that simply defining the factors and responses in an experiment can be a formidable task. Taking this as the typical starting point, it is easy to see why concepts such as statistical design efficiency are even more challenging to a SME who is not acquainted with this terminology. Therefore, some may argue that more sophisticated evaluation metrics, such as design efficiency, are not important to practical problems. I disagree and believe that the authors have listed important criteria and evaluation tools to quantify a design's ability to achieve the overall experimental objectives and, with proper communication, are usually helpful to the SME. The responsibility rests on the statistician's ability to translate and communicate the experimental objectives into precise, quantifiable measures of performance and explain the relative trade-offs among multiple evaluation criteria.

From my experience in experiment planning, I find it helpful to ask the SME what type of statement they would like to make about the experimental results. Obviously, this question reveals the SME's pre-experimental bias, which is beneficial to know, but it also helps to identify the most important criteria for their experimental design. For example, some common answers to this question follow.

1. We want to find the optimum factor levels (operating conditions) to maximize the performance (yield) of our system (process).
2. We want to use a response surface model (equation) to predict a response based on a given combination of factor levels with a specified level of accuracy (prediction quality).
3. We want to understand the relative magnitude and influence of the factors on the response, including how the factors interact.

While all of these questions require a design that supports the development of a response surface model, they also indicate the relative importance of the design characteristics (Box and Draper, 1975) and efficiency measures. For case (1), the SME is not particularly interested in the coefficients of the response surface model or in the model form itself, rather the model is a tool used to find an optimum within the design space or to discover that an optimum may occur outside of the current design space. Depending on the prior knowledge of the optimum location, a design with a low average prediction quality (I-optimal criteria) over the entire design space may be desirable. In this type of experiment, minimizing the experimental resources (efficiency) to find the optimum is likely an important criterion. This case represents a classical response surface methodology application.

Case (2) also points us toward prediction quality; however, there is a requirement to achieve a specified prediction variance. In this case, comparing designs based on statistical efficiency would not be as appropriate, rather we use the unscaled prediction

DOI of the original article: 10.1016/j.jspi.2008.04.004.
E-mail address: peter.a.parker@nasa.gov.

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2008.04.006

variance. From a SME perspective, it is more natural to specify the prediction quality (variance) in engineering units, rather than discussing efficiency that describes the prediction performance per experimental run. For this case, the ability to sequentially build models of increasing order and replication throughout the design space are often important. As a note, including a set of confirmation points, not used to build the response surface model, is a vital component that is frequently overlooked. This application illustrates a calibration experiment requiring a response surface design.

Lastly in case (3), the SME has a primary interest in the coefficients themselves. Clearly, this leads us to D-optimality as an important criteria. If the model is empirical in nature, we should include sufficient information to test for lack of fit. Additionally, the SME may have a requirement for the estimation precision of the model coefficients, and therefore sufficient degrees of freedom are required for estimation and inference. Similar to the unscaled predication variance in case (2), we might focus on unscaled coefficient estimation performance. In this application, the SME usually possesses a physical interpretation of the coefficients and is interested in estimating them to a specified precision to better understand the underlying physics.

In practice, the determination of the appropriate criteria is not as straightforward as in these simple examples. There is usually interest in both the predictive capability of the model and the coefficients; however, the relative importance is dependent on the SME's goals. We recognize that identifying the most important criteria is an iterative process that should not be prematurely stopped until both the statistician and the SME are satisfied in their understanding of the design's performance.

Once the statistician has understood and correctly translated the experimental objectives into performance measures, then the characteristics and methods presented in the article provide powerful tools to evaluate competing designs. In particular, the graphical design evaluation tools provide a wealth of information that illustrates performance throughout the design space, rather than a single number metric. A clear description of the graphical methods to the SME that highlights the impact to their application is vital, otherwise it just becomes another interesting plot. Once the lexicon is established, then competing designs can be easily compared and the SME usually develops a rapid appreciation for the power of these tools. In most situations, the SME needs to present the final experimental design and justification to a review committee of their peers, not statisticians, and these graphical methods are particularly helpful for that purpose.

In conclusion, I think it is important to state that good experimental designs are not selected, instead they are built for a particular application. In my opinion, the building of a good design always involves multiple competing criteria and must be made with respect to a particular experimental application. Design building is an iterative process that incorporates many non-statistical criteria, such as the restrictions of the experimental apparatus, cost, schedule, and administrative factors. By presenting multiple designs, and discussing their relative benefits in the SME's language, enables faster convergence to a final design. Ultimately at the end of this process, we should be confident that we have built a design that answers the right experimental questions, rather than simply selecting a good statistical design. I define a successful design building and evaluation process as one that results in a design that is explainable by the SME to their peers. If this is achieved, then I have found it to be a rewarding experience when the experimental design becomes the SME's design, rather than the statisticians.

References

Box, G.E.P., Draper, N.R., 1975. Robust designs. Biometrika 62, 347–352.
Coleman, D.E., Montgomery, D.C., 1993. A systematic approach to planning for a designed industrial experiment, (with Discussion). Technometrics 35, 1–27.

Journal of Statistical Planning and Inference 139 (2009) 647 – 649

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison" by Christine M. Anderson-Cook, Connie M. Borror, Douglas C. Montgomery

André I. Khuri*

Department of Statistics, University of Florida, PO Box 118545, 103 Griffin-Floyd Hall, Gainesville, FL 32611-8545, USA

The article by Anderson-Cook, Borror, and Montgomery provides an interesting overview of the use of graphical techniques for the evaluation and comparison of response surface designs. They focus on second-degree response surface models. However, they also include discussions concerning robust parameter designs, designs with special structures such as mixture and split-plot designs, in addition to designs for generalized linear models (GLMs).

I applaud the authors' bringing attention to the graphical approach to compare response surface designs. Single-number criteria such as D- and G-efficiency have for long been used to evaluate the "goodness" of a response surface design. This is particularly true in the case of designs for GLMs and nonlinear models in general. As the authors pointed out, D- and G-efficiency, as well as other alphabetic optimality criteria, do not provide complete information about the quality of prediction throughout the experimental region under consideration. In addition, these criteria can, in some cases, yield rather awkward designs as was demonstrated by Dette and Sahm (1997).

The authors focus on two graphical techniques for comparing response surface designs, namely, those based on using *variance dispersion graphs* (VDGs) and *fraction of design space* (FDS) plots. The former was developed by Giovannitti-Jensen and Myers (1989) and the latter was introduced by Zahran et al. (2003). The VDGs plot the maximum, minimum, and the average of the scaled prediction variance (SPV) values over concentric hyperspheres, chosen within the experimental region, against their radii. The FDS plots are obtained by making a correspondence between predetermined SPV values and the proportions of the volumes of the subregions (of the experimental region) within which the SPV function is less than the predetermined values. Both techniques were widely used by the authors and some of their co-workers in a variety of situations.

The authors, however, fail to mention or even make a reference to other graphical techniques, namely, those based on the use of *quantal plots* (QPs) of the prediction variance and the *quantile dispersion graphs* (QDGs). The former was developed by Khuri et al. (1996) for a typical response surface situation based on using the standard linear model, and the latter was introduced by Robinson and Khuri (2003) for the purpose of dealing with GLMs and nonlinear models in general. The first introduction of QDGs, however, was made by Khuri (1997) in an analysis of variance (ANOVA) situation involving the comparison of designs for estimating variance components. In the QPs approach for a response surface design, the entire distribution of the SPV on a given hypersphere inside a region of interest is determined in terms of its quantiles. Plots of such quantiles corresponding to hyperspheres of varying radii clearly depict the effect of the design on the prediction variance. These plots can then be effectively used to compare several candidate designs for fitting polynomial-based response surface models. The QPs were particularly developed to enhance the amount of information that can be gleaned from the use of VDGs. It is obviously true that having knowledge of the entire distribution of the SPV on a given hypersphere is more informative than knowing only its extremes (maximum and minimum) as is the case with the VDGs. Khuri et al. (1996) demonstrate that it is possible to have two designs with almost identical VDGs patterns, but widely different quantile plots, throughout the region of interest. This indicates that the QPs can discriminate between two designs when the corresponding VDGs fail to do so. The use of QPs was later extended to other types of response surface models defined over regions that are not necessarily hyperspherical. Khuri et al. (1999) used the QPs to compare designs for mixture models defined on possibly irregularly shaped constrained regions in the mixture space.

* Tel.: +1 352 376 0002; fax: +1 352 392 5175.
E-mail address: ufakhuri@stat.ufl.edu.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.007

The QDGs approach deals exclusively with nonlinear situations under nonstandard conditions where the design depends on unknown parameters. For example, designs for estimating variance components depend on ratios of unknown variance components. Comparisons among such designs can be easily made using this approach as was demonstrated in Khuri (1997) and Lee and Khuri (1999, 2000). Comparisons of designs for nonlinear models under standard conditions on the error distribution were described in Khuri and Lee (1998). However, the main thrust regarding the use of QDGs has been in the areas of GLMs and response surface models with random effects, as in a split-plot design situation. Robinson and Khuri (2003) used QDGs for evaluating and comparing designs for logistic regression models (see also Khuri and Mukhopadhyay, 2006, who applied the QDGs to GLM situations using Poisson regression). Saha and Khuri (2008) used QDGs to compare designs for response surface models with random block effects. In any of these situations, the design depends on unknown parameters. For example, in the case of a GLM, the design depends on the unknown parameters of the linear predictor. This necessitates the specification of a *parameter space*, $\mathscr{C}$, on the unknown parameters. Then, for a given point in this space, quantiles of a certain criterion function are obtained on concentric hypersurfaces that can be obtained by shrinking the boundary of the region of interest using a shrinkage factor. By repeating this process using several points chosen from $\mathscr{C}$ we can calculate for each p (the proportion corresponding to a selected quantile) and a given hypersurface the minimum and maximum over $\mathscr{C}$ of the pth quantile. Plotting these values against p produces the QDGs for the design under consideration. The criterion function can be the SPV, the mean squared error of prediction, or the variance of a certain estimate of a variance component. The parameter space is obtained by constructing a confidence region on the unknown parameters. Thus the QDGs approach can be used in a wide variety of situations. In each case, it enables the researcher to assess the "goodness" of a design throughout the region of interest. It also provides a clear depiction of the dependence of the design on the unknown parameters of the model. An overview of the QDG approach was given in Khuri (2003).

While the FDS plots account for the proportions of the volumes of the regions in the design space within which the SPV does not exceed certain predetermined values, the QPs can indicate where in the design space the SPV can be large resulting in unreliable predictions. Hence, in this respect, the FDS and QPs complement each other and should therefore be both utilized. On the other hand, the dependence of the design on the unknown parameters of the model (for GLM, nonlinear, or ANOVA models with random effects) is the primary reason for the development of the QDGs where the parameter space $\mathscr{C}$ is constructed in a formal fashion as was mentioned earlier. Furthermore, the dispersion in the quantile values clearly indicates how robust a design is to changes in the unknown parameter values. By contrast, the design dependence problem (on the unknown parameters) is given less prominence in the FDS approach. Here, robustness "to misspecification of the variance ratio, d", for example, in Section 4.3, is only briefly mentioned. The variance ratio is just one unknown parameter. It is not clear what can be done in situations involving several unknown parameters as is the case with GLMs.

Models for split-plot designs as well as those that contain control and noise factors depend on random effects with possibly unknown variance components. Hence, designs for such models also depend on several unknown parameters. In this case, "considering robustness to a variety of model parameter values", as the authors state in Section 4.4, can become a serious and complex problem that can hamper the practical application of the FDS approach.

In addition to GLMs, *generalized linear mixed models* (GLMMs) have received considerable attention in the biological and medical sciences. These models represent an extension of GLMs that includes random effects. An expository treatment of the use of GLMMs in industrial split-plot experiments was given by Robinson et al. (2004). More recently, Robinson et al. (2006) discussed the use of GLMMs for the analysis of robust parameter design problems. It would be quite a challenge to apply any graphical technique for comparing designs in this situation. Obviously, such design comparisons should be based on criteria that can lead to adequate estimation of the process mean as well as the process variance.

The authors make no reference to sequential experimentation. Many applications of response surface methodology are sequential in nature whereby information acquired in one stage of the experiment is used to plan the next stage. It is also possible that variables included in one stage may be dropped in later stages and new variables are introduced into the model. Thus in the sequential approach, the experimenter is not necessarily dealing with a fixed model, but rather, with a multitude of models. This brings attention to the need to design an experiment in stages through design augmentation. It would be of interest to see how the graphical techniques can be effectively used in building up the design in such situations.

Any graphical technique depends on the presumption that the fitted model is correct. Design plots can be useless if they are based on the wrong model. It is well recognized in response surface methodology that the design should give good detectability of model lack of fit. This property was one of the several design properties listed by Box and Draper (1975). Thus it would be desirable for the design to be robust to model misspecification. The authors allude to this problem in Section 4.1. If a chosen design is deemed to lack such robustness, then it would be of interest to know what course of action to take to beef up its robustness. Perhaps this can be accomplished by design augmentation. Furthermore, in situations where the experimenter is willing to tolerate some model bias, it would be necessary to use the mean squared error criterion rather than the SPV as a criterion for comparing designs. Estimation bias can also occur due to using maximum likelihood in GLMs and nonlinear models in general. In such cases, the design will become dependent on unknown parameters and this should be reflected in the proper choice of the graphical technique needed to evaluate the design.

