Sizing Fixed Effects for Computing Power in Experimental Designs

Gary W. Oehlert
School of Statistics
University of Minnesota

Pat Whitcomb
Stat-Ease, Inc.

March 2, 2001

Corresponding author:
Gary W. Oehlert
School of Statistics
313 Ford Hall
224 Church St.
Minneapolis, MN 55455-0493

Voice: 612-625-1557
Fax: 612-624-8868
email: gary@stat.umn.edu

Summary

Power tells us the probability of rejecting the null hypothesis for an effect of a given size and helps us select an appropriate design prior to running the experiment. The key to computing power for an effect is determining the size of the effect. We describe a general approach for sizing effects that covers a wide variety of designs including factorials with categorical levels, response surfaces, mixtures, and crossed designs.

Key words: power, type 2 error, fixed effects, factorials, response surfaces, mixtures

Short Bio Gary W. Oehlert is a Professor in the School of Statistics at the University of Minnesota. He has taught experimental design for 15 years and is the author of the recently published text *A First Course in Design and Analysis of Experiments*. His other research interests include statistical computing, environmental statistics, and brain imaging. Pat Whitcomb is President and founder of Stat-Ease, Inc., the developers of the Design-Expert ® software. Pat began as a chemical engineer, has CQE and PE credentials, and has written and consulted in design of experiments and statistical methods for 20 years.

Introduction

We design and run experiments to answer questions. For example, an orange-flavored soft drink is an emulsion with the essential oil of orange in a carbonated sugar-water solution. In order to give the product a long shelf life, we add an powdered gum acacia as an emulsifier. The raw gum can be treated in a variety of different ways prior to its use in the soft drink. We want to know whether or not to demineralize the gum prior to its use in the soft drink, so we run an experiment to answer our question.

Our question is expressed as the null hypothesis that demineralization has no effect on the emulsification properties of gum acacia. There is also an alternative hypothesis, which describes the state of nature if the null hypothesis is incorrect. For example, we might believe that demineralization could only improve the emulsification properties (a one-sided alternative), or we might believe that demineralization could improve or degrade the emulsification properties (a two-sided alternative).

We use experimental information to decide between the null and alternative hypotheses, rejecting the null hypothesis if we observe data that are consistent with the alternative and sufficiently inconsistent with the null, or failing to reject the null hypothesis if the data are consistent with the null. The usual measure for assessing consistency with the null or alternative hypotheses is the p-value of a significance test. The p-value is the probability, assuming that the null is true, of observing data as extreme or more extreme than the data we did observe. "Extreme" in this context means unlike the null and like the alternative. When the p-value is small enough, we reject the null hypothesis in favor of the alternative.

We can make two kinds of errors that we can make. First, we can reject the null hypothesis when the null is correct; this is a Type 1 error. Second, we can fail to reject the null hypothesis when the null is incorrect; this is a Type 2 error. The Type 1 error rate (usually denoted α) is the probability of rejecting the null hypothesis when the null is true. The Type 1 error rate is completely under the control of the experimentor; if the null is rejected when the p-value is 5% or less, then the Type 1 error rate is 5%. The Type 2 error rate (usually denoted β) is the probability of failing to reject the null hypothesis when the null is false. The Type 2 error rate is more complicated, because β depends on α, the experimental design (including sample sizes), the size σ^2 of experimental errors, and the precise set of parameters in the alternative model. In general, larger sample sizes, smaller experimental errors, or more extreme alternatives decrease the Type 2 error rate.

The power of a test is the probability of rejecting the null hypothesis when the null is false; power is simply $1 - \beta$. An experiment that is so small that we are unlikely to detect effects of interest is wasteful of resources. A too-large (and thus "too powerful) experiment may also be wasteful, because it uses more resources

than are likely to be needed for detecting the alternative. The sweet spot for power is the range from about .8 to .95. Power below this range means that there is a substantial chance of missing effects of interest, and the additional resources spent to increase power above this range give little return for the investment (from a testing perspective). We will not always be able to achieve such power for all null hypotheses of interest.

Mathematical Basis for Power

Most designed experiments are analyzed using a regression-like model. Suppose that we have n observations $y_1, y_2, \ldots, y_n$. The ith observation y_i is associated with p predictive variables $x_{i1}, x_{i2}, \ldots, x_{ip}$, and the expected value of y_i is

$$\mathrm{E}(y_i) = \sum_{j=1}^{p} x_{ij}\gamma_j \quad ,$$

where the γ_j's are unknown parameters. This specifies the mean structure for the data. The predictive variables x_{ij} are determined by the kind of model being fit and are known to the modeler.

In addition to the mean structure, there is also an error structure. The observation y_i equals its expected value plus random variation ϵ_i. The random variation ϵ_i is assumed to have a normal distribution with mean 0 and variance σ^2. We further assume that all the errors ϵ_i are independent of each other.

Let $\boldsymbol{y}$ be the n-vector containing all the responses, $\boldsymbol{y}^{\mathrm{T}} = (y_1, y_2, \ldots, y_n)$; let $\boldsymbol{X}$ be the n by p matrix with i, j element x_{ij}; let $\boldsymbol{\epsilon}$ be the n-vector containing all the errors, $\boldsymbol{\epsilon}^{\mathrm{T}} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$; and let $\boldsymbol{\gamma}$ be the p-vector of parameters, $\boldsymbol{\gamma}_{\mathrm{T}} =$

$(\gamma_1, \gamma_2, \ldots, \gamma_p)$. Then the model is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon} \quad .$$

We assume that the matrix $\boldsymbol{X}$ has full rank.

Consider null hypotheses of the form: $h_0 : \gamma_{k_1} = \gamma_{k_2} = \cdots = \gamma_{k_m} = 0$ for some set of indices $1 \leq k_1 < k_2 < \cdots < k_m \leq p$. This is a test that m of the coefficients in the linear model are zero. More general null hypotheses can be formulated, but this type is sufficient for our purposes. Let the matrix $\boldsymbol{X}$ be rearranged as

$$\boldsymbol{X} = [\boldsymbol{X}_n : \boldsymbol{X}_a] \quad ,$$

where $\boldsymbol{X}_a$ consists of columns $k_1, k_2, \ldots, k_m$ and $\boldsymbol{X}_n$ consists of the remaining columns. Construct $\boldsymbol{\gamma}_a$ and $\boldsymbol{\gamma}_n$ in an analogous way. Matrix $\boldsymbol{X}_a$ consists of those columns of $\boldsymbol{X}$ that have nonzero coefficients γ_a only under the alternative hypothesis.

Example 1. Orange drink

Suppose that we wish to test the effect of demineralizing gum acacia used as an emulsifier on the shelf life of an orange drink emulsion. The experiment will have six treatments formed as the factor/level combinations of demineralization (yes or no) and source of the gum (suppliers 1, 2, or 3). We make twelve emulsions, two for each of the six treatments. Our model will include an overall constant, a main effect for source, and a main effect for demineralization. There are many possible $\boldsymbol{X}$ matrices that we could use; one is shown in Table 1.

[Table 1 about here.]

In this matrix, the coefficient of the first column is the overall constant, the coefficients of the second and third columns are the main effects for suppliers 1 and 2, and the coefficient of the fourth column is the main effect for the first level of demineralization (yes). We use the usual restriction that main effects sum to zero.

When testing the null hypothesis that demineralization has no effect (a coefficient of zero), the null matrix $\boldsymbol{X}_n$ is the first three columns of $\boldsymbol{X}$, and the alternative matrix $\boldsymbol{X}_a$ is the last column of $\boldsymbol{X}$.

Test H_0 with an F-test. First fit the data $\boldsymbol{y}$ to a reduced model that contains only the columns in the matrix $\boldsymbol{X}_n$. Let SSR_n be the residual sum of squares for this fit. Next fit the data $\boldsymbol{y}$ to the full model $\boldsymbol{X}$, and let SSR_f be the residual sum of squares for this fit. The improvement sum of squares is $SS_i = SSR_n - SSR_f$ and has m degrees of freedom; the mean square $MS_i = SS_i/m$. The F-test for H_0 is

$$F = \frac{MS_i}{MS_E} \quad ,$$

where $MS_E = SSR_f/(n-p)$ is the mean square for error from the full model. This observed F-ratio should be compared with an F distribution with m and $n-p$ degrees of freedom to compute the p-value.

Reject the null hypothesis if the observed F-ratio is larger than $F_{\alpha,m,n-p}$, the upper α percent point of the F-distribution with m and $n-p$ degrees of freedom. To compute power, we need the distribution of the F-ratio under the alternative hypothesis. Let $\boldsymbol{\mu}$ be the expected value of $\boldsymbol{y}$:

$$\boldsymbol{\mu} = \boldsymbol{X}\boldsymbol{\gamma} = \boldsymbol{X}_n\boldsymbol{\gamma}_n + \boldsymbol{X}_a\boldsymbol{\gamma}_a = \boldsymbol{\mu}_n + \boldsymbol{\mu}_a \quad .$$

Let $\boldsymbol{\mu}_r$ be the residuals obtained when we fit $\boldsymbol{\mu}$ to the null hypothesis columns $\boldsymbol{X}_n$. When the alternative is true, the F-ratio has a *noncentral F distribution* with

degrees of freedom m and $n-p$ and *noncentrality parameter* ζ, where

$$\zeta = \frac{\boldsymbol{\mu}_r^{\mathrm{T}} \boldsymbol{\mu}_r}{\sigma^2} \quad .$$

Power is thus the probability that a noncentral F distribution with m and $n-p$ degrees of freedom and noncentrality parameter ζ is larger than the cutoff $F_{\alpha,m,n-p}$.

