Journal of the
Royal Statistical Society

SERIES C
Applied
Statistics

Appl. Statist. (2012)
61, *Part* 3, *pp.* 345–401

Optimum design of experiments for statistical inference

Steven G. Gilmour

University of Southampton, UK

and Luzia A. Trinca

Universidade Estadual Paulista, Botucatu, Brazil

[*Read before* The Royal Statistical Society *on Wednesday, November 16th, 2011, the President*, Professor V. S. Isham, *in the Chair*]

Summary. One attractive feature of optimum design criteria, such as *D*- and *A*-optimality, is that they are directly related to statistically interpretable properties of the designs that are obtained, such as minimizing the volume of a joint confidence region for the parameters. However, the assumed relationships with inferential procedures are valid only if the variance of experimental units is assumed to be known. If the variance is estimated, then the properties of the inferences depend also on the number of degrees of freedom that are available for estimating the error variance. Modified optimality criteria are defined, which correctly reflect the utility of designs with respect to some common types of inference. For fractional factorial and response surface experiments, the designs that are obtained are quite different from those which are optimal under the standard criteria, with many more replicate points required to estimate error. The optimality of these designs assumes that inference is the only purpose of running the experiment, but in practice interpretation of the point estimates of parameters and checking for lack of fit of the treatment model assumed are also usually important. Thus, a compromise between the new criteria and others is likely to be more relevant to many practical situations. Compound criteria are developed, which take account of multiple objectives, and are applied to fractional factorial and response surface experiments. The resulting designs are more similar to standard designs but still have sufficient residual degrees of freedom to allow effective inferences to be carried out. The new procedures developed are applied to three experiments from the food industry to see how the designs used could have been improved and to several illustrative examples. The design optimization is implemented through a simple exchange algorithm.

Keywords: *A*-optimality; Blocking; Compound criterion; *D*-optimality; Exchange algorithm; Factorial design; Lack of fit; Pure error; Response surface

1. Introduction

Experiments with complex treatment structures, such as fractional factorial, response surface or mixtures designs, are very common in industrial research and development, as well as in many laboratory-based sciences. In such experiments, variance-based optimality criteria are increasingly used, owing to their availability in software packages, to choose the treatment design and, if appropriate, to arrange the design in blocks. An advantage of criteria such as *D*-optimality, *A*- (or more generally *L*-)optimality, *G*-optimality, etc., is that they have meaningful interpretations in relation to the statistical analysis to be performed on the data from the experiment. For example, a *D*-optimum design minimizes the volume of a joint confidence region for the

Address for correspondence: Steven G. Gilmour, School of Mathematics, University of Southampton, Highfield, Southampton, SO17 1BJ, UK.
E-mail: s.gilmour@soton.ac.uk

0035–9254/12/61345

parameters, an A-optimum design minimizes the average variance of parameter estimates and minimizes the average squared width of the corresponding confidence intervals and so on. Very importantly, multiple objectives can be met through the use of compound criteria, although this is not yet common in practice. For a comprehensive and accessible description of optimality criteria and their applications, see Atkinson *et al.* (2007).

The work that is presented here was motivated mainly by experimental research carried out for the food industry. Fractional factorial and response surface designs are widely used for such experiments, but the relatively large run-to-run variation that is caused by the use of biological materials means that some of the very small designs that are used in some other industrial applications are ineffective. The analysis of data involves inferences based on confidence intervals or hypothesis tests, to ensure that fitted response surfaces are not overinterpreted as showing effects which could be due to random variation. Despite the supposed relationships between optimality criteria and common methods of analysing the data from these experiments, many experimenters still prefer to use standard designs, such as central composite designs (CCDs). One advantage of classical designs is that, unlike most optimum designs, they include replicate points.

In almost all experiments, the statistical inference procedures that are used rely on an internal estimate of the variance between experimental units, which is often called 'error'. In experiments in which the proposed model for treatment effects includes fewer parameters than there are treatments, there are some differences in practice about which estimate of error is used. The default in many statistical computing packages, especially if a general regression program is used, and the procedure that is described in many regression textbooks (which are often aimed at analysing observational data), is to use the residual mean square from the fitted model. However, most textbooks on the design and analysis of experiments recommend separating lack of fit from 'pure error' and using the pure error mean square for inference.

We strongly recommend the latter method of estimating error. However, once this has been accepted, the usual definitions of optimum design criteria no longer have all of the statistical interpretations claimed for them, since the sizes of confidence intervals and regions, or equivalently the power of hypothesis tests, depend not only on the variance matrix of the parameter estimates, but also on the degrees of freedom for pure error. The objectives of this paper are to show how the criteria should be correctly defined in order to have the properties claimed, to show that they can be applied in some of the same ways as the traditional criteria and to illustrate by examples that quite different designs can be optimal under the new criteria.

In Section 2, we discuss the analysis of data and, in particular, the estimation of error for inferential procedures. In Section 3, adjusted definitions of various optimality criteria which have the desired interpretations and examples of their use to choose designs are given. In Section 4 compound criteria which allow a compromise between different experimental objectives are developed and illustrated. Finally, some overall lessons are drawn in Section 5.

The programs that were used to obtain near optimum designs can be obtained from

`http://www.blackwellpublishing.com/rss`

2. Inference from designed experiments

The analysis of data from factorial-type designs typically includes fitting one or more polynomial models and carrying out hypothesis tests on these models, in particular to compare models of different orders and to test whether the model is better than a null model, but also to test individual high order parameters. Once a reasonable model has been found it is interpreted by estimating the parameters. Except in rare cases in which the variance of experimental units, σ^2, can be assumed to be known, all such tests require an estimate of σ^2. We assume that estimating σ^2 is not of interest in itself but is required only for carrying out inference that is related to the treatment comparisons.

For a completely randomized design, in general, we assume that an experiment has been run with t treatments defined by combinations of the levels of the factors and that the responses can be modelled as

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \qquad i = 1, \ldots, t, \quad j = 1, \ldots, n_i, \tag{1}$$

where Y_{ij} is the response from the jth replicate of treatment i, μ_i is the expected response from treatment i, $E(\varepsilon) = \mathbf{0}$ and $V(\varepsilon) = \sigma^2 \mathbf{I}$. We shall refer to this as the *full treatment model*. Experimenters often try to make interpretation easier and more informative by fitting a submodel with

$$\mu_i = \beta_0 + \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}, \qquad i = 1, \ldots, t, \tag{2}$$

where $\mathbf{x}_i$ represents the levels of the factors $X_1, \ldots, X_q$ in treatment i, $\mathbf{f}$ is a $(p-1)$-dimensional function of these levels and $\boldsymbol{\beta}$ is a $(p-1)$-dimensional vector of parameters. In this paper we shall assume that $\mathbf{f}$ represents a polynomial which respects functional marginality, but the same arguments would apply to other linear models.

The first question that we shall address is whether, when performing inference from model (2), we should use the estimate of σ^2 from fitting model (2), or the pure error estimate obtained from fitting model (1). To illustrate the difference that it might make, consider exercise 11.6 of Box and Draper (2007). They analysed data from a three-factor rotatable CCD with four centre points, with one gross outlier (from a factorial point) removed. The pure error estimate of σ^2 is $s^2 = 10.77$ on 3 degrees of freedom, whereas the estimate that is obtained from the second-order model, i.e. pooling the pure error and second-order model lack-of-fit degrees of freedom, is $s_p^2 = 7.03$ on 7 degrees of freedom. The mean square for second-order effects is 30.90 and, using s_p^2, Box and Draper found that the test of the second-order parameters gives a p-value of 0.037 and went on to do further interpretation of this model. However, using s^2, the test of the second-order parameters gives a test statistic of $F = 2.87$, which on 6 and 3 degrees of freedom gives a p-value of 0.208, which would suggest that there is little justification for further interpreting the second-order model. A more clear-cut recommendation could have been given if the design had allowed more than 3 degrees of freedom for pure error.

This is not an isolated example. A few pages earlier in the same book (Box and Draper (2007), pages 668–669), one example has $s^2 = 5.667$ and $s_p^2 = 19.612$, although lack of fit is not significant, and the next example has $s^2 = 220.5$ and $s_p^2 = 81.7$. Although in these two examples the qualitative conclusions are not changed greatly, standard errors could be quite drastically underestimated or overestimated and confidence intervals can be too narrow or too wide. Clearly, it is an important decision which estimate of σ^2 is used, but there is no unanimity among researchers. In many of the textbooks which users of factorial and response surface designs use, there is no acknowledgement of the complexity of the issue. Khuri and Cornell (1996), Dean and Voss (1999), Box and Draper (2007) and Atkinson *et al.* (2007) recommend using s_p^2. Cochran and Cox (1957) and John (1971) recommend using s^2. Cornell (2002) seems to recommend both in different sections and Hinkelmann and Kempthorne (2005) seem to give a different recommendation from Hinkelmann and Kempthorne (2008).

Although most researchers do not discuss the issue and their practice varies, those who do discuss it generally come down in favour of using pure error, i.e. s^2. Among some of the classic texts, Draper and Smith (1998), page 48, state that s^2

> 'will usually provide an estimate of σ^2 which is much more reliable than we can obtain from any other source',

adding that

> 'For this reason, it is sensible when designing experiments to arrange for (replicates)'.

In the context of factorial experiments, Scheffé (1959), pages 126–127, strongly condemned the practice of pooling sums of squares from non-significant interactions because it leads to biased estimation of σ^2 (since pooling of zero effects will only be done when their estimates turn out to be small) and leads to procedures whose statistical properties are not known. He also recommends

> 'designing the experiment so that there will be a sufficient number of d.f. for (pure) error'.

Cox (1958) agrees that σ^2 'is best estimated' by pure error, but is more forgiving and allows the use of s_p^2 if 'only one observation is made on each treatment'. Davies (1956) agrees and Wu and Hamada (2009) tend towards the same view but make a less clear-cut recommendation.

We accept the view that inferential procedures should be carried out by using the unbiased pure error estimator of σ^2, while acknowledging that statistical inference is often not the most important part of the analysis and interpretation of experimental data. The main problem with using s_p^2 is that the biases induced are unmeasurable and the inferences are therefore difficult to interpret. In any particular experiment, if carrying out inference is regarded as being important, then the design should be chosen to make that inference as informative as possible. In the next section, we show how the standard optimality criteria must be modified to do this. Later, we shall consider compound criteria for situations in which inference represents a part, but not the whole, of the means of analysing the data from the experiment.

3. Design criteria for inferential procedures

3.1. Definitions

If the data from an experiment are to be analysed primarily by using confidence intervals or regions and/or hypothesis tests, then the experiment should be designed to ensure that these procedures will be as informative as possible. For simplicity of presentation and to clarify the relationships with standard criteria, we shall assume that the aim is to obtain unbiased confidence intervals or regions of minimal length or volume. The same criteria will maximize the power of hypothesis tests which can be obtained from these confidence intervals or regions, e.g. t-tests of individual parameters or F-tests of sets of parameters. To carry out these procedures the design must allow a sufficient number of degrees of freedom for estimating error. We can, in fact, specify formally the appropriate criterion.

The usual statistical justification for D-optimality is that it minimizes the volume of the joint confidence region for the parameters—see, for example, Atkinson *et al.* (2007), page 135. This is based on the fact that the volume of the confidence region is proportional to $|\mathbf{X}'\mathbf{X}|^{-1/2}$, where $\mathbf{X}$ is the polynomial model matrix, given the treatment design, with ith row $(1\ \mathbf{f}(\mathbf{x}_i)')$, and so the D-criterion minimizes $1/|\mathbf{X}'\mathbf{X}|$. Although this is correct, as noted by Kiefer (1959), 'with σ^2 known or else (pure error degrees of freedom) the same for all designs', in a general form, the volume of a $100(1-\alpha)\%$ confidence region (see Draper and Smith (1998), page 145) is proportional to

$$(F_{p,d;1-\alpha})^{p/2}|\mathbf{X}'\mathbf{X}|^{-1/2},$$

where p is the number of parameters in the model, d is the number of pure error degrees of freedom and $F_{p,d;1-\alpha}$ is the $(1-\alpha)$-quantile of the F-distribution with p numerator and d

denominator degrees of freedom. Thus the D-criterion should be to minimize

$$(F_{p,d;1-\alpha})^p/|\mathbf{X}'\mathbf{X}|.$$

We shall refer to this as the DP(α) criterion. In this paper, we shall use $\alpha=0.05$ for illustration and refer to the criterion simply as DP, but the required confidence level should be considered carefully for each experiment. Despite the above quotation, Kiefer (1959) did not suggest this additional step, since he did not separate lack of fit from pure error.

Similarly, D_S-optimality is intended to minimize the volume of a joint confidence region for a subset of p_2 of the parameters by minimizing $|(\mathbf{M}^{-1})_{22}|$, where $\mathbf{M}=\mathbf{X}'\mathbf{X}$ and $(\mathbf{M}^{-1})_{22}$ is the portion of its inverse corresponding to the subset of the parameters of interest. To take account of pure error estimation correctly, the (DP)$_S$ criterion is to minimize

$$(F_{p_2,d;1-\alpha})^{p_2}|(\mathbf{M}^{-1})_{22}|.$$

This criterion should be used, for example, when a major objective of the experiment is to compare the first-order model with the second-order model. Then the higher order terms will form the subset and minimizing the volume of a confidence region for them will be equivalent to maximizing the power of a test for their existence. Note that if the parameters of interest are the treatment parameters and the nuisance parameter(s) is or are the intercept or the intercept plus block effects, then standard D_S-optimality reduces to D-optimality. With the new criterion, this is no longer true, owing to the reduction in the numerator degrees of freedom. Throughout this paper, we shall use D_S to refer to the situation where all parameters of the polynomial treatment model (excluding the intercept) contribute to the criterion, unless otherwise stated.

L-optimality is intended to minimize the mean of the variances of several linear functions of the parameters, defined by $\mathbf{L}'(\beta_0\ \beta')'$ by minimizing $\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\}$, where $\mathbf{W}=\mathbf{L}\mathbf{L}'$. If $\mathbf{W}$ is diagonal, this reduces to weighted-A-optimality and if all the diagonal elements are equal we obtain A-optimality, whereas, if p_2 of them are equal and the rest are 0, we obtain A_S-optimality. Note that A- and A_S-optimality are scale dependent, i.e. they are not invariant to linear reparameterizations, so, whenever a design is described as A or A_S optimal, we shall state with respect to which scaling. If $\mathbf{L}$ has a single column, the criterion reduces to c-optimality. With L-optimality the property that is claimed for the criterion can be related to point estimation and so the standard criteria are acceptable in this sense. However, the width of a confidence interval for the ith linear function of the parameters is proportional to $\surd\{F_{1,d;1-\alpha}\mathbf{l}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{l}_i\}$, where $\mathbf{l}_i'$ is the ith row of the matrix $\mathbf{L}'$, and so the mean of the squared lengths of such intervals is minimized by minimizing

$$F_{1,d;1-\alpha}\,\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\},$$

which we refer to as the LP-criterion, with the letter P also being used for the special cases, such as AP-optimality. It might be advisable to replace $F_{1,d;1-\alpha}$ with a similar quantity corrected for multiple testing, but this is rarely done in the analysis of data from experiments of this type.

The usual form of G-optimality minimizes the maximum variance of the estimated response over the design region by minimizing $\max_{\mathbf{x}}\{(1\ \mathbf{f}(\mathbf{x})')(\mathbf{X}'\mathbf{X})^{-1}(1\ \mathbf{f}(\mathbf{x})')'\}$. Again this is suitable for point estimation but, if (pointwise) confidence intervals for the mean response or prediction intervals for new units are required, we should define the GP-criterion to be to minimize

$$F_{1,d;1-\alpha}\max_{\mathbf{x}}\{(1\ \mathbf{f}(\mathbf{x})')(\mathbf{X}'\mathbf{X})^{-1}(1\ \mathbf{f}(\mathbf{x})')'\}.$$

The same general idea can be used for any other design optimality criterion which relates to the experiment's ability to allow statistical inference of any type to be performed. Other obvious examples include I-optimality and T-optimality—see Atkinson *et al.* (2007) for these and other criteria.

We make the following comments about these new criteria.

(a) An echo of the general idea can be found in the last section of Fisher (1966), pages 242–245, in the context of sample size calculations. However, Fisher's method is based on fiducial probability, which might have hindered its acceptance. As he wrote, it

> 'is unintelligible only to those who over a long period resisted the cogency of the fiducial argument'.

(b) In experiments in which the only treatment design question is how many (greater than 0) replicates there should be of each treatment, e.g. experiments with unstructured treatments, the number of distinct treatments is constant, so d depends only on the total number of experimental units and hence the new criteria are identical to the standard criteria.

(c) Asymptotically, as $d \rightarrow \infty$, the new criteria converge to the standard criteria. Hence, in very large experiments, the designs that are chosen will be the same.

(d) The concept of continuous design is not meaningful with the new criteria, since the quantiles of the F-distributions are not proportional to n. Consequently, there are no equivalence theorems. Hence, for finding optimum or near optimum designs, there is usually no alternative to either an exhaustive search or a discrete optimization heuristic, such as an exchange algorithm.

(e) The standard versions of most criteria are meaningful in terms of point estimation, but for the D-criterion and its variants the standard version really has no statistical interpretation and should be abandoned.

Although these criteria are new, the idea of including enough degrees of freedom to estimate pure error is common in response surface methodology. Following Dykstra (1959), several researchers have considered partially replicated two-level designs, although optimality is not considered. More recently, Liao and Chai (2004) and Dasgupta and Jacroux (2010) evaluated their partially replicated designs by using the standard D-criterion. However, in all these cases, the choice of the number of pure error degrees of freedom is made informally and separately from optimality considerations.

We have implemented some of the new criteria in a standard exchange algorithm for constructing optimum designs. A candidate set of treatments (combinations of factors' levels) is formed. A random initial design, selected from the candidate set, starts the search and the complete procedure is repeated for a number of different initial designs (tries), as is usual for exchange algorithms. The search proceeds by systematically making exchanges between treatments in the current design and in the candidate set and accepting any exchange that improves the criterion function. Many other types of exchange algorithm exist, e.g. making use of the ideas of excursion (making more than one exchange before evaluating the design) or stochastic ideas, such as accepting with small probability an exchange which makes the design worse. For typical response surface problems, the simple exchange algorithm that we use is as good as any other, although for specific examples it might be possible to do slightly better. The numbers of pure error degrees of freedom are obtained by labelling the treatment combinations and counting the number of treatment levels at each exchange tried. These algorithms will find near optimum designs, but there is no guarantee that they will find the global optimum.

3.2. Examples

To show how the optimum designs using the new criteria differ from those using the standard criteria, we present some illustrative examples. D_S- and A_S-optimum designs are found for polynomial models, the nuisance parameter being the intercept, such that $(\mathbf{M}^{-1})_{22} = (\mathbf{X}'\mathbf{Q}_0\mathbf{X})^{-1}$ where $\mathbf{Q}_0 = \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}'$ and the $\mathbf{X}$-matrix does not include the column of 1s. When aiming at a second-order model in a cubic region of experimentation under the A_S-criterion, we use weights to bring the different effects to the same scale, i.e. linear and interaction terms are given weight 1 and quadratic effects are given weight 0.25. This ensures that the optimum design for each individual parameter gives the same variance for the relevant parameter and can be considered as a simple application of the scaling that was recommended for non-linear models by Atkinson *et al.* (1993). The elements of the diagonal of $\mathbf{W}$ were scaled to add up to 1. In a spherical region, all parameters are given equal weights because they contribute equally to the polynomial approximation to the unknown true response function. We use the notation of subset designs to describe the treatments that are used. For q factors S_r is defined as the subset from the full factorial, in which r factors appear as ± 1 (or plus or minus another constant in a spherical region) and, if $r < q$, $q - r$ factors appear as 0; see Gilmour (2006) and Ahmad and Gilmour (2010) for more details on subset designs. The following tables show D_S-, $(DP)_S$-, A_S-, $(AP)_S$- and, in some cases, compound optimum designs for different numbers of factors and runs. Compound criteria are discussed in Section 4. To discriminate better between the designs, the tables also show pure error, PE, and lack of fit, LoF, degrees of freedom and efficiencies (in percentages) with respect to the best design found under each criterion.

3.2.1. Example 1 ($n = 16$; $q = 3$; $p = 10$)

In Table 1 we show a few designs, under the second-order model, for three factors in 16 runs. Using the D_S- or A_S-criteria resulted in the same design I which, as usual, includes very few points close to the centre of the design region. There are no replicated treatments and thus the design does not allow pure error estimation. It does, however, allow 6 degrees of freedom for lack of fit. Using the pure error versions of the criteria resulted in 6 and 5 degrees of freedom for the $(DP)_S$- and $(AP)_S$-criteria, designs II and III, respectively. Design II is very extreme in the sense that it has no degrees of freedom for checking lack of fit. In both designs there is replication of points in the corners and on the faces of the cube, but no centre points. This is quite different from the usual practice of experimenters. Design IV is $(AP)_S$ $(\alpha = 0.05/9)$ optimal, i.e. we used a Bonferroni adjustment for multiple comparisons. Although designs II and IV are equivalent in terms of their pure error degrees of freedom they have different properties for estimating the effects: the former being better for joint inferences on the treatment parameters; the latter for inferences on individual parameters. For comparison we also evaluated two subset designs, a CCD with two centre points $(S_3 + S_1 + 2S_0)$, allowing 1 degree of freedom for pure error and a modified Box–Behnken design $(S_2 + 4S_0)$, allowing 3 degrees of freedom for pure error. The CCD is quite similar to the $(AP)_S$-optimum design in terms of properties of the information matrix (D_S-eff = 93.15; A_S-eff = 90.75) but poorer for pure error estimation (($(DP)_S$-eff = 1.91; $(AP)_S$-eff = 4.31). The Box–Behnken design is a compromise in terms of pure error and lack-of-fit degrees of freedom, but is poor in terms of variance properties (D_S-eff = 74.94; A_S-eff = 66.34; $(DP)_S$-eff = 41.95; $(AP)_S$-eff = 50.17). The desirability of such compromises will be discussed further in Section 4.

3.2.2. Example 2 ($n = 18$; $q = 3$; $p = 10$)

For the experiment that was described in exercise 11.6 of Box and Draper (2007), which was

Table 1. Optimum† designs and their properties for three three-level factors in 16 runs under the second-order model (example 1)

	Designs for the following criteria:								
	D_S or A_S—I			*$(DP)_S$—II*			*$(AP)_S$—III*		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
	−1	−1	−1	−1	−1	1	−1	−1	−1
	−1	−1	1	−1	−1	1	−1	−1	1
	−1	1	−1	−1	1	−1	−1	1	−1
	−1	1	1	−1	1	−1	−1	1	−1
	1	−1	−1	−1	1	1	−1	1	1
	1	−1	1	1	−1	1	1	−1	−1
	1	1	−1	1	−1	1	1	−1	1
	1	1	1	1	1	1	1	−1	1
	1	−1	0	1	0	−1	1	1	−1
	1	1	0	1	0	−1	1	1	1
	1	0	−1	0	−1	−1	1	0	0
	1	0	1	0	−1	−1	1	0	0
	0	1	1	−1	0	0	0	1	0
	−1	0	0	−1	0	0	0	1	0
	0	−1	0	0	1	0	0	0	1
	0	0	−1	0	0	1	0	0	1
df (PE; LoF)‡		(0; 6)			(6; 0)			(5; 1)	
D_S-eff		100.00			83.09			93.03	
A_S-eff		100.00			66.52			86.27	
$(DP)_S$-eff		0.00			100.00			96.17	
$(AP)_S$-eff		0.00			85.10			100.00	

(*continued*)

mentioned in Section 2, several optimum designs are shown in Table 2. The efficiencies for the Box and Draper design (using the axial points at $\sqrt{3}$, which is slightly different from those actually used) are D_S-eff = 96.36, A_S-eff = 98.68, $(DP)_S$-eff = 42.46 and $(AP)_S$-eff = 63.10. Note that design III is very similar to this design. Again the designs that are built by the new criteria give quite extreme designs which allow little or no testing for lack of fit.

3.2.3. *Example 3: cassava bread ($n=26$; $q=3$; $p=10$)*

Escouto (2000) performed an experiment to formulate a recipe for gluten-free bread based on cassava flour. Gluten-free food is recommended to people with coeliac disease, which is an autoimmune disorder of the small intestine which occurs in genetically predisposed individuals, and there is a large market for gluten-free versions of staple foods. The experiment involved three factors: X_1, the amount of powdered albumen (egg white) with levels 10, 20 and 30 g; X_2, the amount of yeast with levels 5, 10 and 15 g; X_3, the amount of ground cassava flour with levels 45, 55 and 65 g. Other substances such as salt, sugar, vegetable fat, powdered milk, fermented cassava starch and water were maintained at constant levels in all recipes, as were factors that were associated with the mixing and baking processes. A modified CCD, with duplicated factorial points and four centre points $(2S_3+S_1+4S_0)$ in 26 runs was planned, allowing 11 degrees of freedom for pure error. Several organoleptic characteristics of the product were evaluated as response variables. The objective was to find a formulation which would present similar

Table 1 (*continued*)

	Design for criterion $(AP)_S$§§—IV			*Designs for the following compound criteria§:*					
				$\kappa = (0, 0.2, 0, 0.8)$§§, $\kappa = (0.2, 0, 0, 0.8)$—V			*$\kappa = (0, 0.2, 0, 0.8)$—VI*		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
	−1	−1	−1	−1	−1	−1	−1	−1	−1
	−1	1	−1	−1	−1	1	−1	1	−1
	1	−1	−1	−1	1	−1	−1	1	1
	1	1	−1	−1	1	1	1	−1	−1
	−1	0	1	1	−1	−1	1	−1	−1
	−1	0	1	1	−1	1	1	−1	1
	1	0	1	1	1	−1	1	1	−1
	1	0	1	1	1	−1	1	1	−1
	0	−1	1	1	1	1	1	1	1
	0	−1	1	1	1	1	1	1	1
	0	1	1	−1	0	0	−1	−1	0
	0	1	1	1	0	0	−1	0	1
	−1	0	0	0	−1	0	0	−1	1
	−1	0	0	0	−1	0	1	0	0
	0	0	−1	0	0	−1	0	1	0
	0	0	−1	0	0	−1	0	0	−1
df (PE; LoF)‡		(6; 0)			(4; 2)			(3; 3)	
D_S-eff		80.94			95.10			98.68	
A_S-eff		73.09			89.46			96.03	
$(DP)_S$-eff		97.42			78.21			55.24	
$(AP)_S$-eff		93.50			88.89			72.63	

†Where a heading indicates two criteria, the design is optimal for both.
‡Degrees of freedom for pure error and lack of fit.
§The compound criterion is defined in equation (5) in Section 4.1.
§§Confidence level corrected for multiple comparisons.

characteristics to wheat-based white bread. Several alternative designs are presented in Table 3. Again the D_S- and A_S-criteria give identical designs (design I) allowing 9 degrees of freedom for pure error. Using the $(DP)_S$- and $(AP)_S$-criteria resulted in 15 and 12 degrees of freedom for pure error respectively. Using a Bonferroni correction for multiple comparisons, the $(AP)_S$-$(\alpha = 0.05/9)$ optimum design (design IV) gives a design allowing 13 degrees of freedom for pure error. These designs show that, for carrying out inference, it is beneficial to sacrifice a little in terms of the traditional criteria to improve the estimation of σ^2.

