DESIGN OF EXPERIMENTS

1. Introduction

Obtaining valid results from a test program calls for commitment to sound statistical design. In fact, proper design of experiments (DOE) is often more important than sophisticated statistical analysis. Results of a well-planned experiment are often evident from simple graphical analyses. However, the world's best statistical analysis cannot rescue a poorly planned experimental program. The main reason for designing an experiment statistically is to obtain unambiguous answers to questions of major interest at a minimum cost. The need to quantify effects, learn about interactions among variables, model relationships and measure experimental error are some added reasons for designing an experiment statistically.

Many chemists and engineers think of experimental design mainly in terms of standard plans for assigning categorical treatments and/or numerical factor levels to experimental runs, such as two-level factorial and response surface method (RSM) designs. These designs are described in books, such as those summarized in the general references of this article, and catalogued in various reports and articles. Additionally, numerous commercial software packages are available for generating such experimental designs, as well as to aid the experimenter in analyzing the resulting data. Important as such formal plans are, the final selection of test points represents only the proverbial tip of the iceberg, the culmination of a careful planning process. For this reason, particular emphasis on the *process* of designing an experiment is placed here.

1.1. Multifactor Design Versus One-Factor-At-A-Time (OFAT).

Statistically planned experiments are characterized by the proper consideration of extraneous variables; the fact that primary variables are changed together, rather than one factor at a time (OFAT), in order to obtain information about the magnitude and nature of interactions and to gain improved precision in estimating the effects of these variables; and built-in procedures for measuring the various sources of random variation and for obtaining a valid measure of experimental error against which one can assess the impact of the primary variables and their interactions.

Figure 1 contrasts a three-factor two-level design with an OFAT of equivalent precision. The factorial design offers four runs at the high level for A and the same for the low (right versus left faces of cubical experimental region). Similarly, factors B and C also benefit from having four runs at both high and low levels (top versus bottom of cube and back versus front; respectively). Therefore, the OFAT experimenter must provide four runs each at the high levels of each factor versus four at the base line (all low levels at origin) to provide similar power of replication for effect estimation. However, this necessitates a total of 16 runs for OFAT versus only 8 for the two-level factorial. Due to its far more efficient parallel processing two-level factorial DOE trumps the serial OFAT scheme, and its efficiency advantage only becomes more pronounced as the number of factors increase. See reference (1) for further arguments against one-factor-at-at-time (OFAT) in favor of multifactor design of experiments (DOE) and its follow-up article (2), which illustrates how a two-level factorial can reveal a breakthrough interaction.

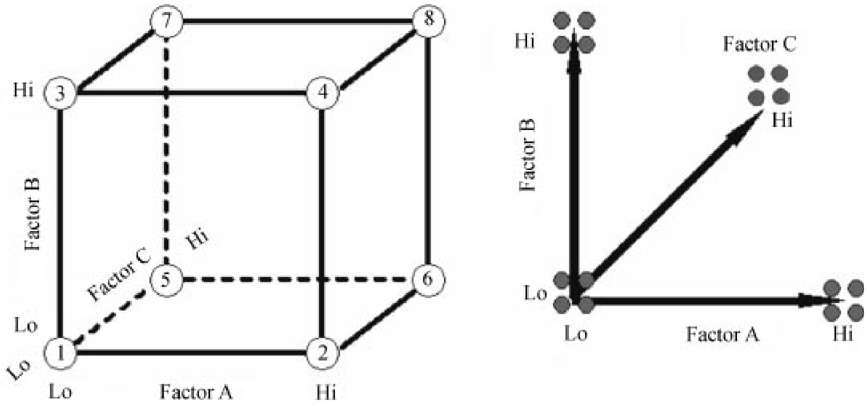

Fig. 1. Comparison of factorial design versus one-factor-at-a-time (OFAT) for contrasting response at low ("Lo") versus high ("Hi") factor levels.

1.2. Working with a Statistician on the Design and Analysis of Experiments.

A well-planned experiment is often tailor-made via computer-aided optimal design to meet specific objectives and to satisfy practical constraints. The final plan may or may not involve a standard textbook design. If possible, a statistician knowledgeable in the design of experiments should be called in early, made a full-fledged team member, and be fully apprised of the objectives of the test program and of the practical considerations and constraints. He or she may contribute significantly merely by asking probing questions. After the problem and the constraints have been clearly defined, the statistician can evolve an experimental layout to minimize the required testing effort to obtain the desired information—or, as a minimum, review a plan that may have been developed by a practitioner. However, designing an experiment is often an iterative process requiring rework as new information and preliminary data become available. With a full understanding of the problem, the statistician is in an improved position to respond rapidly if last-minute changes are required, to help experimenters gain understanding in a sequential manner, and to provide meaningful analyses of the experimental results, including statements about the statistical precision of any estimates.

2. Purpose and Scope of the Experiment

Designing an experiment is like designing a product. Every product serves a purpose; so should every experiment. This purpose must be clearly defined at the outset. It may, for example, be to optimize a process, to estimate the probability that a component operates properly under a given stress for a specified number of years, to maximize robustness of a new chemical formulation to variability in raw materials or end use, or to determine whether a new drug or medical treatment is superior to an existing one. An understanding of this purpose is important in developing an appropriate experimental plan.

In addition to defining the purpose of a program, one must decide on its scope. An experiment is generally a vehicle for drawing inferences about the real world, as expressed by some, usually quantitative, response (or performance) variable. Since it is highly risky to draw inferences about situations beyond the scope of the experiment, care must be exercised to make this scope sufficiently broad. For example, when developing a new product, experimental studies are often conducted on a smaller (eg, lab) scale, with the objective of scaling up successful results to production. This is true whether the “product” is a thermoplastic resin, food product, or a new drug. Thus, if the results are material dependent, the lot(s) used during the experiment must be as representative as possible of what one might expect to encounter in production. If the test program were limited to a single lot of raw material, the conclusions might be applicable only to that lot, irrespective of the sample size. Similarly, in deciding whether or not a processing factor, such as temperature, should be included as an experimental variable to compare different formulations, it must be decided whether the possible result that one formulation outperforms another at a particular temperature would also apply for other temperatures of practical interest. If not, one should consider including temperature as an experimental variable. In any case, one need keep in mind that the statistical inferences one can draw from the experiment apply only to the range of conditions under which the experiment was conducted.