In conclusion, it is hoped that the graphical approach for comparing and evaluating designs will continue to grow. More work is still needed to further its development under a variety of experimental conditions. In this respect, it would be very desirable to have the needed software to implement the graphical techniques which are currently available. This is the only way to make these techniques easily accessible to potential users.

References

Box, G.E.P., Draper, N.R., 1975. Robust designs. Biometrika 62, 347–352.
Dette, H., Sahm, M., 1997. Standardized optimal designs for binary response experiments. South African Statist. J. 31, 271–298.
Giovannitti-Jensen, A., Myers, R.H., 1989. Graphical assessment of the prediction capability of response surface designs. Technometrics 31, 159–171.
Khuri, A.I., 1997. Quantile dispersion graphs for analysis of variance estimates of variance components. J. Appl. Statist. 24, 711–722.
Khuri, A.I., 2003. Current modeling and design issues in response surface methodology: GLMs and models with block effects. In: Khattree, R., Rao, C.R. (Eds.), Handbook of Statistics in Industry, vol. 22. Elsevier Science, Amsterdam, pp. 209–229.
Khuri, A.I., Lee, J., 1998. A graphical approach for evaluating and comparing designs for nonlinear models. Comput. Statist. Data Anal. 27, 433–443.
Khuri, A.I., Mukhopadhyay, S., 2006. GLM designs: the dependence on unknown parameters dilemma. In: Khuri, A.I. (Ed.), Response Surface Methodology and Related Topics. World Scientific, Singapore, pp. 203–223.
Khuri, A.I., Kim, H.J., Um, Y., 1996. Quantile plots of the prediction variance for response surface designs. Comput. Statist. Data Anal. 22, 395–407.
Khuri, A.I., Harrison, J.M., Cornell, J.A., 1999. Using quantile plots of the prediction variance for comparing designs for a constrained mixture region: an application involving a fertilizer experiment. Appl. Statist. 48, 521–532.
Lee, J., Khuri, A.I., 1999. Graphical technique for comparing designs for random models. J. Appl. Statist. 26, 933–947.
Lee, J., Khuri, A.I., 2000. Quantile dispersion graphs for the comparison of designs for a random two-way model. J. Statist. Plann. Inference 91, 123–137.
Robinson, K.S., Khuri, A.I., 2003. Quantile dispersion graphs for evaluating and comparing designs for logistic regression models. Comput. Statist. Data Anal. 43, 47–62.
Robinson, T.J., Myers, R.H., Montgomery, D.C., 2004. Analysis considerations in industrial split-plot experiments with non-normal responses. J. Quality Technol. 36, 180–192.
Robinson, T.J., Wulff, S.S., Montgomery, D.C., Khuri, A.I., 2006. Robust parameter design using generalized linear mixed models. J. Quality Technol. 38, 65–75.
Saha, S., Khuri, A.I., 2008. Comparison of designs for response surface models with random block effects. J. Quality Technol. Quality Management, in press.
Zahran, A., Anderson-Cook, C.M., Myers, R.H., 2003. Fraction of design space to assess prediction capability of response surface designs. J. Quality Technol. 35, 377–386.

Journal of Statistical Planning and Inference 139 (2009) 650 – 652

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison" by Christine Anderson-Cook, Connie Borror, and Douglas Montgomery

John J. Borkowski

Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA

1. Introduction

The authors are to be congratulated for their summary of the various criteria used for the evaluation and comparison of response surface designs. The crux of the design selection problem lies in the choice of competing criteria, and although one design is superior based on one criterion, it may not fare so well in comparison to other designs when other criteria are considered.

The 10-item list suggested by Myers and Montgomery (2002) provide a broad spectrum of characteristics that are desirable of a good experimental design, and, in particular, response surface designs. The question naturally arises: How can the design selection criteria (including the graphical methods) be used to select a design that best satisfies these characteristics? Of course, there is not a simple answer to this question, because tradeoffs or compromises often must be made to choose an acceptable design. Thankfully, the authors provide thoughtful and practical guidance as the various methods and criteria for effective design selection are discussed.

My discussion will focus on the salient points of their discussion with emphasis on the properties of the prediction variance function on which many of the design selection criteria are based. Hopefully my comments will provide some useful supplemental information to further flush out points that I find most interesting and important when evaluating designs.

2. Design optimality criteria

When restricting the response surface problem to response optimization, we want to select a design that will provide a good-fitting model to the data, and, in particular, provide reliable parameter estimates, which then can be used for precise predictions. As the authors state, second-order models are primarily used for this purpose. Other models will be considered when there are restrictions on randomizations in the form of blocking or split-plotting attributable to factors whose levels are hard or costly to change.

Inherently, the prediction variance function should provide the necessary information for an assessment of the stability of predictions throughout the design space. One problem is how to synthesize useful information from a high-dimensional problem (e.g., our design may contain a large number of factors, and, hence, the number of model parameters to be estimated is considerably larger). Single-value criteria (such as the D, A, G, and I optimality criteria) attempt to address estimation or prediction properties of a design via analysis of its variance properties. The authors briefly summarize how these optimality criteria are related to the variance of estimation and prediction.

Although efficient designs can be generated using commercially available software, I believe there are some issues that are not sufficiently addressed in the statistical literature. Because these computer algorithms work with a discretized form of a continuous design space, the researcher must either provide a candidate set to select points from or use some default candidate set associated with that software package. Therefore, the candidate set is dependent on the design space. If the space is a k-dimensional hypercube, it is common to create a L^k factorial set of points with each factor having L equispaced levels, with three levels set at ± 1 and 0 being sufficient assuming a second-order model for a sufficiently large design size. For saturated or near-saturated model designs, however, additional levels will be needed to achieve higher efficiency. In general, it is recommended that L be odd so that mid-level 0 is considered in the candidate set required for a design-generation algorithm.

In the response surface literature, however, much research has been done assuming a hyperspherical design space (e.g., for central composite, Box-Behnken, and hybrid designs). If the design space is truly hyperspherical, then what is an appropriate

E-mail address: jobo@math.montana.edu.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.011

candidate set that supports a highly efficient design? And, how can it easily be generated as input into software? For example, consider the six-parameter second-order model for a two-factor design with a circular design space, and a circular lattice of candidate points given by $x_1 = r\cos(\theta)$ and $x_2 = r\sin(\theta)$ for $r = 0$ to 1 by 0.1 and $\theta = 0$ to $(2t-1)\pi/t$ by $1/t$. Suppose $t = 4$ ($t = 10$), then points $\pi/4$ ($\pi/10$) radians apart are generated for each radius r yielding 91 (211) unique points. Using the Optex procedure in SAS, the best nine-point designs have D-efficiencies of 24.9436 and 24.9173 for $t = 4$ and 10, respectively. Thus, the smaller candidate set produced the more efficient design. However, for an 11-point design, the D-efficiencies are 24.8384 and 25.0151 for $t = 4$ and 10. In this case, the larger candidate set produced the more efficient design. For more experienced practitioners, multiple candidate sets may be considered with the best design retained (or, create one very large candidate set at the expense of increased generation time). For those less experienced, where to begin is not obvious in candidate set generation in hyperspherical spaces. Although the D efficiencies are close in this example, this may not necessarily be the case for higher dimensional problems, or when another criterion (e.g., G-efficiency) is considered. Similar issues also apply to designs in irregularly shaped spaces related to factor-level constraints, such as designs for mixture experiments having upper and lower component constraints, and, in particular, when multiple component constraints exist.

3. Graphical methods

Suppose that two competing designs have similar efficiencies. Which one should be chosen? As the authors' state, single-value design optimality criteria provide only limited information about a design's prediction variance properties. We need additional information to assist in the design selection process. This is where graphical methods can be highly effective and informative tools for evaluating and comparing designs.

I am a strong supporter of the use of variance dispersion graphs (VDGs) and fraction of design space (FDS) plots in design comparison. Despite this personal endorsement, I still have some minor concerns about their use. One concern is the scaling issue: Should we be plotting the scaled prediction variance (SPV) or the unscaled prediction variance (UPV)? I was grateful to see that the authors address this by emphasizing that we are trying to find the best design per unit cost when VDGs or FDS plots are based on the SPV, and that although one design may be superior based on the SPV, it could very well be inferior when considering the UPV.

I also do not want to discount the use of prediction variance quantile (PVQ) plots (Khuri et al., 1996) which provide information about the distributions of SPV or UPV values at specified radii from the center of the design space. Nguyen and Borkowski (2008) combine aspects of VDGs and QDGs into a single plot, called *volatility plots*, that provide information about the distribution of prediction variances throughout the space. Multiple plots based on both the SPV and the UPV functions should certainly be examined in the design comparison and evaluation process.

A second concern is again related to the irregularly shaped design spaces, and, in particular, for mixture designs having component constraints. In these cases, there is not an obvious definition of a VDG given that a plot based on the spherical radius is not intuitive. One graphical display is the prediction variance trace (PVT). PVTs, however, provide limited prediction variance information in certain directions, e.g., for Cox directions (Vining et al., 1993). The shrunken region VDG (Piepel and Anderson, 1993) is a reasonable approach to graphically display the SPV or UPV in an irregularly shaped space. Componentwise VDGs (Borkowski, 2006) are a second alternative.

Although the traditional FDS plots can still be generated for these spaces, one potential drawback is the generation of random sets of points in high-dimensional, highly constrained spaces. An inclusion/exclusion approach can always be used in which random sets of points are generated in a larger region and then only include those points contained in the subspace defined the factor constraints. The practicality of this approach, however, decays with increasing dimensionality as well as with decreasing volume of the subspace relative to the original space (e.g., a mixture experiment with at least 10 components, each with upper and lower bound constraints and some with very narrow ranges, as well as several multiple component constraints). Efficient generation of random points in these type of spaces is a practical issue when it come to generating FDS plots.

My comments suggest the need for continued software development for both standard and nonstandard design space scenarios.

4. Model robustness

A response surface design's robustness to a suite of competing models is another issue that is fortunately receiving more attention. I was pleased to read the authors' discussion of recent approaches to address this problem, and, in particular with respect to the use of computer-intensive methods. For example, strategies utilizing genetic algorithms that have successfully been applied to the design-generation problem for a specific model have recently been extended to the model robustness problem. The use of desirability functions has great potential (e.g., Heredia-Langner et al., 2004) for generating highly efficient designs. Another approach is based on the 'weighted optimality' criterion for which the goal is to maximize a weighted version of D, G, or I-efficiency (Borkowski and Chomtee, submitted for publication) where the weights are assigned to a suite of potential models and are based on the heredity principle (Chipman, 1996; Chipman and Hamada, 1996).

A natural application of the robustness of a design exists with mixture experiments. It is common to fit several competing models once the data have been collected. For example, Scheffé linear, quadratic, and special cubic models may all be fitted, and are then compared (Cornell, 2002; Smith, 2005). If all competing models were incorporated in the design evaluation and comparison

process, it is unclear what resulting design would emerge and how similar or different it would be to commonly implemented designs, such as extreme vertices designs or efficient computer-generated designs based on a single, specified model.

5. Final comments

Although my comments have focused on issues related to optimality criteria, graphical methods, mixture designs, and model robustness, extensions can be made to other topics. For example, how should we deal with model robustness concerns for generalized linear models? How robust is a design to possible factor-level transformations (e.g., what if we are unsure if we should use $\log(x)$ or $\sqrt{x}$ instead of x as a predictor)? Although recent contributions to design evaluation and comparison have been numerous, there are still significant contributions waiting to be made.

Once again, I thank the authors for their timely discussion on design evaluation and comparison. I eagerly look forward to the future contributions of these authors and all of the other researchers that further the development of new methodologies.

References

Borkowski, J.J., 2006. Graphical methods for assessing the prediction capability of response surface designs. In: Khuri, A.I. (Ed.), Response Surface Methodology and Related Topics. World Scientific Press, Singapore, pp. 349–378 (Chapter 14).

Borkowski, J.J., Chomtee, B., submitted for publication. Using weak and strong heredity to generate weighted design optimality criteria for response surface designs. J. Probab. Statist. Sci.

Chipman, H.A., 1996. Bayesian variable selection with related predictors. Canad. J. Statist. 24, 17–36.

Chipman, H.A., Hamada, M.S., 1996. Discussion: factor-based or effect-based modeling? Implications for design. Technometrics 41, 317–320.

Cornell, J.A., 2002. Experiments with Mixtures. Wiley, New York.

Heredia-Langner, A., Ortiz, F., Pignatiello, J.J., Simpson, J.R., 2004. A genetic algorithm approach to multiple-response optimization. J. Quality Technol. 36, 432–450.

Khuri, A.I., Kim, H.J., Um, Y., 1996. Quantile plots of the prediction variance for response surface designs. Comput. Statist. Data Anal. 395–407.

Myers, R.H., Montgomery, D.C., 2002. Response Surface Methodolgy: Process and Product Optimization using Designed Experiments. Wiley, New York.

Nguyen, N.-K., Borkowski, J.J., 2008. New 3-level response surface designs constructed from incomplete block designs. J. Statist. Plann. Inference 138, 294–305.

Piepel, G.F., Anderson, C.M., 1993. Variance dispersion graphs for designs on polyhedral regions. In: 1992 Proceedings of the Section of Physical and Engineering Sciences, pp. 111–117.