For comparison we show in Table 4 the properties of a few subset designs, including the design $(2S_3 + S_1 + 4S_0)$ which was actually used in the experiment. We note that this design allows similar numbers of pure error degrees of freedom to design III but it is poorer with respect to the other properties. A modifed Box–Behnken design $(2S_2 + 2S_0)$ allows the same number of pure error degrees of freedom as the $(AP)_S$- $(\alpha = 0.05/9)$ optimum design but it is also poorer with respect to the other properties. The 3^3-factorial excluding the centre point $(S_3 + S_2 + S_1)$ is quite attractive with respect to its variance properties but allows no degrees of freedom for pure error. Other attempts to construct designs from the S_r subsets are quite inferior with respect to all the properties evaluated.

3.2.4. *Example 4: oil extraction ($n = 40$; $q = 5$; $p = 20$)*

An experiment on the extraction of oil from oilseeds was described by Rosenthal *et al.* (2001).

Table 2. Optimum designs and their properties for three factors in a spherical region ($\alpha = \sqrt{(3/2)}$) in 18 runs under the second-order model (example 2)

	Designs for the following criteria:								
	D_S—*I*			$(DP)_S$—*II*			A_S—*III*		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
	-1	-1	-1	1	-1	-1	-1	-1	-1
	-1	-1	1	1	-1	-1	-1	-1	-1
	-1	1	-1	1	1	-1	-1	-1	1
	-1	1	1	1	1	-1	-1	1	-1
	1	-1	1	0	$-\alpha$	α	-1	1	1
	1	1	-1	0	$-\alpha$	α	1	-1	-1
	1	1	1	0	α	α	1	-1	1
	0	$-\alpha$	$-\alpha$	0	α	α	1	1	-1
	α	0	$-\alpha$	$-\alpha$	0	$-\alpha$	1	1	1
	α	$-\alpha$	0	$-\alpha$	0	$-\alpha$	$-\sqrt{3}$	0	0
	0	0	$-\sqrt{3}$	$-\alpha$	0	α	$\sqrt{3}$	0	0
	0	0	$\sqrt{3}$	$-\alpha$	$-\alpha$	0	0	$-\sqrt{3}$	0
	0	$-\sqrt{3}$	0	$-\alpha$	$-\alpha$	0	0	$\sqrt{3}$	0
	0	$\sqrt{3}$	0	$-\alpha$	α	0	0	0	$-\sqrt{3}$
	$-\sqrt{3}$	0	0	$-\alpha$	α	0	0	0	$\sqrt{3}$
	$\sqrt{3}$	0	0	$\sqrt{3}$	0	0	0	0	0
	0	0	0	$\sqrt{3}$	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
df (PE; LoF)		(1;7)			(8;0)			(3;5)	
D_S-eff		100.00			87.28			98.75	
A_S-eff		96.49			73.56			100.00	
$(DP)_S$-eff		1.61			100.00			43.50	
$(AP)_S$-eff		3.87			89.58			63.94	

(*continued*)

The experiment involved five factors, one of them qualitative (type of enzyme) at two levels (X_1), and a second-order model was expected to fit the data. The actual experiment included 50 runs, 10 of which had no enzyme added and thus two of the other factors were not defined. For illustration we shall consider designs for the 40 runs with enzyme added. For this part of the experiment, the design used was $\frac{1}{2}S_5 + S_2 + 4S_1$ which allows 6 degrees of freedom for pure error. This design and four optimum designs are shown in Table 5. The D_S-optimum design (design I) allows just 1 degree of freedom for pure error, whereas the A_S-optimum design (design III) allows none. The $(DP)_S$-optimum design (design II) has 20 degrees of freedom for pure error, but no degrees of freedom for lack of fit. The effect of this on the other properties of the design is quite large. The $(AP)_S$-optimum design allows 15 degrees of freedom for pure error and 5 for lack of fit and might seem more reasonable to many experimenters.

3.2.5. *Example 5 ($n = 16$; $q = 8$; $p = 9$)*

Table 6 shows designs for eight two-level factors in 16 runs, under the first-order model. This situation differs from the others considered in that the D_S- and A_S-optimum design (design I) has the same information matrix as a regular 1/16 fractional factorial, which does not allow for pure error estimation. The $(DP)_S$-optimum design (design II) is an irregular fraction which

Table 2 (*continued*)

	Design for criterion $(AP)_S$*—IV*			*Designs for the following compound criteria*†:					
				$\kappa=(0.5, 0, 0, 0.5)$—*V*			$\kappa=(0, 0.5, 0, 0.5)$—*VI*		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
	-1	1	1	-1	-1	1	$-\alpha$	$-\alpha$	0
	-1	1	1	-1	-1	1	$-\alpha$	$-\alpha$	0
	1	-1	1	1	-1	-1	$-\alpha$	α	0
	1	-1	1	1	-1	-1	α	$-\alpha$	0
	1	1	-1	1	-1	1	α	α	0
	1	1	1	0	α	$-\alpha$	$-\alpha$	0	$-\alpha$
	1	1	1	0	α	$-\alpha$	$-\alpha$	0	α
	$-\alpha$	$-\alpha$	0	0	α	α	$-\alpha$	0	α
	$-\alpha$	$-\alpha$	0	$-\alpha$	0	$-\alpha$	α	0	$-\alpha$
	0	$-\alpha$	$-\alpha$	$-\alpha$	0	$-\alpha$	α	0	α
	0	$-\alpha$	$-\alpha$	$-\alpha$	α	0	0	$-\alpha$	$-\alpha$
	$-\alpha$	0	$-\alpha$	$-\alpha$	α	0	0	α	$-\alpha$
	$-\alpha$	0	$-\alpha$	α	α	0	0	$-\alpha$	α
	$\sqrt{3}$	0	0	0	0	$\sqrt{3}$	0	α	α
	0	$\sqrt{3}$	0	0	$-\sqrt{3}$	0	0	$-\alpha$	α
	0	0	$\sqrt{3}$	$\sqrt{3}$	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0
df (PE; LoF)		(7;1)			(6;2)			(5;3)	
D_S-eff		90.69			93.50			93.49	
A_S-eff		86.34			88.55			96.58	
$(DP)_S$-eff		95.75			88.55			76.05	
$(AP)_S$-eff		100.00			95.78			94.66	

†The compound criterion is defined in equation (5) in Section 4.1.

allows 7 degrees of freedom for pure error and none for lack of fit. Again this shows that, for inference, it is worth sacrificing a considerable amount in terms of the D-criterion to gain pure error degrees of freedom. The $(AP)_S$-optimum design (design III) is even more extremely irregular, with one factor having 12 runs at one level and four runs at the other. Of course, by sacrificing orthogonality, we lose the ability to perform other types of analysis, such as normal or half-normal plots of effects. If this, rather than formal inference, was regarded as the main purpose of the experiment then, of course, we should use the standard A_S-criterion to reflect this.

3.3. Blocked designs

Since many multifactor designs use moderately large numbers of runs, it is common that they must be run in blocks. For example, in industrial experiments the blocks often correspond to days or shifts.

In a blocked design the full treatment model for the response in unit *j* of block *i* with treatment *k* applied to it may be written as

$$Y_{ij(k)} = \mu_i + \tau_k + \varepsilon_{ij}, \tag{3}$$

where μ_i is the expected response in block i, τ_k is the effect of treatment k, $i=1,\ldots,b$, $j=1,\ldots,n_i$ $(\Sigma_{i=1}^{b} n_i = n)$, $k=1,\ldots,t$, $E(\varepsilon)=\mathbf{0}$ and $V(\varepsilon)=\sigma^2\mathbf{I}$. Block effects may be fixed or random

Table 3. Optimum designs and their properties for three three-level factors in 26 runs under the second-order model (example 3)

	Designs for the following criteria:											
	D_S *or* A_S, $\kappa=(0.5, 0, 0, 0.5)$†, $\kappa=(0, 0.75, 0, 0.25)$†—*I*			$(DP)_S$—*II*			$(AP)_S$, $\kappa=(0.75, 0, 0, 0.25)$† —*III*			$(AP)_S$‡—*IV*		
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1
	−1	−1	1	−1	−1	1	−1	−1	1	−1	−1	1
	−1	1	−1	−1	−1	1	−1	−1	1	−1	−1	1
	−1	1	−1	−1	1	−1	−1	1	−1	−1	1	−1
	−1	1	1	−1	1	−1	−1	1	−1	−1	1	−1
	1	−1	−1	−1	1	1	−1	1	1	−1	1	1
	1	−1	−1	−1	1	1	−1	1	1	−1	1	1
	1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1
	1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1
	1	1	−1	1	−1	1	1	−1	1	1	−1	1
	1	1	−1	1	−1	1	1	−1	1	1	−1	1
	1	1	1	1	1	−1	1	1	−1	1	1	1
	1	1	1	1	1	−1	1	1	−1	1	1	1
	−1	−1	0	1	1	−1	1	1	1	1	1	0
	−1	1	0	1	1	1	1	1	0	1	1	0
	−1	0	1	1	1	1	1	1	0	1	0	−1
	−1	0	1	1	0	0	1	0	1	1	0	−1
	0	−1	1	1	0	0	0	1	1	0	1	−1
	0	1	1	1	0	0	0	1	1	0	1	−1
	1	0	0	0	1	0	−1	0	0	−1	0	0
	1	0	0	0	1	0	−1	0	0	−1	0	0
	0	−1	0	0	1	0	0	−1	0	0	−1	0
	0	1	0	0	0	1	0	−1	0	0	−1	0
	0	0	−1	0	0	1	0	0	−1	0	0	1
	0	0	−1	0	0	1	0	0	−1	0	0	1
df (PE; LoF)		(9; 7)			(15; 1)			(12; 4)			(13; 3)	
D_S-eff		100.00			93.81			98.80			97.87	
A_S-eff		100.00			87.13			97.14			98.25	
$(DP)_S$-eff		86.77			100.00			97.45			99.46	
$(AP)_S$-eff		95.50			93.73			100.00			99.94	

†The compound criterion is defined in equation (5) in Section 4.1.
‡Confidence level corrected for multiple comparison.

but, as discussed in Gilmour and Trinca (2006), in the design phase it is safer to consider them as fixed since the variance components are not known. This ensures that in the most difficult case, with a large block variance component, the design is optimal, whereas, when the block variance component is small, though the design might be suboptimal for this case, it will give better estimation than in the case of a large block variance component. We shall follow this advice here.

As with completely randomized designs, we shall try to fit submodels for τ_k such as

$$\tau_k = \mathbf{f}(\mathbf{x}_k)'\boldsymbol{\beta}, \tag{4}$$

Table 4. Subset designs and their properties for three three-level factors in 26 runs under the second-order model (example 3)

Design	*df (PE; LoF)*	D_S-*eff*	A_S-*eff*	$(DP)_S$-*eff*	$(AP)_S$-*eff*
$2S_3+S_1+4S_0$	(11; 5)	90.89	82.43	86.56	83.16
$S_3+2S_1+6S_0$	(11; 5)	72.68	66.63	69.22	67.22
$S_3+S_2+S_1$	(0; 16)	94.27	92.82	0.00	0.00
$2S_2+2S_0$	(13; 3)	78.71	70.79	79.99	74.13
$S_2+2S_1+2S_0$	(7; 9)	58.81	45.26	44.12	39.56
$S_2+S_1+8S_0$	(7; 9)	55.78	43.89	41.85	38.36

where $\mathbf{f}(\mathbf{x}_k)$ and $\boldsymbol{\beta}$ are defined as in equation (2). Note that $\mathbf{x}_k$ may be written as $\mathbf{x}_{ij}$, i.e. the levels of the q factors that were applied to unit j of block i.

By fitting model (3) we obtain an unbiased estimator of σ^2 under the usual assumption of additive treatment effects, plus a valid randomization. Using the same argument as for unblocked designs, for inferences on the $\boldsymbol{\beta}$-parameters we should use this estimate of error variance and so should use a design which allows its estimation. Note that, as is usually the case, some treatment effects may be confounded with blocks.

The variance matrix of the least squares estimator $\hat{\boldsymbol{\beta}}$ is

$$\mathrm{var}(\hat{\boldsymbol{\beta}})=\sigma^2(\mathbf{M}_\mathrm{B}^{-1})_{22}=\sigma^2(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1},$$

where

$$\mathbf{M}_\mathrm{B}=\begin{pmatrix}\mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{X}\\ \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{X}\end{pmatrix},$$

$\mathbf{X}$ is defined as before (the column of 1s is excluded and so $\mathbf{X}$ has $p-1$ columns), $\mathbf{Q}=\mathbf{I}-\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ and $\mathbf{Z}$ is the $n\times b$ matrix whose columns are indicators for blocks. The subscript B stands for the blocked design.

Criteria for designs allowing for blocks, the usual ones and the new ones proposed in this paper, are defined as before with $\mathbf{M}_\mathrm{B}$ or $\mathbf{X}'\mathbf{Q}\mathbf{X}$ taking the place of $\mathbf{X}'\mathbf{X}$ or $\mathbf{X}'\mathbf{Q}_0\mathbf{X}$, and the appropriate numerator degrees of freedom. In particular, D_S-optimum designs are also D-optimum designs because $|\mathbf{M}_\mathrm{B}|=|\mathbf{X}'\mathbf{Q}\mathbf{X}||\mathbf{Z}'\mathbf{Z}|$ and $|\mathbf{Z}'\mathbf{Z}|$ is constant. However, $(\mathrm{DP})_\mathrm{S}$-optimum designs are not DP-optimum designs, owing to the reduction in numerator degrees of freedom. Whereas the DP-criterion minimizes $(F_{p+b-1,d_\mathrm{B};1-\alpha})^{p+b-1}/|\mathbf{M}_\mathrm{B}|$ the $(\mathrm{DP})_\mathrm{S}$-criterion minimizes $(F_{p-1,d_\mathrm{B};1-\alpha})^{p-1}/|\mathbf{X}'\mathbf{Q}\mathbf{X}|$, where d_B is the pure error degrees of freedom from the blocked design.

An exchange algorithm is also applied to construct near optimum blocked designs. This is a simple extension of the algorithm that was briefly described in Section 3.1. A random initial design starts the search and the complete procedure is repeated for a number of different initial designs. Exchanges which improve the criterion are accepted. For the new criteria the calculation of $d_\mathrm{B}=n-\mathrm{rank}(\mathbf{Z}:\mathbf{T})$, where $\mathbf{T}$ is the $n\times t$ matrix indicator for treatments, is required for obtaining the pure error degrees of freedom, reducing the computational benefits of updating formulae.

3.3.1. Example 6: pastry dough ($n=28; b=7; q=3; p=10$)

This experiment was described by Trinca and Gilmour (1999). The main objective was to discover how some factors that are involved in an extrusion process for mixing dough could be

Table 5. Designs and their properties for five factors, one at two levels and four at three levels, in 40 runs, under the second-order model (example 4; see also Table 8 in Section 4.3.4)

Design $\frac{1}{2}S_5+S_2+4S_1$, *used*					*Designs for the following criteria:*																			
					D_S—*I*					$(DP)_S$—*II*					A_S—*III*					$(AP)_S$—*IV*				
X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
−1	−1	−1	−1	−1	−1	−1	−1	−1	1	−1	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
−1	−1	−1	1	1	−1	−1	−1	1	−1	−1	−1	−1	−1	1	−1	−1	1	1	1	−1	−1	−1	1	−1
1	−1	1	−1	1	−1	−1	1	−1	1	−1	−1	−1	1	−1	−1	1	−1	1	1	−1	−1	−1	1	−1
−1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1	−1	1	−1	−1	1	1	−1	1	−1	−1	1	1	1
−1	1	−1	−1	1	−1	1	1	−1	−1	−1	−1	1	−1	−1	−1	1	1	1	−1	−1	1	−1	−1	−1
−1	1	−1	1	−1	−1	1	1	1	1	−1	−1	1	−1	−1	1	−1	−1	−1	−1	−1	1	−1	1	1
−1	1	1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	1	−1	1	−1	1	1
−1	1	1	1	1	1	−1	−1	−1	1	−1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1	−1	1
1	−1	−1	−1	1	1	−1	−1	1	−1	−1	1	−1	−1	−1	1	−1	−1	1	1	−1	1	1	−1	1
1	−1	−1	1	−1	1	−1	−1	1	1	−1	1	−1	−1	−1	1	−1	1	−1	−1	−1	1	1	1	−1
1	−1	1	−1	−1	1	−1	1	−1	−1	−1	1	−1	1	1	1	−1	1	−1	1	1	−1	−1	−1	−1
1	−1	1	1	1	1	−1	1	−1	1	−1	1	−1	1	1	1	−1	1	1	−1	1	−1	−1	−1	−1
1	1	−1	−1	−1	1	−1	1	1	−1	−1	1	1	−1	1	1	−1	1	1	1	1	−1	−1	1	1
1	1	−1	1	1	1	−1	1	1	1	−1	1	1	−1	1	1	1	−1	−1	−1	1	−1	−1	1	1
1	1	1	−1	1	1	1	−1	−1	−1	−1	1	1	1	−1	1	1	−1	−1	1	1	−1	1	1	−1
1	1	1	1	−1	1	1	−1	−1	1	−1	1	1	1	−1	1	1	−1	1	−1	1	−1	1	1	−1
−1	−1	0	0	0	1	1	−1	1	−1	1	−1	−1	−1	−1	1	1	−1	1	1	1	1	−1	−1	1
−1	1	0	0	0	1	1	−1	1	1	1	−1	−1	−1	−1	1	1	1	−1	−1	1	1	−1	−1	1
1	−1	0	0	0	1	1	1	−1	−1	1	−1	−1	1	1	1	1	1	−1	1	1	1	−1	1	−1
1	1	0	0	0	1	1	1	−1	1	1	−1	−1	1	1	1	1	1	1	1	1	1	−1	1	−1
−1	0	−1	0	0	1	1	1	1	−1	1	−1	1	1	−1	1	1	1	1	−1	1	1	1	−1	−1
−1	0	1	0	0	1	1	1	1	1	1	−1	1	1	−1	−1	−1	−1	1	0	1	1	1	−1	−1
1	0	−1	0	0	−1	−1	−1	−1	0	1	1	−1	−1	1	−1	−1	1	−1	0	1	1	1	1	1
1	0	1	0	0	−1	−1	1	1	0	1	1	−1	−1	1	−1	1	−1	−1	0	1	1	1	1	1

(continued)

Table 5 (*continued*)

	Design $\frac{1}{2}S_5+S_2+4S_1$, *used*					*Designs for the following criteria:*																			
						D_S—*I*					$(DP)_S$—*II*					A_S—*III*					$(AP)_S$—*IV*				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
	−1	0	0	−1	0	−1	1	−1	1	0	1	1	−1	1	−1	−1	−1	−1	0	1	−1	−1	1	−1	0
	−1	0	0	1	0	−1	1	1	−1	0	1	1	−1	1	−1	−1	−1	1	0	−1	−1	−1	1	−1	0
	1	0	0	−1	0	−1	−1	−1	0	1	1	1	1	−1	−1	−1	1	−1	0	−1	−1	1	−1	−1	0
	1	0	0	1	0	−1	−1	1	0	−1	1	1	1	−1	−1	−1	−1	0	−1	1	−1	−1	1	0	−1
	−1	0	0	0	−1	−1	1	−1	0	−1	1	1	1	1	1	−1	−1	0	1	−1	1	−1	1	0	1
	−1	0	0	0	1	−1	1	1	0	1	1	1	1	1	1	−1	1	0	−1	−1	−1	−1	0	1	1
	1	0	0	0	−1	−1	−1	0	−1	−1	1	−1	1	−1	0	−1	0	−1	−1	1	1	−1	0	−1	1
	1	0	0	0	1	−1	−1	0	1	1	1	−1	1	−1	0	−1	0	−1	1	−1	1	0	1	−1	1
	−1	0	0	0	0	−1	1	0	−1	1	1	0	1	−1	1	−1	0	1	−1	−1	−1	0	1	1	0
	−1	0	0	0	0	−1	1	0	1	−1	1	0	1	−1	1	−1	0	1	1	1	−1	0	1	1	0
	−1	0	0	0	0	−1	0	−1	−1	−1	1	−1	0	0	1	−1	1	1	0	0	−1	0	−1	0	1
	−1	0	0	0	0	−1	0	−1	1	1	1	−1	0	0	1	−1	1	0	1	0	−1	0	−1	0	1
	1	0	0	0	0	−1	0	1	−1	1	−1	0	0	1	0	−1	1	0	0	1	−1	0	0	−1	−1
	1	0	0	0	0	−1	0	1	1	−1	−1	0	0	1	0	1	0	1	0	0	−1	0	0	−1	−1
	1	0	0	0	0	1	0	0	0	0	1	0	−1	0	0	1	0	0	1	0	1	1	0	0	0
	1	0	0	0	0	1	0	0	0	0	1	0	−1	0	0	1	0	0	0	1	1	1	0	0	0
df (PE; LoF)	(6; 14)					(1; 19)					(20; 0)					(0; 20)					(15; 5)				
D_S-eff	67.61					100.00					92.33					99.81					95.51				
A_S-eff	60.05					98.38					74.70					100.00					90.01				
$(DP)_S$-eff	40.28					0.93					100.00					0.00					94.48				
$(AP)_S$-eff	50.62					3.08					86.65					0.00					100.00				

Table 6. Optimum designs and their properties for eight two-level factors in 16 runs, under the linear effects model (example 5)

	Designs for the following criteria:															
	D_S *or* A_S*—I*								$(DP)_S$*—II*							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
	−1	−1	−1	−1	1	1	−1	1	−1	−1	−1	1	−1	−1	−1	−1
	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	−1	1	1	1	−1
	−1	−1	1	−1	−1	1	1	−1	−1	−1	1	−1	1	1	1	−1
	−1	−1	1	1	1	−1	−1	−1	−1	−1	1	1	1	1	1	1
	−1	1	−1	−1	1	−1	1	−1	−1	−1	1	1	1	1	1	1
	−1	1	−1	1	−1	1	1	1	−1	1	−1	−1	−1	1	1	1
	−1	1	1	−1	−1	−1	−1	1	−1	1	−1	−1	−1	1	1	1
	−1	1	1	1	1	1	1	1	−1	1	1	−1	1	−1	−1	1
	1	−1	−1	−1	−1	−1	1	1	1	−1	−1	−1	1	1	−1	1
	1	−1	−1	1	1	1	−1	1	1	−1	−1	−1	1	1	−1	1
	1	−1	1	−1	1	−1	1	1	1	−1	1	−1	−1	−1	1	1
	1	−1	1	1	−1	1	1	−1	1	−1	1	−1	−1	−1	1	1
	1	1	−1	−1	−1	1	−1	−1	1	1	−1	1	1	−1	1	−1
	1	1	−1	1	1	−1	1	−1	1	1	−1	1	1	−1	1	−1
	1	1	1	−1	1	1	−1	−1	1	1	1	1	−1	1	−1	−1
	1	1	1	1	−1	−1	−1	1	1	1	1	1	−1	1	−1	−1
df (PE; LoF)					(0; 7)							(7; 0)				
D_S-eff					100.00							88.69				
A_S-eff					100.00							81.08				
(DP)$_S$-eff					0.00							100.00				
(AP)$_S$-eff					0.00							95.25				
	$(AP)_S$, $\kappa = (0, 0.5, 0, 0.5)$†*—III*								$\kappa = (0.5, 0, 0, 0.5)$†*—IV*							
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
	−1	−1	−1	1	−1	−1	1	−1	−1	−1	1	−1	−1	1	1	−1
	−1	−1	−1	1	−1	−1	1	−1	−1	−1	1	−1	1	−1	1	−1
	−1	−1	1	1	1	−1	−1	1	−1	−1	1	1	1	1	−1	1
	−1	−1	1	1	1	−1	−1	1	−1	−1	1	1	1	1	−1	1
	−1	1	−1	−1	−1	−1	−1	1	−1	1	−1	−1	−1	1	−1	1
	−1	1	−1	−1	−1	−1	−1	1	−1	1	−1	−1	1	−1	−1	1
	−1	1	−1	1	1	1	−1	−1	−1	1	−1	1	1	1	1	−1
	−1	1	−1	1	1	1	−1	−1	−1	1	−1	1	1	1	1	−1
	−1	1	1	−1	1	−1	1	−1	1	−1	−1	−1	1	1	−1	−1
	−1	1	1	−1	1	−1	1	−1	1	−1	−1	−1	1	1	−1	−1
	−1	1	1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1
	−1	1	1	1	−1	1	1	1	1	−1	−1	1	−1	−1	1	1
	1	−1	−1	−1	1	1	1	1	1	1	1	−1	1	1	1	1
	1	−1	1	−1	−1	1	−1	−1	1	1	1	−1	1	1	1	1
	1	1	−1	1	1	−1	1	1	1	1	1	1	−1	−1	−1	−1
	1	1	1	1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1
df (PE; LoF)					(6; 1)							(6; 1)				
D_S-eff					93.06							93.06				
A_S-eff					91.14							87.80				
(DP)$_S$-eff					94.27							94.27				
(AP)$_S$-eff					100.00							96.34				

†The compound criterion is defined in equation (5) in Section 4.1.

varied to control the properties of the pastry. Three controllable factors of interest were the flow rate of water into the mix, the initial moisture content in the mix and the screw speed. As the properties of the dough varied from day to day because of uncontrollable factors, the design was divided into seven blocks (days) of four runs each. Table 7 shows some alternative designs, along with the design that was used. The D_S- and A_S-optimum design (design I) allows 2 degrees of freedom for error. The $(DP)_S$-optimum design allows 12 degrees of freedom for error, whereas the $(AP)_S$-optimum design allows 10. Again we note that, for the $(DP)_S$-optimum design, there are no spare degrees of freedom for lack of fit. Because of the extra restrictions required by blocking, we see that the designs which allow for pure error estimation cost more in terms of the traditional criteria than in completely randomized designs. The design that was actually used is clearly inferior to the new designs.