3. Experimental Variables

An important part of planning an experimental program is the identification of the controllable or “independent” variables (also known as factors) that affect the response and deciding what to do about them. The decision as to how to deal with each of the candidate variables can be made jointly by the experimenter and the statistician. However, identifying the variables is the experimenter’s responsibility. (Tools such as the fishbone diagram and brainstorming (3) can be used to facilitate the process.) Controllable or independent variables in a statistical experiment can be dealt with in four different ways, as described next. The assignment of a particular variable to a category often involves a trade-off among information, cost, and time.

3.1. Primary Variables. The most obvious variables are those whose effects on the expected or mean performance of the response variable(s) are to be evaluated directly; these are the variables that, most likely, created the need for the investigation in the first place. Such variables may be quantitative, such as catalyst concentration, temperature, or pressure, or they may be qualitative, such as method of preparation, catalyst type, or batch of material.

Quantitative controllable variables are frequently related to the response variable by some assumed statistical relationship or model. The minimum number of conditions or levels per variable is determined by the form of the assumed model. For example, if a linear (main effects) or two-factor interactive (X_iX_j) relationship can be assumed, two levels (or conditions) may be sufficient; for a quadratic relationship (curvilinear) such as that shown for two factors in the equation below, a minimum of three levels is required.

$$\hat{Y}= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 \quad (1)$$

However, it is recommended that additional points, above the minimum, be included so as to allow assessment of the adequacy of the assumed model (see Response Surface Method (RSM) Designs). Qualitative variables can be broken down into two categories. The first consists of those variables whose specific effects on the mean response are to be compared directly; eg, comparison of the impact on average performance of two proposed preparation methods or of three catalyst types. The required number of conditions for such variables is generally evident from the context of the experiment. Such variables are sometimes referred to as fixed effects or Type I variables.

The second type of qualitative variables are those whose individual contributions to variability or noise in the responses are to be evaluated. The specific conditions of such variables are generally randomly determined. Material batch is a typical example. Usually, one is not interested in the behavior of the specific batches per se that happened to have been selected for the experiment. Instead, one wishes to quantify the variation in performance, as measured by the variance (or, its square root, the standard deviation) caused by differences among batches. The batches used in the experiment are selected randomly (or as close to randomly as is practically feasible) from a large population of batches. It is desirable to have a reasonably large sample (eg, five or more batches) in order to obtain an adequate degree of precision in estimating the variability in response attributable to such variables. These variables are generally referred to as random effects or Type II variables. Differentiation between fixed and random effect variables is an important consideration both in the design of the experiment, and in the analysis of the resulting data.

When there are two or more variables, they might interact with one another, i.e., the effect of one variable upon the response depends on the value of the other variable. Figure 2 shows a situation where two non-interacting

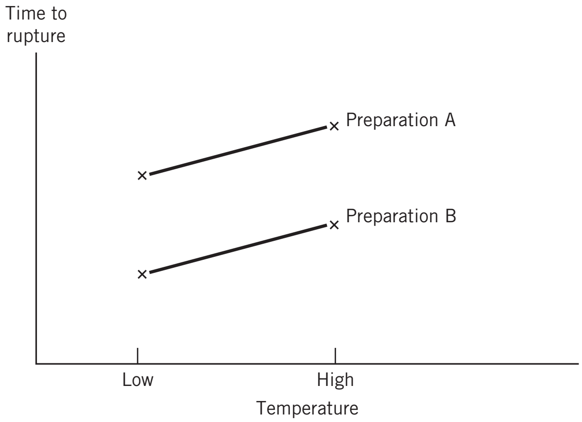

Fig. 2. Situation with no interaction between variables.

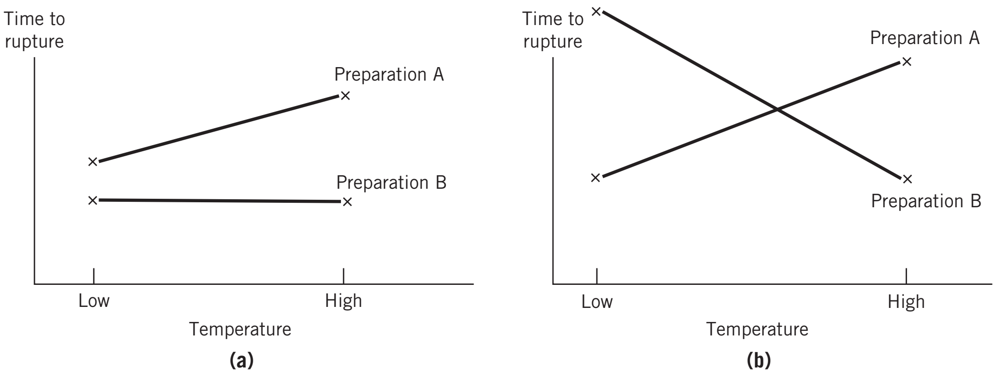

Fig. 3. Situation with interactions between variables, where in (**a**) an increase in temperature is beneficial for preparation A but does not make any difference for preparation B, and (**b**) an increase in temperature raises time to rupture for preparation A but decreases it for preparation B.

variables, preparation type and temperature, independently affect time to rupture, ie, the effect of temperature on time to rupture is the same for both preparation types. In contrast, Figure 3 shows two examples of interactions between preparation and temperature.

An important purpose of a designed experiment is to obtain information about interactions among the primary variables which is accomplished by varying factors simultaneously rather than one at a time. Thus in Figure 2 each of the two preparations would be run at both low and high temperatures using, for example, a full factorial experiment (see Formal Experimental Plans).

3.2. Background Variables and Blocking.

In addition to the primary controllable variables there are those variables, though not of primary interest, that cannot, and perhaps should not, be held constant in the experiment. Such variables are present in real world situations, where they introduce added variability in the response, and where, unlike the experiment, they generally cannot be controlled. They are often referred to as “noise variables”. Typical examples are day-to-day, operator-to-operator, machine-to-machine, and batch-to-batch variability. (Some of these variables, in fact, might be the same as the ones that we had identified in the earlier discussion as Type II primary variables. The difference is that, previously, we were interested in assessing their impact on response variability, per se. Now our major concern is, not so much in their evaluation, but just to ensure that they do not “mess up” our evaluation of the primary variables).

In particular, it is crucial that such background, or noise, variables are separable from (in technical terms, not “confounded” with) the primary variables in the experiment. For example, if preparation A were run only on day 1 and preparation B only on day 2, it would not be possible to determine how much of any observed difference in response between the two preparations is due to normal day-to-day process variation.

Background variables may be introduced into the experiment in the form of experimental blocks. An experimental block represents a relatively homogeneous set of conditions within which different conditions of the primary variables are compared. For example, if one expects day-to-day variation in the response, a day might be taken as an experimental block over which different conditions of the primary variables are compared.