Smith, W.S., 2005. Experimental Design for Formulation. ASA-SIAM Series on Statistics and Applied Probability. Philadelphia.

Vining, G.G., Cornell, J.A., Myers, R.H., 1993. A Graphical Approach for Evaluating Mixture Designs. Appl. Stat. 42, 127–138.

Journal of Statistical Planning and Inference 139 (2009) 653 – 656

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison" by C.M. Anderson-Cook, C.M. Borror, and D.C. Montgomery

Greg F. Piepel

Statistical Sciences, Pacific Northwest National Laboratory Richland, WA, USA

1. Introduction

Anderson-Cook, Borror, and Montgomery (Anderson-Cook et al., 2008) (ABM henceforth) provide a very nice overview of methods for evaluating and comparing response surface designs using *variance dispersion graphs* (*VDGs*) and *fraction of design space plots* (*FDSPs*). Their overviews of VDGs and FDSPs for special situations involving model robustness, robust parameter design, mixture and mixture-process variable experiments, split-plot experiments, and designs for generalized linear models should be especially useful to practitioners faced with such problems.

I focus on three areas in my discussion: (1) the use of scaled prediction variance versus (unscaled) prediction variance on the y-axis of VDGs and FDSPs, (2) the need to consider bias as well as variance properties when evaluating and comparing designs, and (3) the need to use approaches in constructing designs that account for bias as well as variance properties of the designs. I also make a comment about terminology and finish with a brief summary.

2. Scaled prediction variance versus prediction variance

ABM, and indeed most of the articles in the literature for evaluating and comparing experimental designs, display the *scaled prediction variance* (SPV) on the y-axis of VDGs and FDSPs. The SPV is given by

$$\text{SPV} = \frac{N\,\text{Var}[\hat{y}(\boldsymbol{x})]}{\sigma^2} = N\boldsymbol{x}_m'(\boldsymbol{X}_m'\boldsymbol{X}_m)^{-1}\boldsymbol{x}_m \tag{1}$$

where N is the number of design points, $\text{Var}[\hat{y}(\boldsymbol{x})]$ is the variance of the predicted response at the vector of predictor variables $\boldsymbol{x}$, σ^2 is the unknown experimental error variance, $\boldsymbol{x}_m$ is the vector of predictor variables expanded in the form of the model, and $\boldsymbol{X}_m$ is the design matrix expanded in the form of the model. The *prediction variance* (PV) is given by

$$\text{PV} = \frac{\text{Var}[\hat{y}(\boldsymbol{x})]}{\sigma^2} = \boldsymbol{x}_m'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_m \tag{2}$$

with the difference being that the SPV penalizes the PV via multiplying by the number of design points (i.e., SPV $= N \times$ PV). The rationale given by proponents for incorporating this penalty in the SPV is that the "cost" of designs (represented by the number of design points, N) should be accounted for in evaluating and comparing designs containing different numbers of points. Proponents also justify the SPV as a way to assess whether the benefit of additional design points is worth the additional "cost" to perform them. However, I believe that PV rather than SPV should be the first choice to display in VDGs and FDSPs, for the following reasons.

First, it is always a good idea for a practitioner designing an experiment to consider designs of different sizes (i.e., containing different numbers of design points). That way, the practitioner can consider the trade-offs in PV properties of designs containing more points versus the cost of collecting the additional data. Assessing such trade-offs is difficult if the size of the design has been confounded with the PV properties of the design by using SPV on the y-axis of VDGs and FDSPs. Viewed another way, the use of SPV "pre-decides" for the practitioner how changes in PV should be judged as N changes. I argue that VDGs and FDSPs should be prepared with PV on the y-axis, and values of N included in the legend or labels for the graphs. The practitioner can then decide for a given problem whether the decrease in PV seen in VDGs and/or FDSPs (with PV on the y-axis) for larger designs is worth the cost of additional experimental runs necessary to obtain it.

E-mail address: greg.piepel@pnl.gov.

0378-3758/$ - see front matter Published by Elsevier B.V
doi:10.1016/j.jspi.2008.04.008

Second, a practitioner may be willing a select a design with more points (N_2) over one with fewer points (N_1) so that the maximum or average PV (over all or most of the design region) is below a desired value. The practitioner may be willing to select such a design even if $SPV_1 < SPV_2$ over much or all of the experimental region. This situation can occur when the "point of diminishing returns" has been reached, beyond which increasing N will increase the SPV but still decrease the PV. The practitioner may be willing to pay the price of going beyond the point of diminishing returns to achieve maximum or average PV below a desired value. For example, a practitioner may want a design that has $PV \leqslant 1$ ($PV\sigma^2 \leqslant \sigma^2$) over the whole experimental region, which corresponds to being able to predict the response variable with as good or better precision as it can be determined by performing an experimental run and measuring the response variable. Or, an upper limit on PV smaller than 1 may be necessary so that model predictions have uncertainties that meet regulatory, safety, or client requirements. Using PV on the y-axis of VDGs and FDSPs allows a practitioner to easily assess and compare designs relative to the desired PV upper limit of 1 (or other specified value).

Hence, it is my strong opinion that VDGs and FDSPs are more understandable and useful to practitioners with PV rather than SPV on the y-axis. The trade-offs between PV and N (or some other measure of experimental cost) are best evaluated by the practitioner constructing VDGs and FDSPs with PV (rather than SPV) on the y-axis and judging whether the decrease in PV for larger designs is worth the larger number of design points.

3. Need to consider prediction bias properties in evaluating and comparing designs

ABM focus on graphical displays of prediction variance (SPV) to evaluate and compare experimental designs. Considering prediction variance properties in evaluating and comparing designs is important, but bias properties of designs should also be considered. Selecting a design that minimizes prediction variance (i.e., imprecision of predictions) may be of less importance if a fitted model yields substantially biased predictions in one or more subregions of the experimental region. Assessing the mean squared error (MSE = the sum of prediction variance and squared bias) can also be used as a way to consider together the prediction variance and bias properties of a design.

None of the ten characteristics of a good design enumerated by ABM (adapted from Myers and Montgomery, 2002) directly addresses the bias or MSE properties of a design. Two of the characteristics, that a design

1. Result in a good fit of the model to the data.
2. Provide sufficient information to allow a test for lack of fit.

involve goodness/lack of fit, but these are only indirectly related to the bias aspects of a design. Another of the ten characteristics, that a design

10. Provide a good distribution of the variance of the predicted response throughout the design region

directly addresses the need to consider prediction variance properties of a design. I propose that this tenth criterion be revised to state that a design

10. Provide good distributions of the variance and squared bias properties (or MSE = variance + bias^2) of the predicted response throughout the design region.

This revision of the tenth characteristic explicitly addresses the need to assess both variance and bias properties when evaluating and comparing designs.

In Section 4.1, ABM briefly discuss the bias–variance trade-off and review some literature that has addressed using bias or MSE criteria to evaluate or compare designs. One of the references ABM refer to (Anderson-Cook et al., 2007) contains a more extensive review of past literature that has considered bias and MSE in constructing, evaluating, and comparing designs. However, much of this work was published many years ago, and so it seems that deserved attention may be starting to return to considering bias as well as variance properties in evaluating and comparing experimental designs. Hopefully this will translate to capabilities in several mainstream experimental design software packages so that older and newer methods accounting for prediction bias as well as prediction variance can be more easily utilized by practitioners.

4. Other criteria and approaches for constructing designs

Although the focus of the ABM article is on evaluating and comparing experimental designs, I also comment on criteria and approaches for constructing experimental designs. The D-, A-, G-, and I-optimality criteria (discussed by ABM in Section 2.1) are widely used for constructing experimental designs (e.g., in cases where the experimental region is irregular or the number of design points must be different than in standard response surface designs). These criteria are all *variance-optimal criteria* because they involve minimizing variances of model parameter estimates or model predictions. I believe that there ought to be more focus on approaches and optimality criteria for constructing experimental designs with improved bias (or MSE) properties.

One way to do this is to use bias- or MSE-based optimality criteria for constructing experimental designs. Box and Draper (1959, 1963) and Draper and Lawrence (1965) used an average (integrated) MSE criterion to construct designs for one model

(linear or quadratic) when the true model may be larger (quadratic or cubic). Welch (1983) used an MSE-type criterion (denoted by J**) that relied on assuming a maximum bias over the experimental region. DuMouchel and Jones (1994) proposed a Bayesian D-optimal criterion that allows minimal modification of D-optimal design algorithms to develop designs with protection against potential model terms. Allen and Yu (2002), Allen et al. (2003), and Huang and Allen (2005) proposed variations of an expected integrated mean squared error (EIMSE) criterion for designing experiments. The methods discussed by these authors are not widely used or implemented in current, mainstream software for generating experimental designs. However, with modern computing capabilities, the time is right to more widely implement and extend these methods to the full range of response surface design problems discussed by ABM.

Another way to construct designs with improved bias properties is to adopt approaches for constructing experimental designs that provide better bias properties. Ideally, this would be done without giving up too much on variance properties, but this may not always be possible if squared bias may be large relative to variance. Variance-based optimal design criteria tend to "push" design points toward the boundaries of experimental regions, because this tends to reduce the variance of predictions and parameter estimates. Approaches for constructing designs that provide a better distribution of points over the experimental region will tend to have better bias properties and will provide better protection in situations where the form of the model is not known with certainty.

Space-filling designs (sometimes called *uniform designs*) are one way to better spread points over the experimental region. Space-filling and uniform designs have been discussed by too many authors to list them all here, but some key references are Kennard and Stone (1969), Johnson et al. (1990), Wang and Fang (1996), Fang et al. (2000), Santner et al. (2003), and Fang et al. (2006). Space-filling designs tend to have worse variance properties than standard response surface designs or variance-optimal designs near the boundary of the experimental region, but tend to have better variance and bias properties on the interior of the experimental region. Space-filling designs also generally provide a better basis for fitting semi-parametric or non-parametric model forms in cases where polynomial model forms (e.g., quadratic or cubic) are inadequate.

Other design approaches to better spread points over the experimental region were discussed by Piepel et al. (1993). They proposed *central composite analogue designs* (CCADs) and *layered designs* (LDs) for mixture and other experiments on irregularly shaped experimental regions. CCADs and LDs place points on different layers of the experimental region (e.g., an outer layer, an inner layer, and one or more center points). The LD concept is not limited to irregular regions (for example, standard central composite designs are LDs). LDs are especially applicable when there is a smaller experimental region of primary interest, and a larger experimental region of some interest. Piepel et al. (1993) used VDGs and bias dispersion graphs (BDGs) to show for three constrained mixture examples the trade-offs in variance and bias properties for variance-optimal designs (based on the D-, G-, and I-optimality criteria), CCADs, LDs, Welch's MSE-based J** designs, and space-filling designs.

5. Terminology

ABM use variations of the terms *prediction performance*, *prediction quality*, and *predictive capabilities* only in the context of prediction variance. As noted previously, the prediction performance (or quality or capabilities) of a model involves its accuracy (bias) as well as its precision (prediction variance). Hence, future articles that discuss the predictive performance of models (in the context of evaluating and comparing designs, or others) should take care to use terms that clarify whether the variance or bias aspect of prediction performance is being addressed.

6. Summary

The article by ABM does an excellent job summarizing and illustrating the state-of-the art in graphical methods (e.g., VDGs and FDSPs) for evaluating and comparing response surface designs using variance properties of the designs. The special situations ABM discuss include model robustness, robust parameter design, mixture and mixture-process variable experiments, split-plot experiments, and designs for generalized linear models. The article should be especially useful as a guide to practitioners faced with such problems.

I had three main comments. First, I strongly believe that VDGs and FDSPs should generally be produced using unscaled PV rather than SPV. Using PV makes it easier for practitioners to see how prediction variance is decreased by designs with larger numbers of points, and to assess the "cost" of that decrease in a flexible manner. Using SPV confounds "prediction variance" and "cost" information in a pre-defined way, thus making design choice harder for practitioners.

My second and third comments expressed my belief that bias properties as well as variance properties of designs should be accounted for in constructing, evaluating, and comparing experimental designs. The brief discussion in the ABM article, new work in the Anderson-Cook et al. (2007) paper, and previous personal discussions with the authors suggest that they agree with this comment. The ABM article focused on evaluating and comparing designs using graphical assessments of variance properties presumably because that is what the literature the authors reviewed has addressed in the last several years.

I believe that in the next several years we will see a revival of interest in using past methods and developing new methods for constructing, evaluating, and comparing designs by accounting for bias as well as variance properties of designs. Hopefully that will be accompanied by wide implementation of the methods in mainstream experimental design software, so that practitioners can easily apply the methods.

Acknowledgments

I gratefully acknowledge Alejandro Heredia-Langner (a co-worker at Pacific Northwest National Laboratory), who reviewed and commented on a draft of the discussion.