Table 7. Designs and their properties for three three-level factors in seven blocks of four, under the second-order model (example 6; see also Table 9 in Section 4.3.6)

Block	*Designs for the following criteria:*									*Design used—IV*		
	D_S or A_S—I			*$(DP)_S$—II*			*$(AP)_S$—III*					
	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3	X_1	X_2	X_3
1	−1	1	−1	−1	−1	1	−1	−1	1	−1	−1	−1
	−1	−1	1	−1	1	−1	−1	1	−1	−1	1	1
	1	−1	−1	1	−1	−1	1	−1	−1	1	−1	1
	1	1	1	1	1	1	1	1	1	1	1	−1
2	−1	−1	−1	−1	−1	1	1	−1	−1	−1	−1	1
	1	1	−1	−1	1	−1	1	1	1	−1	1	−1
	1	−1	1	1	−1	−1	0	−1	1	1	−1	−1
	0	0	0	1	1	1	1	0	0	1	1	1
3	−1	1	1	−1	−1	1	−1	1	1	−1	1	−1
	1	−1	1	−1	1	−1	1	1	−1	0	−1	0
	1	1	−1	1	−1	−1	−1	−1	0	1	0	0
	0	0	0	1	1	1	0	0	−1	0	0	1
4	−1	1	−1	−1	1	1	−1	1	1	1	−1	1
	−1	−1	0	1	1	−1	1	1	−1	−1	0	0
	1	0	−1	−1	0	0	−1	−1	0	0	1	0
	0	1	1	0	−1	0	0	0	−1	0	0	−1
5	−1	1	1	−1	1	1	−1	1	−1	−1	−1	−1
	1	1	0	1	1	−1	1	−1	−1	1	1	1
	−1	0	−1	−1	0	0	0	−1	1	0	0	0
	0	−1	1	0	−1	0	1	0	0	0	0	0
6	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
	1	−1	0	1	−1	1	1	−1	1	1	1	−1
	−1	0	1	1	1	0	−1	0	1	0	0	0
	0	1	−1	0	0	−1	0	1	0	0	0	0
7	−1	−1	1	−1	−1	−1	−1	−1	−1	−1	1	1
	−1	1	0	1	−1	1	1	−1	1	1	−1	−1
	1	0	1	1	1	0	−1	0	1	0	0	0
	0	−1	−1	0	0	−1	0	1	0	0	0	0
df (PE; LoF)		(2; 10)			(12; 0)			(10; 2)			(7; 5)	
D_S-eff		100.00			88.19			94.87			80.02	
A_S-eff		100.00			78.40			92.48			73.31	
$(DP)_S$-eff		16.36			100.00			99.59			69.01	
$(AP)_S$-eff		29.00			88.65			100.00			70.38	

4. Compound criteria

4.1. Definition

The designs that are produced by the new criteria are quite extreme and experimenters might be reluctant to use designs which are so different from what they are used to. This attitude might be correct sometimes, depending on the objectives of the experiment. We would argue strongly in favour of carefully considering what kinds of data analysis will be used to meet the experimenters' objectives and then carefully choosing the optimality criterion to match that data analysis. If a joint confidence region or a global F-test of the treatment parameters will be the only relevant analysis, then a $(\mathrm{DP})_{\mathrm{S}}$-optimum design should be chosen.

In many experiments several types of data analysis are important, not all of them requiring an estimate of error. In particular, the analysis might involve all or most of the following:

(a) a global F-test of the treatment parameters, for which we should use $(\mathrm{DP})_{\mathrm{S}}$-optimality;
(b) t-tests of the individual treatment parameters, for which we should use weighted-AP-optimality;
(c) point estimation of the individual treatment parameters, for which we should use weighted-A-optimality;
(d) checking for lack of fit of the assumed treatment model and, if appropriate, fitting a few higher order terms.

For the last of these analyses, there is no obvious design optimality criterion to use. Atkinson (1972) proposed a criterion for finding designs which are powerful for detecting lack of fit. This requires a prior estimate of the sizes of the higher order parameters. Jones and Mitchell (1978) relaxed this by using a minimax or average version over a range of parameter values. More recently, Goos *et al.* (2005) combined this criterion with others to produce designs which are both model robust and model sensitive. However, all of these criteria aim specifically at testing for lack of fit, whereas we are also interested in being able to estimate a few higher order parameters and discriminating between models. To avoid having to consider too many different criteria, we use the degree-of-freedom efficiency that was proposed by Daniel (1976), pages 177–178, as a simple way of incorporating all of these requirements. The degree-of-freedom efficiency is the proportion of our experimental resource which is used to estimate the effects of treatments. Clearly this is directly in conflict with our pure error criteria and a good design must be a compromise.

We now combine all of these in a compound criterion, following exactly the methodology that was described by Atkinson *et al.* (2007). We first define the following efficiencies, for the design with treatment model matrix $\mathbf{X}$ which has d degrees of freedom for pure error:

(a) the $(\mathrm{DP})_{\mathrm{S}}$-efficiency,

$$E_1 = \frac{|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|^{1/(p-1)} F_{p-1,d_D;1-\alpha_1}}{F_{p-1,d;1-\alpha_1} |(\mathbf{X}'_{\mathrm{DP}}\mathbf{Q}_0\mathbf{X}_{\mathrm{DP}})|^{1/(p-1)}},$$

where $\mathbf{X}_{\mathrm{DP}}$ is the model matrix for the $(\mathrm{DP})_{\mathrm{S}}$-optimum design, which has d_D degrees of freedom for pure error, and the global F-test will be performed at the $100\alpha_1\%$ level of significance;
(b) the weighted-AP-efficiency,

$$E_2 = \frac{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'_{\mathrm{AP}}\mathbf{X}_{\mathrm{AP}})^{-1}\}F_{1,d_A;1-\alpha_2}}{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\}F_{1,d;1-\alpha_2}},$$

where $\mathbf{X}_{\mathrm{AP}}$ is the model matrix for the weighted-AP-optimum design, which has d_A degrees of freedom for pure error and the individual t-tests will be calculated at the $100\alpha_2\%$ level of significance;

(c) the weighted-A-efficiency,

$$E_3 = \frac{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'_A\mathbf{X}_A)^{-1}\}}{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\}},$$

where $\mathbf{X}_A$ is the model matrix for the weighted-A-optimum design;

(d) the degree-of-freedom efficiency,

$$E_4 = \frac{n-d}{n}.$$

Next we combine these criteria with weights $\kappa_1, \ldots, \kappa_4$ respectively to obtain $E = E_1^{\kappa_1} E_2^{\kappa_2} \times E_3^{\kappa_3} E_4^{\kappa_4}$. After ignoring terms which do not depend on the design to be optimized, this means that we choose a design to maximize

$$\frac{|\mathbf{X}'\mathbf{Q}_0\mathbf{X}|^{\kappa_1/(p-1)}(n-d)^{\kappa_4}}{F^{\kappa_1}_{p-1,d;1-\alpha_1} F^{\kappa_2}_{1,d;1-\alpha_2} \mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{X})^{-1}\}^{\kappa_2+\kappa_3}}. \tag{5}$$

The weights $\boldsymbol{\kappa}$ should be chosen to reflect the relative importance of different aspects of the analysis and, in some experiments, some of the weights might be 0.

4.2. Compound criteria for blocking

The compound criteria that were defined in Section 4.1 can easily be extended to blocked designs. For blocking, $\mathbf{X}'\mathbf{X}$ or $\mathbf{X}'\mathbf{Q}_0\mathbf{X}$ should be replaced by $\mathbf{M}_{\mathrm{B}}$ or $\mathbf{X}'\mathbf{Q}\mathbf{X}$ in the efficiency formulae. Hence, we have the following efficiencies:

(a) the $(\mathrm{DP})_{\mathrm{S}}$-efficiency,

$$E_1 = \frac{|\mathbf{X}'\mathbf{Q}\mathbf{X}|^{1/(p-1)}F_{p-1,\,d_{\mathrm{B}D};1-\alpha_1}}{F_{p-1,d_{\mathrm{B}};1-\alpha_1}|(\mathbf{X}'_{\mathrm{DP}}\mathbf{Q}\mathbf{X}_{\mathrm{DP}})|^{1/(p-1)}},$$

where $\mathbf{X}_{\mathrm{DP}}$ is the treatment model matrix for $\boldsymbol{\beta}$ for the $(\mathrm{DP})_{\mathrm{S}}$-optimum blocked design, which has $d_{\mathrm{B}D}$ degrees of freedom for pure error, and the global F-test will be calculated at the $100\alpha_1\%$ level of significance;

(b) the weighted-$(\mathrm{AP})_{\mathrm{S}}$-efficiency,

$$E_2 = \frac{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'_{\mathrm{AP}}\mathbf{Q}\mathbf{X}_{\mathrm{AP}})^{-1}\}F_{1,d_{\mathrm{B}A};1-\alpha_2}}{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\}F_{1,\,d_{\mathrm{B}};1-\alpha_2}},$$

where $\mathbf{X}_{\mathrm{AP}}$ is the treatment model matrix for $\boldsymbol{\beta}$ for the AP-optimum design, which has $d_{\mathrm{B}A}$ degrees of freedom for pure error and the individual t-tests will be calculated at the $100\alpha_2\%$ level of significance;

(c) the weighted-A_{S}-efficiency,

$$E_3 = \frac{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'_A\mathbf{Q}\mathbf{X}_A)^{-1}\}}{\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\}},$$

where $\mathbf{X}_A$ is the treatment model matrix for β for the A-optimum design;

(d) the degree-of-freedom efficiency,

$$E_4 = \frac{n-b+1-d_\mathrm{B}}{n-b+1}.$$

Ignoring constants, the combined criterion is

$$\frac{|\mathbf{X}'\mathbf{Q}\mathbf{X}|^{\kappa_1/(p-1)}(n-b+1-d_\mathrm{B})^{\kappa_4}}{F^{\kappa_1}_{p-1,d_\mathrm{B};1-\alpha_1}\,F^{\kappa_2}_{1,d_\mathrm{B};1-\alpha_2}\,\mathrm{tr}\{\mathbf{W}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\}^{\kappa_2+\kappa_3}}. \tag{6}$$

4.3. Examples

Designs were built, using various compound criteria, for several of the examples that were presented in earlier sections. Many different compound criteria can give the same optimum design and sometimes they can be the same as the designs that are obtained by using simple criteria. In what follows we discuss some of the more interesting designs that were found.

4.3.1. *Example 1 (n = 16; q = 3; p = 10), continued*

In Table 1 two designs (designs V and VI) are shown which place a high weight on degree-of-freedom efficiency, and the effect of this is very clear. The usual criteria allow no degrees of freedom for pure error and the pure-error-based criteria do not allow any test for lack of fit. The designs that are produced by the compound criteria are less extreme, offering a compromise between pure error and lack-of-fit degrees of freedom. We also note that it is possible to produce better designs than classical designs (e.g. regular subset designs) with similar numbers of degrees of freedom for pure error, e.g. comparing the modified Box–Behnken and design VI. The structure of design V is interesting, containing a full S_3 subset, with duplicates of two points, and six axial points, only two of which are a pair, the others being two replicates of one of each of the other pairs. This structure also seems like a compromise between the very tightly defined structure of subset designs and the apparently messy structures of D-optimum designs. This design, which is optimal for two different weight patterns for the compound criteria, is very similar to the CCD in terms of the variance properties, but it allows much better estimation of pure error. It looks very attractive for practical use.

4.3.2. *Example 2 (n = 18; q = 3; p = 10), continued*

A similar type of compromise can be seen in the compound optimum designs in Table 2 for the experiment in a spherical region. Designs V and VI can be recommended for practical use, even though they are quite different from the standard designs.

4.3.3. *Example 3: cassava bread (n = 26; q = 3; p = 10), continued*

Some of the obvious compound optimum designs, as shown in Table 3, turn out to be the same as some of the other optimum designs that were found. In a situation like this, the criteria are not far from converging to each other, although we still see that the design used can be optimized for different objectives.

4.3.4. *Example 4: oil extraction (n = 40; q = 5; p = 20), continued*

This is another case where the designs in Table 5 give almost all degrees of freedom to lack of fit, for the standard criteria, and to pure error for the new criteria, whereas classical designs, such as that used in the experiment, tend to split the residual degrees of freedom more evenly. Two compound optimum designs are shown in Table 8. Again, we find that they seem to represent a very good compromise, having a reasonable split between pure error and lack-of-fit degrees of freedom and very good variance properties. Design VI, for example, has fairly similar degrees of freedom to those of the design actually used, but it has much lower variances of the treatment parameter estimators. If this consulting problem arose now, this is the design that we would recommend.

4.3.5. Example 5 ($n=16$; $q=8$; $p=9$), continued

Even for the critical case of eight factors in 16 runs that was shown in Table 6, it is possible to obtain some weight patterns for the compound criterion which allow degrees of freedom for both pure error and lack-of-fit estimation. Designs III and IV have the same value of the D_S- and $(DP)_S$-criteria. Design IV looks attractive, since it also seems 'more regular' than the $(DP)_S$- and $(AP)_S$-optimum designs in that each factor is either level balanced or has an imbalance of two runs. Design III has higher weighted-A-efficiency but is highly irregular. For structures of this type, there might be scope for further study of aliasing patterns and their relationship with the new optimality criteria.

4.3.6. Example 6: pastry dough ($n=28$; $b=7$; $q=3$; $p=10$), continued

For the pastry dough experiment, two compound optimum designs are shown in Table 9, which again show an attractive compromise between variance properties and allocation of degrees of freedom. Design VI has the same allocation of degrees of freedom to pure error and lack of fit as the design actually used in the experiment, but it has considerably better estimation properties. In hindsight, this might be the design that we would recommend now.

These examples have shown that the properties of the design can be tailored to specific objectives by choosing appropriate weights (κs) in the compound criterion. As is frequently the case with weighted criteria, in practice the experimenter should try a few different settings of the κs and evaluate the resulting designs to decide which one to use. Some experimenters might appreciate even simpler rules of thumb and, on the basis of the examples given here, along with several others, we could recommend using relative weights of 0, 1 or 4 for analyses which are of no importance, of secondary importance and of primary importance respectively. Small changes in weights (except at 0) do not greatly affect the relative properties of different designs.

5. Discussion

In fitting polynomial models the validity of using higher order terms as an error estimate depends on the validity of the model, whereas the validity of pure error as an error estimate is not dependent on the fit of the model. Classical design criteria were developed assuming that an independent error estimate was available and thus, when that is not so, the usual criteria do not have the properties that they are intended to have. We have proposed modifications to the usual criteria so that the resulting designs take into account the necessity of obtaining a valid estimate of error for proper inferences about the parameters of the model. The corrected criteria, based on the quantiles of appropriate F-distributions for inferences, may result in quite extreme designs that do not allow any lack-of-fit checks. However, by using compound criteria which incorporate

Table 8. Compound optimum designs and their properties for five factors, one at two levels and four at three levels, in 40 runs, under the second-order model (example 4; see also Table 5)

	Designs for the following compound criteria:									
	$\kappa=(0.5, 0, 0, 0.5)$—*V*					$\kappa=(0, 0.5, 0, 0.5)$—*VI*				
	X_1	X_2	X_3	X_4	X_5	X_1	X_2	X_3	X_4	X_5
	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
	−1	−1	−1	−1	−1	−1	−1	−1	−1	1
	−1	−1	−1	1	1	−1	−1	−1	1	−1
	−1	−1	1	−1	1	−1	−1	−1	1	−1
	−1	−1	1	1	−1	−1	−1	1	−1	−1
	−1	−1	1	1	−1	−1	−1	1	1	1
	−1	1	−1	−1	1	−1	1	−1	−1	−1
	−1	1	−1	−1	1	−1	1	−1	1	1
	−1	1	−1	1	−1	−1	1	−1	1	1
	−1	1	−1	1	−1	−1	1	1	1	−1
	−1	1	1	−1	−1	−1	1	1	1	−1
	−1	1	1	−1	−1	1	−1	−1	−1	−1
	−1	1	1	1	1	1	−1	−1	−1	−1
	1	−1	−1	−1	1	1	−1	−1	1	1
	1	−1	−1	−1	1	1	−1	−1	1	1
	1	−1	−1	1	−1	1	−1	1	−1	1
	1	−1	−1	1	−1	1	−1	1	−1	1
	1	−1	1	−1	−1	1	−1	1	1	−1
	1	−1	1	−1	−1	1	−1	1	1	−1
	1	1	−1	−1	−1	1	1	−1	−1	1
	1	1	−1	−1	−1	1	1	−1	1	−1
	1	1	−1	1	1	1	1	1	1	1
	1	1	1	−1	1	1	1	1	1	1
	1	1	1	−1	1	−1	1	1	−1	0
	1	1	1	1	−1	1	1	−1	1	0
	1	1	1	1	−1	1	1	1	−1	0
	1	1	−1	1	1	−1	1	−1	0	−1
	1	−1	1	1	0	−1	1	1	0	1
	−1	1	1	0	1	1	1	1	0	−1
	1	−1	1	0	1	−1	−1	0	−1	−1
	−1	−1	0	−1	1	−1	1	0	−1	1
	1	−1	0	1	1	1	1	0	−1	−1
	−1	0	−1	1	1	−1	0	1	−1	1
	1	0	1	1	1	1	0	1	−1	−1
	−1	−1	−1	0	0	−1	−1	1	0	0
	1	1	−1	0	0	−1	0	−1	−1	0
	−1	1	0	1	0	1	0	−1	0	1
	−1	0	1	−1	0	1	0	0	1	−1
	−1	0	0	0	−1	1	−1	0	0	0
	1	0	0	0	−1	−1	0	0	1	0
df (PE; LoF)			(12; 8)					(9; 11)		
D_S-eff			98.36					98.34		
A_S-eff			92.43					95.84		
$(DP)_S$-eff			89.09					77.22		
$(AP)_S$-eff			98.27					94.53		

Table 9. Compound optimum designs and their properties for three three-level factors in seven blocks of four, under the second-order model (example 6; see also Table 7)

Block	*Designs for the following compound criteria:*					
	$\kappa = (0.5, 0, 0, 0.5)$—*V*			$\kappa = (0, 0.5, 0, 0.5)$—*VI*		
	x_1	x_2	x_3	x_1	x_2	x_3
1	−1	−1	1	−1	−1	1
	−1	1	−1	−1	1	−1
	1	−1	−1	1	−1	−1
	1	1	1	1	1	1
2	−1	−1	−1	−1	−1	−1
	−1	1	1	−1	1	1
	1	−1	0	0	1	−1
	0	0	1	1	0	0
3	−1	−1	−1	−1	−1	−1
	−1	1	1	1	1	−1
	1	0	−1	−1	1	0
	0	1	0	0	0	1
4	−1	−1	1	−1	−1	1
	1	1	1	1	−1	1
	1	0	−1	0	1	1
	0	1	0	1	0	0
5	−1	1	−1	−1	1	1
	1	1	1	1	−1	1
	1	−1	0	−1	0	−1
	0	0	1	0	−1	0
6	1	−1	1	1	−1	−1
	1	1	−1	1	1	1
	0	−1	−1	−1	1	0
	−1	0	0	0	0	1
7	1	−1	1	1	−1	1
	1	1	−1	1	1	−1
	0	−1	−1	−1	0	−1
	−1	0	0	0	−1	0
df (PE; LoF)		(9; 3)			(7; 5)	
D_S-eff		95.71			96.23	
A_S-eff		93.54			95.32	
$(DP)_S$-eff		95.47			82.99	
$(AP)_S$-eff		98.13			91.51	

weights reflecting the objectives of the experiments we construct compromise designs which are efficient in terms of the properties of the information matrix and allow pure error estimation as well as lack-of-fit checking. The criteria are defined for unblocked and blocked designs and the illustrative examples that were investigated lead us to conclude that they will be very useful in practice.

Our conclusion is that the D-criterion, as usually defined, has no place in the design of fractional factorial or response surface experiments, except in the unusual situation that the number of runs is very large. This criterion has no statistical interpretation and should be replaced by DP. The choice between A and AP (or similar) criteria depends on whether the experimenters' objectives will be met mainly through the interpretation of point estimates or through confidence intervals or hypothesis tests on the individual parameters. In most practical situations, meeting the objectives will require several forms of statistical analysis, including hypothesis tests, model checking and simplification and interpretation of the point estimates. We believe that the future of optimum design in practice lies in the application of compound criteria such as those which we have outlined here.

Acknowledgements

This work was carried out under Engineering and Physical Sciences Research Council research grant EP/C541715/1, 'Unifying approaches to design of experiments', and Fundação de Amparo à Pesquisa do Estado de São Paulo grants 2010/0250-08 and 2011/17851-8. Part of the work was completed while the authors were participants in the 'Design of experiments' programme at the Isaac Newton Institute for Mathematical Sciences in 2008. The authors thank the Editor, a screener and six referees for constructive comments, which helped us to sharpen the presentation of the work.

References

Ahmad, T. and Gilmour, S. G. (2010) Robustness of subset response surface designs to missing observations. *J. Statist. Planng Inf.*, **140**, 92–103.

Atkinson, A. C. (1972) Planning experiments to detect inadequate regression models. *Biometrika*, **59**, 275–293.

Atkinson, A. C., Chaloner, K., Herzberg, A. M. and Juritz, J. (1993) Optimum experimental designs for properties of a compartmental model. *Biometrics*, **49**, 325–337.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

Box, G. E. P. and Draper, N. R. (2007) *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd edn. New York: Wiley.

Cochran, W. G. and Cox, G. M. (1957) *Experimental Designs*, 2nd edn. New York: Wiley.

Cornell, J. A. (2002) *Experiments with Mixtures*, 3rd edn. New York: Wiley.

Cox, D. R. (1958) *Planning of Experiments*. New York: Wiley.

Daniel, C. (1976) *Applications of Statistics to Industrial Experimentation*. New York: Wiley.

Dasgupta, T. and Jacroux, M. (2010) Partially replicated fractional factorial designs. *Metrika*, **71**, 295–311.

Davies, O. L. (ed.) (1956) *Design and Analysis of Industrial Experiments*, 2nd edn. London: Oliver and Boyd.

Dean, A. and Voss, D. (1999) *Design and Analysis of Experiments*. New York: Wiley.

Draper, N. R. and Smith, H. (1998) *Applied Regression Analysis*, 3rd edn. New York: Wiley.

Dykstra, O. (1959) Partial duplication of factorial experiments. *Technometrics*, **1**, 63–73.

Escouto, L. F. S. (2000) Desenvolvimento de produto panicável a base de produtos de mandioca visando os hipersensíveis ao glúten. *MSc Dissertation*. Faculdade de Ciências Agronômicas, Universidade Estadual Paulista, Botucatu.

Fisher, R. A. (1966) *The Design of Experiments*, 8th edn. New York: Hafner.

Gilmour, S. G. (2006) Response surface designs for experiments in bioprocessing. *Biometrics*, **62**, 323–331.

Gilmour, S. G. and Trinca, L. A. (2006) Response surface experiments on processes with high variation. In *Response Surface Methodology and Related Topics* (ed. A. I. Khuri), pp. 19–46. New York: World Scientific Publishers.

Goos, P., Kobilinsky, A., O'Brien, T. E. and Vandebroek, M. (2005) Model-robust and model-sensitive designs. *Computnl Statist. Data Anal.*, **49**, 201–216.

Hinkelmann, K. and Kempthorne, O. (2005) *Design and Analysis of Experiments*, vol. 2. New York: Wiley.

Hinkelmann, K. and Kempthorne, O. (2008) *Design and Analysis of Experiments*, vol. 1, 2nd edn. New York: Wiley.

John, P. W. M. (1971) *Statistical Design and Analysis of Experiments*. New York: Macmillan.

Jones, E. R. and Mitchell, T. J. (1978) Design criteria for detecting model inadequacy. *Biometrika*, **65**, 541–551.

Khuri, A. I. and Cornell, J. A. (1996) *Response Surfaces*, 2nd edn. New York: Dekker.

Kiefer, J. (1959) Optimum experimental designs (with discussion). *J. R. Statist. Soc.* B, **21**, 272–319.