A specific example of blocking arises in the comparison of wear for different types of automobile tires. Tire wear may vary from one automobile to the next, irrespective of the tire type, because of differences among automobiles, variability among drivers, and so on. Assume for example that for the comparison of four tire types (A, B, C, and D), four automobiles (1, 2, 3, and 4) are available. A poor procedure would be to use the same type of tire on each of the four wheels of an automobile varying the tire type among automobiles, as in the following tabulation:

Automobile			
1	2	3	4
A	B	C	D
A	B	C	D
A	B	C	D
A	B	C	D

Such an assignment is undesirable because the differences between tires cannot be separated from the differences between automobiles in the subsequent data analysis. Separation of these effects can be obtained by treating automobiles as experimental blocks and randomly assigning tires of each of the four types to each automobile as follows:

Automobile			
1	2	3	4
A	A	A	A
B	B	B	B
C	C	C	C
D	D	D	D

The above arrangement is known as a randomized block design.

The symmetry of the preceding example is not always found in practice. For example, there may be 6 tire types under comparison and 15 available automobiles. Tires are then assigned to automobiles to obtain the most precise comparison among tire types, using a so-called incomplete block design. Similar concepts apply if there are two or more primary variables, rather than one, as was the case in the tire example.

A main reason for running an experiment in blocks is to ensure that the effect of a background variable does not contaminate evaluation of the effects of the primary variables. Blocking, moreover, removes the effect of the blocked variables from the experimental error also, thus allowing more precise estimation of the experimental error and, as a result, more precise estimates of the effects of the primary variables. Finally, in many situations, the effect of the blocking variables on the response can also be readily evaluated, an important added bonus for blocking.

In some situations, there may be more than one background variable whose possible contaminating effect is removed by blocking. Thus in the automobile tire comparison, the differences between wheel positions may be of concern in addition to differences between automobiles. In this case, wheel position might be introduced into the experiment as a second blocking variable. If there are four tire types to be compared, this might be done by randomly assigning the tires of each of the four types according to the following plan, known as a Latin square design:

Wheel position	Automobile			
	1	2	3	4
1	A	D	C	B
2	B	A	D	C
3	C	B	A	D
4	D	C	B	A

In this plan, the effects of both automobile and wheel position are controlled by blocking. However, it should be kept in mind that for the Latin square design, as for various other blocking plans, it is generally assumed that the blocking variables do not interact with the primary variable to be evaluated.

3.3. Uncontrolled Variables and Randomization. A number of further variables, such as ambient conditions (temperature, pressure, etc), can be identified but not controlled, or are only hazily identified or not identified at all but affect the results of the experiment. To ensure that such uncontrolled variables do not bias the results, randomization is introduced in various ways into the experiment to the extent that this is practical.

Randomization means that the sequence of preparing experimental units, assigning treatments, running tests, taking measurements, etc., is determined by chance, based, eg, on numbers provided by a random number generator. The total effect of the uncontrolled variables is thus lumped together as unaccounted variability and part of the experimental error. The more influential the effect of such uncontrolled variables, the larger the resulting experimental error, and the more imprecise the evaluations of the effects of the primary variables. Sometimes, when the uncontrolled variables can be measured, their effect can be removed statistically from experimental error. In addition, note that measurement error on the response variable (and, sometimes, on the controllable variables), is a further component of experimental error.

Background variables could also be introduced into the experiment by randomization, rather than by blocking techniques. Thus in the previous example, the four tires of each type could have been assigned to automobiles and wheel positions completely randomly, instead of treating automobiles and wheel positions as experimental blocks. This could have resulted in an assignment such as the following:

Wheel position	Automobile			
	1	2	3	4
1	B	D	B	D
2	A	D	D	A
3	C	C	A	D
4	B	B	C	A

Both blocking and randomization generally ensure that the background variables do not contaminate the evaluation of the primary variables. Randomization sometimes offers the advantage of greater simplicity compared to blocking. However, under blocking the effect of a background variable is removed from the experimental error (as well as being measurable), whereas under randomization it usually is not. Thus the aim might be to remove the effects of the one or two most important background variables by blocking, while counteracting the possible contaminating effect of others by randomization.

3.4. Variables Held Constant. Finally, some variables should be held constant in the experiment. Holding a variable constant limits the size and complexity of the experiment but, as previously noted, can also limit the scope of the resulting inferences. The variables to be held constant in the experiment must be identified and the mechanisms for keeping them constant defined. The experimental technique should be clearly specified at the outset of the experiment and closely followed.

4. Experimental Environment and Constraints

The operational conditions under which the experiment is to be conducted and the manner in which each of the factors is to be varied must be clearly spelled out.

All variables are not created equal; some can be varied more easily than others. For example, suppose that oven rack position and baking temperature are two experimental variables in a study to determine recommended baking instructions for a new cake mix. A change in oven rack position can be made in a matter of seconds. Changing the oven temperature, however, adds costly time to each experimental run, since additional time is required to reach the target temperature. Also, in many experiments it is easier to change pressure than it is to change temperature—due to the difference in time that it takes to stabilize at the new condition. In such situations, completely randomizing the sequence of testing is impractical. However, basing the experimental plan on convenience alone may lead to ambiguous and un-analyzable results. For example, if the first half of the experiment is run at one temperature and the second half at another temperature, it may not be possible to tell whether the observed difference in response results from the difference in temperature or from some other factors that varied during the course of the experiment such as raw material (cake batter), ambient conditions (humidity), or operator technique. Thus the final experimental plan must be a compromise between cost and information. The experiment must be practical to run, yet still yield statistically valid results.

Practical considerations enter into the experimental plan in various other ways. In many programs, variables are introduced at different operational levels. For example, in evaluating the effect of alloy composition, oven temperature, and varnish coat on tensile strength, it may be convenient to make a number of master alloys with each composition, split the alloys into separate parts to be subjected to different heat treatments, and then cut the treated samples into subsamples to which different coatings are applied. Tensile strength measurements are then obtained on all coated subsamples.

Situations such as the preceding arise frequently in practice and are referred to as split-plot experiments. The terminology results from the agricultural origins of experimental design, eg, a farmer needed to compare different fertilizer types on a plot of land with varying normal fertility. A characteristic of split-plot plans is that more precise information is obtained on the low level variables (varnish coats in the preceding example) than on the high level variables (alloy composition). The split-plot nature of the experimental environment, if present, is important information, both in the planning and in the analysis of the experiment.