References

Allen, T.T., Yu, L., 2002. Low cost response surface methods from simulation optimization. Quality Reliab. Eng. Internat. 18, 5–17.
Allen, T.T., Yu, L., Schmitz, J., 2003. An experimental design criterion for minimizing meta-model prediction errors applied to die casting process design. Appl. Statist. 52, 103–117.
Anderson-Cook, C.M., Borror, C.M., Jones, B., 2007. Graphical tools and a metric for comparing designs if the assumed response surface model has been misspecified, Technical report LA-UR 07-3438, Los Alamos National Laboratory, Los Alamos, NM.
Anderson-Cook, C.M., Borror, C.M., Montgomery, D.C., 2008. Response surface design evaluation and comparison. J. Statist. Plann. Inference, in press, doi:10.1016/j.jspi.2008.04.004
Box, G.E.P., Draper, N., 1959. A basis for the selection of a response surface design. J. Amer. Statist. Assoc. 54, 622–654.
Box, G.E.P., Draper, N., 1963. The choice of a second-order rotatable design. Biometrika 50, 335–352.
Draper, N., Lawrence, W., 1965. Designs which minimize model inadequacies: cuboidal regions of interest. Biometrika 52, 111–118.
Fang, K.T., Lin, D.K.J., Winker, P., Zhang, Y., 2000. Uniform design: theory and application. Technometrics 42, 237–248.
Fang, K.-T., Li, R., Sudjianto, A., 2006. Design and Modeling for Computer Experiments. Chapman and Hall/CRC, Boca Raton, FL.
Huang, D., Allen, T.T., 2005. Design and analysis of variable fidelity experimentation applied to engine valve heat treatment process design. Appl. Statist. 54, 443–463.
Johnson, M.E., Moore, L.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Statist. Plann. Inference 26, 131–148.
Kennard, R.W., Stone, L.A., 1969. Computer aided design of experiments. Technometrics 11, 137–148.
Piepel, G.F., Anderson, C.M., Redgate, P.E., 1993. Response surface designs for irregularly-shaped regions (Parts 1, 2, and 3). 1993 Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association, Alexandria, Virginia, pp. 205–227.
Myers, R.H., Montgomery, D.C., 2002. Response Surface Methodology: Process and Product Optimization using Designed Experiments. second ed. Wiley, New York.
Santner, T.J., Williams, B.J., Notz, W.I., 2003. Space-filling designs for computer experiments. The Design and Analysis of Computer Experiments. Springer, New York. (Chapter 5).
Wang, Y., Fang, K.T., 1996. Uniform design of experiments with mixtures. Sci. China Series A 39, 264–275.
Welch, W.J., 1983. A mean squared error criterion for the design of experiments. Biometrika 70, 205–213.

Journal of Statistical Planning and Inference 139 (2009) 657 – 659

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison"

Peter Goos*

Universiteit Antwerpen, Belgium

Selecting the best possible experimental design for a given situation is not a simple matter. This is because there are a lot of criteria that ought to be taken into account when choosing one out of many alternative design options. In their article, the authors focus on the use of graphical methods for comparing experimental designs. In particular, the article reviews most of the literature on variance dispersion graphs and fraction of design space plots. It is explained that sophisticated variance dispersion graphs have been proposed in the literature for assessing model robustness, for evaluating split-plot designs, for evaluating the impact of measurement error on mixture designs, and for mixture experiments involving process variables.

There is no doubt that variance dispersion graphs and fraction of design space plots are very useful tools for comparing alternative design options. Oftentimes, experimental design options are selected using some design optimality criterion, such as the estimation-based $\mathscr{D}$-optimality criterion and the prediction-based $\mathscr{G}$- and $\mathscr{V}$-optimality criteria (see, e.g., Atkinson and Donev, 1992; Myers and Montgomery, 2002). Rightly so, such an approach is often criticized because the design selection is then based on one-number summaries of the properties of the design options. The $\mathscr{G}$-optimality criterion, for example, favours experimental designs that have the smallest maximum prediction variance over the region of interest without considering the distribution of the magnitude of the prediction variance throughout that region. This shortcoming of the $\mathscr{G}$-optimality criterion is overcome by variance dispersion plots and fraction of design space plots, which provide a detailed picture of the predictive quality of experimental designs throughout the entire region of interest. That software packages such as Design Expert and JMP have implemented similar graphical methods for evaluating the predictive performance of experimental designs should therefore be lauded.

One regrettable aspect of the vast literature on the construction of variance dispersion graphs and fraction of design space plots is the emphasis on the use of scaled prediction variances. For completely randomized designs, the scaled prediction variance *SPV* at a point with model expansion $\mathbf{x}$ is defined as

$$\mathrm{SPV} = N\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x},$$

where N is the number of runs of the experiment and $\mathbf{X}$ is the model matrix, also known as the extended design matrix. I strongly believe that the emphasis should always be on unscaled prediction variances, which are defined as

$$\mathrm{PV} = \frac{\mathrm{var}\{\hat{Y}(x)\}}{\sigma^2} = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}$$

for completely randomized experiments.

I also regret that, in the experimental design literature, many optimal design criteria are defined based on a per observation measure of information. For instance, $\mathscr{D}$-optimal designs are often defined as designs that maximize the determinant of the per observation information matrix, $(\mathbf{X}'\mathbf{X})/N$. In this contribution, I will not elaborate on this version of the $\mathscr{D}$-optimality criterion but focus on the use of scaled prediction variances instead. It should be clear, however, that many of the criticisms on the use of scaled prediction variances below apply to any per observation measure of information or precision.

In the literature, one theoretical and one more practical cost-based justification are given for the use of scaled prediction variances. I will discuss each of these first, and conclude the discussion by drawing the reader's attention to two interesting articles on variance dispersion graphs that were overlooked in the article discussed here.

* Tel.: +32 3 220 40 59; fax: +32 3 220 48 17.
E-mail address: peter.goos@ua.ac.be.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.012

1. Discussion of the theoretical justification for the use of scaled prediction variances

The theoretical justification for the use of scaled prediction variances is derived from a famous theorem in optimal experimental design theory, the general equivalence theorem from Kiefer and Wolfowitz (1960), that establishes a relationship between $\mathscr{D}$-optimal continuous designs and $\mathscr{G}$-optimal continuous designs. The theorem says that the maximum scaled prediction variance equals the number of model parameters, p, for continuous $\mathscr{G}$-optimal designs as well as for continuous $\mathscr{D}$-optimal designs. A continuous design is a theoretical concept that can be seen as a design with an infinitely large number of observations. Although continuous designs are not of direct practical use, they do provide information regarding the geometry required for good experimental designs with finite numbers of runs. The extent to which a practical experimental design (with a finite number of observations) possesses the same geometrical structure as the continuous optimal design can be measured by the $\mathscr{G}$-efficiency, which is defined as the ratio of the number of model parameters, p, to the maximum scaled prediction variance. For such practical experimental designs, the maximum scaled prediction variance usually exceeds p and the $\mathscr{G}$-efficiency is a number between 0 and 1.

This theoretical justification is only valid for the use of scaled prediction variances for the graphical assessment of completely randomized designs. This is because the general equivalence theorem is only valid for completely randomized experiments. The general equivalence theorem therefore does not provide support for the use of scaled prediction variances for evaluating experimental designs with correlated observations, as, for example, split-plot designs, where scaled prediction variances have been used too.

Focusing on the $\mathscr{G}$-efficiency or on the maximum scaled prediction variance leads to counterintuitive recommendations. A simple example can clarify this. The maximum scaled prediction variance when using a completely randomized 2^2 factorial design for estimating a linear main-effects model in two variables equals 3, which is the number of parameters in the model. As a result, the $\mathscr{G}$-efficiency of the 2^2 factorial design is one. A 2^2 factorial design with three of its points duplicated and the other point used only once has a maximum scaled prediction variance of 21/5 and a $\mathscr{G}$-efficiency of 5/7 only. According to the maximum scaled prediction variance or the $\mathscr{G}$-efficiency, the four-point design should be preferred, even though the unscaled prediction variances, the ones that matter for practitioners, are substantially smaller in the entire cuboidal design region for the larger design. This shows that neither scaled prediction variances nor $\mathscr{G}$-efficiencies are practical measures for ranking different design options.

2. Discussion of the practical justification for the use of scaled prediction variances

Recent publications on the comparison of experimental designs suggest using scaled prediction variances because they penalize large experimental designs for being more costly than small designs. Zahran et al. (2003, p. 377), for example, write that the use of unscaled prediction variances is recommended if the cost of experimentation is not a primary consideration and direct comparisons between the expected variances of estimation are desired.

In my view, the use of unscaled prediction variances should be recommended in all practical situations. This is because it is not scaled prediction variances that are of interest to a practitioner, but rather the precision with which an experimental design will allow him or her to make predictions. This precision is directly related to the size of the experiment: larger experiments often lead to smaller prediction variances and thus to a better predictive precision. By looking at unscaled prediction variances, the researcher can evaluate the increase in precision obtained from using a larger experiment. Thus, unscaled prediction variances provide an experimenter with much more useful information than scaled ones.

A problem with using design criteria that penalize large designs in some way or another is that they often suggest selecting experimental designs that are small and/or inexpensive. Such designs often possess a small power for identifying active factors, lead to wide confidence intervals and to imprecise predictions. This has a negative impact on the quality of the decisions made based on the experimental results. The result is that some cost-savings are realized in the short run when an inexpensive experiment is conducted, but also that opportunities for realizing large cost-savings in the long run might be missed. Statistical consultants should therefore stimulate the use of large enough designs. The problem with scaling prediction variances is that this camouflages the poor predictive performances of small designs compared with the larger ones.

It is not difficult to formulate other criticisms to the cost-based justification of the use of scaled prediction variances. A major assumption when using scaled prediction variances is that the size of the experimental design, N, is a reasonable measure of its cost. There are many practical situations where the size of the design is not at all a good proxy for the cost of an experiment. One simple example is a split-plot design involving hard-to-change and easy-to-change factors, where the cost of the experiment is mainly determined by the number of times the hard-to-change factors are reset. Another situation where the design size is not a good indicator of its cost is when the preparation of the experiment is time-consuming. For example, controlling ambient temperature and humidity during the experiment may be required and this often entails a substantial set-up cost, which is independent of the number of runs to be performed. Also, experiments sometimes require a time-consuming preparation of batches of experimental material to be used for all the runs. In such experiments, there may be enough material left for an additional couple of runs at virtually no extra cost. Criticisms such as these inspired Liang et al. (2006) to utilize the cost C of a design as a scaling factor rather than the design size, N.

I do find, however, that scaling by the cost C when evaluating the prediction variances of one or more experimental designs is not a much better thing to do than scaling by the number of experimental runs, N. One argument is that cost-scaled prediction

variances also lead to the selection of relative inexpensive designs. In the case of split-plot designs, it favours designs with few whole plots that allow no or almost no inference about the hard-to-change variables. Also, researchers often face (tight) budget or time constraints, so that the design selection problem is reduced to finding the best possible design for a given size or cost.

Another, more fundamental, reason for not scaling is that the use of scaled prediction variances basically means that a single-number summary of two totally different features of a design, that is the cost and the precision of prediction, is used for evaluating it. As avoiding the reliance on single-number summaries for the selection of experimental designs is exactly one of the reasons for promoting variance dispersion graphs and fraction of design space plots, this is odd.

Instead of combining cost and precision of prediction into a single criterion, I prefer treating these two objectives separately during the process of design selection. This is because the trade-off between cost and precision of statistical inference is case-dependent. That trade-off cannot be made by a statistician alone, which is exactly what the statistician is doing when suggesting the use of scaled prediction variances as a vehicle for making a trade-off. Input from engineers concerning the importance and the tangible benefits (such as cost-savings, increased yield, etc.) of having smaller prediction variances and more precise inferences is required. These benefits, especially when they can be expressed in monetary units, can be compared directly to the cost of running more expensive experimental designs.

3. Some more interesting references

I would like to conclude this discussion by drawing the reader's attention to two articles co-authored by Trinca and Gilmour. In Trinca and Gilmour (1998), they show how to use variance dispersion graphs for evaluating various experimental designs the runs of which are arranged in blocks. In Trinca and Gilmour (1999), they point out that, in many practical situations, the interest is not in predictions at a particular setting of the experimental variables but rather in the predicted differences of responses between different settings of the experimental variables. They show how to construct variance dispersion graphs that are in line with this research objective.

References

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Designs. Clarendon Press, Oxford UK.
Kiefer, J., Wolfowitz, J., 1960. The equivalence of two extremum problems. Canad. J. Math. 12, 363–366.
Liang, L., Anderson-Cook, C.M., Robinson, T.J., 2006. Fraction of design space plots for split-plot designs. Quality Reliability Eng. Internat. 22, 275–289.
Myers, R.H., Montgomery, D.C., 2002. Response Surface Methodology. second ed. Wiley, New York, NY.
Trinca, L.A., Gilmour, S.G., 1998. Variance dispersion graphs for comparing blocked response surface designs. J. Quality Tech. 30, 349–364.
Trinca, L.A., Gilmour, S.G., 1999. Difference variance dispersion graphs for comparing response surface designs with applications in food industry. Appl. Statist. 48, 441–455.
Zahran, A.R., Anderson-Cook, C.M., Myers, R.H., 2003. Fraction of design space to assess prediction capability of response surface designs. J. Quality Tech. 35, 377–386.

Journal of Statistical Planning and Inference 139 (2009) 660 – 661

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Discussion of "Response surface design evaluation and comparison" by Christine Anderson-Cook, Connie Borror and Douglas Montgomery

James M. Lucas

J.M. Lucas and Associates, 5120 New Kent Road, Wilmington, DE 19808, USA

I enjoy the chance to take an overview of response surface methodology (RSM) because it is a major consulting and research area. RSM has been an important part of my life. Reviewing this paper is my second recent chance to take an overview because I just reviewed Box and Draper's (2007) new edition (Lucas, 2007). My perspective is that of an industrial consultant, first as an internal consultant at DuPont and later as an independent consultant and researcher. The authors' statement that the "response surface framework has become the standard approach for much of the experimentation carried out in industrial research, development, manufacturing, and technology commercialization" captures my experience. Almost all of my experiments can be placed in an RSM framework.