Liao, C.-T. and Chai, F.-S. (2004) Partially replicated two-level fractional factorial designs. *Can. J. Statist.*, **32**, 421–438.
Rosenthal, A., Pyle, D. L., Niranjan, K., Gilmour, S. and Trinca, L. (2001) Combined effect of operational variables and enzyme activity on aqueous enzymatic extraction of oil and protein from soybean. *Enz. Microb. Technol.*, **28**, 499–509.
Scheffé, H. (1959) *The Analysis of Variance*. New York: Wiley.
Trinca, L. A. and Gilmour, S. G. (1999) Difference variance dispersion graphs for comparing response surface designs with applications in food technology. *Appl. Statist.*, **48**, 441–455.
Wu, C. F. J. and Hamada, M. (2009) *Experiments*, 2nd edn. New York: Wiley.

Discussion on the paper by Gilmour and Trinca

Anthony C. Atkinson (*London School of Economics and Political Science*)
Our Society has a long tradition of papers for discussion on optimum experimental design, starting with Kiefer (1959). That paper, together with Kiefer and Wolfowitz (1959), laid the foundations of the modern treatment of the subject. Kiefer's careful exposition in Kiefer (1959) includes the important distinction between exact designs (those for a particular n) and approximate designs. The consequence was to free the subject from the necessity of having to consider each n separately. Unfortunately, tonight's authors have had to eschew this simplification and, by the nature of their problem, are forced to find only exact designs.

In section 2A(i) Kiefer (1959) discussed the effect on design of not knowing σ^2, a theme he alludes to at several later places in the paper, but his main interest was in the information matrix for specific models. In tonight's paper the argument is extended to three main aspects of a design:

(a) properties of the information matrix;
(b) degrees of freedom for estimating σ^2 as reflected in quantiles of the F-distribution and
(c) degrees of freedom for lack of fit.

In this list (a) leads to standard criteria such as D-optimality and (a) and (b) lead to tonight's interesting designs.

In example 1, designs II and III unconventionally have replication away from the centre point. If this is seen as a problem, an alternative to D-optimality might be an I- (or V-) optimum design that minimizes the integrated variance of the response over the design region. These designs typically put more weight at the centre of the design region than do D-optimum designs.

In examples 1, 2 and 3 the set of candidate points is the support of the 3^3-factorial. The results of Farrell *et al.* (1968) show that, for a cubical experimental region and second-order model, D-optimum designs are supported on specific subsets S_j of the points of the 3^q-factorial that feature in Table 4. The structure and weights for these D-optimum designs are summarized in section 11.5 of Atkinson *et al.* (2007).

A welcome emphasis in the paper is on the numerical calculation of designs and the assessment of their statistical, rather than geometrical, properties. The authors use an exchange algorithm in which rows of the design matrix are interchanged with rows (design points) in the candidate set. Goos and Jones (2011), section 2.3, argue instead for a co-ordinate exchange algorithm in which each element of the design matrix is the subject of a search. This algorithm does not require a list of candidate points, which may be important for larger problems and finer grids of x-values; points away from S_j may be optimum when n is only a little greater than p.

Although having a large number of degrees of freedom for testing lack of fit is desirable, what is needed is power against alternatives of interest. If the model is augmented with secondary terms, one extreme solution is to consider (locally) T-optimum designs (Atkinson and Fedorov, 1975) that maximize the power of the F-test for departures from the primary model, assuming σ^2 known. Typically, one extra design point is required. The other extreme is formed by D_S-optimum designs for the extra terms, which will add as many extra points as there are secondary parameters. A compromise between these two criteria is the Bayesian procedure of DuMouchel and Jones (1994) which puts a parametrically adjustable amount of information on the secondary terms. In this way the number of extra design points can be controlled. Examples are in chapter 20 of Atkinson *et al.* (2007). Such design criteria could be combined with the authors' procedure for ensuring adequate estimation of σ^2. The designs could either be considered on their own or be combined with designs for parameter estimation.

In the authors' formulation of compound designs there are four κ_j. Of course, there are only effectively three as they can be taken to sum to 1, as indeed they are in the authors' tables. With only one κ (combining two properties) it can be very helpful to look at plots of efficiencies for a series of values of κ. Did the authors find their values of κ by serendipitously searching a grid of values? Might there be good designs for other values of the κ_j?

It is claimed that the discussion at the end of Kiefer (1959) was relatively harmonious. The acrimony that is so evident on the printed page only arose later. I would like to assure the authors that I have enjoyed the opportunity to think about and to respond to their paper. My written comments are, I trust, at least as enthusiastic. It gives me great pleasure to propose the vote of thanks.

Martina Vandebroek (*KU Leuven*)
I would like to start by congratulating the authors on their thought-provoking paper in which they have introduced several innovative design criteria. They encourage people to reflect on the various goals of an experiment and to take these into account when choosing an appropriate design criterion.

The main theme of this paper is that one should use the pure error to estimate the error variance instead of the mean-squared error (MSE). The new design criteria therefore express that the design should assign enough degrees of freedom to estimate the pure error. This then allows for an unbiased estimate of the error variance in case the model is misspecified and it should also reduce the loss in power that one experiences when using the pure error instead of the MSE when the degrees of freedom for pure error are small. That this loss can be substantial is also illustrated in the paper where a central composite design is reanalysed with the pure error and the p-value of the significance test for the second-order terms increases from 0.037 to 0.208.

To investigate how well the new criteria succeed in guaranteeing sufficient power for significance testing and to assess their robustness to model misspecification, I performed several simulations and computed power curves for various situations. Most simulations were done with the designs in Table 1 of the paper. Remember that these 16-run designs were derived for a second-order model in three dimensions. The designs considered are the D_S-optimal design (which is also A_S optimal), the $(DP)_S$- and the $(AP)_S$-optimal designs (which were derived with criteria that are similar to the D_S- and A_S-criteria but use the pure error instead of the MSE), and the compromise design obtained with $\kappa = (0, 0.2, 0, 0.8)$, which gives some weight to the $(AP)_S$-efficiency and most weight to the degrees-of-freedom efficiency. As the first design has no degrees of freedom for pure error, statistical inference is performed with the MSE with 6 degrees of freedom whereas for the other designs the pure error is used with 6, 5 and 3 degrees of freedom. As the designs are not symmetric in the three dimensions, I computed average power curves and generated model coefficients ensuring that all directions were well represented. The power curves then give the average probability of rejecting the null hypothesis of no effect in function of the size of the linear, quadratic and interaction effects. In the case of correct model specification, the average power curves are given in Fig. 1. It is no surprise that the D_S-optimal design with 6 degrees of freedom for MSE outperforms the $(AP)_S$-design with 5 degrees of freedom for pure error and the compromise design with only 3 degrees of freedom for pure error. What is remarkable is that, although the $(DP)_S$-optimal design has the same degrees of freedom for error as has the D_S-optimal design, the $(DP)_S$-optimal design has much less power, which was also reflected in the low A_S-efficiency.

Of course, the real benefits of the new criteria should manifest themselves in the case of model misspecification. I considered several types of misspecification but report here on what I guess is the most realistic scenario, namely missing third-order terms. So data were generated including one or more third-order terms whereas the designs were optimized for a full second-order model. The average power curves in the case that the model is missing the terms $x_1^2x_2$ and $x_1x_2x_3$ are given in Fig. 2. Some unexpected results were obtained: the new designs which produce unbiased estimates of the error variance have indeed more power than the D_S-optimal design but this often occurs at the expense of a seriously inflated type I error. This is caused by the relatively large bias in the factor effect estimates. The D_S-optimal design is quite balanced and therefore its parameter estimates are not highly biased because of the missing third-order terms. For the other designs, however, the factor effects are biased to a larger extent. This leads to rejecting the null hypothesis of no effect much more often than the selected level of significance. The average bias of the linear, quadratic and interaction parameters due to a third-order effect of size 1 is indeed 0.123, 0.286, 0.217 and 0.184 respectively for the D_S-, $(DP)_S$-, $(AP)_S$- and the compromise design and this is apparent in the power curves.

The same effect comes into play when testing for lack of fit. Here the D_S and the (DP_S)-optimal designs are out of competition as the former has no degrees of freedom for pure error and the latter has no degrees of freedom for lack of fit. However, because of the relatively high confounding with higher order terms, the (AP_S)-optimal design and the compromise design do not have much power to detect lack of fit either as a

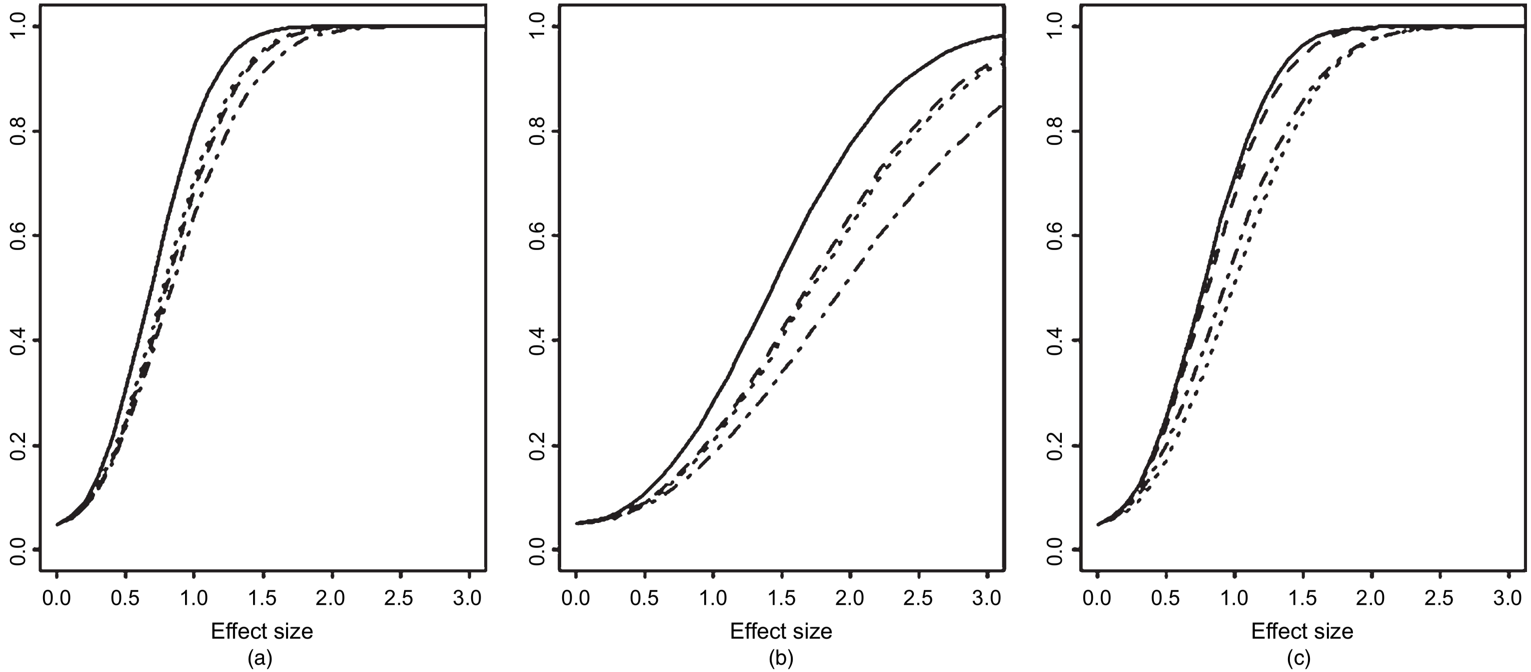

Fig. 1. Average power curves in the case where the model is correctly specified (———, D_S-optimal design; ·······, $(DP)_S$-optimal design; -------, $(AP)_S$-optimal design; · - - - ·, compromise design): (a) linear effects; (b) quadratic effects; (c) interaction effects

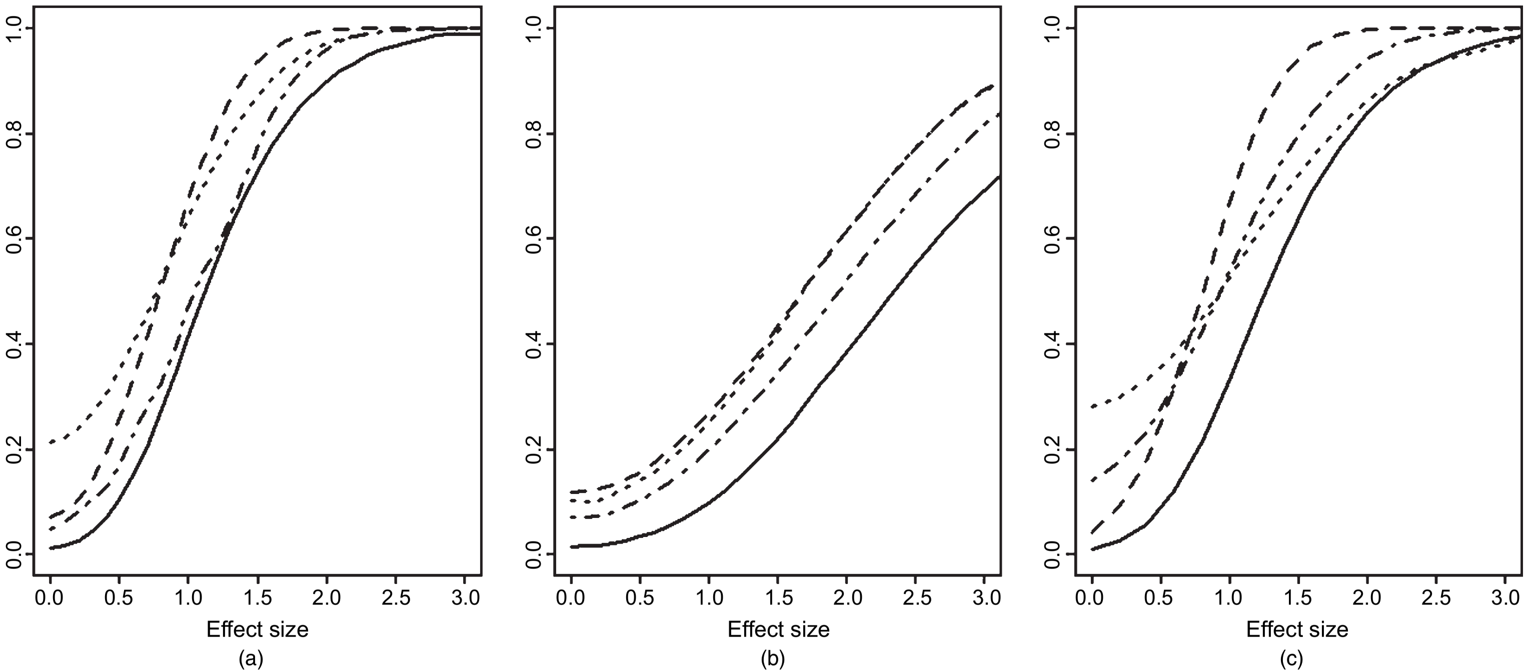

Fig. 2. Average power curves in the case where the model is misspecified (——, D_S-optimal design; · · · · · ·, $(DP)_S$-optimal design; - - - - - - -, $(AP)_S$-optimal design; · - - - ·, compromise design): (a) linear effects; (b) quadratic effects; (c) interaction effects

substantial part of the missing higher order terms is captured by the lower order terms in the model.

It was recognized by the authors that the new criteria impose more imbalance in the design, especially when the number of observations is small, because of the requirement to have replicated observations. As many people have regarded the imbalance of optimal designs compared with classical designs as an important drawback of optimal designs and this problem becomes even more pronounced with the new criteria, it seems useful to strive for at least a minimum level of balance. So, although the authors advise that we abandon the D_S-criterion because it often yields designs with no degrees of freedom for pure error and has no straightforward link with statistical inference when pure error is used, it clearly has the advantage that it imposes more balance in the design and it seems advisable to include it in the compound design criteria anyway.

The new criteria should clearly be used with care in the case of small designs as the imbalance may become too large and the degrees of freedom for pure error may become too small. These disadvantages are less pronounced when dealing with larger designs. With the designs in Table 3 for instance, where 26 runs were used to estimate a second-order model in three dimensions, the D_S-efficiency of the designs obtained with the new criteria is always reasonably large and the degrees of freedom for pure error are always substantial, which avoids to a large extent the strange behaviour that is obtained with the designs in Table 1.

One way to avoid very low degrees of freedom for statistical inference in the case that the design is small is to revert to the common practice of testing first for lack of fit and, if lack of fit is rejected, to pool the pure error and the lack-of-fit error into the MSE. With the new criteria in the paper, a test of lack of fit, however, is not always possible; therefore I propose an extra component that can be included in the design criteria: $F_{n-p-d-1,d,1-\alpha}$, the percentile needed for testing lack of fit. It will guarantee degrees of freedom for lack of fit and for pure error. As there is clearly need for sufficient balance, and as the degrees of freedom for MSE do not depend on the design, a compound criterion that combines the percentile for lack-of-fit testing and the original D_S-criterion looks very promising. Further simulations are needed to find out whether and under which conditions this and other compromise designs outperform the more standard designs.

This paper clearly gave me food for thought as well as inspiration to develop new design criteria and I am convinced that it will inspire others as well. I am therefore pleased to second the vote of thanks.

The vote of thanks was passed by acclamation.

Martijn P. F. Berger (*Maastricht University*)

This very interesting paper has prompted me to comment on three issues: ‘pool or not pool’, ‘new criteria’ and ‘efficiency comparison’.

I must start by saying that factorial designs with hardly any replications are rare in the biomedical and social science research that I have seen. If anyone came up to me with data from such a small experiment, I would ask them to return when the experiment has been replicated many times.

Having said this, I am not convinced about never pooling negligible sources of variation to obtain an estimate of σ^2. The main argument of the authors seems to be that using s_p^2 induces unmeasurable bias. But, if the assumption is reasonable that different sources of variation have equal E(MS), I do not see what is wrong with pooling them to obtain enough error degrees of freedom for a more powerful test of the effects of main interest. It should be kept in mind that the pure error estimator of σ^2 is also a pooled estimator depending on the assumption of constant variances. In my experience, these variances can be relatively heterogeneous in experiments with only a few replications.

The authors have a strong argument for proposing DP- and AP-criteria when σ^2 needs to be estimated. However, I do not understand why they seem to be willing to abandon this strong argument of a better statistical interpretation so easily by introducing compound optimality criteria which, despite their attractiveness, depend on subjective weight choices and in general have no straightforward interpretation.

I liked the efficiency comparison of the designs. Although the authors dismiss the D-optimality design criterion for fractional factorial or response surface designs, it is striking that the DP- and AP-designs remain so highly D_S efficient, and that the only reason why a D_S-optimal design is low on DP- and AP-efficiency seems to be its small error degrees of freedom, which in itself is not a desirable situation and can be helped by adding replications. I wonder whether the authors would still have the same feelings about the D-criterion for designs with a sufficiently large number of runs and error degrees of freedom.

Asymptotically the new criteria reduce to the standard ones as $d \rightarrow \infty$. It would be interesting to know how many replications are needed in practice for the new optimal designs to coincide with the standard designs.

J. P. Morgan (*Virginia Tech, Blacksburg, and Isaac Newton Institute, Cambridge*)
The authors address a genuine tension between lack of fit and variance estimation, but not just that. They build a formal bridge between practitioners who pragmatically insist on replicated data points, and design workers' emphasis on summaries of estimator precision. The not uncommon disjuncture between practice and variance-based optimality is why many consider design to be partly art. In this view A- and other optimalities are simply idealized starting points for pragmatic design. Optimality becomes a design end point only if optimality criteria account for design decisions that would otherwise be left to *ad hoc* judgements. Inculcating pure error degrees of freedom in a formal design assessment does this.

Lack of fit and pure error are among 14 design considerations enumerated by Box and Draper (1975). Traditional variance-based criteria are legitimately criticized for giving no weight to many of these issues, including the need for a heterogeneity check.

Much optimal design work assumes that both first-order (through a posited linear or non-linear model) and second-order (through constant variance by stratum) moment structures are known. Whereas lack-of-fit testing acknowledges possible first-order deficiencies, standard optimal designs avoid replication, sacrificing capacity to address the second-order assumption. The authors' work gains leeway for homogeneity assessment, begging the question of whether this, also, can be rigorously incorporated in design selection criteria. If so it will entail balancing detection power against magnitude of departure relative to seriousness of its consequences.

Eibl *et al.* (1992) provide an interesting example. They explored effects of six factors on paint coat thickness by using four replicates of a 2^{6-3}-fraction. Checking for heterogeneity, one finds that the largest pairwise ratio of the eight pure variances exceeds 18, and that the p-value for Levene's test is 0.014. Given this, one doubts the validity of any single error estimate. Nonetheless, proceeding with s_p^2, the authors arrived at what proved in practice to be useful conclusions.

We should be asking, for both first- and second-order moments, what magnitude of model departure warrants concern. Much is asked of small designs of the type considered in this paper. They have been successful in practice, not because the models are strictly correct, but because the magnitudes of important effects are often sufficiently large to overpower model deficiencies. We nonetheless need designs that can alert us to consequential model defects, with selection of such designs having rigorous, rather than *ad hoc*, justification. Thanks go to the authors for results that push in this direction.

R. A. Bailey (*Queen Mary, University of London*)
I thank the authors for reminding us that inference from designed experiments is about choice of models as well as about estimating parameters. In their general set-up there are two models, M_1 and M_2, with M_2 contained in M_1. If there are n data, then there are $n - \dim(M_1)$ degrees of freedom for pure error, and $\dim(M_1) - \dim(M_2)$ degrees of freedom for lack of fit. This approach respects marginality (Nelder, 1977, 1994), and avoids the problems caused by routinely pooling small mean squares (Janky, 2000).

However, I do not think that they go far enough. Designed experiments typically have far more potential linear models for the expectation of responses. For example, Fig. 3 shows the collection of models, and their relationships, in the comparatively simple case of a factorial experiment with two three-level quantitative factors.

How do we analyse data? We start with a family of models and their marginality relations, like those in Fig. 3. We use F-tests to choose the smallest model which explains the data adequately. Then we estimate the parameters of the chosen model. We need an optimality criterion which takes account of the whole process.

I have two small comments about the designs for experiments conducted in blocks (Section 3.3). It is not clear whether the authors' algorithm includes blocks from the beginning or whether it first finds the best treatment design and then blocks it. The latter method can produce designs which are far from optimal.

If blocks are random, the authors recommend that we should construct the design as if they are fixed. An alternative is to find a design which is good throughout a plausible range of the ratio of stratum variances (Bailey, 1999).

Timothy Waite and David Woods (*University of Southampton*)
We thank the authors for their paper, which has made us think more deeply about optimality criteria and their justification.

Theoretical justifications
Traditional D-optimality and the authors' new DP-optimality both have statistical justifications *when the submodel (2) is true*, as the confidence regions on which they are based are both valid for $\boldsymbol{\beta}$ in this case.

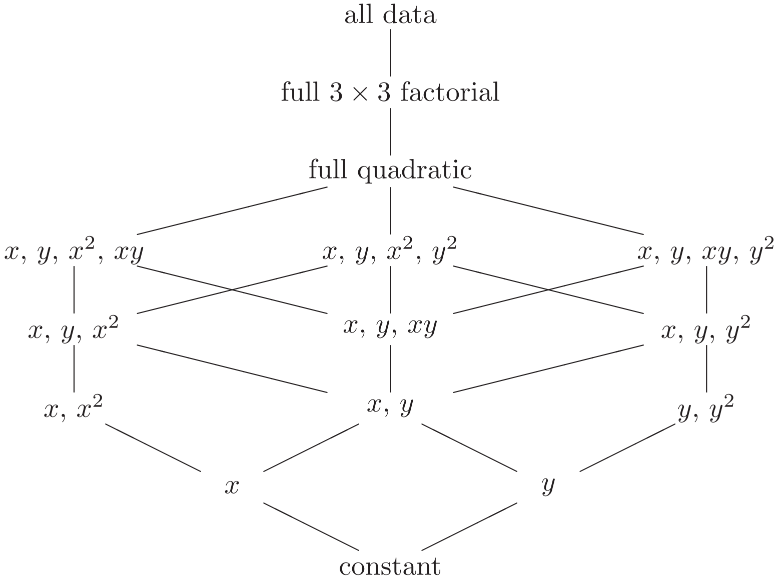

Fig. 3. Collection of expectation models for a factorial experiment with two three-level quantitative factors

Moreover, the confidence region from using the pooled error estimate will typically be smaller than that obtained by using the pure error estimate, and so inference will be more efficient in the former case.

In practice, we can never be completely confident that the submodel (2) is correct. However, suppose that we can make the weaker assumption that the full treatment model (1) holds. Then

$$R_{\text{p.e.}} = \{\boldsymbol{\beta} : (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \leqslant ps^2 F_{p,d;1-\alpha}\},$$

whose volume is minimized by a DP-optimal design, is in fact a valid $100(1-\alpha)\%$ confidence region for the vector $\boldsymbol{\beta}_s$ of *pseudotrue values* of the submodel parameters. These are the values of $\boldsymbol{\beta}$ in submodel (2) which are 'least bad' (Davison (2003), pages 147–148) and are given by

$$\boldsymbol{\beta}_s = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\, E(\mathbf{Y}),$$

where the expectation is taken with respect to the true (full treatment) model. This robustness of the inference based on pure error gives us a more rigorous reason to adopt DP- and similar criteria.