4.1. Prior Knowledge. The experimenter should draw on subject matter expertise when selecting the response variables, design variables, and variable ranges for experimentation. When selecting the response variable, for example, there may be prior knowledge regarding the need for data transformation. For selection of variable ranges, prior knowledge is often available concerning the expected response outcome at certain experimental conditions. For example, some combinations of conditions might be known to yield poor results or might not be attainable with the equipment available, or worse yet, could result in endangering the plant. Furthermore, all proposed conditions in the experiment need to make sense. For example, a misguided proposed experiment had a condition that required a resistance of 50 ohms for a circuit without resistors. Clearly, unreasonable or hazardous conditions must be omitted from the experimental design, irrespective of whether they happen to coincide with a standard statistical pattern. Thus the experiment must be adjusted to accommodate reality and not the reverse. This can be achieved by adjusting variable ranges to obtain a feasible design space, or alternatively, choosing an experimental design plan that accommodates asymmetric constraints in the design space (see section 7.4, Optimal Designs).

4.2. Response Variables. A clear statement is required of the performance characteristics or dependent variables to be evaluated as the experimental response. Even a well-designed experiment fails if the response cannot be measured properly. Frequently, there are a number of response variables; for example, tensile strength, yield strength, percent elongation, etc. It is important that standard procedures for measuring each variable be established and documented. Sometimes the response is on a semi-quantitative scale; e.g., material appearance may be in one of five categories, such as outstanding, superior, good, fair, and poor. For measurement of sensory attributes such as taste, a nine-point scale is commonly used. In these cases of subjective ratings, it is particularly important that standards that define (and perhaps illustrate) each of these categories are developed initially, especially if judgments are to be made at different times and perhaps by different observers. Consideration need also be given to ensuring that the response variables are measured consistently, and in a well-defined manner, throughout the course of the experiment.

4.3. Types of Repeat Information. The various ways of obtaining repeat results in the experiment need to be specified. Different information about repeatability is obtained by (*1*) taking replicate measurements on the same experimental unit, (*2*) cutting a sample in half at the end of the experiment and obtaining a reading on each half, and (*3*) taking readings on two samples prepared independently of one another, eg, on different runs, at the same target conditions. Often, there is greater homogeneity among replicate measurements on the same sample than among measurements on different samples. The latter reflect the random unexplained variation of repeat runs conducted under identical conditions. A skillfully planned experiment imparts information about each component of variability, if such information is not initially available, and uses a mixture of replication and repeat runs to yield the most precise information for the available testing budget. The way in which such information is obtained must also be known to perform a valid analysis of the results.

4.4. Preliminary Estimates of Repeatability. Initial estimates of repeatability should be obtained before embarking on any significant test program. Such information may be available from previous testing; but if not, a variance component study should be conducted. Its purpose is to validate that all critical experimental variables have been identified. Additionally, this study will quantify the experimental noise that can be expected due to (*1*) an inability to reproduce target conditions of critical variables (known as reproducibility) and (*2*) non-homogeneity of the samples or measurement error (known as repeatability). The first step in conducting the variance component study is to collect samples from preliminary runs at different times, under supposedly identical conditions. Each of these samples is then tested multiple times (or in the case of a destructive test method, homogenous subsamples are tested from each of the original samples). A variance component analysis of the resulting data decomposes and quantifies the amount of variability in each response variable due to differences between experimental runs versus test variation. If the results show that the "run-to-run" variability is high, then one might conclude that the important variables that affect the results have not been identified, and further research may be needed before the proposed experiment can commence. Alternatively, if the measurement variability is unsatisfactory, then it may be appropriate to improve precision by taking multiple measurements of the response variable and use the average of these measurements as the prime response in the subsequent analyses of the experimental results. The process of establishing the precision of a measurement process in advance of conducting the experiment has become known as gage R&R (reproducibility and repeatability) studies.

4.5. Consistent Data-Recording Procedures. Clear procedures for recording all pertinent data from the experiment must be developed and documented. These should include provisions not only for recording the values of the measured responses and the desired experimental conditions, but also the experimental conditions that actually occurred, if these differ from those planned. If they deviate significantly from the set point. It is generally preferable to use the values of the actual, rather than the aimed-at, conditions in the statistical analysis of the experimental results. For example, if a test was supposed to have been conducted at 150°C but was mistakenly run at 148.3°C, the latter temperature would be used in the analysis. (In addition, it is also often instructive to conduct an analysis based upon the planned variable settings so as to compare the estimated experimental error standard deviations. The difference in the resulting two estimates reflects the variability in the response that is introduced due to not meeting the desired settings.)

In experimentation with industrial processes, equilibrium should generally be reached before the responses are measured. This is particularly important when complex chemical reactions are involved. The values of other variables that might affect the responses should also be recorded, if possible. For example, although it may not be possible to control the ambient humidity, its value should be measured if it might affect the results. In addition, variations in the factors to be held constant, special happenings (eg, voltage surges), and other unplanned events should be recorded. The values of such covariates can be factored into the statistical analysis (possibly by performing a so-called *covariance analysis*), thereby reducing the unexplained variability or experimental error. If the covariates do indeed have an effect, this leads to more precise evaluations of the primary variables. Alternatively, such covariates may be related to the unexplained or residual variation that remains after the statistical analysis of the experimental results, using graphical or other techniques.

4.6. Sizing the Experiment Design to Provide Adequate Power. Due to the high cost of processing at all levels of development – from bench-scale to manufacturing, the number of runs devoted to a given block tends to be overly restricted. Thus it is vital to estimate the size of the experiment design needed to provide adequate power.

If the experimenter can define the objective in terms of measurable responses, the statistical power for a candidate design can be calculated from two key parameters:

- the difference in response (symbolized by the Greek letter delta: Δ_y) that, at a minimum, is important to detect for each response: This is the *signal*.
- an estimate of experimental error in terms of its overall standard deviation (symbolized σ_y): This is the *noise*.

The signal-to-noise ratio (Δ_y/σ_y) – the higher the better – drives power. Ideally, power should exceed 80%. If not, the simplest remedy is to increase the number of experimental runs. For more details on power, how it's calculated and what to do when it comes up short, see ref. 4.