The authors and I agree that "the choice of a design for fitting a first-order model is a relatively clear-cut decision." I would have explicitly added Plackett–Burman (screening) designs to the first-order designs they mentioned because I have successfully used them over my entire career. I feel that the choice of a second-order design has also been known for many years. DuPont has an excellent RSM short course that was also sold outside DuPont in the 1970s and 1980s. (This course, while profitable, was not central to the company's mission; so outside offerings were discontinued.) For fitting second-order models composite and Box–Behnken designs were used. "Which Response Surface Design is Best" (Lucas, 1976) and my letter to the editor (Lucas, 2007) concluded with "the designs that have been used by practicing statisticians have very high efficiencies. While more efficient designs are possible, either they remain to be discovered or they require significantly more experimental points. "Classical" designs will still be used in most applications." In addition to composite designs and Box–Behnken designs I also noted that Hoke (1974) designs performed well but could not comment on the Roquemore (1976) designs as they were published in the same issue as my 1976 paper. I do note that there is little indication that Hoke and Roquemore designs have been widely used in spite of their good statistical properties. I agree that a variance dispersion graph (VDG) provides more information than a single-number criterion such as the *G*-efficiency, my favorite single-number criterion; however, I also note that my (Lucas, 1978) observation about *G*-efficiency still holds and that "no example has been given of a design having good *G*-efficiency yet poor performance by other criteria." The work using VDGs has increased our understanding of design performance but has caused little change in the designs that are used because practitioners have been using excellent quadratic RS designs. Because VDGs do give more information than a single-number criterion they could be useful in examining designs in nonstandard situations such as RS designs with constraints or with model terms of higher than quadratic order. Can they be made easy enough to use for this situation? Can they be easily implemented using standard statistical software such as JMP and Minitab? Ridge Analysis (Hoerl, 1959; Box and Draper, 2007) is useful for examining the response contour for models of quadratic order or less. RS models with higher-order terms are sometimes used; a specific example is the use of a special cubic for mixture models. The mathematics behind VDG could be used to provide a ridge analysis for this situation. This would be a useful development.

A deeper discussion of the example in Fig. 1, where differences in the VDG were shown while the *G*-efficiency is nearly the same, can be enlightening. The example was for two factors but the following discussion holds for any number of factors. The *G*-optimal design, for a quadratic model with p terms in a spherical region, places a fraction $1/p$ points in the center and a fraction $(p-1)/p$ of the points on the surface so the design is rotatable. For a 2-factor design a pentagon with a center-point is a *G*-optimal design. Designs that underweight the center of the region as do the hexagon and the CCD with one center-point will have a higher variance of prediction at the center and over much of the middle of the region. It is poor statistical practice to underweight the center of a spherical region; the weight there should always be at least $1/p$. *G*-efficiency is much better at showing this poor performance than is *D*-efficiency as a low center-point weight can make the D/G efficiency ratio arbitrarily large (Lucas, 1977). For

E-mail address: JamesM.Lucas@world.att.net.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.013

a CCD the highest *G*-efficiencies are usually achieved with 2-center-points though a few more center-points can be recommended to get a flat variance of prediction near the center of the region, to achieve the *V*-optimal design, or for estimating experimental error. When the experimental region is a hypercube, there is much less reason for additional center-points. A CCD with star-point distance $(\alpha) = 1$ (a face centered cube) is often the design of choice. Because the star-points are so close to the center of the region the highest *G*-efficiencies are achieved with zero center-points. Experimental error can be estimated by replication of any experimental points such as a (proper or improper) fraction of the factorial points.

Experiments with hard-and easy-to-change factors have been a major part of my consulting and research during my career as an independent consultant. Split-Plot designs are the designs of choice for this situation. Here I feel that the major emphasis should be on design development rather than on design evaluation. For example, the development of quadratic RS designs when there is one or more hard-to-change (HTC) factors is still an open question. Parker et al. (2007) and the references cited in the paper have provided a worthwhile baseline for this problem. For main effect or main effect plus two-factor interaction models my former students and I have candidate designs in preparation or in the publishing queue (Anbari and Lucas, 2008a, b; Webb and Lucas, 2008). For all numbers of factors we have found "supper efficient" blocking structures that dominate a CRD on a cost and on a *G*-efficiency basis. These designs have a *G*-efficiency > 100% relative to a CRD. The cost models use the excellent cost model described by the authors that we first used in Anbari and Lucas (1994). A very succinct fractional factorial example with one or two HTC factors is the 2^{6-1} for estimating a 22 term main effect plus two-factor interaction model. The fraction uses resolution *V* (intentionally not *VI* so the design is not maximum resolution or minimum aberration) with $I = ABCDE$. For one HTC factor (*A*) use four blocks and the block confounding relationship $I = A = BCF = ABCF = BCDE = ADEF = DEF$. For a second HTC factor (*B*) nest factor *B* in each *A* block. This design is super-efficient; it can be shown to have:

- *G*-efficiency $= 22(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)/(22\sigma_0^2 + 16\sigma_1^2 + 20\sigma_2^2) > 1.0$

where σ_0^2 is the residual variance, σ_1^2 is the variance associated with changing factor *A* and σ_2^2 variance associated with changing factor *B*.

References not listed in Anderson-Cook et al.:

References

Anbari, F.T., Lucas, J.M., 1994. Super-efficient designs: how to run our experiment for higher efficiency and lower cost. In: Proceedings of the ASQ 48th Annual Quality Congress.

Anbari, F.T., Lucas, J.M., 2008a. Designing and running super-efficient experiments: optimum blocking with one hard-to-change factor. J. Quality Tech. 40, 31–45.

Anbari, F.T., Lucas, J.M., 2008b. Fractional factorial split-plots: minimum aberration or optimum blocking, in preparation.

Box, G.E.P., Draper, N.R., 2007. Response Surfaces, Mixtures, and Ridge Analysis. Wiley, New York.

Hoerl, A.E., 1959. Optimum solution of many variables equations. Chem. Eng. Progr. 55 (11), 69–78.

Hoke, A.T., 1974. Economical second-order designs based on irregular fractions if the 3^n factorial. Technometrics 17, 373–384.

Lucas, J.M., 1976. Which response surface designs are best. Technometrics 18, 411–417.

Lucas, J.M., 1977. Design efficiencies for varying numbers of center-points. Biometrika 64, 145–147.

Lucas, J.M., 1978. Discussion of *D*-optimal fractions of three-level designs. Technometrics 20, 381–382.

Lucas, J.M., 2007. Review of "Response Surfaces, Mixtures, and ridge Analysis" by G.E.P. Box and N.R. Draper. J. Quality Tech. 39 (4), 392–394.

Parker, P.A., Kowalski, S.M., Vining, G.G., 2007. Construction of balanced equivalent estimation second-order split-plot designs. Technometrics 49, 56–65.

Webb, D.F., Lucas, J.M., 2008. Designing and running super-efficient experiments: Optimum blocking with multiple hard-to-change factors, in preparation.

Journal of Statistical Planning and Inference 139 (2009) 662 – 668

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Optimum and other response surface designs. Comments on "Response Surface Design Evaluation and Comparison" by Anderson-Cook, Borror and Montgomery

Anthony C. Atkinson

Department of Statistics, London School of Economics, London WC2A 2AE, UK

1. Introduction

It is a pleasure to comment on this paper that contains a wealth of recent references on the assessment and comparison of response surface designs. Despite the length of my contribution, space precludes me from commenting on many interesting aspects.

One purpose of these comments is to provide references to related work, particularly on optimum designs. I establish the customary notation for optimum designs in Section 2. Section 3 contains comments on a number of points raised by the authors. In Section 4, I give a few small response surface designs with high symmetry, some of which are new, that experimenters may find useful. Maybe the authors would like to evaluate these designs.

2. Optimum experimental design

2.1. Exact and continuous designs

The main tools that the authors use are based on the properties of the standardized prediction variance, defined by the authors in their Section 2.1. The importance of this quantity comes, as the authors say, from the theory of optimum design, with the target maximum value determined by the equivalence between continuous *D*- and *G*-optimum designs. To relate their work more closely to parallel work on optimum design, it is helpful to establish the standard notation of that literature.

Consider an experiment in which there are m factors and write the linear model as

$$\mathrm{E}(y) = F\beta.$$

Here y is the $N \times 1$ vector of responses, β is a vector of p unknown parameters to be estimated by least squares and F is the $N \times p$ *extended* design matrix. The ith row of F is $f^{\mathrm{T}}(x_i)$, a known function of the m explanatory variables that can take values in the design region $\mathscr{X}$. For the second-order model with $m = 2$

$$\mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2, \tag{1}$$

and

$$f^{\mathrm{T}}(x_i) = (1 \quad x_{1i} \quad x_{2i} \quad x_{1i}^2 \quad x_{2i}^2 \quad x_{1i}x_{2i}).$$

This notation can be extended to compound design criteria, Section 3.5, in which the jth model has extended design matrix F_j.

The design is determined by the experimental values of x and by the number of replications n_i or, equivalently, the weights $w_i = n_i/N$ at each x_i. The design can be written as a measure ξ that puts weight w_i at x_i, that is

$$\xi = \begin{Bmatrix} x_1 & x_2 \ldots x_n \\ w_1 & w_2 \ldots w_n \end{Bmatrix},$$

E-mail address: a.c.atkinson@lse.ac.uk.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.014

with ξ_N the exact design in which $w_i = n_i/N$. The theory was developed by Kiefer whose work on optimum design is collected in Brown et al. (1985). Use of the continuous design, to which the equivalence theory applies, removes dependence of the design on N. Any practical design for N trials will be exact with integer replication n_i at each design point. Exact and continuous D-optimum designs may be identical, but they usually are not in response surface work.

The least squares estimator of the parameters is

$$\hat{\beta} = (F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}y,$$

where the $p \times p$ matrix $F^{\mathrm{T}}F$ is the information matrix for β. For the design measure ξ we consider instead the standardised information matrix

$$M(\xi) = F^{\mathrm{T}}WF, \tag{2}$$

where W is a diagonal matrix with elements w_i.

The variance of the predicted response $\hat{y}(x)$ for an exact design is conveniently written as

$$d(x, \xi_N) = \frac{N\mathrm{var}\{\hat{y}(x)\}}{\sigma^2} = f^{\mathrm{T}}(x)M^{-1}(\xi_N)f(x), \tag{3}$$

the quantity plotted for several designs in the authors' Fig. 1. For the continuous design (3) becomes

$$d(x, \xi) = f^{\mathrm{T}}(x)M^{-1}(\xi)f(x).$$

2.2. The general equivalence theorem and design efficiency

Exact D-optimum designs maximize $|M(\xi_N)|$ and exact G-optimum designs minimize the maximum over $\mathcal{X}$ of $d(x, \xi_N)$ given by (3). We write this maximum for some ξ as

$$\bar{d}(\xi) = \max_{x \in \mathcal{X}} d(x, \xi).$$

The equivalence theorem for D- and G-optimality holds for continuous designs. If ξ^* is a continuous D-optimum design, then $\bar{d}(\xi^*) = p$, the number of parameters of the linear model. Exact D-optimum designs may have values of $\bar{d}(\xi_N^*)$ slightly larger than p.

The D-efficiency of a design ξ is

$$D_{\mathrm{eff}} = \left\{ \frac{|M(\xi)|}{|M(\xi^*)|} \right\}^{1/p}. \tag{4}$$

The comparison of information matrices for designs that are measures removes the effect of the number of observations, while taking the pth root of the ratio of the determinants in (4) results in an efficiency measure which has the dimensions of a variance, irrespective of the dimension of the model.

The G-efficiency of the design ξ is

$$G_{\mathrm{eff}} = \bar{d}(\xi^*)/\bar{d}(\xi) = p/\bar{d}(\xi).$$

A recent book-length treatment of optimum experimental design, including both theory and SAS code for the construction of designs, is Atkinson et al. (2007) which also provides numerous references to the literature on optimum design.

3. Detailed comments

3.1. Graphical displays

The distinction between exact and continuous optimum designs is nicely made by two-factor designs for a circular region as illustrated in the variance-dispersion graph of the authors' Fig. 1.

D-optimum response surface designs are described by Farrell et al. (1968) for three types of design region. For spherical regions the D-optimum design puts a weight $2/\{(m+1)(m+2)\}$ at the centre point, with the rest of the weight distributed uniformly around the circumference of the design region. For $m = 2$ the region is a circle and the weight at the centre is $\frac{1}{6}$. Thus one continuous and exact D-optimum design consists of points in a regular pentagon with one at the centre. In Fig. 1 the authors have plotted the properties of a hexagonal design. This has weight $\frac{1}{7}$ at the centre and so is not quite the continuous D-optimum design. This is shown by the value of $d(x, \xi)$ being slightly more than 6 at the centre of the region and slightly lower than 6 at the edge.