Figure 1 illustrates these distributions. The solid curve shows the density of the F-ratio when the null distribution is true, and the vertical line is the 5% cutoff value. The dashed line is the density of the F-ratio under one particular alternative (that is, a noncentral F). The area under the dashed line to the right of the cutoff is power.

[Figure 1 about here.]

Example 2. Orange drink (continued)

Suppose that the effect of demineralization is to increase the shelf life by 10 days; then $\boldsymbol{\gamma}_a = 5$. Assume further that σ is 20 days. Then we need to regress

$$\boldsymbol{\mu}_a = \boldsymbol{X}_a \boldsymbol{\gamma}_a = \begin{bmatrix} 1 \\ 1 \\ -1 \\ \vdots \\ 1 \\ -1 \\ -1 \end{bmatrix} 5 = \begin{bmatrix} 5 \\ 5 \\ -5 \\ \vdots \\ 5 \\ -5 \\ -5 \end{bmatrix}$$

on $\boldsymbol{X}_n$. For this design, the residuals $\boldsymbol{\mu}_r = \boldsymbol{\mu}_a$, because the columns of the matrix $\boldsymbol{X}$ are orthogonal. The noncentrality parameter is

$$\zeta = \frac{\boldsymbol{\mu}_r^{\mathrm{T}} \boldsymbol{\mu}_r}{\sigma^2} = \frac{12 \times 25}{20^2} = .75.$$

The F test for demineralization has 1 and 8 degrees of freedom; the upper 5% point of an F with 1 and 8 degrees of freedom is 5.318. The probability of rejecting the null hypothesis when $\zeta = .75$ is the probability that a noncentral F with 1 and 8 degrees of freedom and noncentrality parameter .75 exceeds 5.318. This probability is .119, which is our power (and is pretty low).

Power increases with larger noncentrality parameters. The noncentrality parameter becomes larger when either $\boldsymbol{\mu}_r^{\mathrm{T}}\boldsymbol{\mu}_r$ increases or when the error variance σ^2 becomes smaller. The term $\boldsymbol{\mu}_r^{\mathrm{T}}\boldsymbol{\mu}_r$ gets bigger when the alternative parameter $\boldsymbol{\gamma}_a$ becomes bigger or when the sample sizes increase (so that there are more terms being added together). Blocking designs or changed experimental technique can be used to decrease σ.

Here is an alternative way to think about the noncentrality parameter: $\boldsymbol{\mu}_r/\sigma$ is the residuals when $\boldsymbol{\mu}_a/\sigma = \boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ is regressed on $\boldsymbol{X}_n$. Thus ζ is the residual sum of squares when $\boldsymbol{\mu}_a/\sigma$ is regressed on $\boldsymbol{X}_n$. The ratio $\boldsymbol{\mu}_a/\sigma$ can be thought of expressing the "alternative" portion of the mean in multiples of standard deviations.

In summary, the steps for finding power are:

1. Choose the Type 1 error rate α.
2. Find the design matrices $\boldsymbol{X}$, $\boldsymbol{X}_n$, and $\boldsymbol{X}_a$.
3. Find the degrees of freedom for the F test and determine the critical value for the F test.
4. Choose $\boldsymbol{\gamma}_a/\sigma$, the alternative parameters for which we wish to compute power.

5. Compute the noncentrality parameter ζ as the sum of squared residuals of $\boldsymbol{\mu}_a/\sigma = \boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ regressed on $\boldsymbol{X}_n$.

6. Compute power as the probability that a noncentral F with this noncentrality and these degrees of freedom exceeds the central F critical value.

Conceptually, the most difficult step is 4, choosing the alternative for which you wish to computer power. Historically, the most difficult computational step is the last, as software for computing noncentral F probabilities has not been widespread, and tables of the noncentral F are rather sparse. Modern software (for example, [1], [2], [3], [4], and [5]) can compute noncentral F probabilities or powers directly.

Sizing Effects

Computing the noncentrality parameter is the key to computing power. We compute power for a specific size of alternative, so we must be able to specify interesting sizes for effects. Here "interesting" may have several interpretations. For example, an effect may be interesting because it is

- what we anticipate will occur,
- the smallest effect of practical use, or
- an important special case.

In this paper we suggest a unified approach for describing the sizes of effects across a broad range of experimental designs, including response surfaces, mixture designs, and factorials. We call this *sizing an effect*.

The noncentrality parameter depends on $\boldsymbol{\mu}_r/\sigma$, which is a function of $\boldsymbol{X}_n$, $\boldsymbol{X}_a$, $\boldsymbol{\gamma}_a$, and σ. Matrices $\boldsymbol{X}_n$ and $\boldsymbol{X}_a$ depend on the specific experimental design we are using, so they should not be involved in a generic description of the size of an effect. For example, the power for estimating a linear effect of time with observations at 75, 100, and 125 seconds will differ from that with observations at 90, 100, and 110 seconds, even if the linear effect (coefficient in response units per second) is the same in the two designs. So the power for detecting an effect of a given size depends on the experimental design, but the way we describe the size of an effect should not depend on the design.

Many response surface or mixture designs work in coded units, pseudocomponents, or the like. For example, response surfaces are usually centered at 0 with steps up and down from the center coded to 1. Our descriptions of effect sizes are based on coded units for response surfaces and pseudocomponents for mixture designs, rather than on the original units. This means that our descriptions of effect sizes also depend on how we do the coding, so that an effect of size 1 standard deviation in one design is equivalent to an effect of size 1 standard deviation in a second design *only if* the codings for the two designs are the same.

The description of effect size is generally fairly easy when the alternative matrix $\boldsymbol{X}_a$ contains only a single column. In that case, there is typically only a single value of $\boldsymbol{\gamma}_a/\sigma$ that yields an effect of a given size. When there is more than one column in $\boldsymbol{X}_a$, there may be infinitely many different parameter sets $\boldsymbol{\gamma}_a/\sigma$ that lead to an effect of a given size. In that case, we evaluate power by a *principle of conservatism*: report as power for a given size effect the smallest possible power among all those effects with the given size.

Response surface models

Work in coded units, with 0 at the center of the design, and unit steps for each design variable. Evaluate the size of an effect by looking at its range over the square where the design variables run from –1 to 1. (This will be a cube or hypercube with more than two design variables.) Divide this range by σ to determine the size of the effect in multiples of standard deviations. For example, consider a linear effect with coefficient 5 and $\sigma = 4$. The effect ranges from –5 to +5 over the evaluation cube giving a range of 10; dividing this by $\sigma = 4$ give us a linear effect of size 2.5 standard deviations. Now consider a pure quadratic effect with the same coefficient of 5. This quadratic effect will range from 0 to 5 over the evaluation cube, and dividing by 4 gives a quadratic effect of size 1.25 standard deviations.

Think of $\boldsymbol{X}_e$ as being the term of interest observed all across the evaluation region. The size of an effect is the range of $\boldsymbol{X}_e\boldsymbol{\gamma}_a/\sigma$, or equivalently, $\boldsymbol{\gamma}_a/\sigma$ times the range of $\boldsymbol{X}_e$. Table 2 gives these ranges for standard polynomial effects.

[Table 2 about here.]

For example, a linear effect of 2 standard deviations has $\boldsymbol{\gamma}_a/\sigma = 1$, and a pure quadratic effect of 2 standard deviations has $\boldsymbol{\gamma}_a/\sigma = 2$.

Note that the evaluation region for sizing an effect (the cube or $\boldsymbol{X}_e$) is divorced from the actual experimental design that is used ($\boldsymbol{X}_n$ and $\boldsymbol{X}_a$). We can thus compare power for different designs in the context of the same effect size, provided both designs use the same transformation into coded variables. Effect size gives us $\boldsymbol{\gamma}_a/\sigma$, from which we can compute $\boldsymbol{\mu}_a/\sigma = \boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$.

[Figure 2 about here.]

Example 3. Rotatable central composite

Consider the rotatable, two-factor, central composite design with five center points shown in coded units in Figure 2. Factor A is time, factor B is temperature, and the response is the yield of the process (in percent). The second-order model includes a constant, first-order terms A and B, and second-order terms AB, A^2, and B^2. The design matrix for this model is given in Table 3.

[Table 3 about here.]

where the columns are in the same order as the terms were listed above.

Suppose that we wish to test the null hypothesis that the B^2 effect is zero, that is, no pure quadratic effect of temperature. Then the first five columns of the design matrix are $\boldsymbol{X}_n$ and the last column is $\boldsymbol{X}_a$. We thus have $\boldsymbol{\mu}_a/\sigma = \boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$, where

$$\boldsymbol{X}_a^{\mathrm{T}} = [1, 1, 1, 1, 0, 0, 2, 2, 0, 0, 0, 0, 0]\,.$$

On the evaluation region, a pure quadratic effect ranges from 0 to 1, so that this is a one SD effect when $\boldsymbol{\gamma}_a/\sigma = 1$ and a two SD effect when $\boldsymbol{\gamma}_a/\sigma = 2$.

To compute power we need to size our effect. Similar experiments in the past have had a error standard deviation of about 5%, so we anticipate the same standard deviation in this experiment. A change in yield of 5% is important economically, so we will use a 5% change as the size of our alternative. This is a 1 SD effect. To compute power for a 1 SD sized effect we need the noncentrality parameter, computed as the residual sum of squares when we regress $\boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ on the columns in $\boldsymbol{X}_n$. The SSR is 6.957. The .05 critcal value for an F with 1 and 7 degrees of freedom is 5.59, and the probability that a noncentral F with 1 and 7 degrees of freedom and noncentrality parameter 6.957 exceeds 5.59 is .621.

Thus the power for this 1 SD effect is .621. For a 2 SD effect, the noncentrality parameter is 27.83, and the power is .994.

Modern software relieves us of the computational burden. Figure 3 shows the power evaluation output for this problem from the Design-Expert® package [1].