Compromise designs
If we believe that the full treatment model (1) holds and are interested in whether the submodel (2) is an appropriate approximation, a compromise design incorporating model uncertainty might be a more realistic comparator for the $(\mathrm{DP})_{\mathrm{S}}$-design rather than the naive D_{S}-optimal design. Let us assign probability γ that model (2) is true, assuming that model (1) is true otherwise. Then we can find the compromise design ξ_{C} which maximizes the *compound objective function*

$$\Phi(\xi) = \gamma \log\{|(\mathbf{M}^{-1})_{22}|^{-1/p_2}\} + (1-\gamma)\log(|\mathbf{M}_\mu|^{1/t})$$

where $\mathbf{M}_\mu$ is the information matrix for model (1), and

$$|\mathbf{M}_\mu|^{1/t} = \prod_{i=1}^{t} n_i^{1/t}.$$

See Läuter (1974) and Atkinson and Cox (1974) for related criteria.

The optimal design for $\gamma = 0$ is the balanced design for the full treatment model, and for $\gamma = 1$ we obtain the D_{S}-optimal design. Clearly for $0 < \gamma < 1$, the second term in $\Phi(\xi)$ will act as a penalty which encourages replication.

We found designs for example 1. For $0 < \gamma \leqslant 0.67$, $\xi_{\mathrm{C}} = \xi_{(\mathrm{DP})_{\mathrm{S}}}$, and, for $\gamma = 0.7$, $\xi_{\mathrm{C}} = \xi_{(\mathrm{AP})_{\mathrm{S}}}$. For $\gamma = 0.8, 0.9$, ξ_{C} is highly D_{S} efficient (99%) but with df(PE) = 2. Thus explicit acknowledgement of model uncertainty in the framework of traditional criteria can result in designs similar to those proposed by the authors and allows a compromise between experiment objectives.

Mervyn Stone (*Ruislip*)
This paper foresees a healthy future for its compound design criteria, thereby defying the now historical predictions that the Third Millennium would be Bayesian *for everything*. The authors refer to *run-to-run variation* but would, it seems, treat each run as a one-off experiment to be addressed by 'classical' inference tools. They do not consider the (strictly un-Bayesian) empirical Bayes approach, which would use those tools to estimate a 'prior' Π for the full parameterization (including $\boldsymbol{\beta}$) of any sequence of runs, and deliver a multipurpose 'posterior' for $\boldsymbol{\beta}$ from the results of the chosen design for each run. A unique design criterion can be defined as the gain ΔI in expected Shannon information I about $\boldsymbol{\beta}$. ΔI is invariant under 1–1 transformation of $\boldsymbol{\beta}$ and may appeal to any classical practitioner who cannot decide on the weights $\kappa_1, \kappa_2, \kappa_3$ and κ_4.

Consider the simple empirical Bayes case of an error variance σ^2 assumed constant (either known or reliably estimated from 'pure' replication in a large number of runs) rather than varying from run to run (and taken into Π). Suppose that $\mathbf{Y}$ is $N(\mathbf{X}\boldsymbol{\beta}, \mathbf{I}\sigma^2)$ and that $\boldsymbol{\beta}$ is $N(\mathbf{a}, \mathbf{A}\sigma^2)$. The prior value of I is $\int \log\{\pi(\boldsymbol{\beta})\} \mathrm{d}\Pi(\boldsymbol{\beta})$ and the expected posterior value is $\int \log\{\pi(\boldsymbol{\beta}|\mathbf{Y})/\pi(\boldsymbol{\beta})\} \, \mathrm{d}\Pi(\boldsymbol{\beta}|\mathbf{Y}) \, \mathrm{d}P(\mathbf{Y})$ where π is density and P is the marginal distribution of $\mathbf{Y}$ under $\boldsymbol{\Pi}$. By Stone (1959), the gain in I is $\Delta I = \frac{1}{2} \log |\mathbf{I} + \mathbf{A}\mathbf{X}^{\mathrm{T}}\mathbf{X}| = \frac{1}{2} \log(S)$, say. The determinant S shows the influence of the estimated 'prior' and also reminds us that 'ignorance priors' have no place in experimental design, since such priors would here dictate an infinity in $\mathbf{A}$ that could dominate the choice of $\mathbf{X}$.

Example
Suppose that $E(\mathbf{Y}|\boldsymbol{\beta}) = \alpha + \beta x + \gamma x^2$ and $\mathbf{A} = \mathrm{diag}(u, v, w)$. With $n = 4$, design A is the three-point 1–2–1 design on $x = 0, 1, 2$. The superficially riskier design B, 2–0–2, is willing to leave the centre point unobserved. My algebra gives the respective S-values as $1 + 4u + 6v + 18w + 8uv + 36uv + 8vw + 8uvw$ and $1 + 4u + 8v + 32w + 64uw + 16uv$. The difference in favour of design A is $2w(4uv + 4v - 14u - 7) - 2v - 8uv$, which is *negative* for w small. Sufficiently small w/v may be seen as knowledge that γ is sufficiently close to 0 for curvature to be ignored relative to linearity; if $u = v = 10$, this happens when $w/v < 82/386 \approx 1/5$. However, if $4v < (14u + 7)/(u + 1)$, there is *no* value of w that would favour 1–2–1 over 2–0–2—which is food for thought (for significance testers) when w/v is large.

Frank Critchley (*The Open University, Milton Keynes*)
It is a pleasure to welcome this paper, with its emphasis on optimal design in practice, in particular, its argument for, where possible, separating lack of fit from true error. Overall, the paper illustrates strongly the truism that the way a problem is formulated affects—focuses, limits, (potentially) prejudges, . . . —the form or outcomes of its subsequent analysis. This is certainly so for the compound criteria of the paper and raises the following questions.

(a) What are the effects—intended, or otherwise—of different possible choices of such criteria?
(b) In particular, what are the effects of the multiplicative form used?
(c) In this case, positive power transforms mapping (0, 1) to (0, 1), is there—if not a full implied utility function—at least an implied set of non-linear scales on which trade-offs between (transformed) efficiencies can be made? If so, are these interpretable, in some way?
(d) Collecting the exponents into a vector $\boldsymbol{\kappa}$, what more can be said—either theoretically or practically—about how a $\boldsymbol{\kappa}$-optimal design varies with $\boldsymbol{\kappa}$? Where are the (discrete) discontinuities? Is some form of sensitivity analysis possible?

Finally, I have two questions about possible extensions.

(a) Throughout, the paper assumes a common variance across observations. In common with earlier discussants, I wonder how far heterogeneity could be accommodated.
(b) The methodology presented is invariant to overall scale changes in $\mathbf{X}$. If or where appropriate, is (some form of) affine invariance achievable?

It will be clear that I warmly welcome both tonight's paper and the wide discussion and further developments that it will undoubtedly bring. I thank you for it.

Martin Ridout (*University of Kent, Canterbury*)
I have enjoyed reading this interesting and carefully argued paper. One issue that the authors do not discuss is robustness of their designs to missing values. For the new design criteria that are considered here, a missing value may change the pure error degrees of freedom as well as modifying the X-matrix.

In small designs, even a single missing value can cause the design to break down in the sense that the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ becomes singular. In Table 1, for example, design IV breaks down if any of the first four (corner) points is missing and design II breaks down if any of the design points 5, 8, 15 or 16 is lost.

Even in larger designs, loss of a single point can have a substantial effect. For example, in Table 5, design IV is $(\mathrm{AP})_{\mathrm{S}}$ optimal but, if design point 10 is lost, the $(\mathrm{AP})_{\mathrm{S}}$-efficiency drops to 0.793, below the corresponding minimal $(\mathrm{AP})_{\mathrm{S}}$-efficiency of 0.852 for design V (Table 8), which arises when point 37 is missing. A modification of design IV that replaces the original design points 34 and 36 by replicates of points 1 and 10 is more robust. The $(\mathrm{AP})_{\mathrm{S}}$-efficiency of the full design is 0.992 and the $(\mathrm{DP})_{\mathrm{S}}$-efficiency is 0.954, which is slightly better than design IV. The lowest $(\mathrm{AP})_{\mathrm{S}}$-efficiency following loss of a single point is 0.865.

So I wonder whether it would be useful to incorporate some measure of efficiency in the presence of missing values into composite design criteria, for situations where missing values might occur and where it is difficult to repeat missing experimental runs. In several of the designs in this paper, the loss of efficiency from a single missing value is much greater for a small subset of points than for the majority of points and designing to minimize the maximum loss (e.g. Ahmad and Gilmour (2010)), although natural, might be unduly cautious. From a practical point of view, it may be worth alerting the experimenter to any critical design points, in case it is possible to take extra precautions to avoid losing these runs.

Ben Torsney (*University of Glasgow*)

The authors are to be commended for raising the profile of experimental design and to do so by extending, not regressing, the subject. They do themselves an injustice to suggest 'that the *D*-criterion has no place in the design of ... experiments'. This is for two main reasons.

(a) On the positive side they have extended standard criteria to df(PE;LoF)-dependent criteria which are on a par with parameter-dependent non-linear local design criteria.

(b) On the negative side, the new criteria rely heavily on the assumption of normality, whereas the standard criteria can be viewed as focusing on the properties of ordinary least squares estimation, although they have interpretations under normality. There is no avoiding this. If there is no normality, there are no degrees of freedom. If it is argued that we can be relaxed about the issue, then surely this also applies to standard criteria. Conversely if normality is seen to be crucial, but is in doubt, then it may be safer to resort to standard criteria.

Moreover there may be further extended roles for these criteria; for example, optimize one subject to a 'start-up' design which ensures a given df(PE;LoF) structure. This compares with a design which is optimal for a second stage, given a first-stage design; see Covey-Crump and Silvey (1970). Designing subject to constraints is not new.

The authors' designs can be viewed as made up of two component designs: one for pure error or lack of fit and one for parameter inference. One might opt for a *D*-optimal design for the latter, subject to a given design for the former, or maybe vice versa. For example, in the $(\mathrm{AP})_{\mathrm{S}}$-design III of Table 1, do the six singly replicated design points define a constrained optimal design, given the five doubly replicated designs? Possibly the two component designs can be chosen jointly optimal for a given number of runs for each—another compound criterion.

Finally it is the variance term which has given rise to the authors' new criteria. What extensions might there be for mixed effects models with several variance terms? Of course in generalized linear models, there is no variance term. So standard design criteria can reign supreme for the time being. I hope that this is of as much comfort to the authors as it is to me! I thank them for their stimulating paper.

The following contributions were received in writing after the meeting.

Christine M. Anderson-Cook and Lu Lu (*Los Alamos National Laboratory*)

We commend the authors on an excellent paper and for challenging the commonly accepted assertion that *D*-optimality is the appropriate metric for constructing optimal designs. We agree that this single-minded strategy often leads to undesirable solutions, unless augmented (often arbitrarily) to compensate for other properties. Devoting design resources to obtaining a model-independent estimate of error variance is an important priority which has not been consistently articulated in the literature.

The new metrics proposed arise naturally from common design analyses and seek to balance efficient estimation of model parameters, unbiased estimation of error variance and the ability to assess lack of fit. They help to broaden the focus away from exclusively *D*-optimality in a helpful way. Lu and Anderson-Cook (2012) propose alternative metrics for considering the same three objectives in design construction. Our

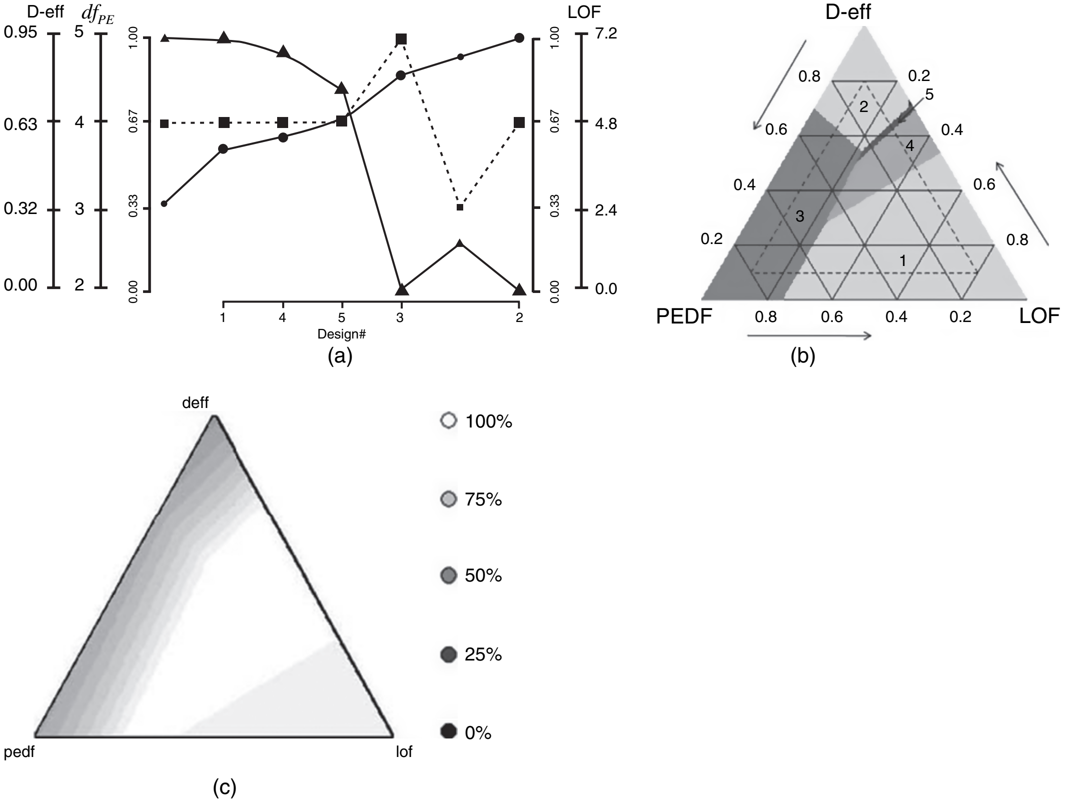

Fig. 4. Set of graphical tools for facilitating decision making by showing the trade-off between criteria values for promising designs, robustness of best designs to different weight combinations of the three criteria and relative efficiency of an individual design's desirability compared with the best possible for different weightings: (a) trade-off plot (●, *D*-efficiency; ■, pure error degrees of freedom; ▲, lack of fit); (b) mixture plot; (c) synthesized efficiency plot

metrics are simpler than those proposed by the authors but also seek to capture these three key elements of a good design, which we contend represent a common set of priorities for many screening design situations.

The compound criteria designs proposed by Gilmour and Trinca nicely moderate the extremes found by focusing on a single objective. We understand that any combination of weights could have been used for the optimization, but we consider one or two particular choices to oversimplify what is needed. A key aspect of design construction and selection is how to combine the metrics! We favour keeping the metrics separate and using a Pareto front optimization approach (Lu *et al.*, 2011) to assemble a suite of candidate designs which are best for different weightings of the metrics. In our experience, specification of relative weightings of different aspects of goodness is usually quite subjective and imprecise. Although computationally somewhat more demanding than optimizing a single objective, the Pareto approach allows for this user uncertainty to be flexibly explored. Decision making when balancing different criteria feels most natural when a set of promising designs is assembled for comparison and then their relative merits are evaluated numerically and graphically. Fig. 4 shows several graphics developed in Lu and Anderson-Cook (2012) and Lu *et al.* (2011) to help experimenters to examine trade-offs between candidates (Fig. 4(a)), to identify best designs for different weight combinations (Fig. 4(b)) and to evaluate the synthesized efficiency of any potential final design relative to the best among all competitors (Fig. 4(c)).

We wholeheartedly endorse considering multiple aspects of design goodness but feel that there are several good options for how to proceed from this starting point.

Alexis Boukouvalas and Dan Cornford (*Aston University, Birmingham*)

We thank the authors for a thought-provoking paper. Our work has been focusing on optimal design for estimation of covariance parameters of Gaussian processes (Boukouvalas, 2011) and is based on the work of Zhu and Stein (2005).

In the correlated error case, there is no general result under infill asymptotics to prove the optimality of using the log-determinant of the Fisher information matrix as a design criterion. In our work we have empirically observed that, under complex noise models, the approximation error of the Fisher information matrix to the log-determinant of the covariance of maximum likelihood parameter estimates increases as the ratio of single to multiple replicated points is increased.

Do the authors believe that their paper can be extended to this domain and to help to explain the increase in the approximation error? Do they believe that this is due to biased estimation of the covariance parameters?

Chris Brien (*University of South Australia, Adelaide*)
I applaud the authors for developing the new optimality criteria. They are essential when using software packages to search for designs and the proposed compound criteria are particularly appealing for their ability to deal with multiple objectives.

However, I believe that there are two distinct issues underlying Section 2:

(a) the need for a pure error estimate and
(b) the desirability of using a pooled *versus* a pure estimate of the error variance.

In terms of (a), a good estimate of pure error is required for testing lack of fit, at least, and the new 'P'-criteria help to ensure that it is available. As for (b), an objective approach exists in the form of deciding between 'never-pool', 'sometimes-pool' and 'always-pool' procedures for estimating the error variance. Janky (2000) has provided a technical review of the area and Hines (1996) a review of the practice. The always-pool procedure is generally not considered viable.

Relevant to the paper is the pooling of fixed effects with a pure error estimate, which both Mead *et al.* (1975) and Hines (1996) addressed. Rather than the never-pool option that the authors settle on, the recommendations of Mead *et al.* (1975) appear, on balance, to have worthwhile benefits. Their investigation of the size and power of the competing procedures suggests that, in the context of the paper,

(i) the sometimes-pool procedure should only be used if the lack-of-fit degrees of freedom are about equal to or considerably larger than the pure error degrees of freedom and
(ii) the never-pool procedure is recommended if the lack-of-fit degrees of freedom are smaller than the pure error degrees of freedom, and the latter are reasonably large.

The authors generally commend designs with approximately equal pure error and lack-of-fit degrees of freedom. On the basis of Mead *et al.* (1975), the sometimes-pool procedure, with a preliminary test for lack of fit that uses $\alpha = 0.50$, can be expected to provide useful gains in power for such designs. Design VI for example 4 and design VI for example 6 fit into this category. Similarly, the sometimes-pool procedure would be used for exercise 11.6 of Box and Draper (2007) discussed in Section 2; because the p-value for the lack-of-fit test is 0.73 and so greater than 0.50, the use of s_p^2 is indicated. However, the pooled estimate would never be used with design IV for example 5 because of the very small lack-of-fit degrees of freedom.

Elvan Ceyhan (*Koç University, Istanbul*)
I congratulate the authors for this interesting paper on optimal experimental design, especially, for developing compound criteria which incorporate multiple objectives.

Kiefer (1975) introduced the concept of universal optimality and defined the conditions for a design to be universally optimal. Kiefer's optimality conditions were generalized by Yeh (1986) for binary block designs. Kiefer's optimality criteria include many criteria such as D- and A-optimality among others. One wonders whether the criteria introduced in the paper (such as DP and AP) are Kiefer universally optimal. The authors write 'quite different designs can be optimal under the new criteria'. This is expected, because all traditional criteria agree on several design classes according to Kiefer's optimality, which suggests that optimality is more robust to the changes in optimality criteria, compared with changes in the model.

The authors use Bonferroni adjustment, to correct for multiple testing; other correction methods (e.g. Šidák's correction) could also be employed, as the Bonferroni method is known to cause a reduction in power. Let β be the level of significance for each of n tests; for experimentwise error rate α, the Bonferroni method suggests $\beta_B = \alpha/n$, whereas Šidák's method suggests $\beta_S = 1 - (1-\alpha)^{1/n}$. However, Šidák's method requires independence of tests, and the difference seems negligible (for example, with $n = 9$ as in example 1 and $\alpha = 0.05$, $\beta_B \approx 0.0056$ and $\beta_S \approx 0.0057$). In this context a single-step correction is more sensible, and a step-down approach like Holm's (1979) correction may not be immediately applicable.

To form a compound criterion, perhaps the *desirability function* of Harington (1965) can also be employed (with some modifications) by transforming the functions to a common scale in [0,1]. Then the transformed functions are combined and optimized as the overall metric. Desirability functions are popular in response surface methodology and have previously been applied in experimental design (see, for example, Rafati and Mirzajani (2011) and Azharul Islam *et al.* (2009)). For example, the efficiency functions in Section 4.1, E_i, $i=1,2,3,4$, can be transformed to functions d_i whose range is [0,1] (and the value of d_i increases as the desirability of the function increases). For example, if a response is to be maximized, the desirability function

$$d_i(E_i)=\begin{cases} 0 & \text{if } d_i(E_i)<A_i, \\ \left\{\dfrac{d_i(E_i)-A_i}{B_i-A_i}\right\}^{w_i} & \text{if } A_i \leqslant d_i(E_i)<B_i, \\ 1 & \text{if } d_i(E_i)\geqslant B_i, \end{cases} \tag{7}$$

can be used, where A_i, B_i and w_i are chosen by the researcher. So, the goal is to optimize $D=(d_1d_2d_3d_4)^{1/\Sigma w_i}$, which is very similar to equation (5) in the text. For desirability functions where the response is 'minimized' or the 'target is best' is needed, see, for example, Derringer and Suich (1980).

Marion J. Chatfield (*GlaxoSmithKline, Stevenage*)
In the pharmaceutical industry experimental design is used to develop, understand and validate processes. Optimal design is becoming increasingly important, as it allows the designer to cater for aspects such as constrained regions, specific models and flexibility in the number of runs.

I welcome this paper as another step towards providing optimal designs which meet the practical needs of industry because

(a) it is usually desirable to have a pure error estimate (excepting screening of many factors) and
(b) better composite criteria could produce appropriate optimal designs with reduced designer time and knowledge.

Industrial application usually seeks a pure error estimate, albeit often with poor precision. Although Bayesian analysis is rarely applied (and beyond most scientists), informally an experimenter compares their prior expectations with estimated fixed effects, model lack of fit and the variability observed. The pure error estimate is used to assess lack of fit against, and to signal when variation is higher than expected (indicating a problem with the experimentation). Classical designs which include repeat points (unlike D-optimal designs) are popular, though repeats are typically centre points. The ability to spread repeats through the design region, as observed by using the DP(α) criterion, would be desirable, providing more protection against missing or rogue observations and an 'average' level of the background variation (which is useful in the case of heterogeneity).

'Efficiency', in industry, includes the time to produce the design, to plan and implement the experimental work, to analyse and interpret the results, the quality and use of the information gained. Classical designs are often used, especially by trained scientists, as they seem 'safe' designs, incorporating many desirable properties (Box and Draper, 1975) and are relatively quick to design and analyse (detailed evaluation not required and easily analysed in commercial software). Appropriate composite criteria making optimal design more accessible and encompassing are desirable.

I have the following specific comments.

(a) Often in small designs lack of fit is combined with pure error to improve power, given prior expectation that variation is small and inference is just one objective. The consequent unknown bias in the 'error' estimate is perceived a reasonable risk to reduce resource. The authors' strong recommendation to base inference on pure error, and not taking the risk of bias in exercise 11.6 of Box and Draper (2007), seems practically too severe.
(b) Compound criteria including lack-of-fit estimating some higher order parameters (e.g. DuMouchel and Jones (1994) and Jones and Nachtsheim (2011)) would be useful.

D. R. Cox (*Nuffield College, Oxford*)
A common criticism of the theory of optimal experimental design is that the considerations involved in designing an investigation are usually too varied to be captured successfully in a simple criterion. The authors are to be congratulated on their careful and enlightening study of incorporating one more aspect, error estimation, into the formal theory.

Estimates of variance based on small numbers of degrees of freedom are fragile things, capable of producing spuriously high or low apparent precision and their use quite sensitive to the often, but not always, unimportant normality assumption. How often is it that there is no external information about error variance? In traditional agricultural field trials, where typically more degrees of freedom are available for error estimation, I believe the practice was to make informal comparison with the coefficient of variation of yield anticipated from experience. In the present context estimates of variance, individually imprecise, might become available from a chain of related studies; what would the authors recommend in such cases? One possibility would be a partially (empirical) Bayes approach in which estimation of the current variance would be from its posterior distribution, whereas the primary parameters would, unless there were good reason otherwise, be regarded as unknown constants.

Do the authors have any comments on experiments with a Poisson- or binary-distributed response, the role here of error estimation being the assessment of overdispersion?

David Draper (*University of California, Santa Cruz*)
The subject of this interesting paper—optimal experimental design for industrial research—would seem to be a good task for Bayesian decision theory, for which there are two approaches, involving

(a) non-adaptive and
(b) adaptive designs.

In the cassava bread example, for instance, in case (a), consider initially a single organoleptic characteristic such as 'pleasing taste', and let $T > 0$ be the mean value of this characteristic for wheat-based white bread in the population of potential customers for the new cassava bread product. Each choice of the control variables $X = (X_1, X_2, X_3)$ has associated with it a population mean θ_X on the pleasing taste scale, which is estimated in an unbiased manner by each design point with control variable values equal to X; let θ collect all the θ_X-values (there are $3^3 = 27$ of them in the authors' formulation). The action space $\mathcal{A}$ consists of vectors $(a_1, a_2, \ldots)$ of non-negative integers $n_{a_i} = (n_1, \ldots, n_{27})_{a_i}$ keeping track of the numbers of runs made under action a_i at each of the 27 control variable settings. Any sensible utility function $U(a, \theta)$ for this problem would have two components:

(i) a problem relevant measure of the distance $D_\theta \geqslant 0$ between θ and T, and
(ii) a measure of the total cost $C_a > 0$ of all the runs made under strategy a.

These would need to be combined into a single real-valued utility measure by trading off accuracy against cost, e.g. through $U(a, \theta) = -\{\lambda D_\theta + (1-\lambda) C_a\}$ for a $\lambda \in [0, 1]$ that is appropriate to the problem context; the optimal design then maximizes expected utility, where the expectation is over uncertainty quantified by the posterior distribution for θ. In the adaptive case (b), the Bayesian decision theoretic approach might go like this: pick an X, make a run with the control variables set to that X, update the current posterior for θ with the new data on θ_X, thereby generating a new posterior for D_θ, choose a new X where the uncertainty about θ_X is largest and continue in this manner till you exhaust your experimentation budget. It would be interesting to compare this approach with that of the authors, which appears to depend to a disturbing extent on

> 'whether the experimenters' objectives will be met mainly through the interpretation of point estimates or through confidence intervals or hypothesis tests',

none of which would seem to be relevant when the problem is considered decision theoretically.