Table 1
D- and *G*-efficiencies of some small response surface designs, $m = 4$, cubical region

Design	*N*	*D*-efficiency	*G*-efficiency
hoke4a1	15	0.8024	0.3913
hoke4a2	15	0.8024	0.3913
D-opt	15	0.8747	0.4522
bd4	16	0.3114	0.0011
Roq416b	16	0.2033	0.0606
D-opt	16	0.8939	0.4167
hoke4a5	19	0.8696	0.3145
D-opt	19	0.9449	0.6900
CCD	25	0.9113	0.7798
D-opt	25	0.9773	0.6891

As the authors comment, all three designs in Fig. 1 are rotatable. If they are not, it is informative to consider the minimum and maximum variance at each radius, as in the original plots of Box and Behnken (1960), and perhaps the average variance as well. It is also necessary to decide how to treat designs over a cubical region. For example, should the calculations for a radius greater than one include only those parts of $\mathscr{X}$ for that radius or all points on the circle or sphere?

The fraction of design space plot overcomes these problems. However, there are a few outstanding questions. One, mentioned by the authors at the end of Section 3, is that the plots are for measures, so information on the number of trials is lost when designs for different *N* are being compared. I discuss design optimality and costs in Section 3.3.

A final point about these useful graphical tools is that, on both Figs. 1 and 2 it has been felt necessary to include lines of efficiency. An alternative would be to plot efficiency rather than variance.

3.2. The comparison of small designs

Towards the end of Section 3 the authors commend the designs of Hoke (1974) and Roquemore (1976) when *N* is hardly larger than *p*. The numerical comparisons of Atkinson and Tobias (2008) indicate that great care is needed in the choice of these small response surfaces designs. Some that have been suggested in the literature can be surprisingly inefficient, both for cubical and spherical regions.

Table 1 lists the *D*- and *G*-efficiency, for a cubical region, of the best of some small designs for $m = 4$ taken from the catalogue of Borkowski (`http://www.math.montana.edu/~jobo/cr/designs.txt`). See also Borkowski and Valeroso (2001). In addition the table includes *D*-optimum designs found by searching over the 81 points of the 3^4 factorial.

For $N = 15$ the two Hoke designs have identical properties, but have relatively 10% lower *D*- and *G*-efficiency than the exact *D*-optimum design. These three designs share with the others in the table the feature that the *D*-efficiencies are much higher than the *G*-efficiencies. For $N = 16$, on the other hand, the *D*-optimum design far outperforms the Roquemore design and the Box-Draper design that has a *G*-efficiency of 0.01%, caused by many of the design points being shrunk away from the edges of the design region. The Hoke design for $N = 19$ has a *D*-efficiency that is again relatively 10% lower than that of the exact *D*-optimum design, while the ratio of *G*-efficiencies is much lower at less than one half. Only for $N = 25$ is there a design that challenges the *D*-optimum design: the CCD with one centre point has, of course, a lower *D*-efficiency than the *D*-optimum design, but it does have a higher *G*-efficiency. This seems to be the only design in these comparisons where the authors' graphical methods would be needed in helping to decide about a design.

Designs are available for some other values of *N*, for example those of Hoke, which have uniformly poor *G*-efficiencies. In general, except for a few values of *N*, there seems no sensible alternative to numerically constructed *D*- (or *G*- or I-) optimum designs.

3.3. Costs

Towards the end of Section 3 the authors discuss the inclusion of costs in experimental design. Fedorov and Leonov (2005) give examples of optimum design incorporating costs and discuss some theory. If the cost of experimenting at x_i is $c(x_i)$ and the total cost of the experiment is constrained to be not greater than *C*, continuous optimum designs are found as in Section 2.1 but now with weights $w_i = n_i c(x_i)/C$. See also Atkinson et al. (2007, p. 149). This alternative normalization of designs, originally suggested by Elfving (1952), provides an equivalence theorem and so should allow modification of the authors' plots to include costs.

3.4. Split-plot designs

After many years of comparative neglect, there is now a rapidly expanding literature on the design of split-plot experiments. A book-length treatment of blocked and split-plot designs is Goos (2002). The optimum designs of Goos and Vandebroek (2001) and Goos and Vandebroek (2004) allow for the different costs of changing the levels of the two kinds of factor present in these experiments.

3.5. Compound designs

In Section 4.1 the authors consider problems of model robustness that arise because the model is uncertain. For the jth of h models let the D-efficiency be

$$D_j^{\text{eff}} = \left\{ \frac{|M_j(\xi)|}{|M_j(\xi_j^*)|} \right\}^{1/p_j}. \tag{5}$$

A compound D-optimum design reflecting the weights κ_j of interest in each model then maximizes

$$\Phi(\xi) = \prod_{j=1}^{h} \{D_j^{\text{eff}}\}^{\kappa_j} = \left\{ \frac{|M_j(\xi)|}{|M_j(\xi_j^*)|} \right\}^{\kappa_j/p_j}.$$

On taking logs we see that this is the same as maximizing

$$\sum_{j=1}^{h} \frac{\kappa_j}{p_j} \log |M_j(\xi)| \tag{6}$$

and that

$$d(x, \xi) = \sum_{j=1}^{h} \frac{\kappa_j}{p_j} d_j(x, \xi). \tag{7}$$

An equivalence theorem then applies to this compound design criterion. If ξ^* is the design maximizing (6)

$$\bar{d}(\xi^*) = \sum_{j=1}^{h} \kappa_j.$$

The standard algorithms for the construction of optimum designs can then be used, which may be faster than the genetic algorithm of Heredia-Langner et al. (2004). The efficiency of the exact compound designs can be calculated, as can the individual efficiencies (5) and (7) used to provide the authors' plots. Plots for each component $d_j(x, \xi)$ are also a possibility, which would be an extension of the authors' Fig. 3. An advantage of this form of compound design over the additive form of Biedermann et al. (2005) is that calculation of ξ^* does not require knowledge of the individual ξ_j^*. However these designs are, of course, required for the calculation of D-, but not G-, efficiencies.

In practice designs may need to be calculated for several κ_j to find satisfactory component efficiencies. For small h plots of the component efficiencies as a function of the κ_j provide a convenient method of choosing a design. Examples are in Chapter 21 of Atkinson et al. (2007) together with extensions to other forms of compound design, such as DT-optimality, in which interest is in both model discrimination and parameter estimation.

3.6. Model checking

A disadvantage to the use of compound D-optimum designs is that the design has to have sufficient points of support to fit all models, even the largest. In Section 4.1 the authors mention potential third-order models. If this were a faint possibility, many design points would have small weights w_i and any exact design would require an excessive value of N. An alternative is the Bayesian model checking design introduced by DuMouchel and Jones (1994) in which prior information about the terms needed for model checking reduces the number of points of support. In the case of checking a second-order model for third-order terms, as this prior information decreases the design would move smoothly from that for the second-order model with one or a few extra design points, to the D-optimum design for the full third-order model. Examples and an equivalence theorem are in Chapter 20 of Atkinson et al. (2007).

3.7. Generalized linear models

In the generalized linear model let $\mathrm{E}(y_i) = \mu_i$, with link function $g(\mu_i) = \eta_i$ and linear predictor $\eta_i = \beta^{\mathrm{T}} f(x_i)$. The error distribution is a member of the one-parameter exponential family with $\mathrm{var}(y_i) = \phi V(\mu_i)$. Maximum likelihood estimation of β reduces to iteratively re-weighted least squares with, in the notation of the authors' Section 4.4,

$$v_i = V^{-1}(\mu_i) \left(\frac{d\mu_i}{d\eta_i} \right)^2.$$

The information matrix for a design ξ, analogous to (2) is thus

$$M(\xi) = F^{\mathrm{T}} WVF, \tag{8}$$

where W and V are both diagonal matrices, the latter with elements v_i. An equivalence theorem then applies to the variance

$$d(x, \xi) = vf^{\mathrm{T}}(x) M^{-1}(\xi) f(x),$$

which extends the authors' definition of the SPV.

The structure of optimum response surface designs is not as clear as that for linear models, since the designs are only locally optimum, depending on the values of the parameters β. However, if the effects β are sufficiently small, the means μ_i will not vary greatly and so the weights v_i will be sensibly constant. Then the information matrix (8) reduces to that for the linear model. Designs for the linear model will therefore be efficient for zero or small effects.

This argument is due to Cox (1988). It seems that the results of Zahran et al. (2007) mentioned in Section 4.4 quantify Cox's result for small effects in first-order models. For second-order models, Fig. 22.6 of Atkinson et al. (2007) shows how the D-optimum regression design on the points of the 3^2 factorial is gradually distorted as the parameters of a binomial model move away from zero. Their Fig. 22.9 shows that for, a gamma model with a Box-Cox link with parameter $\lambda = 0.5$, the distortion of the 3^2 factorial is much less as the parameters change than it is for the binomial model. For the log link the regression design is optimum whatever the parameters of the linear model.

4. Some symmetrical optimum response surface designs

An advantage of the methods associated with optimum designs is that they lead to algorithms for the construction of new designs. This section presents a few interesting D-optimum exact designs tabulated by Atkinson and Tobias (2008) that also have good G-efficiency. It might be informative to use fraction of design space plots to compare these designs with others, such as the corresponding G-optimum designs.

The results of Farrell et al. (1968) for cubical regions show that the continuous D-optimum designs are supported on those points of the 3^m factorial with 0, $m-1$ and m non-zero co-ordinates. For the response surface model (1) in two factors these are the points of the 3^2 factorial and $p = 6$. When the design region is continuous and bounded by the square with vertices ± 1, Fig. 12.1 of Atkinson et al. (2007) shows that D-optimum designs for $N = 6, 7$ and 8 contain trials not at the points of the 3^2 factorial. However, for $N = 9, 13$ and 14, Atkinson and Tobias (2008) find that the exact D-optimum designs are supported at the factorial points. Fig. 1 shows the designs and Table 2 their D- and G-efficiencies.

All three designs are symmetrical in x_1 and x_2. For $N = 9$ we have the 3^2 factorial, for $N = 13$ the design consists of the 3^2 factorial with replicated corner points, that may also be thought of as the combination of a 2^2 and a 3^2 factorial; the design for $N = 14$ adds a second centre point to the design for $N = 13$. The D-optimum designs for neighbouring values of N lack such symmetry. An advantage of the designs for $N = 13$ and 14 is that the four or five replicated points provide model-free estimates of error variance on 4 or 5 degrees of freedom.

For $m = 3$, there are now $p = 10$ parameters. Atkinson and Tobias (2008) give the properties of D-optimum exact designs for N from 10 to 30. For all $N \neq 10$ the optimum designs when searching over the 5^3 grid are supported at the points of the 3^3 factorial.

The designs for $N = 14, 21$ and 30 are displayed in Fig. 2. The designs are shown as 3^2 factorials at each of the levels of a factor arbitrarily labelled x_1. What is immediately noticeable in these designs is the strong symmetric structure. In all three cases the design for x_2 and x_3 is repeated at the high and low levels of x_1. Further, at all levels of x_1 the designs in x_2 and x_3 have a highly

1	1	1	2	1	2	2	1	2
1	1	1	1	1	1	1	2	1
1	1	1	2	1	2	2	1	2

Fig. 1. Square region, $m = 2$. Designs with good properties for $N = 9, 13$ and 14: $N = 9$, the 3^2 factorial; $N = 13$, a combination of the 3^2 and 2^2 factorials and, $N = 14$, the design for $N = 13$ with an extra centre point.

Table 2
D- and G-efficiencies of D-optimum exact designs of Fig. 1 for the second-order model in two factors

N	D-efficiency	G-efficiency
9	0.9740	0.8276
13	0.9977	0.8718
14	0.9944	0.9606

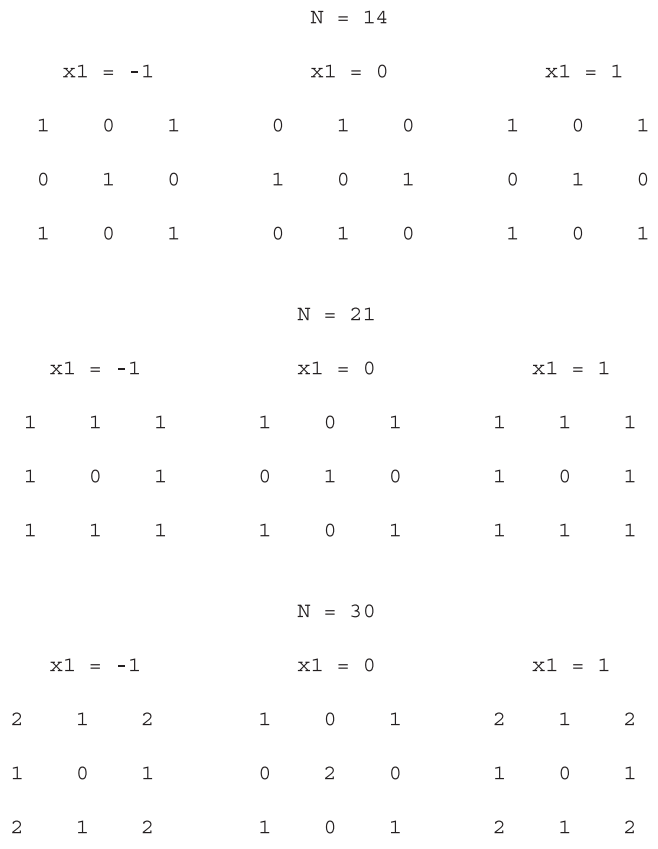

Fig. 2. Square region, $m = 3$. Designs with good properties for $N = 13$, 21 and 30 supported on the points of the 3^3 factorial. For each N the figure provides the number of points for each 3^2 factorial in x_2 and x_3 at the three levels of x_1. The labelling of the factors is arbitrary.