5. Stage-wise Experimentation

Contrary to popular belief, a statistically planned experiment does not require all testing to be conducted at one time. As the eminent statistician, George Box has pointed out repeatedly [eg, (5)], the design of experiments is a catalyst for the general scientific learning process. Thus, much experimentation should be sequential, involving, for example, stages of 8–20 runs. This permits changes to be made in later tests based on early results and allows preliminary findings to be reported. For example, an experiment to improve the properties of a plastic material involved such variables as mold temperature, cylinder temperature, pressure, ram speed, and material aging. The experiment was conducted in three stages. After the first stage, the overall or main effects of each of the variables on the experimental responses were evaluated; after the second stage, interactions between pairs of variables were analyzed, and after the third stage nonlinear effects were assessed. Each stage involved about a month of elapsed time, and management was periodically apprised of progress. If unexpected results had been obtained at an early stage, eg, poor results at one of the selected ram speeds, the general plan for the later stages of the experiment might be changed.

Whether or not to conduct an experiment in stages depends on the program objectives and the specific situation; a stage-wise approach is recommended when units are made in groups or one at a time and a rapid feedback of results is possible. Running the experiment in stages is also attractive in searching for an optimum response, because it might permit moves closer to the optimum from stage to stage. On the other hand, a single-stage experiment may be desirable if there are large start-up costs at each stage or if there is a long waiting time between the start of the experiment and the measurement of the results. This is often the case in many agricultural experiments and also when the measured variable is product life.

If the experiment is conducted in stages, precautions must be taken to ensure that possible differences between the stages do not invalidate the results. Appropriate procedures to compare the stages must be included, both in the test plan and in the statistical analysis. For example, some baseline standard test conditions, known as controls, may be included in each stage of the experiment.

A very sophisticated stage-wise approach to designed experimentation called *evolutionary operation* (also known as "EvOp") is detailed by Box and Draper (6). EvOp searches continually for the current optimal conditions on an operating manufacturing process.

6. Other Considerations

Many other questions must be considered in planning the experiment:

1. What is the most meaningful way to express the controllable or independent variables? For example, should current density and time be taken as the experimental variables, or are time and the product of current density and time the real variables impacting the mean response? Judicious selection of the independent variables often reduces or eliminates interactions between variables, thereby leading to a simpler experiment and analysis. Also interrelationships among variables need to be recognized. For example, in an atomic absorption analysis, there are four possible variables: air-flow rate, fuel-flow rate, gas-flow rate, and air/fuel ratio, but there are really only two independent variables.

2. What is a proper experimental range for the selected quantitative controllable variables? Assuming a linear relationship between these variables and performance, the wider the range of conditions or settings, the better, usually, are the chances of detecting the effects of the variable. However, the wider the range, the less reasonable is the assumption of a linear or other simple relationship between the experimental variables and the response variables. Also, one generally would not want to conduct experiments appreciably beyond the range of physically or practically useful conditions. The selection of the range of the variables depends in part on the ultimate purpose of the experiment: is it to learn about performance over a broad region (eg, to establish a "response library", that provides information of what to expect under a variety of possible experimental conditions), or is it to search for an optimum condition? A wider range of experimentation would be more appropriate in the first case than in the second.
3. What is a reasonable statistical model, or equation form, to approximate the relationship between the independent variables and each response variable? The more complex the assumed model, the more runs are usually required in the experiment in order to fit the model. We shall consider this topic further in the discussion of response surface method (RSM) designs.
4. What is the desired degree of precision of the statistical estimates and conclusions based upon the analyses of the experimental results? The greater the desired precision, the larger is the required number of experimental runs. Statistical precision is, most frequently, quantified by a statistical confidence interval. Such an interval expresses, eg, the uncertainty in the estimated mean value of the response variable for a specified set of conditions, in the estimated coefficients of a fitted statistical model, or in the estimated experimental error standard deviation.
5. Are there any benchmarks of performance? If so, it might be judicious to include these conditions in the experiment in order to compare the results with those from past experience (and seek an explanation so as to remove the causes of differences larger than what one might expect from random variation).
6. What statistical techniques are required for the analysis of the resulting data, and can these tools be rapidly brought to bear after the experiment has been conducted (see section 8, Statistical Tools)?

7. Formal Experimental Plans

The test plan is developed to best meet the goals of the program. This might involve one of the standard plans developed by statisticians and practitioners. Such plans are described in varying detail in numerous texts on the design of experiments. As already suggested, in practice, one frequently would use combinations of such plans in a stagewise approach–eg, a factorial experiment conducted in blocks or a central composite design using a fractional factorial base.

7.1. Blocking Designs. As previously described, blocking is used to remove the effect of extraneous variables from experimental error. Well-known blocking designs include randomized block designs and balanced incomplete block designs to remove the effects of a single extraneous variable, Latin and Youden square designs to remove the effects of two extraneous variables, and Greco-Latin and hyper-Latin square plans to remove the effects of three or more extraneous variables.

7.2. Factorial and Fractional Factorial Designs. These very popular designs apply for two or more primary independent variables. Factors are varied simultaneously rather than one at a time, so as to obtain information about interactions among variables and to obtain a maximum degree of precision in the resulting estimates. In complete factorial plans, all combinations of conditions of the independent variables are run. For example, a $3 \times 3 \times 2 \times 2$ complete or full factorial design requires running all 36 combinations of four variables at three, three, two, and two conditions, respectively.

A fractional factorial design is often used when there are a large number of combinations of possible test points, arising from many variables or many conditions per variable or both, and it is not possible or practical to run all combinations. Instead, a specially selected fraction is run. For example, a 2^{5-1} fractional factorial plan is one where there are five variables each at two conditions, normally resulting in a total of 32 possible combinations, but only a specially selected one-half (2^{-1}), or 16, of these combinations are actually run. The selection of test points is based upon the specific information that is desired from the experiment. The optimal 2^{5-1} fraction can independently estimate the impact of all main effects and two-factor interactions, but it is assumed that higher order interactions (ie, interactions among three or more factors) are negligible. This particular design is characterized as *resolution V*. The larger the number of primary variables, the greater is the degree of fractionation that is possible.

Fractional factorial plans are especially useful for screening purposes when it is desired to find those variables that have the greatest impact (over the specified experimental region). Medium-resolution designs that estimate main effects aliased only with interactions of three or more factors are ideal for this purpose. These are characterized as *resolution IV*. An additional advantage of full factorial and fractional factorial designs is that by providing a comprehensive scanning of the experimental region they can often identify, without formal analyses, a small number of test conditions that appear to provide especially favorable results. The region around these conditions would then be explored further in subsequent experimentation.

The most frequently used fractional factorials are for situations in which all variable are at two conditions (or levels). Test plans (providing specific test points) for this situation, together with their properties, are provided in most texts on the design of experiments and by computer programs for generating experimental designs. A specialized family of two-level fractional factorial designs, known as Plackett-Burman (P-B) plans, permits a large number of variables to be evaluated in a small number of test runs and is discussed in many standard texts. However, these P-B designs provide very low resolution – confounding main effects with two-factor interactions, thus they must be used with caution. Designs fractionated to this extent are characterized as *resolution III*. They are useful for verification and ruggedness testing (7).