Wenceslao González Manteinga (*Universidade de Santiago del Compostela*) **and Emilio Porcu** (*University of Castilla la Mancha, Ciudad Real*)
We congratulate the authors for this beautiful paper, where some new optimality criteria for optimal design are proposed and where the necessity of compound criteria to take into account multiple objectives rather than inferential purposes only is emphasized.

(a) The optimality methods enucleated by the authors correspond to different philosophies of weighting the positive definite matrix $\mathbf{X}'\mathbf{X}$ or its inverse, as well as to different metrics (determinant, trace and max). We wonder how this comparison can be done for processes being correlated over time or space, and where the index j in equations (1) and (2) would represent the index set of the observations.
(b) Another interesting point of discussion would be to consider more sophisticated models (see, for example, Dette and Wong (1999)) where the variance is a function of the mean. This assumption is common in applied works (see, for example, Jobson and Fuller (1980)) where some functional relationships between the variance of the process and the parameter vector have been assumed.

Table 10. *I*-optimal 16-run design for a full quadratic model in three factors (example 1)

X_1	X_2	X_3
−1	−1	−1
−1	−1	1
−1	0	0
−1	1	−1
−1	1	1
0	−1	0
0	0	−1
0	0	0
0	0	0
0	0	1
0	1	0
1	−1	−1
1	−1	1
1	0	0
1	1	−1
1	1	1

(c) There are probably some issues related to computational burdens that are associated with some efficiency criteria, since some of them imply computation of the inverse of some matrix whose dimension can be large. Is there any trade-off between statistical efficiency and computational complexity for this kind of models?

(d) It would be interesting to study optimality criteria that can allow us to take into account hypothesis testing on the structure of the mean μ_i in equation (2).

(e) Finally, an extension of potential interest may be that of a completely or partially unknown function $\mathbf{f}$, corresponding respectively to non-parametric or semiparametric modelling.

Peter Goos (*Universiteit Antwerpen and Erasmus Universiteit Rotterdam*)
I compliment the authors on presenting an innovative view on optimal experimental design. However, given the focus on response surface designs, I would have expected them to pay more attention to the prediction-oriented I-optimality criterion. The I-optimality criterion seeks designs that minimize the average prediction variance, which, for a completely randomized design, is proportional to

$$\text{average variance} = \frac{\int_{\chi} \mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}}{\int_{\chi} \mathrm{d}\mathbf{x}},$$

with χ representing the experimental region. If a pure error estimate is used and the interest is in prediction intervals, this I-optimality criterion can be easily modified to the IP(α) criterion

$$\mathrm{IP}(\alpha) = \frac{\int_{\chi} F_{1,d;1-\alpha}\,\mathbf{f}'(\mathbf{x})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{f}(\mathbf{x})\,\mathrm{d}\mathbf{x}}{\int_{\chi} \mathrm{d}\mathbf{x}},$$

in the spirit of the paper. This quantity can be computed as

$$\mathrm{IP}(\alpha) = \frac{F_{1,d;1-\alpha}}{\int_{\chi} \mathrm{d}\mathbf{x}} \mathrm{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}\} = \frac{F_{1,d;1-\alpha}}{\int_{\mathcal{X}} \mathrm{d}\mathbf{x}} \mathrm{tr}\{\mathbf{M}(\mathbf{X}'\mathbf{X})^{-1}\},$$

Table 11. I-optimal 26-run design for a full quadratic model in three factors (example 3)

X_1	X_2	X_3
−1	−1	−1
−1	−1	0
−1	−1	1
−1	0	−1
−1	0	1
−1	1	−1
−1	1	0
−1	1	1
0	−1	−1
0	−1	1
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	0	0
0	1	−1
0	1	1
1	−1	−1
1	−1	0
1	−1	1
1	0	−1
1	0	1
1	1	−1
1	1	0
1	1	1

where $\mathbf{M}$ is the moments matrix $\int_{\mathbf{x}\in\chi}\mathbf{f}(\mathbf{x})\,\mathbf{f}'(\mathbf{x})\,\mathrm{d}\mathbf{x}$. This final expression for the IP(α) criterion has exactly the same form as the LP-criterion in the paper.

Interestingly, the I-optimality criterion itself generally produces attractive designs for completely randomized experiments, in that they usually strike a reasonable balance between degrees of freedom for pure error estimation and lack of fit. The I-optimal design for example 1 in the paper, shown in Table 10, has 1 degree of freedom for pure error (due to two centre point replicates) and 5 degrees of freedom for lack of fit. The design is therefore more useful than the (DP)$_{\mathrm{S}}$-optimal design for that example, or the (AP)$_{\mathrm{S}}$-optimal design corrected for multiple comparisons, because these designs leave no degrees of freedom for lack of fit. The I-optimal design for example 3 in the paper, shown in Table 11, has 5 degrees of freedom for pure error (due to six centre point replicates) and 11 for lack of fit. The I-optimal design for example 5 in the paper, shown in Table 12, has 4 degrees of freedom for pure error (due to five replicates of the point $(-1,0,0,0,0)$) and 16 for lack of fit. So, also for examples 3 and 5, the I-optimal design provides a better balance between degrees of freedom for pure error and for lack of fit than the (DP)$_{\mathrm{S}}$-criterion. In conclusion, the I-optimality criterion yields completely randomized designs that allow for pure error estimation and testing for lack of fit. It would be interesting to verify whether the IP(α) criterion results in better designs than the classical I-optimality criterion, and to embed that criterion in compound criteria.

It would be interesting to extend the present work to multistratum response surface experiments, where more than one variance component needs to be estimated, and where, in general, the non-orthogonality of the designs makes it less straightforward to determine degrees of freedom for the global F-test, confidence intervals and prediction intervals, and even for tests on individual model parameters.

Linda M. Haines (*University of Cape Town*)

I congratulate the authors on a stimulating paper which encourages traditional optimal designers to move out of their equivalence theorem mind set and to explore meaningful criteria through computation.

Table 12. *I*-optimal 40-run design for five factors, one at two levels and four at three levels, using a full quadratic model (example 5)

X_1	X_2	X_3	X_4	X_5
−1	−1	−1	−1	0
−1	−1	−1	1	−1
−1	−1	−1	1	1
−1	−1	0	0	1
−1	−1	1	−1	−1
−1	−1	1	−1	1
−1	−1	1	1	−1
−1	0	−1	−1	−1
−1	0	0	0	0
−1	0	0	0	0
−1	0	0	0	0
−1	0	0	0	0
−1	0	0	0	0
−1	0	1	1	1
−1	1	−1	−1	1
−1	1	−1	1	−1
−1	1	−1	1	1
−1	1	0	0	−1
−1	1	1	−1	−1
−1	1	1	−1	1
−1	1	1	1	0
1	−1	−1	−1	1
1	−1	−1	0	−1
1	−1	0	−1	−1
1	−1	0	1	0
1	−1	1	0	0
1	−1	1	1	1
1	0	−1	0	1
1	0	−1	1	0
1	0	0	−1	1
1	0	0	0	0
1	0	0	1	−1
1	0	1	−1	0
1	0	1	0	−1
1	1	−1	−1	−1
1	1	−1	0	0
1	1	0	−1	0
1	1	0	1	1
1	1	1	0	1
1	1	1	1	−1

I have one particular comment which relates to $(\mathrm{DP})_\mathrm{S}$-optimality. Specifically, the authors observe that the $(\mathrm{DP})_\mathrm{S}$-criterion on its own is 'quite extreme' and I therefore wondered whether other simply formulated single criteria would produce more pleasing designs. To this end I explored the idea of maximizing two such criteria, $\ln|X^\mathrm{T} Q_0 X| + \ln(d)$, which combines information on the parameter estimates with information on the estimate of pure error in a natural way, and an attenuated version, $\ln|X^\mathrm{T} Q_0 X| + \ln(d) + \ln(n - p - d)$, which additionally incorporates information on the estimation of lack of fit. The results that were obtained from implementing an exchange algorithm to construct the requisite designs for examples 1 and 2 with 1 million random starts are summarized in Table 13. Clearly the designs are not unique. This is not unexpected since three-dimensional symmetry considerations hold but it is interesting to note that the numbers of $(\mathrm{DP})_\mathrm{S}$-optimal designs are inordinately large. The problem of which design to use in practice could be

Table 13. Number of optimal designs and pure error degrees of freedom for examples 1 and 2†

Criterion	*Results for example 1*		*Results for example 2*	
	N	*d*	*N*	*d*
$\ln\lvert X^{\mathrm{T}} Q_0 X\rvert$	24	0	8	1
$\ln\lvert X^{\mathrm{T}} Q_0 X\rvert - (p-1)\ln(F_{p-1,d;1-\alpha})$	6341	6	1064	8
$\ln\lvert X^{\mathrm{T}} Q_0 X\rvert + \ln(d)$	24	3	24	5
$\ln\lvert X^{\mathrm{T}} Q_0 X\rvert + \ln(d) + \ln(n-p-d)$	24	3	12	3

†*N* is the lower bound on the number of designs.

addressed by introducing a two-stage design procedure, i.e. by selecting designs from the class of $(\mathrm{DP})_{\mathrm{S}}$-optimal designs which are optimal with respect to secondary criteria. In addition symmetry-based classes of optimal designs could be identified by invoking group theoretic arguments and, more specifically in the present case, the theory of molecular symmetry and point groups. Search procedures could then be restricted to non-isomorphic groups. To return to the main point of my exercise, however, the two simple criteria that I have introduced do indeed provide optimal designs with fewer repeated points than their $(\mathrm{DP})_{\mathrm{S}}$-optimal counterparts and may be worth further consideration.

I have two other smaller comments. First I wondered whether it would be advantageous, computationally, to use candidate-set-free co-ordinate exchange algorithms in the construction of the optimal designs, particularly when p is large or when n is large and benchmark near-continuous designs are sought. I also noted that results from experiments based on repeated points can be fitted by using linear mixed models. However, I am not clear what advantages in terms of criterion formulation and design construction, if any, might accrue from this insight.

Heinz Holling (*University of Muenster*) **and Rainer Schwabe** (*Otto von Guericke University, Magdeburg*)
The authors provide interesting access into the definition of meaningful design criteria based on the needs for statistical inference, particularly for the construction of confidence intervals. This attempt is in full agreement coincidence with the general requirement ‘Estimates of treatment effects should be accompanied by confidence intervals, whenever possible’ and the requirement ‘Statistical analysis is generally based on the use of confidence intervals’ for equivalence trials stated in European Medicines Agency (1998), pages 27 and 18. Therefore, this approach shows great promise.

However, the restriction to variance estimates based on pure error rather than on residuals presupposes that the underlying model assumed may not be correct. In that case the variance estimate based on residuals may be biased, as the authors state. But then also the bias in the parameter estimates and in the prediction should be taken into account, which should result in completely different criteria based on the mean-squared error instead of the variance. Moreover, in this case of lack of fit it may become unclear what the meaning of the estimated parameters is, in particular, if the deviation of the model assumed is essential. Therefore, criteria based on the quality of estimation of the response function or for prediction of further observations like the integrated mean-squared error and the G-criterion suitably adjusted for confidence or prediction intervals are to be preferred over criteria based directly on the performance of the parameter estimates. As a side remark, the proposed GP-criterion, which is good for (pointwise) confidence intervals, must be adjusted to $F_{1,d,1-\alpha}[1+\max_x\{(1\ \mathbf{f}(x)'(\mathbf{X}'\mathbf{X})^{-1}(1\ \mathbf{f}(x)')'\}]$ for prediction intervals. This may result in even more replications for the corresponding optimal design. In contrast with the above emphasis on replications, non-parametric regression would be more adequate than fitting a low dimensional response surface, if there may be a substantial lack of fit, and equidistributed designs turn out then to be optimal for pointwise estimation of the response (Rafajlowicz and Schwabe, 2003),

As a conclusion, under severe uncertainty with respect to the model it may be doubtful to strive for optimal designs, and safeguarding against bias in the variance estimate may propagate bias in the point estimate of the response. Hence, an optimal design is only as good as the adequacy of the underlying model.

Bradley Jones (*SAS Institute, Cary*) **and Dibyen Majumdar** (*University of Illinois at Chicago*)
We congratulate the authors on a scholarly and useful paper. The idea of incorporating the need for the

estimation of the error variance in the optimality criterion has merit. Forcing replicate runs into the optimal design is an interesting idea. The use of a compound criterion for design will give the practitioner flexibility to tailor the design to the specific goals of the experiment.

The level of significance of the test seems to be a tuning parameter of the algorithm. We wonder whether using a larger value for the level of significance might result in designs that do not require so many replicate runs. The authors do not discuss the details of their algorithm except for saying that it is an exchange algorithm. However, it seems that the number of replicate runs is a free variable in the optimization. We think that it would be useful to set this number to a range of values and to do several optimizations with respect to this constraint. The properties of resulting designs would allow the practitioner to make trade-offs between the utility of extra degrees of freedom for pure error and other criteria of design quality.

The authors have stated a preference for using pure error degrees of freedom, rather than lack-of-fit degrees of freedom, for the estimation of the error variance and further testing of hypotheses. It is true that lack-of-fit estimates are only unbiased if higher order effects not in the *a priori* model are zero. However, design optimality theory relies heavily on the model assumed and, if the authors are concerned about the bias in the error variance estimate due to misspecification of the model, then it seems reasonable that they should also be concerned about the bias in the estimators of the treatment parameters. The fact that their optimality criterion tends to use up the extra degrees of freedom for replication has consequences. First, tests for lack of fit will lack power. Second, little information will be available for estimating a higher order effect if it exists. Third, if indeed the *a priori* model is correct, forcing replicate runs will reduce the efficiency of inference.

Despite these concerns the paper makes a significant contribution in raising the issue of error estimation in design optimality. We see the need for an evenly divided approach to the assignment of pure error and lack-of-fit degrees of freedom, coupled with model robustness considerations, in the choice of design.

Joseph B. Kadane (*Carnegie Mellon University, Pittsburgh*)

An experimental design is a gamble, in that one does not know what data will ensue, and hence what inferences will result. Thus experimental design is a statistical decision problem, characterized by several inputs: the set of allowable designs to choose between, the purpose(s) to be served by the observations once collected and a belief about the nature of the observations not yet available.

To discuss optimal design, it is necessary to be precise about these inputs.

The paper is particularly strong in its insistence on the second input, which can be thought of as specification of a utility (or, equivalently, loss) function. It is less strong on being explicit about the third component: the probability model. If one intends to use model (2) to estimate β, why would one want to use model (1) to estimate σ^2 (called the 'true error' in the paper)? I think that it would be more faithful to the beliefs of the experimenter to draw observations towards the subspace of model (1) specified by model (2). The extent to which such smoothing is done should depend on how strongly model (2) is believed as a special case of model (1). There are well-studied models that have this behaviour. A compromise view of σ^2 results.

Although my taste in inference is different from those of the authors, I recognize their paper as a contribution towards more accurately finding good designs that better reflect their purposes and the underlying beliefs about the data-generating process.

Joachim Kunert (*Technische Universität Dortmund*)

The choice between the various methods of estimating the error is not always as clear cut as Gilmour and Trinca state in their paper. There can be situations when it is not appropriate to base inference on the pure error and when lack of fit must be used instead.

One case where this is true is if we want to make predictions of the response under conditions that have not been used in the experiment. Specifically, consider sensory experiments with consumers. Assume that we have performed an experiment with 11 consumers to compare A and B, where each consumer evaluates both A and B twice. Further assume that consumers 1–6 consistently give 10 out of 10 points to A and 8 out of 10 to B, whereas consumers 7–11 always rate A as 8 and B as 10. We observe that the average difference between A and B is 2/11 and the pure error estimate is 0; hence A is significantly better than B. Most likely, however, we are not interested in these 11 consumers only but would like to predict the preferences for a larger set of consumers. If one consumer consistently prefers product A to B, this does not mean that other consumers will also prefer A to B. It is well known, therefore, that for this experiment we should not base our inference on the pure error. Instead, we should introduce a random interaction between assessors and products into the model and use the interaction sum of squares to estimate the variance.

Another class of examples where pure error might not be always appropriate is fractional factorial 2^{k-l}-designs. If we want to predict the response at a factor combination which has not been used in the design, the pure error might grossly underestimate prediction variance. The sum of squares for lack of fit, however, will contain information about the size of possible interactions which we did not have in the model.

Thus there can be good reasons why some textbooks recommend use of the residual mean square. Increasing only the degrees of freedom of pure error has some disadvantages. In particular, it may reduce information about the lack of fit of our model. Increasing the degrees of freedom for the lack of fit can help to detect mistakes in the way that the experiment was run or in the choice of the model for the analysis.

Jesús López-Fidalgo (*Universidad de Castilla-La Mancha*)
A typical criticism of optimal experimental design theory is that it usually provides not enough different points to estimate the variance and to make inferences. This paper launches an interesting and useful idea on experimental design for inferences through a proper estimation of the error. The real examples considered give an added value to the paper.

Although the main objective of the paper is not focused on discriminating between models there is a possible link between the approach of this paper and optimal designs for discriminating between models (1) and (2). For this approach see for example Atkinson and Fedorov (1975) and López-Fidalgo *et al.* (2007) where they use T-optimality and Kullback–Leibler distance optimality as an extension of the first to non-normal models. If there is interest in estimating the p parameters of model (2) then $t \geqslant p$. This means that model (2) is somehow nested into model (1). Let $\boldsymbol{\mu}^{(0)} = (\mu_1^{(0)}, \ldots, \mu_t^{(0)})'$ be nominal values for the means, where each mean μ_i is repeated n_i times in the vector. Thus, considering model (1) as the true model the criterion consists of maximizing

$$\begin{aligned} T(t, n_i) &= \min_{\beta}\{(\boldsymbol{\mu}^{(0)} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{\mu}^{(0)} - \mathbf{X}\boldsymbol{\beta})\} \\ &= (\boldsymbol{\mu}^{(0)} - \mathbf{X}\boldsymbol{\beta})'(\boldsymbol{\mu}^{(0)} - \mathbf{X}\boldsymbol{\beta}) \\ &= \boldsymbol{\mu}^{(0)\prime}(\mathbf{I} - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X})\boldsymbol{\mu}^{(0)}. \end{aligned}$$

This is the non-centrality parameter of the likelihood ratio test for the lack of fit of model (2) when model (1) is considered to be the true model. In this formula $\hat{\boldsymbol{\beta}}$ are the maximum likelihood estimates of model (2) assuming that $\boldsymbol{\mu}^{(0)}$ are the responses. The most disturbing point here is the need for nominal values for the means. Note that this vector depends in any case on the design. This means that the general function $T(t, n_i)$ claims nominal values for each particular possible design point.

Hugo Maruri-Aguilar (*Queen Mary University of London*)
Response surface methodology is often performed in stages, with the first stage being concerned about screening unimportant factors in a first-order model. A second stage may involve a more complex model, such as a second-order model, possibly with interactions; see Box *et al.* (2005). Assume that the design region remains unchanged for the second stage. Could the methodology that is described in the paper be adapted for its use in a sequential design approach?

Although one of the criteria $(\mathrm{DP})_{\mathrm{S}}$ is tailored to distinguish between nested models, it would seem that some adaptation would be required, as the list of active factors may not be known in advance. A block design with separate models for each block as in equation (3) might also be used. However, it would still present the same inconvenience of not knowing active factors in advance.

A simpler approach for this sequential approach would be to search for a $\mathrm{DP}(\alpha)$ design in which the part of the design matrix $\mathbf{X}$ corresponding to the first stage remains fixed, while the search for the optimal design is only carried out for the new design points and a bigger model. This approach could be refined to design a block optimal design in which the first block has already been fixed. What would the authors suggest in this case?

I would also like to join those who thanked the authors. This is a thought-provoking, most welcome advance in response surface modelling.

Jorge Mateu, Francisco J. Rodriguez-Cortés and Jonatan A. González (*University Jaume I, Castellón*)
The authors are to be congratulated on a valuable contribution and thought-provoking paper in the context of design of experiments. They develop compound criteria by taking into account multiple objectives. We would like to bring the authors' attention to a particular problem that could be benefit from this strategy.

Spatial point processes describe stochastic models where events that are generated by the model have an associated random location in space (Diggle, 2003; Illian *et al.*, 2008). Second-order properties provide information on the interaction between points in a spatial point pattern. The K-function $K(r)$ is one such second-order characteristic that can be expressed as an expectation or the number of events within a distance r of an arbitrary event.

A common problem in the analysis of spatial point patterns arising in some disciplines, such as neuroanatomy (Diggle *et al.*, 1991), neuropathology or dermatology, has to do with the identification of differences between several replicated patterns or groups of patterns, and the K-function plays a crucial role. This function could be significantly affected by a set of exogenous factors, and a linear fixed effects model could be implemented in a classical way as $K(r)-\mathbf{X}\beta+\varepsilon$. As usual, we may assume (although with some caution) that the within-group errors are independent and identically normally distributed with zero mean. In this context, we aim at maximizing the goodness of fit while checking the lack of fit by considering an optimum design. We wonder whether the compound criterion that is proposed by the authors and described in Section 4.1 and equation (5) can be adapted in this context to gain information on the (minimum) sampling size of the point patterns, which is a particular topic that remains open in the spatial literature.

This idea can also be extended to a linear mixed model case, in which $K(r)=\mathbf{X}\beta+\mathbf{Z}\mathbf{u}+\varepsilon$ where $\mathbf{u}$ is a vector of random effects. Fixed effects describe the behaviour of the entire population, and random effects are associated with individual experimental units sampled from the population. Instead of the second-order K-function, we could also assume that the conditional intensity is log-linear or comes through a mixed model (Bell and Grunwald, 2004) of the form $\log\{\lambda(x_0;x)\}=\mathbf{X}_{(x_0,x)}\beta+\mathbf{Z}_{(x_0,x)}\mathbf{u}+\varepsilon_{(x_0,x)}$ where $\mathbf{X}_{(x_0,x)}$ is a fixed effect design vector at location x_0 and $\mathbf{Z}_{(x_0,x)}$ is the random-effects design vector at x_0. We pose the question whether a correct extension of the compound criterion in equation (5) by adding a covariance matrix of the random effects could be used to treat optimal designs in the spatial contexts considered here.

Werner G. Müller and Milan Stehlík (*Isaac Newton Institute, Cambridge, and Johannes Kepler University, Linz*)

We thank the authors for bringing up the important issue of estimating error and its influence on optimum design. However, we are unsure that their mistrust in the mean model at the same time justifies their strong reliance on other invariances of the underlying distributions.

In fact their motivating example, exercise 11.6 of Box and Draper (2007), is particularly well suited for illustrating our point. Box and Draper first eliminated observation $y_6=86.6$ owing to a significant lack-of-fit test. Then, as the authors point out, the test for second-order parameters based on the pure error does not reject the null hypothesis whereas the test based on pooled error does. However, if we do not drop observation y_6 the conclusions completely reverse. Moreover, if we vary its value as is done in Fig. 5 this reversion stays the same for a great range of values that would not be detected by the lack-of-fit test—those to the left of the vertical line. Thus we wonder whether, for design purposes, it may not be advantageous to enter model uncertainties explicitly into the criterion as for instance in Liu and Wiens (1997) or Bischoff and Miller (2006).

Another issue concerns the use of compound designs later in the paper. After their conception in Läuter (1976) these have been successfully employed for many purposes (see for example McGree *et al.* (2008) or Müller and Stehlík (2010)). The challenge in their efficient use, however, remains in the proper selection of the weights κ involved. We would have appreciated a more detailed discussion of the consequences of the particular choices for κ that were taken in the paper and the sensitivity of the design to other choices.

Kalliopi Mylona (*University of Antwerp and University of Southampton*)

The authors introduce modified optimality criteria for the construction and evaluation of fractional factorial designs and response surface designs. The new criteria are applied to several cases with respect to the number of runs and factors and a comparison is provided of the optimal designs constructed by the new approach with the optimal designs constructed by traditional approaches. I believe that the results convincingly demonstrate the effectiveness and the usefulness of the new criteria, especially the results of the compound criteria which offer substantial flexibility to the experimenter to tailor the design to the experimenter's objectives. I was wondering whether, instead of a compound criterion, a sequential optimization of the corresponding criteria for the construction of the optimal designs would be possible. In this way, there would be no need to define weights each time.

Furthermore, in the case of blocked designs, the authors consider the block effects as fixed. However, it would be interesting to investigate to what extent the results would be affected if the blocked effects were considered as random and how effective the new criteria would be. In this case, a Bayesian adaptation of the criteria to deal with the problem of the unknown variance components could be possible and helpful.