[Figure 3 about here.]

Now suppose that we are not working with the full second order model, but instead have previously determined that the pure quadratic effect in time should not be used in the model. Then we have a model with terms A, B, AB, and B^2 (plus the constant), and we again wish to test the null hypothesis that the B^2 effect (quadratic in temperature) is zero, a test with 1 and 8 degrees of freedom. This is the same null hypothesis as before, but now the null model only contains four degrees of freedom (the first four columns of $\boldsymbol{X}$) instead of five. The 1 SD effect is still when $\boldsymbol{\gamma}_a/\sigma = 1$; changing the null model does not change the size of an effect. We again get the residual sum of squares when $\boldsymbol{\mu}_a/\sigma$ is regressed on the (now four column) null matrix $\boldsymbol{X}_n$; this SSR is 7.077, and the power is now .646.

[Figure 4 about here.]

Example 4. Nonrotatable central composite design

Consider now an alternate design for fitting the second order response surface in time and temperature from Example 3. This new design is a nonrotatable, two-factor, central composite design with five center points using axial points that are at a distance 1 from the center. This design is shown in coded units in Figure 4 and can be recognized as a 3^2 with replicated center points; it is also the same as

the design in Example 3 except that the axial points have been pulled back to the cube. The second order $\boldsymbol{X}$ matrix for this design is given in Table 4.

[Table 4 about here.]

We still wish to test the null hypothesis that there is no quadratic effect of temperature (B^2), we're just using a different design. We can compare the two designs by comparing their powers for testing the same null and alternative hypothesis. A one SD sized effect for this quadratic in temperature occurs when $\boldsymbol{\gamma}_a/\sigma = 1$, the same as for the rotatable design.

To compute the noncentrality parameter we need the SSR when the $\boldsymbol{\mu}_a/\sigma$ is regressed on the other five columns; this noncentrality parameter is 2.762. The power for testing with this one SD effect is only .301, much less than the power in Example 3.

Example 5. Two-degrees-of-freedom

Consider now testing the null hypothesis that there is no pure quadratic effect of either time or temperature, that is, that A^2 and B^2 have zero coefficients. This is a two-degree-of-freedom null hypothesis. We are still assuming that $\sigma = 5\%$, and we still wish to test against the alternative that the two pure quadratic terms together cause a 5% change in the response, a 1 SD effect. The alternative contribution to the scaled response mean is

$$\boldsymbol{\mu}_a/\sigma = A^2\boldsymbol{\gamma}_{a1}/\sigma + B^2\boldsymbol{\gamma}_{a2}/\sigma$$

The evaluation region for this effect is the square $(-1,1) \times (-1,1)$, and on this region $\boldsymbol{\mu}_a/\sigma$ has a range of 1 when

$$|\boldsymbol{\gamma}_{a1}/\sigma| + |\boldsymbol{\gamma}_{a2}/\sigma| = 1.$$

Any pair of parameters $\boldsymbol{\gamma}_{a1}, \boldsymbol{\gamma}_{a2}$ with $|\boldsymbol{\gamma}_{a1}/\sigma| + |\boldsymbol{\gamma}_{a2}/\sigma| = 1$ gives a 1 SD effect.

To compute the noncentrality parameter, we first regress the columns of the alternative portion of the matrix $\boldsymbol{X}_a$ on the null portion of the design matrix $\boldsymbol{X}_n$ and get a matrix of residuals $\boldsymbol{X}_r$. Using the design in Example 4, the last two columns form $\boldsymbol{X}_a$, the first four columns form $\boldsymbol{X}_n$, leading to $\boldsymbol{X}_r$ as shown in Table 5.

[Table 5 about here.]

The noncentrality parameter ζ is

$$\begin{aligned}
\zeta &= (\boldsymbol{\gamma}_a/\sigma)^{\mathrm{T}} \quad \boldsymbol{X}_r^{\mathrm{T}}\boldsymbol{X}_r \quad (\boldsymbol{\gamma}_a/\sigma) \\
&= (\boldsymbol{\gamma}_{a1}/\sigma, \boldsymbol{\gamma}_{a2}/\sigma) \begin{bmatrix} 3.2308 & 1.2308 \\ 1.2308 & 3.2308 \end{bmatrix} \begin{pmatrix} \boldsymbol{\gamma}_{a1}/\sigma \\ \boldsymbol{\gamma}_{a2}/\sigma \end{pmatrix}
\end{aligned}$$

Now introduce the notation $\lambda_i = \boldsymbol{\gamma}_{ai}/\sigma$; our conservative noncentrality parameter is the minimum of

$$(\lambda_1, \lambda_2) \begin{bmatrix} 3.2308 & 1.2308 \\ 1.2308 & 3.2308 \end{bmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix}$$

over the set where $|\lambda_1| + |\lambda_2| = 1$. The sign of $\boldsymbol{\mu}_a$ is irrelevant, so we may choose $\lambda_1 > 0$ without loss of generality. Then either $\lambda_2 = 1 - \lambda_1$ or $\lambda_2 = -1 + \lambda_1$. In the first case, we have

$$\begin{aligned}
\zeta &= 3.2308\lambda_1^2 + 2 \times 1.2308\lambda_1(1-\lambda_1) + 3.2308(1-\lambda_1)^2 \\
&= [3.2308 - 2 \times 1.2308 + 3.2308]\lambda_1^2 + 2[1.2308 - 3.2308]\lambda_1 + 3.2308 \\
&= 4\lambda_1^2 - 4\lambda_1 + 3.2308
\end{aligned}$$

This is minimized at $\lambda_1 = .5$. In the second case,

$$\begin{aligned}\zeta &= 3.2308\lambda_1^2 + 2 \times 1.2308\lambda_1(-1+\lambda_1) + 3.2308(-1+\lambda_1)^2 \\ &= [3.2308 + 2 \times 1.2308 + 3.2308]\lambda_1^2 - 2[1.2308 + 3.2308]\lambda_1 + 3.2308 \\ &= 8.9231\lambda_1^2 - 8.9231\lambda_1 + 3.2308\end{aligned}$$

which is again minimized at $\lambda_1 = .5$. The minimum in the first instance is $8.9231/4 = 2.2308$, whereas the minimum in the second instance is $(6.4616 - 2.4616)/4 = 1$. Thus the conservative noncentrality parameter is 1.

The F test has 2 and 7 degrees of freedom, and the 5% critical value is 4.737. The (minimum) power for testing the null hypothesis of no pure quadratic effects of time and temperature against the alternative that they are a 1 SD effect is the probability that a noncentral F with 2 and 7 degrees of freedom and noncentrality parameter 1 exceeds 4.737, or .140.

Mixture designs

The components of a k-component mixture are nonnegative and sum to a fixed total T. For example, three components of a detergent mixture are water, alcohol, and urea, and they must constitute 9% of the complete detergent. One possible mixture is 4% water, 3% alcohol, and 2% urea. We refer to this representation of the components as the *actual* variables x_{ai}. The components may also be normalized (by division by T) so that they sum to 1; we refer to this representation of the components as the *real* variables $x_{ri} = x_{ai}/T$. The above example would be .444 water, .333 alcohol, and .222 urea when expressed in real variables. Of

course, if the total T of the actuals is already 1, then the reals and the actuals are the same. In some problems, there is a minimum allowable value for each component; let d_i be the minimum for component i in actuals. The *pseudocomponents* are $x_{pi} = (x_{ai} - d_i)/(T - \sum_i d_i)$; these pseudocomponents are nonnegative and sum to 1. In our detergent example, the minima for the three components are 3%, 2%, and 2%, for a $\sum_i d_i$ of 7%. Our example in pseudocomponents is water $(4\% - 3\%)/(9\% - 7\%) = .5$, alcohol $(3\% - 2\%)/(9\% - 7\%) = .5$, and urea $(2\% - 2\%)/(9\% - 7\%) = .0$. If the minima d_i are all 0, then the pseudocomponents are the same as the real components.

The pseudocomponents form a standard representation of the allowable design region and coincide with the full k-simplex of nonnegative numbers that sum to 1. Thus the natural evaluation region for mixture effects is the simplex of the pseudocomponents. We use the full simplex in pseudocomponents as the evaluation region for sizing effects even if the design we use lies on a subset of the simplex. Once again, divorce the evaluation region for evaluating effects from the actual experimental design that is implemented.

We will discuss sizing of effects for the so-called Scheffé parameterization of the mixture model. All variables occur in the Scheffé parameterization, but not all terms in these variables are needed due to the constraint that the variables add to 1. Because of the special structure of mixture models, a first-order coefficient of zero corresponds to a zero response when the mixture is pure in that component; thus we do not ordinarily test for first-order effects being zero. We discuss testing first-order effects below.

The size of a second- or higher-order effect is its range (measured in standard deviations) over the evaluation region. This is the range of $\boldsymbol{X}_e\boldsymbol{\gamma}_a/\sigma$, or equiv-

alently, $\boldsymbol{\gamma}_a/\sigma$ times the range of $\boldsymbol{X}_e$. The range of $\boldsymbol{X}_e$ over the full simplex is determined by term type, as shown in Table 6.

[Table 6 about here.]

Thus a quadratic term with coefficient equal to two standard deviations is a .5 SD effect, and a special cubic term with coefficient equal to 27 standard deviations is a 1 SD effect.

[Figure 5 about here.]

Example 6. An augmented simplex-lattice

Consider using the three-component mixture design shown in Figure 5 to study how the detergent formulation affects turbidity; this design is an augmented simplex-lattice with four replicated points. The variables are A—water, B—alcohol, and C—urea. The second order model includes terms A, B, C, AB, AC, and BC; the design matrix for this model is shown in Table 7.

[Table 7 about here.]