Detailed discussions of fractional factorial designs are provided in books by Box, Hunter and Hunter and Montgomery (see *General References*).

7.3. Response Surface Designs. These designs apply when one is dealing principally with quantitative independent variables (the x's), such as temperature and pressure, that one wishes to relate to the response variable(s) (y or y's), assuming some statistical model. When the form of this model is known from physical considerations, then one would want to use that model form (if it is not too complicated). These are characterized as *mechanistic* models. As already suggested, when the form of the relationship is not known, as is frequently the case, one instead approximates it by a polynomial (Taylor series approximation) model. These are characterized as *empirical* models. A first-order linear model for k independent variables is expressed by the simple relationship

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \varepsilon \tag{2}$$

where ε represents the model error, which results from any unexplained behavior in y, and is usually assumed to be normally distributed with constant variance. This model error can result from experimental noise, exclusion of important x's in the model, or the need for higher order terms in the model. The results of the experiment are used to estimate the unknown β's using least-squares regression analysis (see Statistical Tools).

However, the model that is most frequently used in practice is the second-order model, also known as a *quadratic*

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_j + \sum_{j=1}^{k} \beta_{jj} x_j^2 + \sum_{i} \sum_{<j=2}^{k} \beta_{ij} x_i x_j + \varepsilon \tag{3}$$

which includes linear (x_j), curvature (x_j^2), and interaction ($x_i x_j$) terms. An intermediate model might include the interaction (or cross-product) terms, but not the quadratic ones.

The choice of an appropriate model merits considerable reflection, based on both physical and empirical considerations. As already suggested, a linear model may provide a reasonable approximation over a narrow range, but a second-order, or more complex, relationship might be called for over a broader range. Also, often based upon physical considerations, an improved, or simpler, representation might be gained by transforming the original y's to some other scale (such as taking logarithms or reciprocals). In addition, transformations of the x's can be beneficial.

Two-level full and fractional factorial designs can be used to explore response surface models that include two-factor interactions. Three level full factorial and some fractional factorial designs apply for second-order models with squared terms. However, these may require more experimental points than are feasible from a practical viewpoint. For this and other reasons, specialized designs have been developed for response surface methods (RSM) analysis, especially for second-order models. Most popular among these are the so-called

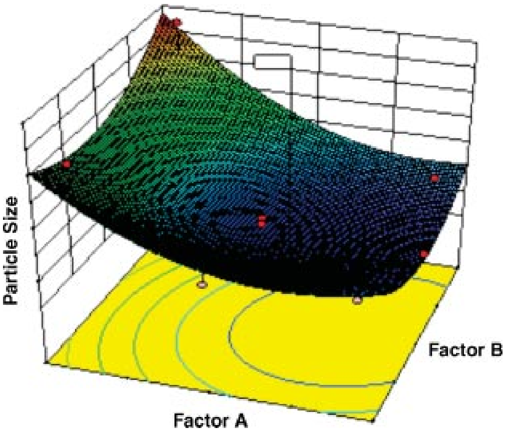

Fig. 4. Response surface from an experiment on nanomilling of a pharmaceutical-grade starch (8).

central composite designs (invented by Box and Wilson) and the Box-Behnken designs.

Ultimately the goal of RSM is to produce a map of the response, for example the one shown in Figure 4, which represents the true surface well enough to accomplish the objective for the optimization – in this case the minimization of particle size.

Response surface method (RSM) designs are discussed in many books on experimental design, and in detail in the book by Anderson and Whitcomb and in greater depth in the texts by Myers, Montgomery, and Anderson-Cook and (see *General References*).

7.4. Optimal Designs. Optimal designs are crafted for specific predictive models based on specific statistical criteria. Due to their computational-intensity they are generally software-generated.

The best-known design of this type is the "D" optimal, which seeks to maximize the determinant of the information matrix X'X of the design. This criterion minimizes the volume of the joint confidence region on the regression coefficients. Because they focus on precise estimates of the effects, D-Optimal designs work well for screening experiments on general (multilevel) factorials.

Optimal designs are of particular value in situations when central composite, Box-Behnken (or other "symmetrical-type") RSM design is not practical or desirable. Thus, in various applications, some points in the design space might be uninteresting, or inappropriate. This might be because, based upon prior knowledge, the results at that condition are fully predictable, or, perhaps, known to be inferior. Another possibility might be that a particular experimental test condition might be potentially hazardous. A good example is a reaction where the specific combination of high time and high temperature creates a runaway condition. Sometimes, such undesired points can be avoided by an appropriate redefinition of the experimental variables. If this is not possible, the resulting design region might be asymmetrical. For this case, optimal designs might be especially useful. A good criterion for this purpose is the “IV” (or “I”), which makes use of an integrated variance criterion that minimizes the average variance for predicted responses throughout a region of interest. An IV-optimal design tends to place more runs more uniformly throughout the experimental region than D-optimal.

In addition, optimal designs are employed for situations in which, based upon process knowledge, one wishes to fit a nonstandard model (i.e., one that assumes something other than a first- or second-order polynomial model with normally distributed errors), or in which one is limited to running a specified number of test points. Additional detail is provided in the book by Myers, Montgomery and Anderson-Cook (see General References).

7.5. A Strategy of Experimentation for Process Improvement. Figure 5 charts a tried-and-true strategy of experimentation that begins with a number of unknown factors to be quickly screened via medium-resolution (“Res”) fractions of two-level factorial. During this discovery phase, factors known to affect the process are temporarily held fixed. After throwing the trivial many factors off to the side (preferably by holding them fixed or blocking them out), the experimental program enter a phase where interactions get characterized by high-resolution designs. This requires higher-resolution, or possibly full, two-level factorial designs. Center points – mid-level settings of all numerical factors – provide a measure of curvature.