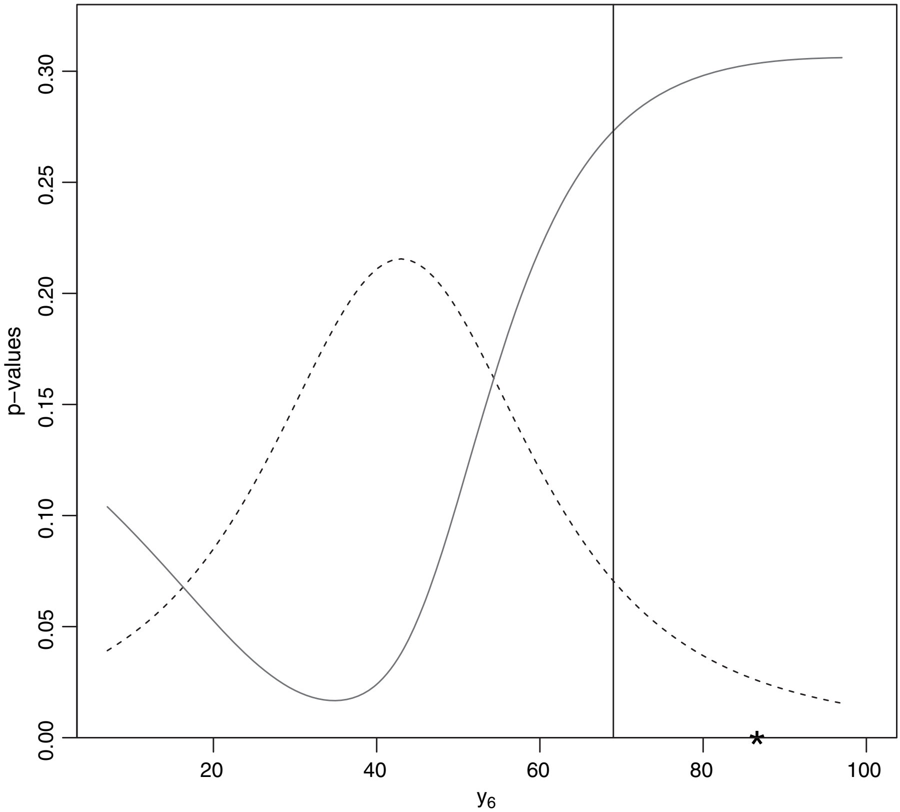

Fig. 5. Box and Draper data: p-values for an F-test of second-order parameters based on pure error (-------) and pooled error (———) (|, threshold value for a lack-of-fit test; ⋆, true value of y_6)

William Notz (*Ohio State University, Columbus*)

I congratulate the authors on a stimulating and insightful paper. Many popular designs are justified on intuitive grounds. Finding optimality criteria for which such designs are optimal or nearly optimal provides mathematical insight. It also allows us to generate designs with desirable properties in other settings.

Gilmour and Trinca seek criteria that produce designs that provide an estimate of pure error and allow a test for lack of fit. As I read the paper, the following observations occurred to me.

(a) The authors (rightly) assume that in practice models are approximations. Hence model bias will always be present. The comment by Box and Draper (1987), 'Essentially, all models are wrong, but some are useful', is relevant.

(b) Designs that consider model bias or designs that produce replicated observations might be competitors to those in the paper. How might the following designs perform?

   (i) There are several references, beginning with Box and Draper (1959), which discuss optimal designs when model bias is present. Among them are Atkinson and Fedorov (1975) and Jones and Mitchell (1978).

   (ii) Continuous optimal designs for regression often place unequal weights on points of support. When implementing such designs, multiple observations at some support points are necessary to approximate the weights in the optimal continuous design. Such replication will allow for an estimate of pure error.

(c) Another approach to employing multiple-design criteria is to adopt a Pareto optimal approach.

(d) In his work on D-optimality, I believe that Kiefer assumed that the model is correct. Hence the issue of pure error *versus* the standard estimate of error does not arise. Perhaps motivated by the work of Box and Draper, Kiefer later considered the problem of optimal design criteria when model bias is present. But Kiefer always began with a formal mathematical model.

(e) Is it too strong a statement to say that classical D-optimality has no statistical interpretation? Kiefer (1959) gave several interpretations but made additional (perhaps unrealistic) assumptions about the model. In other papers (Kiefer and Wolfowitz, 1959) Kiefer is careful to discuss the shortcomings of 'optimal' designs. Any statistical interpretation of an optimality criterion is a consequence of the model assumed. Perhaps the issue is whether the model assumptions are realistic. A lack of clarity about the model assumed leads to misuses of optimality criteria, perhaps motivating Box and Draper's concerns about the use of 'alphabetic optimality'.

Luc Pronzato (*Centre National de la Recherche Scientifique and Université de Nice-Sophia Antipolis*)
The authors reconsider classical criteria for optimal design in the framework of statistical inference for linear regression models with homoscedastic errors. Denote $\hat{\boldsymbol{\theta}}^n$ the least squares estimator and $\mathbf{M}_n$ the information matrix for n observations,

$$\mathbf{M}_n = \sum_{i=1}^{t} n_i (1\ \mathbf{f}(\mathbf{x}_i)')'(1\ \mathbf{f}(\mathbf{x}_i)')$$

in model (1)–(2) with parameters $\theta = (\beta_0\ \beta')'$. Although confidence ellipsoids for θ all have the same shape as the ellipsoids $\mathscr{E}_n(c) = \{\boldsymbol{\theta} \in \mathbb{R}^p : (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^n)'\mathbf{M}_n(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^n) \leqslant c\}, c > 0$, their sizes, at a given level of confidence, depend on whether the error variance σ^2 is known or estimated, and then on the way it is estimated. The authors should be complimented for drawing attention to this point: since different designs allow different degrees of freedom d for the estimation of σ^2, the value of d affects the size of confidence regions and should thus enter the definition of the design criterion. It seems to me, however, that some underlying assumptions, or approximations, have been overlooked and that usual criteria for optimal design based on the shape of the ellipsoids $\mathscr{E}_n(c)$ deserve a fairer treatment. In particular, I find the paper too abrupt in its conclusion about D-optimality.

First, note that the estimation of σ^2 through residuals, with $d = n - p$ degrees of freedom for n observations and p parameters, remains perfectly legitimate in all circumstances where the model is taken as valid, which leaves some room for the use of classical design criteria. Secondly, the confidence regions that are suggested, $\mathscr{E}_n(c)$ with c proportional to $F_{p,d;1-\alpha}$ ($d = n - t$ in the model (1)–(2)), are exact *when the errors are normally distributed*. They are approximate in other circumstances, the degree of approximation depending on how large n and p are. When the errors are normal, or sufficiently close to normality to allow the use of $c \propto F_{p,d;1-\alpha}$, it means that $\hat{\boldsymbol{\theta}}^n$ is normal, or approximately normal, and using designs based on geometrical properties of the covering ellipsoids $\mathscr{E}_n(c)$ makes sense for point estimation. In particular, a D-optimal design that minimizes their volume seems quite reasonable to me. Also note that a D-optimal experiment minimizes the entropy (Shannon and Renyi of order q for any $q > 0$) of the normal (or approximately normal) distribution of $\hat{\boldsymbol{\theta}}^n$. Although the case of large n is not considered in the paper, it is worthwhile to note that the asymptotic normality of $\hat{\boldsymbol{\theta}}^n$ is obtained under rather weak conditions, even when the dimension p is allowed to grow to ∞ with n; see Huber (1973), proposition 2.2. Moreover, even if $\hat{\boldsymbol{\theta}}^n$ is far from being normal (so that the confidence regions in the paper may not be valid), its variance–covariance matrix is still given by $\sigma^2\mathbf{M}_n^{-1}$, and a D-optimal experiment maximizes the volume $V(t) = \text{vol}\{\mathbf{a} \in \mathbb{R}^p : \text{var}(\mathbf{a}'\hat{\boldsymbol{\theta}}^n) \leqslant t\}$ for any $t > 0$ (since $V(t) \propto \det^{1/2}(\mathbf{M}_n)$), which makes sense as a measure of precision of the (point) estimation of θ. Of course, these arguments in favour of classical criteria based on geometrical properties of $\mathscr{E}_n(c)$ remain valid when the parameters of interest form a subset of $\boldsymbol{\theta}$, with D_S-optimality as a special case; see Silvey and Titterington (1973).

Timothy J. Robinson (*University of Wyoming, Laramie*)
Kudos goes to the authors for addressing the inherent flaws when using the D-criterion for identifying an 'optimal' design and for highlighting the need to consider multiple-design criteria simultaneously. The new criteria offer improved alternatives to the standard D-criterion. Software packages readily determine 'optimal' designs based on *single-design criteria*. The nomenclature 'optimal design', however, is unfortunate because the criterion is often not clearly explained and optimal designs can perform poorly for other important criteria. As Box and Draper (1975) pointed out, a good experimental design should simultaneously exhibit many desirable traits.

My comments focus on the philosophical notion of bundling criteria into a single index.

Taguchi's popularization of robust parameter design in the 1980s challenged practitioners to consider the response mean and variances simultaneously. Taguchi bundled these into the signal-to-noise ratio.

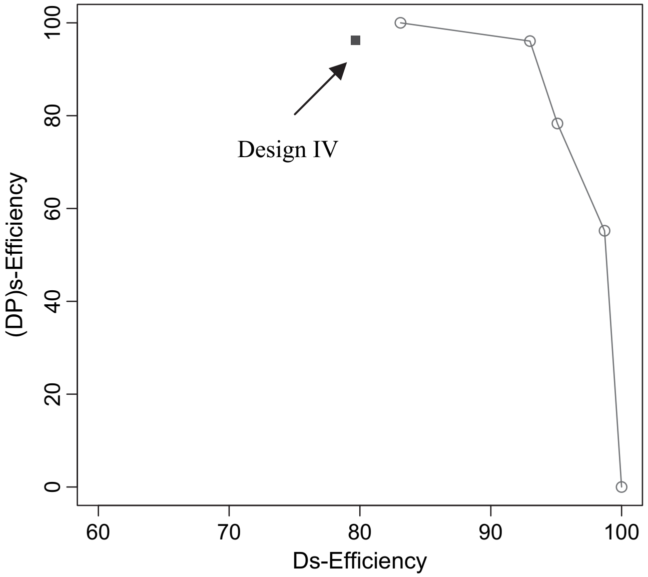

Fig. 6. Pareto front based on the six designs in Table 1; design IV ($D_S = 80.94$; $(DP)_S = 97.42$) is Pareto dominated by design II ($D_S = 83.09$; $(DP)_S = 100$); using a Pareto exchange algorithm which is not based on predefined weight choices, the resulting front would probably be comprised of a reasonably large set (i.e. more than five designs) and then designs could be compared in terms of robustness to weightings and trade-offs between criteria

Box (1988) suggested that one could learn more studying the mean and variance separately. For design optimization, instead of bundling several criteria with prespecified weightings into 'combined criteria', one could utilize a Pareto approach. Lu *et al.* (2011) describe a Pareto exchange algorithm for multiple-design criteria. A design is said to *Pareto-dominate* another design if all of its criteria values are at least as good and at least one criterion is strictly better than its competitors. A design is *Pareto optimal* if and only if no other design dominates it. Once the set of Pareto optimal designs has been identified, final design selection based on the Pareto optimal set involves directly examining trade-offs between criteria. Thus, choice of weighting schemes occurs after generation of the Pareto optimal set.

To illustrate, consider the D_S and $(DP)_S$-efficiencies for designs I–VI in Table 1. Of the designs listed in Table 1, five represent the Pareto optimal set (Fig. 6). Design IV is dominated by the others and can be eliminated from consideration. The steep slope of the front suggests that gains in D_S-efficiency require disproportionate sacrifice for $(DP)_S$-efficiency. Although the front in Fig. 6 is based on the six contenders in Table 1, if a more complete search is done, improved designs could potentially be identified and a more complete set of competing choices found. Lu *et al.* (2011) propose methods for finding the Pareto front and then illustrate how to examine trade-offs between criteria by using a variety of weightings. Bundling criteria into a single index inhibits exploration of trade-offs, whereas the Pareto approach offers transparency in the decision-making process.

Byran J. Smucker (*Miami University, Oxford*)

In a provocative, potentially transformative paper, Gilmour and Trinca propose a significant shift in the optimum design of experiments. In particular, they maintain that inference for designed experiments should be based on an estimate of error that is independent of the fitted model, instead of the more commonly used mean-squared error. If this premise is accepted, the authors argue without equivocation that classical optimal design criteria must be reconstituted, because the classical criteria implicitly assume an independent error estimate.

At heart, this is a philosophical paper. In advancing their case for pure error and against the residual error, the authors rely primarily on an argument from authority, even while admitting that there is not a strong consensus among the authorities. The most substantive statistical argument is from Sheffé (1959) who recommended against pooling sums of squares from non-significant interactions. However, this does not directly address the issue, because if a model is fitted—with the implicit assumption that unmodelled effects are absent—the analyst could avoid Sheffé's criticism by using the mean-squared error without pooling the sums of squares of the small modelled effects.

Given the lack of consensus and the absence of an open-and-shut argument in favour of pure error, the forcefulness of the authors' conclusions seems unjustified. For instance, they say that

> 'the *D*-criterion, as usually defined, has no place in the design of fractional factorial or response surface experiments',

except when the run size is large. Perhaps this is true, though one could imagine experiments for which a criterion related to classical *D*-optimality would be appropriate (e.g. supersaturated designs). Regardless, the authors' arguments in favour of pure error would be strengthened by an empirical study demonstrating directly its inferential advantages.

This commenter is in agreement with the broad point made to close the paper: designs based on compound criteria offer an effective and flexible alternative to single-criterion optimal designs. Furthermore, the authors are to be commended for re-examining a design issue that is not often considered. At the very least, this paper will cause practitioners and researchers alike to think more deeply about the embedded assumptions in the design techniques that they use. At most, it could represent a watershed moment in the practice of optimal design.

Júlio Silvio de Sousa Bueno Filho (*Universidade Federal de Lavras*)
The paper from Steven and Luzia brings fresh news to the subject. They improve on hints from the classics (Kiefer, 1959; Fisher, 1966) to uncork a design criterion for inference based on pure error estimates. This gives a way to find designs that combine different properties. Modified criteria and also compound criteria are meaningful under exact optimality. Potential uses of the criteria are illustrated by using search algorithms.

This new way to look at design inferential properties avoids *ad hoc* assumptions on comparing new designs with some well-known standard designs like the central composite and Box–Behnken types of designs as well as blocked designs.

I wonder how this kind of reasoning could be extended to other types of inference like variance components estimation and ranking of treatments that are random samples from a population. This could be related to comment (b) on page 350, but in this case other design questions might be involved.

Congratulations and many thanks go to the authors for this delightful and invigorating paper.

Milan Stehlík (*Johannes Kepler University, Linz*) **and Luboš Střelec** (*Mendel University, Brno*)
We congratulate the authors for opening a challenging world of designing under uncertainty.

Statistical models are often purely theoretical constructions that approximate the stochastic behaviour of a system but have otherwise nothing to do with the real data mechanism. In our case it would be much more infomative to use more facts from subject science, i.e. food chemistry. Therefore, alongside the importance of outliers that is discussed here by Müller and Stehlík, there is a further issue concerning the distributional deviations from the *F*-distribution. On page 347 there is a discussion saying that

> 'if we use s^2, the test of the second-order parameters gives a test statistic of $F=2.87$, which on 6 and 3 degrees of freedom gives a *p*-value of 0.208, which would suggest that there is little justification for further interpreting the second-order model'.

However, the values y_i, which are realizations of $y=\gamma K_2/(1+\Sigma K_i x_i')$, are not only positive but strictly over some positive constant (see their measured values in Table III of Carr (1960) or realize their chemical meaning, i.e. k_i are equilibrium absorption constants and x_i are partial pressures of hydrogen, *n*-pentane and isopentane). Thus the proper test should be based on truncated normals. Here for $C=2$ the *p*-value is 0 for the above-mentioned $F=2.87$ on 6 and 3 degrees of freedom. Fig. 7 shows the *p*-value for various truncations $C>0$.

Finally, we recommend the use of robust test class tests, introduced in Stehlík *et al.* (2011), which may suggest a specific truncation level $C>0$. Especially, mean–median types of tests have much better power than classical Jarque–Bera tests, especially test versions which have medians in the denominators. Even better are trimmed versions of the robust test class of tests, where trimming in the denominators is crucial.

Basic research in statistics for food chemistry is a very important issue, which may prevent many subsequent issues in the analysis and give us insight into proper statistical practice for the food industry. For

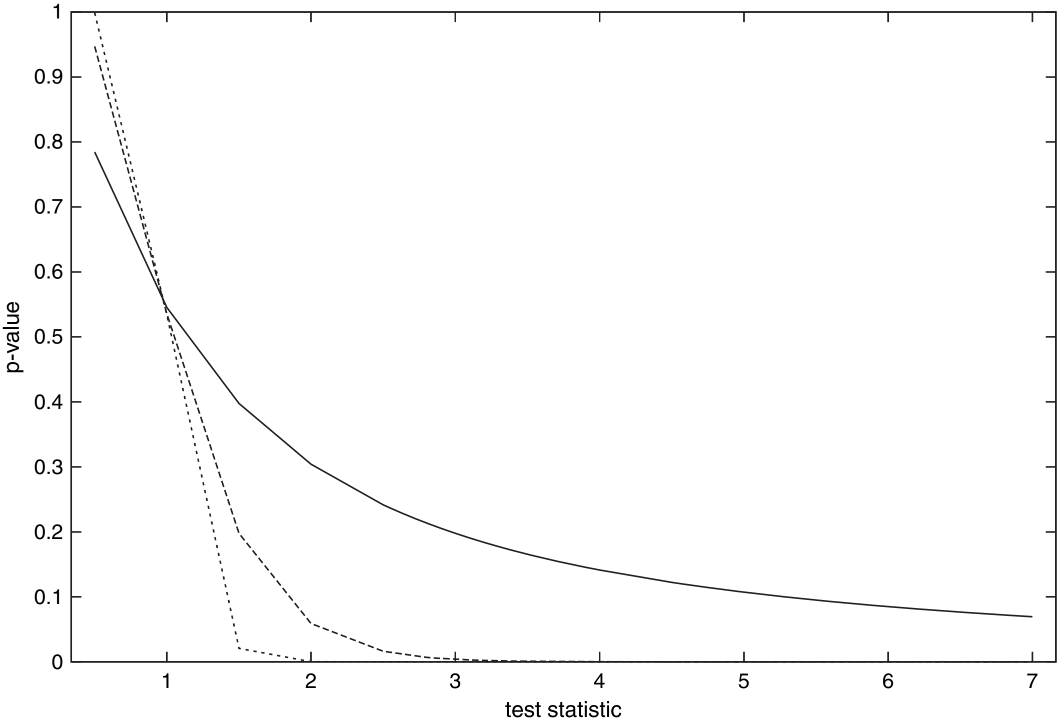

Fig. 7. p-values, $F(6,3)$: ———, F; - - - - - - -, $F_{\text{trunc}\,c_1}$; · · · · · · ·, $F_{\text{trunc}\,c_2}$

example see American Chemical Society (2010), the materials of the Division of Agricultural and Food Chemistry, where the first author participated at the American Chemical Society in 2010. We believe that more attention should be paid to this area generally.

David M. Steinberg (*Tel Aviv University*)

In this interesting paper, Steven Gilmour and Luzia Trinca point out that criteria for deriving and evaluating experimental designs may be inadequate when statistical inference is a major goal. Gilmour and Trinca focus on factorial experiments with more potential design sites than runs. In my collaborations, inference has been a major objective primarily in simple comparative experiments in which the sample size was more than sufficiently large to permit replication. The standard design question is to set the sample size to achieve desired power. On small and medium-sized factorial experiments, inference was not at the forefront.

The designs that were derived by Gilmour and Trinca are heavily influenced by their clear preference for a pure error estimate of σ^2. There are certainly benefits to having a pure error estimate, but I do not share the authors' absolute endorsement. Allotting many degrees of freedom to replication can come at a substantial price. First, there will be little information to assess lack of fit. Second, the design will depend strongly on an assumed model. Third, the parameter estimates may themselves be biased. Gilmour and Trinca recognize and discuss the first two issues, but not the third. Consider an experiment with seven two-level factors and 16 runs. Following example 5 in the paper, the authors' criteria would select a replicated 2^{7-4} fractional factorial (with resolution III) rather than the standard 2^{7-3}-design (with resolution IV). Their recommendation reflects concern that a model-based estimate of σ^2 will be biased by the presence of interactions or by unjustified removal of large interaction contrasts. If any two-factor interactions *are* present, the authors' design will suffer from biased estimates of the main effects. The subsequent inference will not be valid. I prefer the 2^{7-3}-design and the model-based variance estimate.

Design optimality criteria are now widely available in software packages and will be used with increasing frequency. How should we use them? Gilmour and Trinca are correct that specific criteria and associated construction algorithms may ignore some important goals. Sometimes it is helpful to modify the criteria to reflect additional goals. More often, though, I think a better alternative is to work with standard criteria, regarding the designs as a first iteration that can be modified or augmented to meet additional needs. We should make efforts to educate those who use packages on the benefits and the limitations of algorithmically generated designs. Inclusion of some diagnostics describing the inferential quality of a design could be a useful step in that direction.

Eleanor Stillman (*University of Sheffield*)

I congratulate the authors on an interesting and accessible paper. Primarily they remind us of the vital importance of encapsulating within our design criterion precisely what is of interest to us. This is a salutary lesson to us all and the examples are a useful teaching tool.

Presumably the problems that they demonstrate are most acute with D-optimality because of its tendency to use unreplicated extreme points. Criteria such as V-optimality, which typically induce greater replication, might produce designs scoring acceptably on multiple objectives without modification. The issue of requiring replication to enable variance estimation via sample variances rather than residuals has been addressed by Goos *et al.* (2001). Its use in a compound criterion for simultaneous mean and variance estimation is discussed in Emmett (2010).

A comparison of the paper's compound approach with one addressing the subelements of interest sequentially would also be interesting. Can they quantify the benefits in using a design optimizing their compound criterion over constructing a sequence of experiments which firstly select an appropriate model, then estimate its parameters and so on?

I was particularly interested to see the authors' recognition that, in practice, users frequently seek to remedy perceived deficiencies in proposed designs by making adjustments, often in an *ad hoc* manner. This accords with my own experience (Martin *et al.*, 2003), but they do not highlight the key point that users almost always do this by adjusting the design, not the criterion. Ease of use of the criterion is clearly, therefore, something that must be taken into account; users need to be guided through the steps of reflecting on potential uses of the experimental data and assessing their relative importance. Although the paper reports relative insensitivity to the choice of weights in the compound criterion, it might be an interesting exercise to elicit and incorporate expert opinion on this weighting for a number of practically important scenarios. Compound criteria including probabilistic mechanisms for introducing a 'degree-of-interest weighting' (McGree and Eccleston, 2008; Emmett *et al.*, 2011) might also be explored.

Pi-Wen Tsai (*National Taiwan Normal University, Taipei*)

I congratulate the authors on this interesting paper which emphasizes the need to consider how we are intending to analyse the data before planning the experiments. It focuses on the fundamental question of error variance estimation and proposes several modified optimality criteria that combine the degree of freedom for pure error with the efficiency of parameter estimation. Convincing examples are given to show that different designs might have different optimal properties under different criteria. The authors also provide an interesting discussion about how to design an experiment which respects experimental objectives. When the experimental objectives are many, the standard criterion is no longer enough, and some compound criteria are needed.

The proposed $(\mathrm{DP})_\mathrm{S}$- and $(\mathrm{AP})_\mathrm{S}$-criteria, which are modifications of the classical D_S- and A_S-criteria respectively, involve some quantiles of the F-distribution to correct for the effect of the estimation of pure error on sizes of the confidence regions. The magnitude of the effect of the quantile is larger for $(\mathrm{DP})_\mathrm{S}$ than it is for $(\mathrm{AP})_\mathrm{S}$. I suspect that this is why the designs constructed on the basis of the $(\mathrm{DP})_\mathrm{S}$-criterion often have larger numbers of degrees of freedom for pure error than those based on the $(\mathrm{AP})_\mathrm{S}$-criterion. In one or two examples, the numbers of degrees of freedom for the pure error are quite small. I would be cautious about doing statistical inference based on such error estimates. Mead (1988) suggested

> 'to obtain a good estimate of error it is necessary to have at least 10 degrees of freedom, and many statisticians would take 12 or 15 degrees of freedom as their preferred lower limit. . . '.

I think that it might be useful to suggest a 'best limit' of the number of degrees of freedom for pure error rather than finding an optimal design which provides only 3 or 4 degrees of freedom. If the experimenters are worried about the estimate of the error, I think that the best suggestion statisticians could give is to ask them to increase the number of experimental runs or to simplify the model that they are intending to fit. If the best number of degrees of freedom for pure error can be specified, we might use some standard optimality criterion to obtain an optimal design for the given model, and supplement it with some extra repeated points for error estimation.

The **authors** replied later, in writing, as follows.

We thank the discussants for their contributions. We have little space, but we hope that discussions will continue. Where possible, we have grouped the comments.

Bayesian decision theory

We are open to the viewpoint (Draper and Kadane) that Bayesian decision theoretic methods would be more appropriate than classical inference, but we chose our words carefully in considering how 'experimenters' objectives *will* be met'. We should not fool ourselves into believing that Bayesian decision theory can simultaneously be both simple and useful. We would have to take account of all the uncertainties and to give subjective probabilities or distributions to each of them. This seems overwhelming. Chatfield reminds us that one cost of running an experiment is the (human) time taken to come up with a design. Decision theory might be more useful at a supraexperiment level, when many experiments of the same type are run.

Estimation

Empirical Bayes estimation of $\boldsymbol{\beta}$ (Stone) is unlikely to make much difference, since we usually have little prior knowledge. For the fairly small experiments that we consider, non-parametric regression (Holling and Schwabe, and Manteinga and Porcu) is unrealistic, but our methods could be adapted for such models where they are useful.

Although, when more than one support point is replicated, the pure error estimate (never pooling) can also be regarded as being pooled (Berger), the randomization justification seems to us a much stronger reason for using this analysis than an assumed constant variance model.

Vandebroek and Brien defend the commonly used method of sometimes pooling. We tend to agree with Janky (2000) that

> 'The idea of a "reasonable" level of size inflation is philosophically inconsistent with a formal hypothesis testing environment'.

Janky also showed that the gains in power are often small. If one adopts this procedure with $\alpha = 0.5$ (Brien), how would one design the experiment? The relevant criterion function would be, for example, $\{(F_{p,d;1-\alpha})^p + (F_{p,n-p;1-\alpha})^p\}/|\mathbf{X}'\mathbf{X}|$ which would give similar results to $(\text{DP})_\text{S}$, but with occasionally slightly fewer pure error degrees of freedom.