This is a design in the pseudocomponents, so (1,0,0) corresponds to 5% water, 2% alcohol, and 2% urea.

Suppose first that we wish to test that there is no alcohol/urea interaction effect on turbidity, or, equivalently, that the coefficient of BC is zero. Previous studies of turbidity have had error σs of about 150, so we will base our power computations on that. Turbidity changes are not important until they are about 250, so we will check the power for a $250/150 = 1.67$ SD effect. BC is a quadratic term, so a multiplier of 6.67 corresponds to a 1.67 SD effect. The noncentrality parameter is

the error sum of squares when 6.67 times the BC column of $\boldsymbol{X}$ is regressed on the other columns: 2.869. The F test for BC has 1 and 8 degrees of freedom, so the power will be the probability that a noncentral F with 1 and 8 degrees of freedom and noncentrality 2.869 exceeds the upper 5% point of a central F with 1 and 8 degrees of freedom (5.318). The power is .321.

Example 7. A constrained mixture design

Now consider what happens to power in a mixture when the mixture components are constrained. In this example, we have a three component mixture with the total in actuals equal to 9%. Suppose that the actuals are constrained by the following: $5.8\% \leq A + B$, $5.8\% \leq A + C$, and $5.2\% \leq B + C$. Reexpressed in the pseudocomponents, these constraints are $.4 \leq A + B$, $.4 \leq A + C$, and $.6 \leq B + C$. This constrains the design space to be a subset of the full simplex in the pseudocomponents, and we must use only points that meet these constraints.

[Figure 6 about here.]

Figure 6 is a sample design that meets the constraints.

Note that the coding for this design into pseudocomponents is the same as for the unconstrained design. Thus we may compare power between the constrained and unconstrained designs. If the constraints we add imply different lower bounds on the variables than what we had before, we would get a new coding into pseudocomponents (because there are new minima), and the powers for the constrained and unconstrained designs would not be directly comparable due to the different codings into pseudocomponents.

The design matrix $\boldsymbol{X}$ for this design and the second order model is shown in Table 8.

[Table 8 about here.]

Suppose again that we wish to test that BC (alcohol by urea) has a zero coefficient when, in fact, it is a 1.67 SD effect. The multiplier for a 1.67 SD BC mixture effect is 6.67, so the noncentrality parameter will be the residual sum of squares when 6.67 times the last column of $\boldsymbol{X}$ is regressed on the first 5 columns; this SSR is .552. The test has 1 and 8 degrees of freedom, so the power is the probability that a noncentral F with 1 and 8 degrees of freedom exceeds the upper 5% point of a central F with 1 and 8 degrees of freedom (5.318); the power is .101, not much bigger than the minimum possible power of .05.

Basically two things conspire to make the power of constrained mixtures small. First, the standard error of the coefficient estimates is inflated as the range of the variables in the design shrinks and the regression variables become more colinear. This makes it more difficult to detect any effects that are there. Second, as we constrain the design more and more (but keeping the same coding into pseudo-components!) the range of any effect observed over the constrained region gets smaller and smaller. Thus the variability that we are trying to detect is getting smaller as well.

Note that even though it is difficult to detect that individual coefficients are nonzero, the least squares fit is still an unbiased fit of the surface. Furthermore, the probability that the overall model will be found significant is considerably greater than the power for detecting individual terms in the model.

Testing first-order effects in mixtures

An inert mixture component has a first-order coefficient equal to the average of the first-order coefficients of the active components. Thus the appropriate null hypothesis for a first order mixture effect is that its coefficient is equal to the average of the other first-order coefficients, not that its coefficient is zero. We can put this test into our framework by a suitable modification of $\boldsymbol{X}_n$.

Suppose that we wish to test the null hypothesis that A has no first-order effect in a k-component mixture. The usual null matrix $\boldsymbol{X}_n$ contains the columns for the other first order effects (and any higher order effects that may be included in the model). The modified null matrix $\tilde{\boldsymbol{X}}_n$ replaces each first-order column by that column plus the A column divided by $k-1$. The alternative matrix $\boldsymbol{X}_a$ is not modified. We then proceed as before, using $\tilde{\boldsymbol{X}}_n$ in place of $\boldsymbol{X}_n$.

Example 8. Testing a first-order mixture effect

We wish to test that water (A) has no first-order effect when we fit the first-order model using the unconstrained simplex design of Example 6. The matrix $\tilde{\boldsymbol{X}}_n$ is found by replacing B with B + A/2 and with C + A/2, as shown in Table 9.

[Table 9 about here.]

A 1 SD first-order effect has $\boldsymbol{\gamma}_a/\sigma = 1$ (because the A column ranges from 0 to 1). To get the noncentrality parameter we need to regress $\boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ on $\tilde{\boldsymbol{X}}_n$ and get SSR, here 1.691. Power is the probability that a noncentral F with this noncentrality parameter and 1 and 11 degrees of freedom exceeds the percent point of the central F with 1 and 11 degrees of freedom (4.844); this power is

.221.

Two-series factorials

Two-series factorials are factorials with k factors, each factor having just two levels. We model two-series factorials using main effects, two-factor interactions, three-factor interactions, and so on. The main effect of a factor adds δ_1 to the expected response at the low level of the factor and δ_2 to the expected response at the high level, with the constraint that $\delta_1 + \delta_2 = 0$, or equivalently, $\delta_2 = -\delta_1$. The range of a main effect across all levels is $2|\delta_1|$, so the size of a main effect is $2|\delta_1|/\sigma$ standard deviations.

We define the size of an interaction effect in a similar way. An interaction effect in a two-series factorial is a quantity, κ_{ijk} for example, that is either added or subtracted from the expected response at each factor level combination. Thus the range of the interaction effect is $2|\kappa_{ijk}|$, so the size of the interaction effect is $2|\kappa_{ijk}|/\sigma$ standard deviations.

If we code the low levels of the factorial at -1 and the high levels at $+1$, then this view of the effect size for two-series factorials is exactly equivalent to how we size effects for response surface designs. Two-factor and higher interactions in the factorial correspond to two-way and higher cross products of the quantitative factors in the response surface design. Nothing new needs to be done.

General factorials

General factorials have factors with nominal (categorical) levels that are not arranged along a numerical scale. When factors have more than two levels, then main effects and interactions have more than one degree of freedom, and we cannot use the trick we employed in two-series factorials and embed the problem in a response surface. We must instead define effect size in a different way.

Consider a three-way factorial with treatment means μ_{ijk} expanded into main effects and interactions as

$$\mu_{ijk} = \mu + \rho_1 + \tau_j + \omega_k + \rho\tau_{ij} + \rho\omega_{ik} + \tau\omega_{jk} + \rho\tau\omega_{ijk}$$

Each of these main effects (for example, τ_j) or interactions (for example, $\rho\omega_{ik}$) sums to zero across any subscript. A main effect is of size 1 if one pairwise difference $\rho_i - \rho_{i'}$ equals 1 and all other pairwise differences are no more than 1 (in absolute value). A two-factor interaction is of size 1 if a quartet difference $\rho\tau_{ij} - \rho\tau_{i'j} - \rho\tau_{ij'} + \rho\tau_{i'j'}$ is one and all other quartet differences are no more than one (in absolute value). A three-factor interaction is of size 1 if one octet difference $\rho\tau\omega_{ijk} - \rho\tau\omega_{i'jk} - \rho\tau\omega_{ij'k} + \rho\tau\omega_{i'j'k} - \rho\tau\omega_{ijk'} + \rho\tau\omega_{i'jk'} + \rho\tau\omega_{ij'k'} - \rho\tau\omega_{i'j'k'}$ is one and all other octet differences are no more than 1 (in absolute value). Higher-order terms are defined similarly.

Each main effect or interaction can also be expressed in matrix notation as the range of some matrix $\boldsymbol{X}_A$, $\boldsymbol{X}_{BC}$, etc. The main effect of A is of size 1 SD if the range of $\boldsymbol{X}_A\boldsymbol{\gamma}_A/\sigma$ is 1, and the interaction BC has size 1 SD if the range of $\boldsymbol{X}_{BC}\boldsymbol{\gamma}_{BC}/\sigma$ is 1. When testing the BC interaction, for example, the matrix $\boldsymbol{X}_{BC}$ is the alternative matrix $\boldsymbol{X}_a$. We obtain the noncentrality parameter ζ as the SSR for $\boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ regressed on the null matrix $\boldsymbol{X}_n$. The principle of conservatism

states that we should use the smallest possible ζ among all those corresponding to effects of a given size.

There are two common choices for the null matrix $\boldsymbol{X}_n$. The *principle of hierarchy* for general factorials states that we should only consider using hierarchical models. Thus the null model for a term should be the largest hierarchical model that does not contain the term to be tested. This is called a type 2 test in SAS. Alternatively we may use as null matrix all the columns of $\boldsymbol{X}$ that are not in $\boldsymbol{X}_n$; this is standard parametric and called a type 3 test in SAS. We use hierarchical testing as a default, and only consider nonhierarchical testing when we are certain that the contrasts implied by those tests are appropriate.

[Figure 7 about here.]

Example 9. An unbalanced 3^2

We wish to study the effects of supplier (1, 2, or 3) and gum treatment (raw gum, demineralized gum, or Pasteurized gum) on the shelf life of orange drink emulsions and use a three by three factorial with the (1,raw) and (2,demineralized) treatments having four replications and all other factor/level combinations having a single replication. Figure 7 illustrates the design. The terms in the model are the overall mean, main effects of supplier (A, 2 df), main effects of gum (B, 2 df), and the AB interaction (4 df). The $\boldsymbol{X}$ matrix is shown in Table 10.

[Table 10 about here.]

Other parameterizations and $\boldsymbol{X}$ matrices are possible, but this one is convenient for our definition of effect sizes.