Table 3
D- and *G*-efficiencies of *D*-optimum exact designs of Fig. 2 for the second-order model in three factors

N	*D*-efficiency	*G*-efficiency
14	0.9759	0.8929
21	0.9833	0.8571
30	0.9995	0.9697

symmetric structure, for example a 2^2 factorial with a centre point for $N = 14$. Apart from their good properties an advantage of such designs is that they are easy to specify and so to have performed correctly by unskilled personnel.

Table 3 gives the *D*- and *G*-efficiencies of the designs of Fig. 2. The *D*-efficiencies increase steadily from 0.9759 to 0.9995. However, the *G*-efficiencies of the designs are much lower than these values; as low as 0.8571 when $N = 21$.

Atkinson and Tobias (2008) also give the properties of *D*-optimum designs for $m = 4$ for N up to 40. Because of the sparseness of this number of points on the 81 points of the 3^4 factorial, none of the designs has an obvious structure. A challenge is to try to find designs that have good *D*-efficiency whilst also having good properties as analysed by the authors' plots.

A final comment is that in Sections 2.1 and 3 the authors mention *D*- *G*- and I- (or V)-optimum designs and their construction. Are there cases in which the authors' graphical procedures lead to a different choice of design from that resulting from consideration of these three efficiencies of a design?

References

Atkinson, A.C., Tobias, R.D., 2008. Optimal experimental design in chromatography.
Atkinson, A.C., Donev, A.N., Tobias, R.D., 2007. Optimum Experimental Designs, with SAS. Oxford University Press, Oxford.
Biedermann, S., Dette, H., Zhu, W., 2005. Compound optimal designs for percentile estimation in dose-response models with restricted design intervals. In: Ermakov, S., Melas, V., Pepelyshev, A. (Eds.), Proceedings of the 5th St Petersburg Workshop on Simulation. NII Chemistry University Publishers, St Petersburg, pp. 143–148.

Borkowski, J.J., Valeroso, E.S., 2001. Comparison of design optimality criteria of reduced models for response surface designs in the hypercube. Technometrics 43, 468–477.
Box, G.E.P., Behnken, D.W., 1960. Some new 3 level designs for the study of quantitative variables. Technometrics 2, 455–475.
Brown, L.D., Olkin, I., Sacks, J., Wynn, H.P. (Eds.) 1985. Jack Carl Kiefer Collected Papers III. Wiley, New York.
Cox, D.R., 1988. A note on design when response has an exponential family distribution. Biometrika 75, 161–164.
DuMouchel, W., Jones, B., 1994. A simple Bayesian modification of *D*-optimal designs to reduce dependence on an assumed model. Technometrics 36, 37–47.
Elfving, G., 1952. Optimum allocation in linear regression theory. Ann. Math. Statist. 23, 255–262.
Farrell, R.H., Kiefer, J., Walbran, A., 1968. Optimum multivariate designs. In: Proceedings of 5th Berkeley Symposium, vol. 1. University of California Press, Berkeley, CA, pp. 113–138.
Fedorov, V.V., Leonov, S.L., 2005. Response-driven designs in drug development. In: Berger, M., Wong, W.-K. (Eds.), Applied Optimal Designs. Wiley, New York, pp. 103–136 (Chapter 5).
Goos, P., 2002. The Optimal Design of Blocked and Split-plot Experiments. Springer, New York.
Goos, P., Vandebroek, M., 2001. Optimal split-plot designs. J. Quality Technol. 33, 436–450.
Goos, P., Vandebroek, M., 2004. Outperforming completely randomized designs. J. Quality Technol. 36, 12–26.

Journal of Statistical Planning and Inference 139 (2009) 669 – 670

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A discussion of "Response surface design evaluation and comparison"

Timothy J. Robinson*

Department of Statistics, University of Wyoming, Laramie, WY 82071-3332, USA

1. Choosing an experimental design is an art

The authors have presented a timely article on response surface design evaluation and comparison. During the past five years, software companies such as Design-Expert, MINITAB, and SAS JMP have incorporated user-friendly modules for the selection of optimal designs. These algorithms generally require the user to specify a model, select the number of experimental runs, indicate which factors, if any, have levels that are hard to change, and then specify an alphabetic optimality criterion. Given this information, the software provides an *optimal* design complete with a variance dispersion plot and fraction of design space plot to the user. Unlike any other time in history, the practitioner has the capability of generating optimal designs and evaluating competing designs in the blink of an eye. Instead of forcing a given problem into a catalogued design, the practitioner can customize a design for the specific situation at hand.

Although the luxury of design construction that is afforded to us by software companies is handy, one must resist the temptation of thinking that the experimental design stage of problem solving is a stage that only needs casual thought. Indeed the authors are correct to remind us of the significance of Box and Draper's (1975) emphasis that a good response surface design must possess a suite of desirable properties (in this manuscript the authors note 10 properties). Although important in many ways, using alphabetic optimality criteria along with variance dispersion plots to come up with an optimal design only addresses properties 1, 4, 7, and 10 in the authors list. The design that is optimal for one criterion is often not even close to the design that is optimal for other criteria. The question of what constitutes a *good* response surface design depends upon the circumstances of the research question being posed and the experimental apparatus. Regardless of the situation, trade-offs exist when considering desirable properties of a response surface design. For instance, for a design to have good detectability of lack-of-fit (property 2) and provide for an estimate of *pure* error (property 4), the design must involve a sufficient number of experimental points. However, if many active design factors exist, in order to run a small experiment (in the spirit of property 7), the experiment may become saturated, thus prohibiting the capability of detecting lack-of-fit or providing an estimate of pure error. As Box and Draper point out, "the art of the good practicing statistician is his ability to assess the needs of a given situation and then to choose the design which comes as close as possible to meeting them". Anderson-Cook, Borror and Montgomery are to be commended for emphasizing the fact that designing an experiment requires critical thinking and not simply the casual clicks of a mouse required by the software.

Perhaps no topic in experimental design has received more attention in the last decade than that of industrial split-plot designs. Many important tools have been developed to aid the practitioner in selecting an optimal split-plot design. It is important to note, however, that much of the literature thus far on optimal split-plot designs has focused upon one and perhaps as many as two of Box and Draper's criteria when proposing an optimal design algorithm. The split-plot model and analysis is somewhat more complicated than it is for the completely randomized design. As such, Box and Draper's list of desirable attributes needs to be expanded (see Parker et al., 2008 for a case study). Future research in industrial split-plots may involve design selection that incorporates numerous criteria simultaneously. It should also be noted that precious little in the split-plot literature addresses design selection that is robust to the presence of outliers, designs which are robust to model misspecification, designs that are robust to errors in control of design levels, designs that are robust to missing data, and diagnostics of model fit.

2. Optimal designs for robust parameter design settings

Since the popularization of robust design by Taguchi (1987) and Taguchi and Wu (1985), much has been written concerning the use of the dual response in the analysis of the combined array. The combined array, put forth by Welch et al. (1990) and Shoemaker et al. (1991), was introduced as an alternative to the product or crossed array design concept suggested by Taguchi

* Tel.: +1 307 766 5108; fax: +1 307 766 3927.
E-mail address: tjrobin@uwyo.edu.

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.010

because the latter often requires a large number of trials. Borkowski and Lucas (1997) proposed the use of mixed resolution designs for the robust parameter design setting. Although mixed resolution designs have desirable characteristics for the robust design setting, there are many situations in which a user cannot afford to run a design that is of the size required by the mixed resolution design. Anderson-Cook, Borror and Montgomery correctly point out that in these settings, the evaluation and comparison of combined array designs have generally involved the use of optimality criteria such as *D*-, *G*-, and *I*-optimality. While these optimality criteria certainly reflect a given design's capability to precisely estimate the parameters in the response model as well as the entire response as a whole, they may not be appropriate criteria for certain interests in robust design.

Anderson-Cook, Borror and Montgomery discuss the need for a design to precisely estimate the vector of slopes, $\mathbf{l}(\mathbf{x})$. When a response model is composed primarily of terms influencing the process mean, the traditional *D*-, *G*-, and *I*-optimality criteria will naturally focus more on mean estimation than on estimating the process variance. Consequently, the very entity that is the focus of robust parameter design, the process variance, may not be estimated well. Borror et al. (2002) explore the use of variance dispersion plots for comparing competing designs in terms of their ability to estimate the variance of the estimated slope. Myers et al. (2007) propose new optimality criteria which focus upon a design's ability to adequately estimate the process variance as well as weighted combinations of the process mean and variance. They also develop and illustrate graphical techniques for comparing competing designs when these new optimality criteria are utilized.

3. Extending response surface methods to complex systems

Complex systems are those systems which involve collections of multiple processes, components, and nested sub-systems where the overall system is difficult to understand (for a more complete definition, see Science, 1999). These systems permeate the very environment and infrastructures that we all live in. Examples include munitions stockpiles, nuclear power facilities, ecological systems, electric power grids, cyber-security systems. Predicting the likelihood of proper functionality of these systems (full system reliability) is a critical component of system understanding. While response surface methods have led to a great deal of success in many types of systems in terms of quality improvement, lowering overall operational costs, and increasing process reliability, these methods have not been utilized within the realm of complex systems. Response surface techniques (design and analysis) tend to focus on a single source of data (data from the statistical design) and how best to collect new combinations of inputs to improve estimation and prediction. Clearly many sources of data can be collected within the context of a complex system ranging from full system tests, sub-system tests, component tests, maintenance or repair data, expert knowledge, computer simulation, accelerated tests, leveraged data from related systems, etc. Extending response surface principles (particularly design principles) to account for the multiple sources of data in the meta-analyses used with complex systems would be an important advance in the field of statistics.

In complex systems there are invariably different costs associated with measurements taken at the full system, sub-system, and component levels. In the design of a new system, or when determining allocation strategies for future resources, it will be important to account for such things as the relative costs of different types of data, the amount of existing data, the amount of uncertainty in existing data, the reliability of the component or sub-system, and structure of the complex system as a function of the components. Methods for optimizing information gain per fixed unit of cost will allow for ideal use of new resources to improve understanding of the system's reliability. Invariably, the quality of estimation and prediction of system models will need to be balanced with the cost of different data collection strategies.

Much of the work in response surface methods assumes that the functional relationship between the dependent variable of interest and the independent factors is well approximated by a lower order polynomial model. Complex systems often involve highly nonlinear relationships between inputs and outputs and often, models are hierarchical in nature due to the nesting of sub-systems within the full system. As a result, the mathematics involved with data from complex systems is fundamentally different than that used with conventional applications of response surface methods.

As it becomes increasingly more important to consider the sustainability of complex systems, it is perhaps an opportune time to advance response surface methods to these problems, particularly when it comes to the collection of data for both the monitoring and future design of these systems. The reliability of any system is directly related to an understanding of system capacity. By developing design strategies to precisely estimate system capacity, information can be obtained as to eventual system improvement and increased system efficiency.

References

Borkowski, J.J., Lucas, J.M., 1997. Designs of mixed resolution for process robustness studies. Technometrics 37, 63–70.
Borror, C.M., Montgomery, D.C., Myers, R.H., 2002. Evaluation of statistical designs for experiments involving noise variables. J. Quality Tech. 34, 54–70.
Box, G.E.P., Draper, N.R., 1975. Robust designs. Biometrika 62, 347–352.
Myers, W.R., Myers, W.H., Robinson, T.J., 2007. A structural approach to design optimality in robust parameter design. Technical Report, Virginia Tech Department of Statistics.
Parker, P.A., Anderson-Cook, C.M., Robinson, T.J., Liang, L., 2008. Robust split-plot designs. Quality Reliability Eng. Internat. 24, 107–121.
Science, 1999. Complex Systems (special issue), Vol. 284, pp. 1–212.
Shoemaker, A.C., Tsui, K.L., Wu, C.F.J., 1991. Economical experimentation methods for robust design. Technometrics 33, 415–427.
Taguchi, G., 1987. System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost. Quality Resources, White Plains, NJ.
Taguchi, G., Wu, Y., 1985. Introduction to off-line quality control. Central Japan Quality Control Association (available from American Supplier Institute, Dearborn, MI).
Welch, W.J., Yu, T.K., Kang, S.M., Sacks, J., 1990. Computer experiments for quality control by parameter design. J. Quality Tech. 22, 15–22.

Journal of Statistical Planning and Inference 139 (2009) 671 – 674

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Rejoinder for "Response surface design evaluation and comparison"

Christine M. Anderson-Cook[a,*], Connie M. Borror[b], Douglas C. Montgomery[c]

[a] *Statistical Sciences, Los Alamos National Laboratory, USA*
[b] *Mathematical Sciences and Applied Computing Department, Arizona State University West, USA*
[c] *Department of Industrial Engineering, Arizona State University, USA*

First of all, we would like to thank all of the discussants for the care and thoughtfulness that they have taken in preparing their comments. The topic of design evaluation and comparison is one that has room for many diverse opinions and emphases, and hearing the thoughts of so many of the leading researchers in this area has been a treat for us as authors. In retrospect, our original paper may have been a bit too ambitious in the topics that we endeavored to present, and the discussants helped fill in some of our missing details to help round out a broader perspective of design evaluation.