We are surprised that many (Berger, Waite and Woods, Chatfield, Jones and Majumdar, Kunert, Notz, Pronzato, Schwabe and Holling, and Smucker) are in favour of always pooling, noting that it is valid if higher order terms can be assumed to be negligible *a priori*. To obtain valid inferences, one must really be willing to believe this assumption, even if the data collected cast doubt on it. We would not recommend this for formal inference, but we would not object for less formal methods of analysis (Kunert). We cannot emphasize enough that the inefficiency introduced by replicates arises not just if the assumed model is correct (Jones and Majumdar), but only if it is *known to be correct*. In the words of Holling and Schwabe, the use of pure error 'presupposes that the assumed underlying model *may* not be correct'. This is the argument that we rely on, not authority (Smucker). We are not willing to assume absolutely that a convenient, but *ad hoc*, empirical model is true on the authority of the statistician.

Sometimes σ^2 can be estimated from previous experiments or experience of the process (Stone and Cox). We have not considered empirical Bayes methods, but they might be useful in well-controlled processes where we can be confident that history will accurately reflect what will happen in the experiment. In our experience, experimenters are reluctant to use historical estimates of variance and, when encouraged to consider how many observations they are worth, rarely go above 4 or 5. This could be incorporated into our criteria by replacing d by $d + d_\text{P}$ where d_P measures the worth of the prior estimate in terms of pseudodegrees of freedom. If d_P is large, the criteria will converge to the standard criteria.

Chatfield suggests that experimenters should compare internal estimates with prior expectations informally, since an unusual estimate might be an indication of something important. From this viewpoint, the new criteria seem reasonable.

The future of D

Our statement that there is no place for D-optimality provoked some disagreement. This does not mean that D-optimum designs are bad in all situations, but that the definition of D should be more general, leading to *DP in the types of experiment that we are discussing*. Several discussants (Waite and Woods, Stone, Torsney, Notz and Smucker) disagree with our statement, but not, it seems to us, with its meaning, rather pointing out cases where our logic leads back to the standard criterion.

Only Pronzato defends the usual definition, arguing that it gives a good summary of a design's properties for *point estimation*. Although the mean (leading to A) seems like a more natural summary, this might be a matter of convention. If we accept Pronzato's geometrical interpretation, it actually clarifies the main point of our paper. The choice between standard and pure error criteria (or their relative weights) depends on whether (or to what extent) point estimation and inference are important. Whether the metrics should be D, A, etc. is a separate question.

Convergence to standard criteria

When the new criteria converge to the standard criteria (Berger) deserves further study, but our results suggest that they start to become close when we have more runs than the number of support points in the continuous optimum design. This makes sense, as beyond this point the standard criteria will tend to include replicates (Notz). Kiefer's universal optimality (Ceyhan) is intrinsically an asymptotic concept.

Increasing α gives fewer replicates (Jones and Majumdar), as can be seen by comparing the designs with and without the Bonferroni correction. However, we do not see any reason to design using a significance level that would not be considered in the analysis, though $\alpha = 0.1$ might be a good default.

Non-normality

The new criteria are not exact without normality (Torsney, Cox, Pronzato, and Stehlík and Střelec). However, degrees of freedom are defined by the differences between the dimensions of different models (Bailey) and only the link to the volume of the confidence region depends on normality. This is also true for the standard criteria, which also depend on the linear predictor, so the new criteria are more robust. Even if we do not believe normality, a design with optimal pure error degrees of freedom under normality is better than a design with no pure error degrees of freedom.

In generalized linear models (Torsney), the new criteria would be the same as the standard criteria. Models with unknown dispersion parameters (Cox) could be studied along similar lines to our paper. Continuous, but non-normal, distributions can be useful (Stehlík and Střelec). It would be interesting to extend our criteria to these cases.

Alternative criteria and algorithms

In our examples, Šidák's method (Ceyhan) gives the same results as the Bonferroni correction.

Torsney, Jones and Majumdar, Steinberg and Tsai suggest the use of standard criteria subject to given numbers of degrees of freedom, whereas Haines suggests combining the two. Using the former for a range of pure error degrees of freedom gives an alternative algorithm, but we would still want to calculate the new criteria to decide how many pure error degrees of freedom are needed—this gives the correct trade-off for inference, rather than an *ad hoc* adjustment (Stillman). Designs could be constructed in two parts, first choosing a standard optimum design, and then optimally replicating some of the points (Torsney), but these might be suboptimal. For very large problems (Manteinga and Porcu), this might be the only feasible method. Haines's results show that the effect of the F-distribution is non-linear: 1 degree of freedom is less than half as good as 2, whereas 50 are more than half as good as 100.

I-optimum designs tend to include replicated (centre) points (Atkinson, Goos and Stillman), but we prefer to choose a design to do the right thing, rather than hoping that a design chosen for the wrong purpose will turn out to do so. I concentrates on estimation of the response and this is dominated by the intercept β_0. I chooses replicate centre points because they are in the centre, not because replication is needed. We found designs for example 3 with candidate points made up of $(-1, -\frac{1}{3}, \frac{1}{3}, 1)$ for each factor. The I-optimum design that we found has no pure error degrees of freedom but has three points with all factors at $\pm\frac{1}{3}$, whereas the $(\mathrm{DP})_\mathrm{S}$-optimum design has 14 pure error degrees of freedom and no points with all factors at $\pm\frac{1}{3}$, and the D_S-optimum design has four and none respectively. Goos's first two designs are respectively $S_3 + S_1 + 2S_0$ and $S_3 + S_2 + 6S_0$. It is misleading to call these 'more useful' than the $(\mathrm{DP})_\mathrm{S}$- and $(\mathrm{AP})_\mathrm{S}$-optimum designs because they give lack-of-fit degrees of freedom. If we wanted to check for lack of fit, we would use a compound criterion; we would only use $(\mathrm{DP})_\mathrm{S}$ or $(\mathrm{AP})_\mathrm{S}$ if inference was the *only* analysis contemplated.

IP-optimum designs can be easily found by using our algorithm (Goos). We did not emphasize prediction inference, since it is invalid under randomization, but difference-based prediction criteria, which were briefly mentioned by Trinca and Gilmour (1999), might be useful.

A co-ordinate exchange algorithm (Atkinson and Haines) is worth considering for larger problems, but it will sometimes be better and sometimes worse than the point exchange. The computational saving is less with the new criteria, since degrees of freedom must be recalculated after each exchange and the co-ordinate exchange algorithm needs more exchanges.

Small designs

Whereas several discussants (Vandebroek, Berger, Cox and Tsai) urge caution in using our criteria for small designs, we would urge caution in using inference in these cases. We emphasize again that inference is often not the main point of experiments (Steinberg). The resource equation of Mead (1988) (Tsai) is also useful. With very small designs, there is a danger of asking for too much. The new criteria make this clearer: comparing examples 1 and 3, the relative amounts of information are $26/16 = 1.625$ under continuous D_S-efficiency, 1.666 under D_S-efficiency and 1.857 under $(\mathrm{DP})_S$-efficiency. Formalizing this into sample size calculations (Mateu, Rodríguez-Cortes and González) might be worth exploring.

Block structures

In Trinca and Gilmour (2000), despite the loss of efficiency (Bailey), we chose treatment designs and blocked them separately to preserve the many desirable properties of classical designs. The compound criteria make this less beneficial and here we choose the treatment design and blocking together. The use of fixed block effects for design protects against the worst case, although the use of prior estimates (Bailey and Mylona) is an alternative.

Extensions to split-plot and other multistratum structures requiring mixed models (Torsney, Goos, Haines, Mateu and colleagues and Bueno Filho) are under development. Conceptually, it is less clear that one would ever want a joint confidence region for all the parameters.

A sequential adaptive design (Draper, Maruri-Aguilar and Stillman) has obvious benefits. Each stage in a group sequential scheme can be dealt with by using our methods, or the empirical Bayes modification suggested above if a common σ^2 across stages can be assumed.

Other models

To extend the new criteria to correlated errors (Boukouvalas and Cornfield, Mateu and colleagues, and Manteinga and Porcu) the crucial question is whether biased estimation of the covariance parameters is the problem, but we do not know the answer. One practical problem is that unbiased estimation of many covariance parameters would need many more replicate points.

Models with random treatment effects (Bueno Filho) raise new problems which we have not studied. Kunert's sensory experiments seem to fall into this category, with the consumers being noise factors with random effects. If there are large treatment by consumer interactions, we would not be interested in the treatment main effects. Otherwise, we see no reason to avoid criteria that are developed along the same lines as those in our paper.

Alternatives

Desirability functions (Ceyhan) are a generalization of compound criteria and, with $A_i = 0$ and $B_i = 1\ \forall i$, they are identical. The generalization allows efficiencies below A_i to be declared unacceptable or all efficiencies above B_i to be declared equally good. This might be useful, though we would prefer a more informal approach, adjusting the weights to achieve what is desired.

Optimizing one criterion, subject to bounds on the others (Mylona) is very similar to compound criteria and seems a matter of taste. We do not find the setting of bounds any more natural than the choice of weights.

Pareto optimization (Anderson-Cook and Lu, Notz and Robinson) is an attractive alternative to combining metrics into a compound criterion and we hope that both approaches will see increased application in the future. Only experience will tell when each is most useful. The choice of weights being subjective and imprecise can be seen as an advantage. It encourages users to 'play around' with weights to find designs with different properties and to study these properties separately (they are bundled together only to generate interesting designs). This can raise new issues, leading to additional metrics in a new compound criterion, which were not thought of initially. This kind of interactive construction requires speed, which is more difficult with Pareto optimization, especially with large numbers of component criteria (which also make interpreting the Pareto front difficult). Ideally, one might follow up the interactive phase of finding designs and properties with a Pareto optimization to make sure that nothing has been missed.

Choice of weights

The multiplicative form of weights (Critchley) ensures the scale invariance. Interpretability depends on whether the concept of efficiency is interpretable. Although this is not clear, there is at least considerable experience in using it. We have not considered a stronger affine invariance (Critchley), but invariance considerations and interpretability tend to push us in opposite directions.

We did not study the choices of weights in great detail and there might be other good designs which we have missed (Atkinson, Critchley, and Müller and Stehlík), but experience suggests that nothing drastic

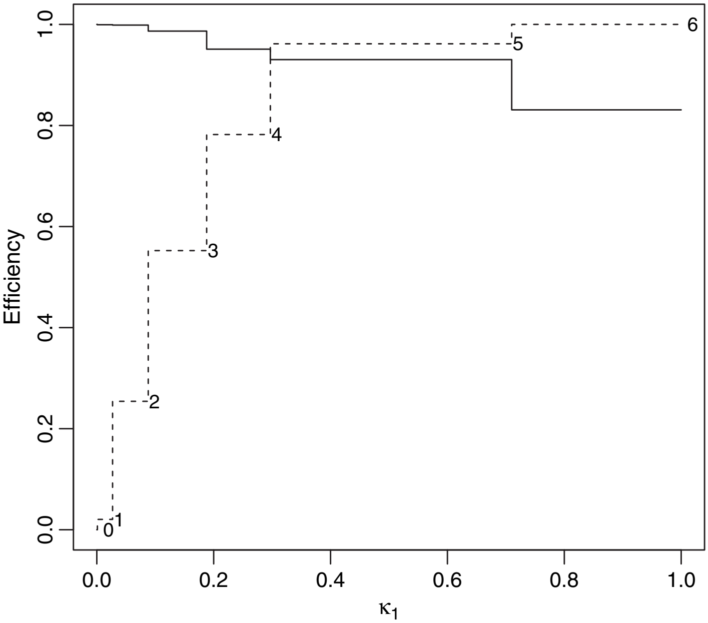

Fig. 8. Criterion efficiencies for optimum designs in example 1 as weights change: ———, D_S; - - - - - -, $(DP)_S$

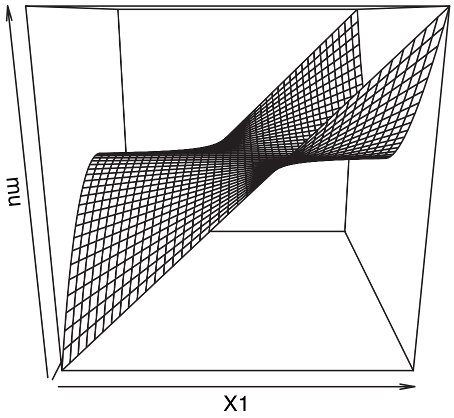

Fig. 9. Response surface with third-order, but no second-order, effects

happens and we think that there is not much of interest missed. Fig. 8 shows a typical pattern, with seven different optimum designs. Haines's results indicate that there is even robustness to the exact criterion that is used, as she obtains similar designs by using d directly, instead of $F_{p-1,d;1-\alpha}$.

We do not see any real problem with the subjectivity of the weights (Berger, and Müller and Stehlík), since designing experiments *should* be subjective. The criteria generate useful designs which should be studied in detail before being used.

Lack of fit

In the presence of third-order terms the $(DP)_S$-optimum design has more inflated size, but higher power, than the D_S-optimum design (Vandebroek, Jones and Majumdar). Which is preferable? There are no low order effects, but non-negligible third-order effects, which implies an underlying response surface like that shown in Fig. 9. It might be better to discover the wrong effect than to discover no effect.

The new designs lack power for detecting lack of fit (Vandebroek, Jones and Majumdar, Kunert, Müller and Stehlík, and Steinberg), which is another good argument for using the compound criteria. We used degree-of-freedom efficiency to account for lack of fit. This does not explicitly avoid aliasing higher order terms with primary terms (Jones and Majumdar, Holling and Schwabe, and Steinberg). Other criteria suggested by discussants try to do this. Mean-square error criteria (Notz, and Holling and Schwabe) focus on prediction, rather than inference. The Bayesian procedure of DuMouchel and Jones (Atkinson and Chatfield) is one possibility and something like the criterion of Jones and Nachtsheim (2011), which explicitly tries to avoid aliasing, might tackle this problem more directly (Chatfield).

We are not convinced that this is the correct way to deal with lack of fit at the design stage, however. What good are unbiased estimators of main effects if there are several large interactions, which mean that any interpretation of main effects is misleading (Holling and Schwabe)? It is the size of departures from assumptions that is important (Morgan), rather than their existence. We prefer designs which can fit as many models as possible. Degree-of-freedom efficiency is a simple way. Others are Q_{B} (Tsai *et al.*, 2006), the percentile of the F-distribution (Vandebroek), or to obtain power against specific alternatives explicitly (Müller and Stehlík). Combining one of these with $(\mathrm{DP})_{\mathrm{S}}$, A_{S} and $(\mathrm{AP})_{\mathrm{S}}$ might be the way to go.

Waite and Woods make an important contribution by discussing the relevance of the pseudotrue parameter values. This provides a direct justification for the DP-criterion in the case where the assumed model cannot be guaranteed to be correct. This fits in directly with the usual response surface philosophy that polynomial models are always approximations and we want the fitted surface to be as close to the true surface as possible. This seems more relevant than biased estimation of specific parameters.

Additional properties

Vandebroek suggests that level balance is useful and that the D-criterion imposes this. This is not true in general, e.g. in saturated two-level designs, and if balance is desirable, e.g. for the sake of model selection, it would be better to use it directly in the compound criterion.

Morgan, Critchley, Chatfield, Stillman, and Manteinga and Porcu suggest checking for, or allowing for, variance heterogeneity. Careful consideration would have to be given to the alternatives to constant variance. In particular, would the variance depend on the mean response (Manteinga and Porcu), or the levels of treatment factors?

Model selection by using F-tests (Bailey and Maruri-Aguilar) could get out of hand, with many $(\mathrm{DP})_{\mathrm{S}}$- and A_{S}-criteria needed for all models and all nesting relationships between them. Are these formal hypothesis tests? If not, then perhaps including Q_{B} would be enough. For non-nested models, T (López-Fidalgo) is needed but is difficult to implement with many models.

Robustness to missing values (Ridout and Chatfield) might be achieved by including the maximum loss. This is more complicated than in Ahmad and Gilmour (2010), since missing values will affect each aspect of the analysis in different ways. Hence, some weight would have to be given to a missing value robust version of each individual criterion used in the compound criterion.

How many of these different criteria are actually needed in a compound criterion and how many will come out as by-products of others (Morgan and Chatfield) are fertile ground for further research.

References in the discussion

Ahmad, T. and Gilmour, S. G. (2010) Robustness of subset response surface designs to missing observation. *J. Statist. Plann Inf.*, **140**, 92–103.

American Chemical Society (2010) *Cornucopia*, spring.

Atkinson, A. C. and Cox, D. R. (1974) Planning experiments for discriminating between models (with discussion). *J. R. Statist. Soc.* B, **36**, 321–348.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS*. Oxford: Oxford University Press.

Atkinson, A. C. and Fedorov, V. V. (1975) The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57–70.

Azharul Islam, M., Sakkas, V. and Albanis, T. A. (2009) Application of statistical design of experiment with desirability function for the removal of organophosphorus pesticide from aqueous solution by low-cost material. *J. Haz. Mater.*, **170**, 230–238.

Bailey, R. A. (1999) Choosing designs for nested blocks. *List. Biometr.*, **36**, 85–126.

Bell, M. and Grunwald, G. K. (2004) Mixed models for the analysis of replicated spatial point patterns. *Biostatistics*, **5**, 633–648.

Bischoff, W. and Miller, F. (2006) Optimal designs which are efficient for lack of fit tests. *Ann. Statist.*, **34**, 2015–2025.

Boukouvalas, A. (2011) Emulation of random output simulators. *PhD Thesis*. University of Aston, Birmingham. (Available from `https://wiki.aston.ac.uk/foswiki/pub/AlexisBoukouvalas/WebHome/thesis.pdf`.)

Box, G. E. P. (1988) Signal-to-noise ratios performance criteria, and transformations. *Technometrics*, **30**, 1–17.

Box, G. E. P. and Draper, N. (1959) A basis for the selection of a response surface design. *J. Am. Statist. Ass.*, **54**, 622–653.

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

Box, G. E. P. and Draper, N. R. (1987) *Empirical Model-building and Response Surfaces*. New York: Wiley.

Box, G. E. P. and Draper, N. R. (2007) *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd edn. New York: Wiley.

Box, G. E. P., Hunter, J. S. and Hunter, W. G. (2005) *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd edn. Hoboken: Wiley-Interscience.

Carr, N. L. (1960) Kinetics of catalytic isomerization of n-pentane. *Industrl Engng Chem.*, **52**, 391–396.

Covey-Crump, P. A. K. and Silvey, S. D. (1970) Optimal regression designs with previous observations. *Biometrika*, **57**, 551–566.

Davison, A. C. (2003) *Statistical Models*. Cambridge: Cambridge University Press.

Derringer, G. and Suich, R. (1980) Simultaneous optimization of several response variables. *J. Qual. Technol.*, **12**, 214–219.

Dette, H. and Wong, W. K. (1999) Optimal designs when the variance is a function of the mean. *Biometrics*, **55**, 925–929.

Diggle, P. J. (2003) *Statistical Analysis of Spatial Point Patterns*. London: Arnold.

Diggle, P. J., Lange, N. and Benes, F. (1991) Analysis of variance for replicated spatial point patterns in clinical neuroanatomy. *J. Am. Statist. Ass.*, **86**, 618–625.

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

Eibl, S., Kess, U. and Pukelsheim, F. (1992) Achieving a target value for a manufacturing process: a case study. *J. Qual. Technol.*, **24**, 22–26.

Emmett, M. (2010) Design of experiments with multivariate response. *PhD Thesis*. University of Sheffield, Sheffield. Unpublished.

Emmett, M., Goos, P. and Stillman, E. C. (2011) A weighted prediction-based selection criterion for response surface designs. *Qual. Reliab. Engng Int.*, **27**, 719–729.

European Medicines Agency (1998) ICH Topic E 9: statistical principles for clinical trials—step 5. European Medicines Agency, London. (Available from `http://www.emea.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf`.)

Farrell, R. H., Kiefer, J. and Walbran, A. (1968) Optimum multivariate designs. In *Proc. 5th Berkeley Symp. Mathematical Statistics*, vol. 1, pp. 113–138. Berkeley: University of California Press.

Fisher, R. A. (1966) *The Design of Experiments*, 8th edn. New York: Hafner.

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

Goos, P., Tack, L. and Vandebroek, M. (2001) Optimal designs for variance function estimation using sample variances. *J. Statist. Planng Inf.*, **92**, 233–252.

Harington, J. (1965) The desirablity function. *Industrl Qual. Control*, **21**, 494–498.

Hines, W. G. S. (1996) Pragmatics of pooling in ANOVA tables. *Am. Statistn*, **50**, 127–139.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, **6**, 65–70.

Huber, P. (1973) Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

Illian, J., Penttinen, A., Stoyan, H. and Stoyan, D. (2008) *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester: Wiley.

Janky, D. G. (2000) Sometimes pooling for analysis of variance hypothesis tests: a review and study of a split-plot model. *Am. Statistn*, **54**, 269–279.

Jobson, J. D. and Fuller, W. A. (1980) Least squares estimation when the covariance matrix and parameter vector are functionally related. *J. Am. Statist. Ass.*, **75**, 176–181.

Jones, E. R. and Mitchell, T. J. (1978) Design criteria for detecting model inadequacy. *Biometrika*, **65**, 541–551.

Jones, B. and Nachtsheim, C. J. (2011) Efficient designs with minimal aliasing. *Technometrics*, **53**, 62–71.

Kiefer, J. (1959) Optimum experimental designs (with discussion). *J. R. Statist. Soc.* B, **21**, 272–319.

Kiefer, J. (1975) Construction and optimality of generalized Youden designs. In *A Survey of Statistical Design and Linear Models* (ed. J. N. Srivastava), pp. 333–353. Amsterdam: North-Holland.

Kiefer, J. and Wolfowitz, J. (1959) Optimum designs in regression problems. *Ann. Math. Statist.*, **30**, 271–294.

Läuter, E. (1974) Experimental design in a class of models. *Math. Operforsch. Statist.*, **5**, 379–398.

Läuter, E. (1976) Optimal multipurpose designs for regression models. *Math. Operforsch. Statist.*, **7**, 51–68.

Liu, S. and Wiens, D. (1997) Robust designs for approximately polynomial regression. *J. Statist. Planng Inf.*, **64**, 369–381.

López-Fidalgo, J., Tommasi, C. and Trandafir, P. C. (2007) An optimal experimental design criterion for discriminating between non-normal models. *J. R. Statist. Soc.* B, **69**, 231–242.

Lu, L. and Anderson-Cook, C. M. (2012) Rethinking the optimal response surface design for a first-order model with two-factor interactions, when protecting against curvature. *Qual. Engng*, to be published.

Lu, L., Anderson-Cook, C. M. and Robinson, T. J. (2011) Optimization of designed experiments based on multiple criteria utilizing a Pareto frontier. *Technometrics*, **53**, 353–365.

Martin, R. J., Platts, L. M., Seddon, A. M. and Stillman, E. C. (2003) The design and analysis of a mixture experiment on glass durability. *Aust. New Zeal. J. Statist.*, **45**, 19–27.

McGree, J. M. and Eccleston, J. A. (2008) Probability-based optimal design. *Aust. New Zeal. J. Statist.*, **50**, 13–28.

McGree, J. M., Eccleston, J. A. and Duffull, S. B. (2008) Compound optimal design criteria for nonlinear models. *J. Biopharm. Statist.*, **18**, 646–661.

Mead, R. (1988) *The Design of Experiments: Statistical Principles for Practical Application.* Cambridge: Cambridge University Press.

Mead, R., Bancroft, T. A. and Han, C.-P. (1975) Power of analysis of variance test procedures for incompletely specified fixed models. *Ann. Statist.*, **3**, 797–808.

Müller, W. G. and Stehlík, M. (2010) Compound optimal spatial designs. *Environmetrics*, **21**, 354–364.

Nelder, J. A. (1977) A reformulation of linear models (with discussion). *J. R. Statist. Soc.* A, **140**, 48–76.

Nelder, J. A. (1994) The statistics of linear models: back to basics. *Statist. Comput.*, **4**, 221–234.

Rafajlowicz, E. and Schwabe, R. (2003) Equidistributed designs in nonparametric regression. *Statist. Sin.*, **13**, 129–142.

Rafati, H. and Mirzajani, F. (2011) Experimental design and desirability function approach for development of novel anticancer nanocarrier delivery systems. *Pharamazie*, **66**, 31–36.

Scheffé, H. (1959) *The Analysis of Variance*. New York: Wiley.

Silvey, S. and Titterington, D. (1973) A geometric approach to optimal design theory. *Biometrika*, **60**, 21–32.

Stehlík, M., Fabián, Z. and Střelec, L. (2012) Small sample robust testing for Normality against Pareto tails. *Communs Statist. Simuln Computn*, to be published.

Stone, M. (1959) Application of a measure of information to the design and comparison of regression experiments. *Ann. Math. Statist.*, **30**, 55–70.

Trinca, L. A. and Gilmour, S. G. (1999) Difference variance dispersion graphs for comparing response surface designs with applications in food technology. *Appl. Statist.*, **48**, 441–455.

Trinca, L. A. and Gilmour, S. G. (2000) An algorithm for arranging response surface designs in small blocks. *Computnl Statist. Data Anal.*, **33**, 25–43; erratum, **40** (2002), 475.

Tsai, P.-W., Gilmour, S. G. and Mead, R. (2006) Three-level main-effects designs exploiting prior information about model uncertainty. *J. Statist. Planng Inf.*, **137**, 619–627.

Yeh, C.-H. (1986) Universal optimality of block designs. *Biometrika*, **73**, 701–706.

Zhu, Z. and Stein, M. L. (2005) Spatial sampling design for parameter estimation of the covariance function. *J. Statist. Planng Inf.*, **134**, 583–603.