We first wish to test the effect of supplier (A). As mentioned in Example 2, the error standard deviation for shelf life is $\sigma = 20$ days. The experiment is being run under harsh conditions, so 20 days is also a fairly large change, and we would like to know how likely we are to detect the situation where some pair of suppliers differs by 20 days in shelf life. This is a 1 SD effect.

The $\boldsymbol{X}_a$ matrix is the second and third columns of $\boldsymbol{X}$, and the $\boldsymbol{\gamma}_a/\sigma$ corresponding to 1 SD effects for the various pairwise differences are listed in Table 11.

[Table 11 about here.]

The noncentrality parameter is found as the SSR for regressing $\boldsymbol{X}_a\boldsymbol{\gamma}_a/\sigma$ on $\boldsymbol{X}_n$. The null matrix $\boldsymbol{X}_n$ for hierarchical testing is columns 1, 4, and 5 of $\boldsymbol{X}$ (the constant and main effect of gum treatment), and the resulting ζ values are given in the table above. The minimal ζ is 1.875, so the power for testing a 1 SD sized main effect of supplier is the probability that a noncentral F with 2 and 6 degrees of freedom and noncentrality parameter 1.875 exceeds the upper 5% point of a central F with 2 and 6 degrees of freedom (5.14); this power is .147.

If we wish to do nonhierarchical testing, then the null matrix $\boldsymbol{X}_n$ is all the columns of $\boldsymbol{X}$ except 2 and 3; the $\boldsymbol{\gamma}_a$s for 1 SD effects do not change, but the ζ values do change: 1.727, 1.727, and 2. For nonhierarchical testing, the power of .139 is obtained using the noncentrality parameter 1.727.

Consider next testing the supplier by gum treatment interaction; this test has 4 and 6 degrees of freedom. The $\boldsymbol{X}_a$ matrix is the last four columns of $\boldsymbol{X}$, and the $\boldsymbol{\gamma}_a/\sigma$ corresponding to the various quartets and 1 SD sized effects are listed in Table 12.

[Table 12 about here.]

The null matrix $\boldsymbol{X}_n$ is the first five columns of $\boldsymbol{X}$ for both testing methods, and the resulting ζ values are given in the table above. The minimal ζ is 1, so the power for testing a 1 SD sized interaction is the probability that a noncentral F with 4 and 6 degrees of freedom and noncentrality parameter 1 exceeds the upper 5% point of a central F with 4 and 6 degrees of freedom (4.53); this power is .078.

[Figure 8 about here.]

Example 10. A less unbalanced 3^2

Let's consider a different three by three factorial for testing the supplier and gum treatment effects on orange drink shelf life. We use the same number of runs, but in this design the runs are assigned so that (1,no treatment), (2,demineralization), and (3,Pasteurization) treatments have one replication each, and all other factor/level combinations have two replications as shown in Figure 8. The terms in the model are the same as before: the overall mean, main effects of supplier, main effects of gum treatment, and their interaction. The $\boldsymbol{X}$ matrix is shown in Table 13.

[Table 13 about here.]

Again test the main effect of supplier and compute power for 1 SD sized effects using hierarchical testing. The $\boldsymbol{X}_a$ matrix is the second and third columns of $\boldsymbol{X}$, and the $\boldsymbol{\gamma}_a/\sigma$ corresponding to the various pairwise differences and 1 SD sized effects are the same as before as shown in Table 14.

[Table 14 about here.]

The null matrix $\boldsymbol{X}_n$ for hierarchical testing is columns 1, 4, and 5 of $\boldsymbol{X}$ (the constant and main effect of B), and the resulting ζ values are given in the table above. For this allocation of the runs, the noncentrality parameter is 2.4 leading to a power of .177 (again testing with 2 and 6 degrees of freedom). For nonhierarchical testing, the ζ values are all 2.25, and the power is .168.

Now test the interaction; this test has 4 and 6 degrees of freedom. The $\boldsymbol{X}_a$ matrix is the last four columns of $\boldsymbol{X}$, and the $\boldsymbol{\gamma}_a/\sigma$ corresponding to the various quartets and 1 SD sized effects are shown in Table 15.

[Table 15 about here.]

The null matrix $\boldsymbol{X}_n$ is the first five columns of $\boldsymbol{X}$ for both testing methods, and the resulting ζ values are given in the table above. The minimal ζ is 1.35, so the power for testing a 1 SD sized interaction is the probability that a noncentral F with 4 and 6 degrees of freedom and noncentrality parameter 1.35 exceeds the upper 5% point of a central F with 4 and 6 degrees of freedom (4.53); this power is .088.

In all cases the powers are larger for this design than for the more unbalanced design in Example 9 (albeit not much larger). In addition, there is less variation among the ζ values for the different pairs and quartets, again showing the better balance in the second design.

Crossed designs

Crossed designs contain both mixture variables and continuous nonmixture process variables. Both the mixture and process variables need to be coded, with the

mixture variables coded to pseudocomponents and the process variables coded to steps of ± 1. The evaluation space for sizing effects in a crossed design is the "product" of the simplex of pseudocomponents and the (–1,1) cube for the process variables; that is, anything from the simplex combined with anything from the cube. As before, the size of an effect is the range of the effect over the entire evaluation region. Any effect that involves only mixture variables has its size determined as those variables range over the full simplex of pseudocomponents. There are no effects that involve only process variables, because any effect can be multiplied by the sum of the mixture variables (which sum is 1).

Effects involving both mixture and process variables can usually be written as a product of an effect involving just mixture variables and an effect involving just process variables. Find the minimum and maximum for both the mixture and process terms, and find the four products of the form mixture minimum or maximum times process minimum or maximum. The range of the crossed $\boldsymbol{X}_e$ is the range of these four products, as shown in Table 16.

[Table 16 about here.]

Example 11. A crossed design

Example 6 introduced a three-component mixture design in water, alcohol, and urea which studied the effects of these components on turbidity. What really matters to some users is not how clear the detergent is but how long the suds last. Let's keep the three mixture factors and add two process factors: the temperature of the wash water (factor D, low or high) and the grease load of the dishes (factor E, low or high). We need a crossed design with three mixture factors (A, B, and C) and two process factors (D and E). Figure 9 illustrates such a design.

[Figure 9 about here.]

For simplicity, we consider only the model that includes first order mixture effects (A, B, C) and first order process effects (AD, BD, CD, AE, BE, and CE). The design matrix is shown in Table 17.

[Table 17 about here.]

Note that there are no pure process terms.

Suppose that we wish to test for the absence of a urea-grease interaction, that is, that the coefficient for the CE term is zero. The standard deviation for the lifetime of the suds is 5 minutes, and a change of 10 minutes is really big, so we would like to be able to detect this 2 SD effect. The range for CE is 2, so $\gamma_a/\sigma = 1$ in a 2 SD effect. Thus the noncentrality parameter will be the SSR when the column for CE (the last column) is regressed on the other columns; this SSR is 1.837. The test has 1 and 31 degrees of freedom, and the power is only .260.

Discussion

This paper has been concerned with a unified way for describing the sizes of effects in fixed-effects models. Once we have the size of an effect, we can compute the power in a fairly routine way. The theory behind these computations has been described in many excellent texts on linear models, including [6], [7], and [8]. Most texts on statistical methods or experimental design seem to describe

power computations only for the simplest possible models (often only the one-way classification) and then only for balanced designs. This focus on simple designs tends to blur the roles of the size of effects and the design used to test the effects. We feel that the distinction between size of effects and design helps when comparing designs, particularly more complex designs.

We have presented that effects are sized as their range (in multiples of standard deviations) over a standard evaluation region. This satisfies the mathematician in us, but it still leaves us with the practical problem of specifying an actual physical size when computing power in the real world. It is easy just to look at "standard" sizes, such as 1 or 2 SD effects, but the strength of power computations comes when we do them for the actual sizes that interest us. For this there is no substitute for knowledge of the goals, expectations, background, and process in the experiment; without these we cannot hope to choose an appropriate alternative $\boldsymbol{\mu}_a$ and size it relative to σ.

Power is a useful tool for evaluating an experimental design, but it is not the be-all and end-all tool. Power tells us how likely we are to reject a null hypothesis when a particular alternative is true; nothing more, nothing less. Large power does not imply that a particular model will fit well, and low power does not imply that a particular model will fit poorly. There are two aspects to a model fitting well. First, our model must describe the mean and error structure of the data correctly, and secondarily, we might hope that there is little variation in the errors relative to the variation in the mean structure (high R^2). Power is computed *assuming* that the error structure is known and correct and that the alternative model encompasses the full mean structure; it does not allow us to assess whether a model fits particular data. For that we use the extensive model diagnostics and graphics that

are available to us today. We can also have any combination of high or low power, and high or low R^2. For example, colinear explanatory variables can lead to low power even in high R^2 situations, and huge sample sizes can lead to high power even in low R^2 situations. The moral here is to interpret power as power, not as something else.

We have used two principles in our computations for power. The first was the principle of conservatism: if there are multiple ways to achieve an alternative of a given size, report as power for that size the smallest of all the powers. This pessimistic approach assures us that if we report power .6, then we will have power at least .6 for any alternative achieving the quoted size. For highly unbalanced designs and multiple-degree-of-freedom effects, there can be quite a range between the lowest and highest powers among effects of a given size.

The second principle was the principle of hierarchy. Hierarchy suggests that models including a high order term should also include terms beneath that high order term. For example, a factorial model with an AB interaction should also have the main effects of A and B, or a response surface model with a quadratic effect in X should also include the linear effect in X. Hierarchy has been debated extensively in the literature, particularly for factorials, and we are not going to resolve the issue here. We suggest hierarchical testing as a default, for reasons explained in [9].