1. The bigger picture of design evaluation

Jones highlights that the scope of what we have discussed under the heading of "design evaluation" only makes sense once the key science aspects of the problem have been appropriately considered and decided. How many of us know of experiments that answered the wrong question? How efficiently the design did this clearly does not matter! We jokingly said that design evaluation and comparison is all about fine-tuning after all of the important stuff has been settled.

Parker gave some valuable insights into the process of eliciting experiment objectives and how analysis results will be used. He describes how these graphical tools can help facilitate better discussion about the goals of the experiment. This will help statisticians and practitioners develop common language and understanding about what is important for particular studies, while quantifying trade-offs between different competing criteria.

2. Additional graphical tools

In the collection of tools for looking at the ranges and locations of prediction variances, we were remiss in pointing out several alternative graphical approaches which should be considered.

The prediction variance profiler shown in Jones' discussion is an important tool. The graphical tools we have presented all seek to reduce the often high dimensional design space into something more manageable which can be summarized on a 2- or 3-dimensional plot. One major advantage of the profiler is the ability to use it interactively to see the behavior at *any* location in the design space—in this way there is no reduction of dimensionality of the design space, just the flexibility to investigate it any way that is desired. This clearly is another powerful tool that should be included in the repertoire.

The quantal plots, quantile dispersion graphs (QDG) and volatility plots discussed by Khuri and Borkowski are other useful alternatives, which give the user more flexibility and other ways of summarizing the characteristics of prediction variance, by compressing the high dimensional space into a manageable presentation. In particular, the QDGs for situations where model parameters are unknown represent an efficient way of gaining understanding about the ranges of values that might be observed across the combined design space and parameter ranges.

Khuri's idea of considering how these tools could be adapted to sequential experimentation is one that deserves additional thought. Currently our thinking has been focused on a single experiment and selecting a good design at that stage. In the context of a series of experiments building on knowledge acquired in previous stages, tools that would allow us to look ahead and

* Corresponding author. Tel.: +1 505 606 0347.
E-mail address: c-and-cook@lanl.gov (C.M. Anderson-Cook).

0378-3758/$ - see front matter
doi:10.1016/j.jspi.2008.04.009

consider sequences of designs which are desirable for the whole progression would be a welcomed addition over stage-wise greedy approaches.

We wholeheartedly agree with Borkowski's comments that continued software development for many types of design regions and models would be highly desirable to encourage widespread use of these tools. Although not the primary focus of the paper, design creation based on various criteria has many interesting aspects, not the least of which is how to incorporate alternative search algorithms (such as genetic algorithms) to obtain candidate set free solutions.

3. To scale or not to scale

The issue of whether scaled ($SPV = N\mathbf{x}'^{(m)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^m$) or unscaled ($UPV = \mathbf{x}'^{(m)}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}^m$) prediction variance should be used was the most frequently and passionately commented upon topic from the discussants. Jones added a third alternative, the square root of the UPV, which has the advantage of being in the original units of the response. While we chose to illustrate most of the plots by using the SPV, we feel that there is good reason to consider each metric for quantifying the goodness of a design. First, it is important to note that if we are comparing designs of the same size, then the relative ranking and performance of the designs will not be affected by our choice of scaled or unscaled. In this case, as Piepel and Goos noted, the practitioner will undoubtedly prefer the UPV as this summary gives a direct summary proportional to what can be expected after data are collected. While closer to what the practitioner is actually interested in, there is still a level of abstraction between the UPV and what he/she is really interested in, namely what the width of a confidence interval will be for the response at a given location. This results from the scaling of all of these metrics to remove the unknown σ^2 from $\mathrm{Var}(\hat{y}(x_m))$. Alternately, by considering SPV we can gain some theoretical assessment of the design through G-optimality which Lucas emphasizes is a highly desirable metric summarizing the goodness of the design. While there may be some debate over whether G- or I-optimality is a better single number summary of the distribution of the prediction variance (since we as statisticians more traditionally focus on central tendencies rather than extremes), protecting against the worst case prediction in the design space is one that will continue to be popular.

In cases where comparisons between designs of different sizes are being made, the choice of SPV versus UPV is very important. The UPV will allow for the most direct assessment of the how much improvement is gained with the use of additional resources. The SPV does assume a particular relationship between the cost of experimental units and how they should be traded-off against improvement in prediction variance, and while this relationship will not universally be true, it is a common choice for how to quantify the incremental cost of expanding an experiment. Whether we want to construct a single measure of prediction performance, in the spirit of Derringer and Suich (1980), or do the balancing of cost and prediction variance separately would appear to be a subjective choice. We feel that there are good reasons for both measures to continue to be used in different situations. It was interesting to note how the SPV was the natural metric for Atkinson, while UPV was more natural for Jones, Piepel, and Goos.

To further illustrate the differences, Figs. 1 and 2 show FDS plots for UPV, SPV and $\sqrt{\mathrm{UPV}}$ for the new 2- and 3-factor designs proposed in Atkinson's discussion. Both of the suggested smallest designs (9 observations for 2 factors, and 14 observations for 3 factors) are closely related to the face center cube (central composite designs with a single or no center run, respectively). From the UPV plots, we can see the incremental reduction in UPV as we increase the sample size. For the 2-factor designs with smaller designs nested in the larger ones, we can see the advantage of the additional points being added. The SPV plots allow us to extract the G-efficiency listed in Atkinson's Tables 2 and 3, and also to see that with the smallest design we are able to predict well across most of the region, and have diminishing returns for the added observations, except at the maximum of the region where the additional points continue to yield a marked improvement. As expected the rankings using the UPV and $\sqrt{\mathrm{UPV}}$ are identical, but there is a small amount of flattening of the curves with the square root version of the plots because of the change of scale.

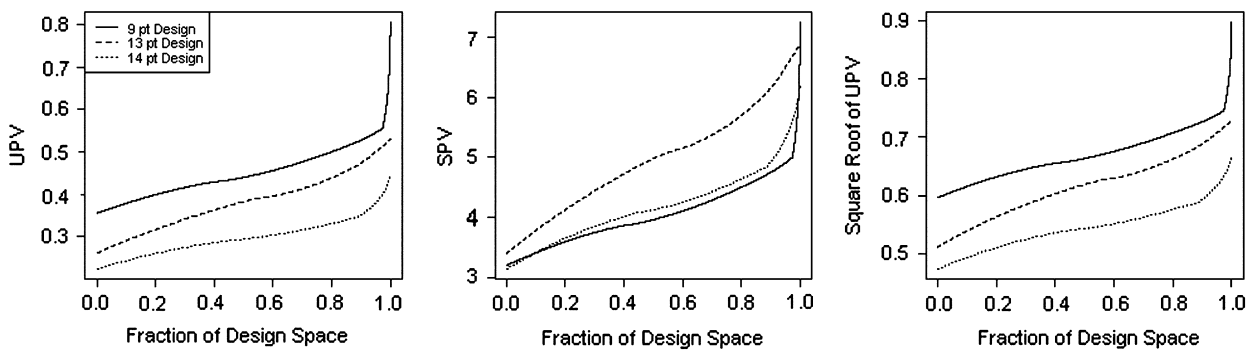

Fig. 1. Comparison of Atkinson's 2-factor designs of size 9, 13, and 14 for a square region.

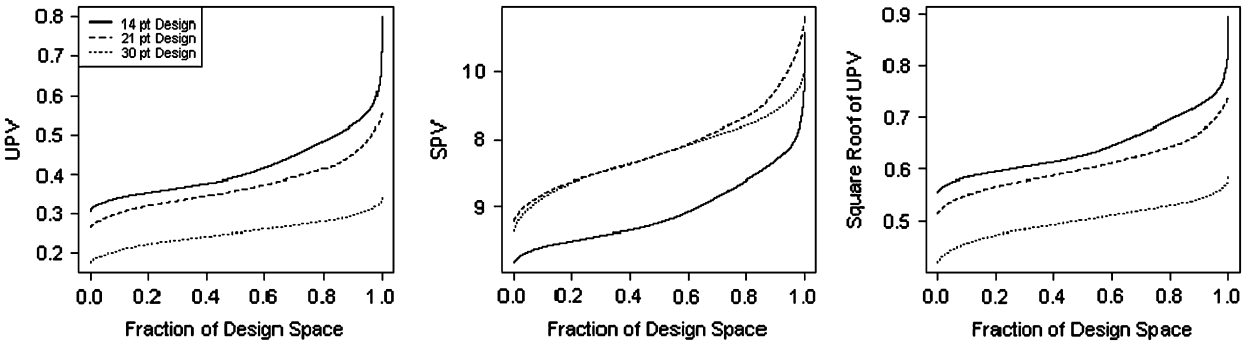

Fig. 2. Comparison of Atkinson's 3-factor designs of size 14, 21, and 30 for a cuboidal region.

4. Mean squared error (variance and bias)

Piepel highlights that our choice of terminology of "prediction performance, etc." focusing on just prediction variance inherently assumes that we are indeed using the correct model to characterize the underlying relationship between input factors and the response. We certainly agree that it is a dangerous assumption to not consider model misspecification as an important factor in assessing and comparing designs. An important and substantial body of research has been listed in the discussion by Piepel. We agree that methods for a combined assessment of prediction variance and expected squared bias would be a beneficial addition to the general practice of comparing designs. One of the complications of using bias is that an alternative (more complicated) model and the approximate size of the missing term coefficients must be specified for the assessment. For example, suppose that we assume that a first order model involving three factors is needed to characterize a particular relationship, then the calculation of prediction variance (SPV or UPV) for a particular design can be immediately performed. However, to assess the bias considerations, we have to specify what type of model to protect against: first order with interactions, second order, third order, or just some specially selected individual terms from one or more of those categories. Typically, this will be quite difficult for practitioners to articulate, and the bounding of the potential size of those terms coefficients will be additionally challenging. Having said that, it is an important consideration to make our design assessments and comparisons more realistic, and even more research would be beneficial to further understand the implications of including bias.

5. Miscellaneous comments

Robinson presents some important considerations for the robust parameter design setting. Given the special nature of these designs and the ensuing analysis to identify key interactions and relationships between control and noise factors, various terms in the model have different roles to play. Balancing prediction variance (and bias) throughout the design space with adequate estimation of the vector of slopes associated with the control-by-noise interactions should be emphasized. Recent work has shown that often substantial gains in overall performance can be achieved by jointly considering these objectives, rather than optimizing just prediction variance.

Lucas highlights the many recent developments in the area of split-plot experiments. Awareness of the needs for these types of experiments continues to grow in industry, and the advantages of these approaches are also being better understood. We disagree with Lucas that the major emphasis should be on development of new designs rather than a design evaluation. We feel that since there are many competing criteria which are being proposed and used to create new designs, it will remain important to have good tools available to assess them. The paper by Parker et al. (2008) provides some good discussion of various qualitative and quantitative metrics for comparison of designs.

We agree with Robinson that a new type of design assessment may be needed for examining meta experiments where multiple sources of data and experiments-within-experiments are considered. In these cases, the design space is much more complex as there are typically choices between different types of data which can be collected and then which particular combinations of inputs within a particular type. We see this as an emerging area of research with many potential applications.

Borkowski raises some interesting points about how design assessment may need to be further expanded for generalized linear models and for cases where transformations are considered. He also suggests the need for expansion of the comparison of designs for mixture experiments where multiple models are explored.

Jones presents a context of experimentation that may be increasingly common: modeling very complex relationships between large numbers of input factors and a response. In these situations, a Bayesian approach which leverages prior knowledge and sparsity of effects will allow for supersaturated designs to be considered in place of impractically large designs. Assessment tools will be needed to determine reasonable bounds on how to balance economy of design with adequate estimation and prediction.

Atkinson challenges us with the question of whether using graphical methods has changed the recommendation of what designs to select from what would have been recommended by looking at just D-, G- and I-efficiencies. Certainly, the graphical plots lead to a deeper understanding of the trade-offs between looking at each of the criteria, but in addition, we have encountered several cases where looking at the plots has altered our interpretation of the results and led to a re-ordering of our design rankings. In the second example involving split-plot designs in Liang et al. (2006), and examining minimum resolution designs for 6–10 factors in Li et al. (2007), graphical methods highlighted that for some designs a tiny proportion of the region with large prediction variance values drives disappointing G-optimality, while most of the region is predicted much better than other competing designs.

We are delighted to have been able to spawn such vigorous and diverse discussion about design of experiments. The list of design characteristics to consider when creating and assessing an experiment can be daunting in its complexity and also an opportunity to focus on what is really important for a particular study.

References

Derringer, G., Suich, R., 1980. Simultaneous optimization of several response variables. J. Quality Technol. 12, 214–219.
Li, J., Liang, L., Borror, C.M., Anderson-Cook, C.M., Montgomery, D.C., 2007. Comparing Prediction Variance Performance for Variations of the central Composite Design using Graphical Summaries. Los Alamos National Laboratory Technical Report LA-UR 07-8119.
Liang, L., Anderson-Cook, C.M., Robinson, T.J., 2006. Fraction of design space plots for split-plot designs. Quality Reliab. Eng. Internat. 22, 275–289.
Parker, P.A., Anderson-Cook, C.M., Robinson, T.J., Liang, L., 2008. Robust split-plot designs. Quality Reliability Eng. Internat. 24, 107–121.