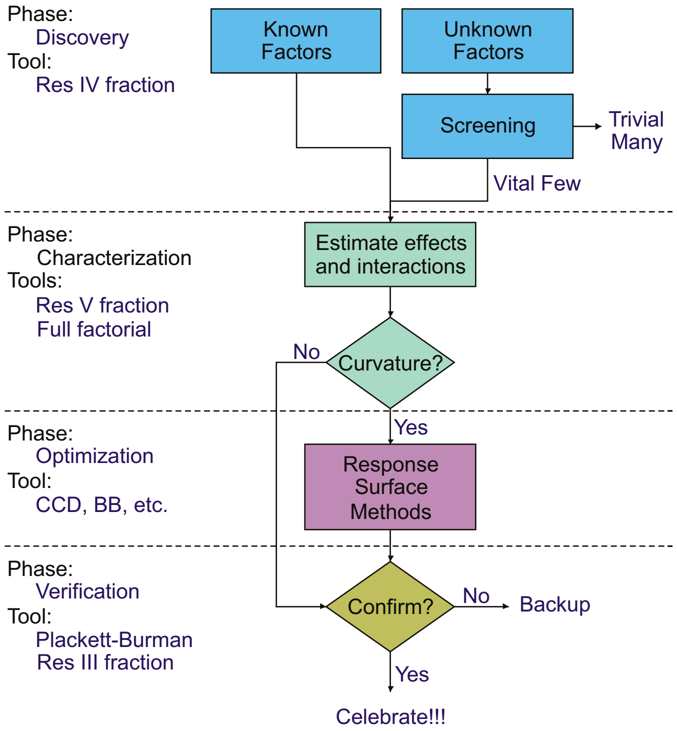

Fig. 5. Strategy of experimentation.

If curvature is significant and important, response surface methods (RSM) come into play. The typical tools used for RSM are the central composite design (CCD) and Box-Behnken design (BBD). If constraints make these standard templates impractical, optimal designs can be employed.

The last phase of this strategy for experimentation provides verification that the proposed factor settings will achieve the process objectives. Low resolutions fractions of standard two-level designs are acceptable at this stage. Plackett-Burman designs also find use for such purposes (see referenced ASTM standard, for example).

7.6. Mixture Designs. These designs arise when the variables under consideration are the ingredients of a product that must add to a constant and the response is a function of proportions. For example, in an experiment to learn how to optimize the baking of a cake, the relative impact of each of the ingredients (baking powder, shortening, flour, sugar, milk, water, etc) might need to be assessed (9). Often, an experimenter will want to study *both* mixture and non-mixture variables. In the previous example, one may wish to assess the effect of varying baking time and temperature, in addition to the relative percentages of the cake ingredients. These designs are referred to as *mixture-process designs*. Both mixture and mixture-process designs are especially relevant in the evaluation of chemical processes. Detailed expositions are provided by Cornell and also by Smith (see *General References*).

7.7. The Taguchi Approach. An eminent Japanese engineering professor and advocate for quality improvement, Genichi Taguchi, has been an important proponent of planned experimentation in product design, development, and evaluation. In fact, the influence of Taguchi and his followers has been an important factor in popularizing the use of experimental design in industry. Taguchi has especially promoted minimal-run, low-resolution "orthogonal arrays", a family of experimental designs derived from standard fractional factorial plans.

Taguchi has also stressed the role of experimental design in identifying the process conditions that show the greatest consistency, or are most robust, in the face of variability in manufacturing conditions and customer use. The goal is to achieve similar and desirable product performance despite uncontrollable variation due to such noise variables as incoming raw materials, and the ambient conditions under which a product is to be used. For example, an automobile battery needs to work well, irrespective of whether a car is driven in Florida or Minnesota.

In particular, Taguchi has introduced experimental plans that help identify the conditions (ie, settings of the primary variables) that minimize the variability in the noise variables, thus elevating the role of such variables from nuisance factors whose impact needs merely to be neutralized by appropriate blocking, etc, in the experimental plan.

Many classical blocking designs, such as incomplete block and Latin square designs assume that there is no interaction between the primary variables and the noise variables and that the variability generated by the noise variables is the same for all conditions of the primary variables. These plans can, therefore, not be used for identifying the conditions of the primary variables that minimize variability. Instead, Taguchi has proposed "robust parameter designs" for this purpose, including the concept of "inner and outer array" plans. In such designs, the noise variables (outer array) are evaluated for a series of conditions of the primary variables (inner array). Typically, the inner array is in the form of a fractional factorial plan, while the outer array is a full factorial plan. These designs have been criticized due to the fact that, although they provide the desired information about the noise variables at various conditions of the primary variables, they tend to provide limited information about interactions among primary variables (due to the fractionization of the inner array). Also, though appropriate for some "split-plot" situations, these plans frequently lead to a large number of test conditions. Instead, some "combined array" experimental plans have been proposed to meet Taguchi's basic goal, but in a more efficient manner; see the book by Montgomery (see *General References*) for a more detailed discussion.

Taguchi has also proposed various methods of data analysis and other concepts, some of which are controversial. These include the concept of combining the mean and variability in the response variable in the form of the "signal/noise ratio," which is to be maximized. This concept has been criticized on the grounds that it is generally more informative to evaluate the mean response and the variability separately, rather than combining them into a single metric.

For a detailing of Taguchi methods see the text by Roy (see *General References*).

8. Statistical Tools for the Analysis of Designed Experiments

We have so far in this article made little mention of statistical tools for the analysis of experiments for two reasons. First, as stated at the outset, we feel that the proper design of the experiment is more important than its analysis. Second, various books on experimental design, including those in the General References, provide ample descriptions of these methods. Thus, we provide only a brief summary. In particular, the analysis of most designed experiments involves a combination of three major types of tools: the analysis of variance, regression analysis, and graphical analysis. We will comment briefly on each. In so doing, we reiterate that in stage-wise experimentation such analyses are interspersed with the experimental runs and, in fact, provide important inputs on how to proceed in the next stage of testing.

The analysis of variance ("ANOVA") is, probably, the most frequently used tool for analyzing the results of a designed experiment. It is a formal statistical method that breaks down the total observed variability in a response variable into separate parts attributable to each of the sources of variation, based on an assumed statistical model. Thus, the analysis of variance isolates the variation due to the main effects of each of the experimental variables and their interactions and assess the "statistical significance" of each of these sources of variation. This provides useful information about the relative importance in the experiment of different effects and their interactions.

The results of an analysis of variance should, however, not be over-interpreted because "statistical significance" does not necessarily imply practical importance. In particular, a statistically significant result is one that is unlikely to have occurred due to chance alone. It does not, per se, measure the magnitude of the impact of the associated effect. For Type II variables, the analysis of variance also provides insightful estimates, in the form of so-called “variance components”, of the variability attributable to individual sources of variation.

Regression analysis allows one to fit a relationship between a series of experimental variables (x’s) and a response variable (y), which is especially relevant for data from response surface designs. For such designs, regression analysis allows one, for example, to estimate the unknown β’s in an assumed model, such as the previously discussed second-order model, together with their associated confidence bounds. This will permit prediction of an expected response and/or variability for that response, also with statistical confidence bounds, for a specified combination of the experimental variables, based upon the assumed model. In fact, regression analysis has been referred to as “the workhorse of statistical data analysis”.