We have described effect sizes in terms of coded units. When doing a hierarchical test, the power in coded or uncoded units is the same, provided that the same alternative $\boldsymbol{\mu}_r/\sigma$ is used (coding/uncoding changes $\boldsymbol{X}_a$ and $\boldsymbol{\gamma}_a$, but not $\boldsymbol{\mu}_a$). The power for nonhierarchical tests can differ greatly between coded and uncoded units, because there is typically much more colinearity among terms when

expressed in uncoded units. Another way of expressing this is that the condition number of the $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ matrix is much better in coded units than uncoded units.

Finally, here are a few lessons that we learn from the examples above and should keep in mind when computing power.

1. The size of an effect depends on the coding for the variables.

2. The size of an effect does *not* depend on the points in the experimental design or the other terms in the model.

3. The power for detecting an effect of a given size *does* depend on the experimental design and the other terms are in the model.

4. Mixture designs often have small power, due in part to correlation among the predicting variables. This problem generally becomes worse when the design is constrained.

5. Effects with more numerator degrees of freedom have less power for the same size noncentrality parameter.

References

1. Stat-Ease, Inc. (2000) *Version 6 User's Guide, Design-Expert® Software*, Minneapolis, MN.

2. SAS Institute Inc. (1999) *SAS® Language Reference: Dictionary, Version 7-1*, Cary, NC.

3. MathSoft, Inc. (1999) *S-Plus® 5 for UNIX Guide to Statistics*, Seattle, WA.

4. Ihaka, R. and Gentleman R. (1996) “R: A Language for Data Analysis and Graphics,” *Journal of Computational and Graphical Statistics*, 5, 299–314.

5. Oehlert, G. W. and Bingham, C. B. (2000) “MacAnova Reference Manual,” Technical Report #618, School of Statistics, University of Minnesota, `ftp://www.stat.umn.edu/macanova/docs/referenc.pdf`.

6. Scheffé, H. (1999) *The Analysis of Variance* Wiley: New York.

7. Searle, S. R. (1987) *Linear models for unbalanced data* Wiley: New York.

8. Hocking, R. (1996) *Methods and Applications of Linear Models: Regression and the Analysis of Variance* Wiley: New York.

9. Oehlert, G. W. (2000) *A First Course in Design and Analysis of Experiments* Freeman: New York.

Figures

Figure 1: Null (solid) and alternative (dashed) distributions for the F-ratio.

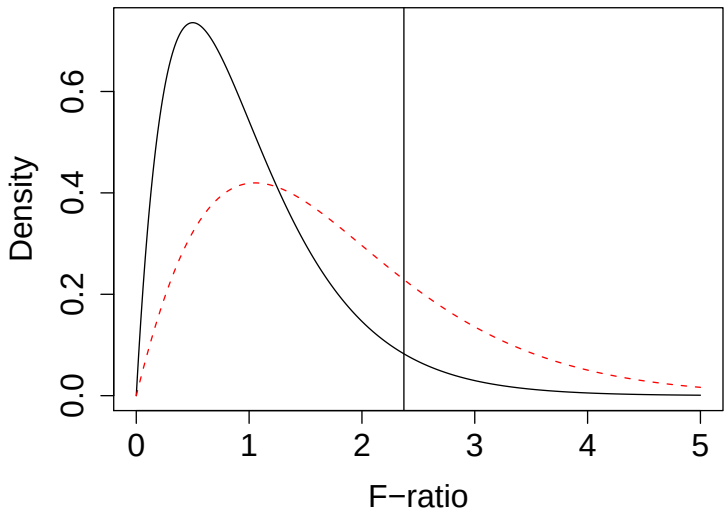


Figure 2: The rotatable, two-factor, central composite design of Example 3.

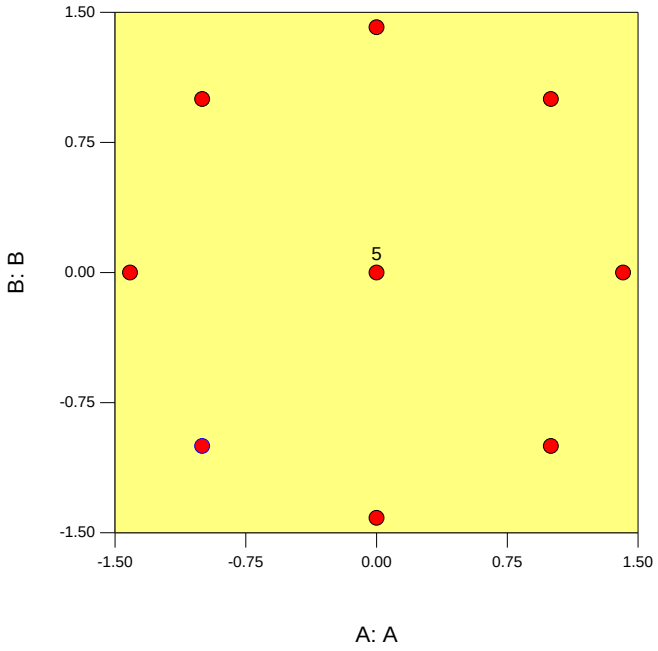


Figure 3: Power evaluation output from Design-Expert® for the second order model in the central composite design of Example 3.

2 Factors: A, B

Design Matrix Evaluation for Response Surface Quadratic Model

No aliases found

Degrees of Freedom for Evaluation

Model	5
Residuals	7
Lack 0f Fit	*3*
Pure Error	*4*
Corr Total	12

				Power at 5 % alpha level for effect of		
Term	**StdErr****	**VIF**	**Ri-Squared**	**1/2 Std. Dev.**	**1 Std. Dev.**	**2 Std. Dev.**
A	0.35	1.00	0.0000	9.4 %	23.2 %	68.1 %
B	0.35	1.00	0.0000	9.4 %	23.2 %	68.1 %
A^2	0.38	1.02	0.0170	20.8 %	62.1 %	99.4 %
B^2	0.38	1.02	0.0170	20.8 %	62.1 %	99.4 %
AB	0.50	1.00	0.0000	7.2 %	14.0 %	40.8 %

****Basis Std. Dev. = 1.0**

Figure 4: The nonrotatable, two-factor, central composite design of Example 4.

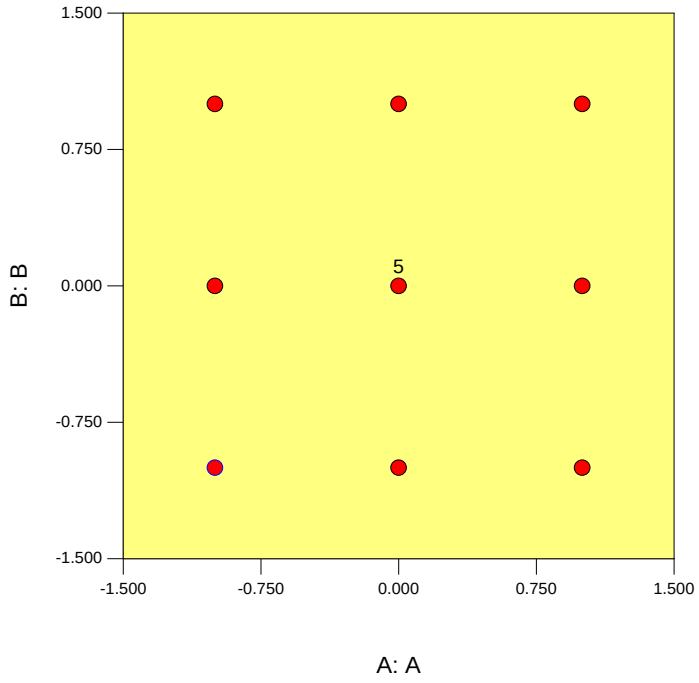


Figure 5: The augmented simplex-lattice design of Example 6.

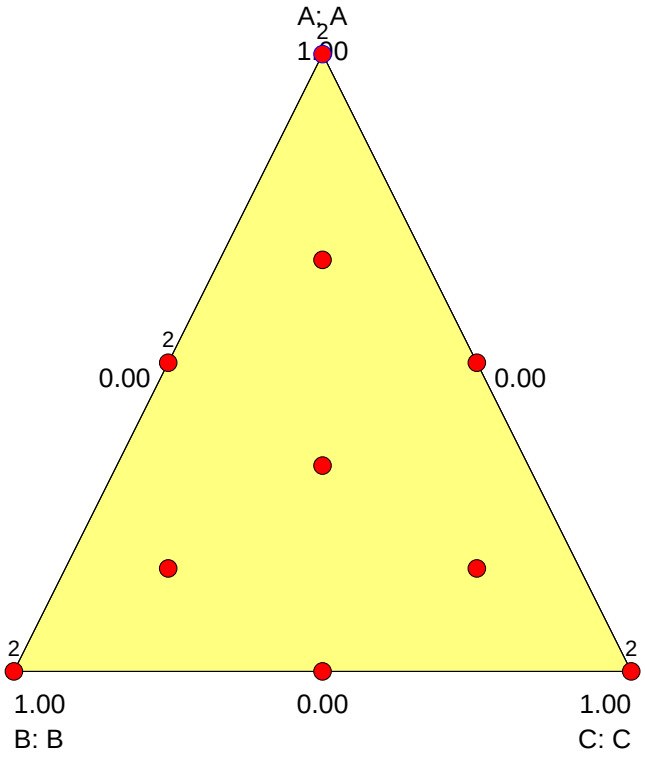


Figure 6: The three-component constrained mixture design of Example 7.

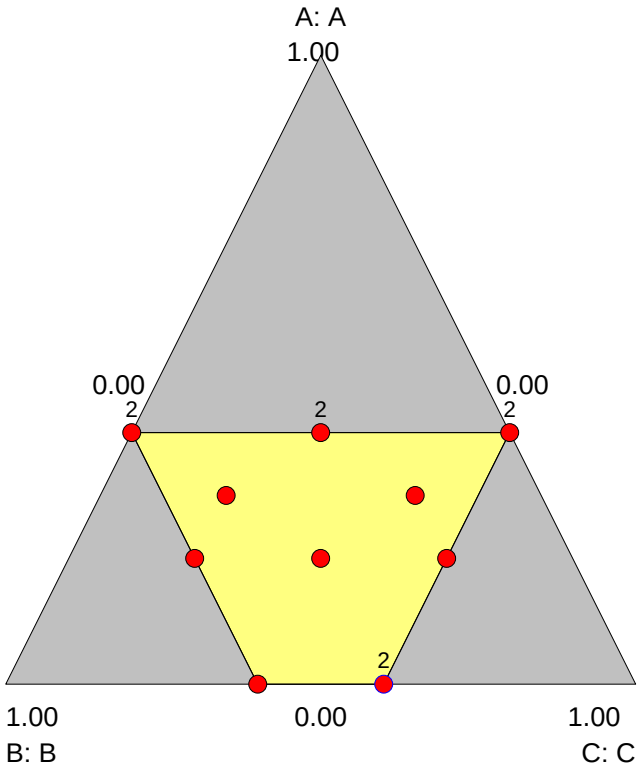


Figure 7: The unbalanced 3^2 factorial of Example 9.

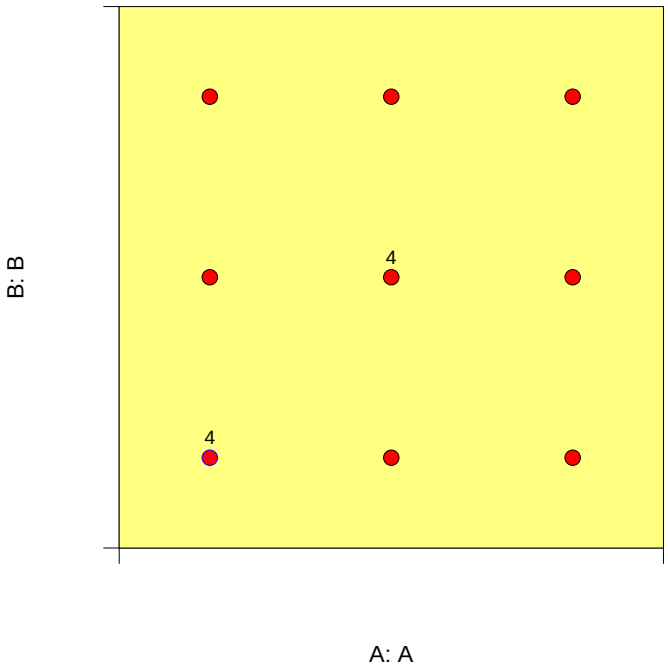


Figure 8: The unbalanced 3^2 factorial of Example 10.

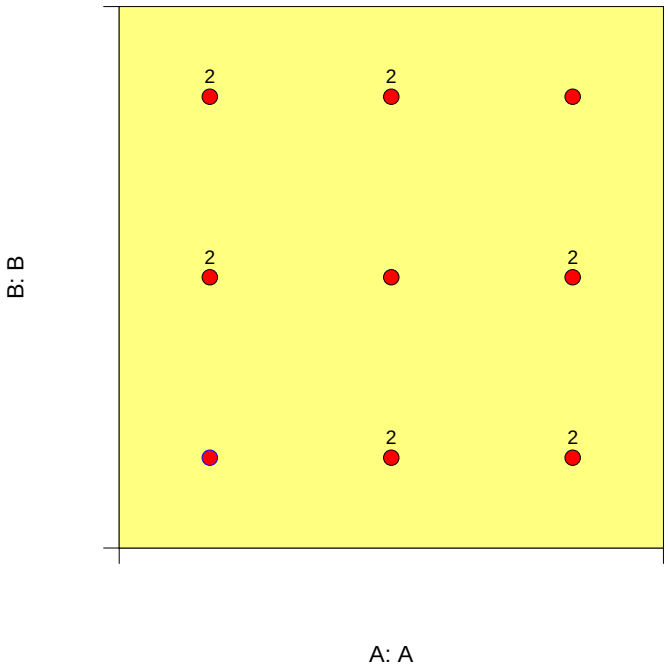
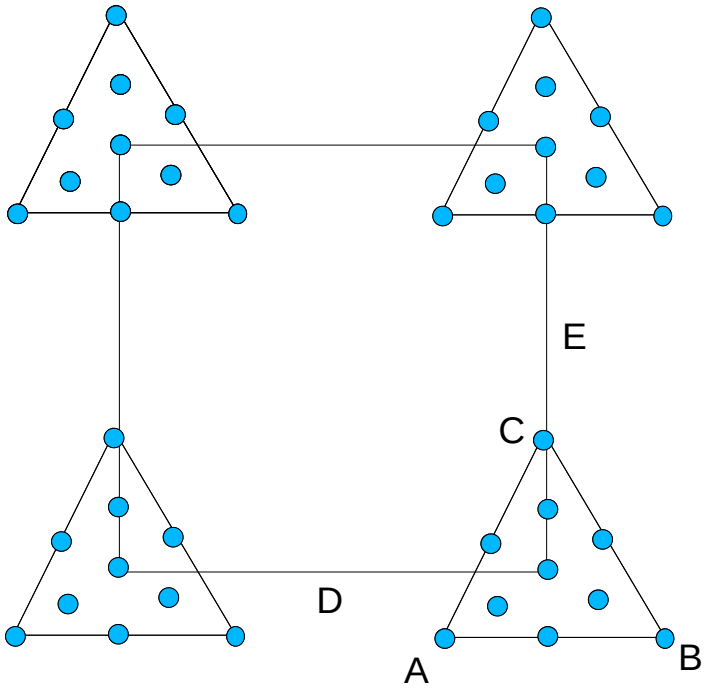


Figure 9: The crossed design with three mixture factors and two process factors of Example 11.



Tables

Table 1: Design matrix for the factorial of Example 1.

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & -1 \\ 1 & 1 & 0 & -1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & 0 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}$$

Table 2: Ranges of standard polynomial effects over the evaluation cube.

Type	Minimum	Maximum	Range
Linear A	–1	1	2
Quadratic A^2	0	1	1
Quadratic AB	–1	1	2
Cubic A^3	–1	1	2
Cubic A^2B	–1	1	2
Cubic ABC	–1	1	2
Quartic A^4	0	1	1
Quartic A^3B	–1	1	2
Quartic A^2B^2	0	1	1
Quartic A^2BC	–1	1	2
Quartic ABCD	–1	1	2

Table 3: Design matrix for the rotatable central composite design of Example 3.

$$\begin{bmatrix}
1.000 & -1.000 & -1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & -1.000 & -1.000 & 1.000 & 1.000 \\
1.000 & -1.000 & 1.000 & -1.000 & 1.000 & 1.000 \\
1.000 & 1.000 & 1.000 & 1.000 & 1.000 & 1.000 \\
1.000 & -1.414 & 0.000 & 0.000 & 2.000 & 0.000 \\
1.000 & 1.414 & 0.000 & 0.000 & 2.000 & 0.000 \\
1.000 & 0.000 & -1.414 & 0.000 & 0.000 & 2.000 \\
1.000 & 0.000 & 1.414 & 0.000 & 0.000 & 2.000 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000
\end{bmatrix}$$

Table 4: Design matrix for the nonrotatable central composite of Example 4.

$$\begin{bmatrix} 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Table 5: Residual matrix $\boldsymbol{X}_r$ for Example 5.

$$\begin{bmatrix} 0.538 & 0.538 \\ 0.538 & 0.538 \\ 0.538 & 0.538 \\ 0.538 & 0.538 \\ 0.538 & -0.462 \\ 0.538 & -0.462 \\ -0.462 & 0.538 \\ -0.462 & 0.538 \\ -0.462 & -0.462 \\ -0.462 & -0.462 \\ -0.462 & -0.462 \\ -0.462 & -0.462 \\ -0.462 & -0.462 \end{bmatrix}$$

Table 6: Size of polynomial effects for mixture designs using the Scheffé parameterization.