After a model is fitted, it is frequently desired to exercise it to find an optimum condition, or set of experimental conditions. This would be the set of conditions that minimizes or maximizes the response variable when one is considering the average response, or the condition that minimizes the variability (or some desirable combination of the two); the book by Myers, Montgomery and Anderson-Cook (see *General References*) provides further discussion.

We strongly encourage graphical analyses to supplement the more formal statistical analyses, and, on occasion, to take their place. For example, half-normal probability plots (see the book by Montgomery in *General References*) provide information similar to a formal analysis of variance concerning the relative importance of the individual sources of variation, but in graphical form.

9. Multiple Response Variables

The preceding discussion has assumed, by and large, that we are dealing with a single response variable, although, in practice, this frequently is not the case. As already mentioned, one may be interested in both the average and the variability of the response variable (leading Taguchi to propose the controversial signal/noise ratio, to combine these two variables into one for the purpose of analysis). Moreover, often one is concerned with two or more performance variables, such as, for example, the viscosity, tensile strength, ductility, etc, of a plastic material. Multiple response variables complicate the analyses as well as the stage-wise development of the experimental plan, especially if we are seeking an optimum experimental region (versus just developing a response library). A number of software packages that feature response surface methods provide tools for numerical optimization of multiple responses based on maximizing a desirability function. This is detailed by Myers, Montgomery and Anderson-Cook (see *General References*) in their chapter on multiple response optimization.

10. Some Historical Background

Sir Ronald Fisher is generally regarded as the “founding father” of the design of experiments, and wrote an early book on the subject (10). As previously indicated, many of the initial applications were in agriculture. Application to industry, in general, and the chemical industry, in particular, was recognized after World War II, and was spurred on by important work by George Box, Cuthbert Daniel, and Stuart Hunter, among others. Starting in the late 1970s, the previously mentioned work by Genichi Taguchi and his associates generated extensive interest in experimental design in industry (as well as some controversy). Most recently, the Six Sigma initiative (11) has resulted in focusing a great deal of attention by some large corporations on the design of experiments. In fact, the statistical design of experiments is one of the key items in the Six Sigma "toolbox" especially in Design for Six Sigma (DFSS).

11. Acknowledgments

This article is an update of the work of Gerald J. Hahn and Angela N. Patterson, who at the time of its writing were colleagues at General Electric. We are grateful for their wise words, which we changed as little as possible. As famed statistician and quality guru W. Edwards Deming preached, when a system is stable and running at a high level of quality, it does not pay to tamper with the process.

BIBLIOGRAPHY

"Design of Experiments" in *ECT* 3rd ed., Vol. 7, pp. 526–538, by G. J. Hahn, General Electric Co.; in *ECT* 4th ed., Vol. 7, pp. 1056–1071, by Gerald J. Hahn, General Electric Company; "Design of Experiments" in *ECT* (online), posting date: December 4, 2000, by Gerald J. Hahn, General Electric Company; in *ECT* 5th ed., Vol. 8, pp. 383–410, by Gerald J. Hahn and Angela N. Patterson, General Electric.

CITED PUBLICATIONS

1. M. J. Anderson, "Trimming the FAT out of Experimental Methods," *Optical Engineering*, (June 2005).
2. M. J. Anderson, "Trimming the FAT: Part II," *Optical Engineering*, (September 2005).
3. W. F. Adams and co-workers *Handbook for Experimenters*, Stat-Ease, Inc., Minneapolis, Minn. 2009.
4. M. J. Anderson and P. J. Whitcomb, *DOE Simplified, Practical Tools for Experimentation*, 2nd ed., Productivity Press, New York, 2007 (1st ed., 2000).
5. G. E. P. Box, *J. Quality Technol.*, **31**(1), 16-29 (1999).
6. G. E. P. Box and N. R. Draper, *Evolutionary Operation: A Statistical Method for Process Improvement*, John Wiley & Sons, New York, 1998.
7. E1169-07, *Standard Practice for Conducting Ruggedness Tests*, ASTM International, West Conshohocken, Pa, 2007.
8. N. I. Bukhari and co-workers "Statistical Design of Experiments on Fabrication of Starch Nanoparticles – A Case Study for Application of Response Surface Methods," April 2008, www.statease.com/pubs/doe_for_starch_milling.pdf.

9. M. J. Anderson and P. J. Whitcomb, "A Primer on Mixture Design: What's In It for Formulators? 2009, www.statease.com/pubs/MIXprimer.pdf.
10. R. A. Fisher, *The Design of Experiments*, Oliver and Boyd, London, 1935.
11. T. Pyzdek and P. A. Keller, *The Six Sigma Handbook*, 3rd ed., McGraw-Hill, New York, 2009.

GENERAL REFERENCES

This section provides a listing of books that mainly deal with the application of experimental design to scientific, industrial, and general situations. Books directed principally at educational, psychological, or related applications, and those dealing mostly with the theory of experimental design, or principally with the analysis of experimental data, are omitted. Other than the first few books, most presume that the reader has had at least one introductory course in statistics. In addition, such journals as *Quality Engineering*, *The Journal of Quality Technology*, and *Technometrics* (in increasing order of complexity) periodically carry articles about experimental design.

M. J. Anderson and P. J. Whitcomb, *RSM Simplified, Optimizing Processes Using Response Surface Methods for Design of Experiments*, Productivity Press, New York, 2005.

G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters, Design, Innovation and Discover*, 2nd ed., John Wiley & Sons, Inc., New York, 2005 (1st ed., 1978)

G. E. P. Box and N. R. Draper, *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd ed., John Wiley & Sons, Inc., New York, 2007 (1st ed., 1987)

J. A. Cornell, *Experiments with Mixtures, Designs, Models and Analysis of Mixture Data*, 3rd ed. (1st ed., 1981)

D. C. Montgomery, *Design and Analysis of Experiments*, 7th ed., John Wiley & Sons, Inc., New York, 2009 (1st ed., 1976).

R. H. Myers, D. C. Montgomery, and C. M. Anderson-Cook, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed., John Wiley & Sons, 2009 (1st ed., 1995).

R. K. Roy, *Design of Experiments using Taguchi Approach: 16 Steps to Product and Process Improvement*, John Wiley & Sons, Inc., New York, 2001.

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

Mark J. Anderson
Patrick J. Whitcomb
Stat-Ease, Inc.