Type	Minimum	Maximum	Range
Quadratic AB	0	1/4	1/4
Special cubic ABC	0	1/27	1/27
Cubic AB(A-B)	–3/32	3/32	3/16
Special quartic A^2BC	0	1/64	1/64

Table 7: Design matrix for the augmented simplex-lattice design of Example 6.

$$\begin{bmatrix}
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.500 & 0.500 & 0.000 & 0.250 & 0.000 & 0.000 \\
0.500 & 0.000 & 0.500 & 0.000 & 0.250 & 0.000 \\
0.000 & 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.500 & 0.500 & 0.000 & 0.000 & 0.250 \\
0.000 & 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\
0.667 & 0.167 & 0.167 & 0.111 & 0.111 & 0.028 \\
0.167 & 0.667 & 0.167 & 0.111 & 0.028 & 0.111 \\
0.167 & 0.167 & 0.667 & 0.028 & 0.111 & 0.111 \\
0.333 & 0.333 & 0.333 & 0.111 & 0.111 & 0.111 \\
1.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\
0.000 & 1.000 & 0.000 & 0.000 & 0.000 & 0.000 \\
0.500 & 0.500 & 0.000 & 0.250 & 0.000 & 0.000
\end{bmatrix}$$

Table 8: Design matrix for the constrained mixture design of Example 7.

$$\begin{bmatrix} 0.00 & 0.40 & 0.60 & 0.00 & 0.00 & 0.24 \\ 0.20 & 0.60 & 0.20 & 0.12 & 0.04 & 0.12 \\ 0.40 & 0.30 & 0.30 & 0.12 & 0.12 & 0.09 \\ 0.40 & 0.00 & 0.60 & 0.00 & 0.24 & 0.00 \\ 0.40 & 0.60 & 0.00 & 0.24 & 0.00 & 0.00 \\ 0.00 & 0.60 & 0.40 & 0.00 & 0.00 & 0.24 \\ 0.20 & 0.20 & 0.60 & 0.04 & 0.12 & 0.12 \\ 0.20 & 0.40 & 0.40 & 0.08 & 0.08 & 0.16 \\ 0.30 & 0.20 & 0.50 & 0.06 & 0.15 & 0.10 \\ 0.30 & 0.50 & 0.20 & 0.15 & 0.06 & 0.10 \\ 0.40 & 0.30 & 0.30 & 0.12 & 0.12 & 0.09 \\ 0.00 & 0.40 & 0.60 & 0.00 & 0.00 & 0.24 \\ 0.40 & 0.00 & 0.60 & 0.00 & 0.24 & 0.00 \\ 0.40 & 0.60 & 0.00 & 0.24 & 0.00 & 0.00 \end{bmatrix}$$

Table 9: Modified null matrix $\tilde{\boldsymbol{X}}_n$ from the mixture design in Example 8.

$$\begin{bmatrix} 0.500 & 0.500 \\ 0.750 & 0.250 \\ 0.250 & 0.750 \\ 1.000 & 0.000 \\ 0.500 & 0.500 \\ 0.000 & 1.000 \\ 0.501 & 0.501 \\ 0.751 & 0.251 \\ 0.251 & 0.751 \\ 0.500 & 0.500 \\ 0.500 & 0.500 \\ 0.000 & 1.000 \\ 1.000 & 0.000 \\ 0.750 & 0.250 \end{bmatrix}$$

Table 10: Design matrix for the unbalanced 3^2 design of Example 9.

$$\boldsymbol{X} = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\ 1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\ 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Table 11: Alternative coefficients and noncentrality parameters for a main effect in the 3^2 design of Example 9.

i	i'	γ_{a1}/σ	γ_{a2}/σ	ζ
1	3	.5	0	1.875
2	3	0	.5	1.875
1	2	.5	–.5	2.250

Table 12: Alternative coefficients and noncentrality parameters for an interaction in the 3^2 design of Example 9.

i	i'	j	j'	$\boldsymbol{\gamma}_{a1}/\sigma$	$\boldsymbol{\gamma}_{a2}/\sigma$	$\boldsymbol{\gamma}_{a3}/\sigma$	$\boldsymbol{\gamma}_{a4}/\sigma$	ζ
1	3	1	3	.5	0	0	0	1.286
1	3	2	3	0	.5	0	0	1.000
2	3	1	3	0	0	.5	0	1.000
2	3	2	3	0	0	0	.5	1.286
1	2	1	3	.5	0	–.5	0	1.286
1	2	2	3	0	.5	0	–.5	1.286
1	2	1	2	.5	–.5	–.5	.5	1.643
1	3	1	2	.5	–.5	0	0	1.286
2	3	1	2	0	0	.5	–.5	1.286

Table 13: Design matrix for the 3^2 design of Example 10.

$$\begin{bmatrix}
1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\
1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\
1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\
1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1 \\
1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 \\
1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\
1 & -1 & -1 & 1 & 0 & -1 & 0 & -1 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
1 & -1 & -1 & 0 & 1 & 0 & -1 & 0 & -1 \\
1 & 1 & 0 & -1 & -1 & -1 & -1 & 0 & 0 \\
1 & 0 & 1 & -1 & -1 & 0 & 0 & -1 & -1
\end{bmatrix}$$

Table 14: Alternative coefficients and noncentrality parameters for a main effect in the 3^2 design of Example 10.

i	i'	γ_{a1}/σ	γ_{a2}/σ	ζ
1	3	.5	0	2.4
2	3	0	.5	2.4
1	2	.5	–.5	2.4

Table 15: Alternative coefficients and noncentrality parameters for interaction in the 3^2 design of Example 10.

i	i'	j	j'	γ_{a1}/σ	γ_{a2}/σ	γ_{a3}/σ	γ_{a4}/σ	ζ
1	3	1	3	.5	0	0	0	1.35
1	3	2	3	0	.5	0	0	1.65
2	3	1	3	0	0	.5	0	1.65
2	3	2	3	0	0	0	.5	1.35
1	2	1	3	.5	0	–.5	0	1.65
1	2	2	3	0	.5	0	–.5	1.65
1	2	1	2	.5	–.5	–.5	.5	1.65
1	3	1	2	.5	–.5	0	0	1.65
2	3	1	2	0	0	.5	–.5	1.35

Table 16: Polynomial effect ranges for process by mixture crossed designs.

(Mixture)×Process	Minimum	Maximum	Range
(A)*mean	0	1	1
(A)*e	–1	1	2
(A)*ef	–1	1	2
(A)*e^2	0	1	1
(A)*e^3	–1	1	2
(A)*e^2f	–1	1	2
(A)*efg	–1	1	2
(AB)*mean	0	1/4	1/4
(AB)*e	–1/4	1/4	1/2
(AB)*ef	–1/4	1/4	1/2
(AB)*e^2	0	1/4	1/4
(AB)*e^3	–1/4	1/4	1/2
(AB)*e^2f	–1/4	1/4	1/2
(AB)*efg	–1/4	1/4	1/2
(ABC)*mean	0	1/27	1/27
(ABC)*e	–1/27	1/27	2/27
(ABC)*ef	–1/27	1/27	2/27
(ABC)*e^2	0	1/27	1/27
(ABC)*e^3	–1/27	1/27	2/27
(ABC)*e^2f	–1/27	1/27	2/27
(ABC)*efg	–1/27	1/27	2/27
(AB(A-B))*mean	–3/32	3/32	3/16
(AB(A-B))*e	–3/32	3/32	3/16
(AB(A-B))*ef	–3/32	3/32	3/16
(AB(A-B))*e^2	–3/32	3/32	3/16
(AB(A-B))*e^3	–3/32	3/32	3/16
(AB(A-B))*e^2f	–3/32	3/32	3/16
(AB(A-B))*efg	–3/32	3/32	3/16

Table 17: Design matrix for the crossed design of Example 11.

$$
\begin{bmatrix}
1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\
1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & -1.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & -1.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -1.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & -1.00 & 0.00 & 0.00 & 1.00 \\
0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 & 1.00 \\
0.00 & 0.50 & 0.50 & 0.00 & -0.50 & -0.50 & 0.00 & -0.50 & -0.50 \\
0.00 & 0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 & -0.50 & -0.50 \\
0.00 & 0.50 & 0.50 & 0.00 & -0.50 & -0.50 & 0.00 & 0.50 & 0.50 \\
0.00 & 0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 & 0.50 & 0.50 \\
0.50 & 0.00 & 0.50 & -0.50 & 0.00 & -0.50 & -0.50 & 0.00 & -0.50 \\
0.50 & 0.00 & 0.50 & 0.50 & 0.00 & 0.50 & -0.50 & 0.00 & -0.50 \\
0.50 & 0.00 & 0.50 & -0.50 & 0.00 & -0.50 & 0.50 & 0.00 & 0.50 \\
0.50 & 0.00 & 0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 & 0.50 \\
0.50 & 0.50 & 0.00 & -0.50 & -0.50 & 0.00 & -0.50 & -0.50 & 0.00 \\
0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 & -0.50 & -0.50 & 0.00 \\
0.50 & 0.50 & 0.00 & -0.50 & -0.50 & 0.00 & 0.50 & 0.50 & 0.00 \\
0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 & 0.50 & 0.50 & 0.00 \\
0.67 & 0.17 & 0.17 & -0.67 & -0.17 & -0.17 & -0.67 & -0.17 & -0.17 \\
0.67 & 0.17 & 0.17 & 0.67 & 0.17 & 0.17 & -0.67 & -0.17 & -0.17 \\
0.67 & 0.17 & 0.17 & -0.67 & -0.17 & -0.17 & 0.67 & 0.17 & 0.17 \\
0.67 & 0.17 & 0.17 & 0.67 & 0.17 & 0.17 & 0.67 & 0.17 & 0.17 \\
0.17 & 0.67 & 0.17 & -0.17 & -0.67 & -0.17 & -0.17 & -0.67 & -0.17 \\
0.17 & 0.67 & 0.17 & 0.17 & 0.67 & 0.17 & -0.17 & -0.67 & -0.17 \\
0.17 & 0.67 & 0.17 & -0.17 & -0.67 & -0.17 & 0.17 & 0.67 & 0.17 \\
0.17 & 0.67 & 0.17 & 0.17 & 0.67 & 0.17 & 0.17 & 0.67 & 0.17 \\
0.17 & 0.17 & 0.67 & -0.17 & -0.17 & -0.67 & -0.17 & -0.17 & -0.67 \\
0.17 & 0.17 & 0.67 & 0.17 & 0.17 & 0.67 & -0.17 & -0.17 & -0.67 \\
0.17 & 0.17 & 0.67 & -0.17 & -0.17 & -0.67 & 0.17 & 0.17 & 0.67 \\
0.17 & 0.17 & 0.67 & 0.17 & 0.17 & 0.67 & 0.17 & 0.17 & 0.67 \\
0.33 & 0.33 & 0.33 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 & -0.33 \\
0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33 & -0.33 & -0.33 & -0.33 \\
0.33 & 0.33 & 0.33 & -0.33 & -0.33 & -0.33 & 0.33 & 0.33 & 0.33 \\
0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33 & 0.33
\end{bmatrix}
$$