Sequential Experimentation

Introduction

R.A. Fisher, the originator of statistical experimental design, once remarked that "The best time to design an experiment is after it's done". Wry, but certainly true: the more one knows about what factors are important, what extra-experimental sources of variation might **bias** results, which measurements are difficult to make (i.e., are noisy/imprecise) and which are not – in short, the more one knows about the system under study – the better one can focus an experiment to learn more about it.

There is nothing particularly profound in this, of course. Even the justly maligned but widely used practice of studying many factors by holding all but one constant and varying just this one (OFAT: one-factor-at-a-time experimentation) acknowledges this truism. Each experimental factor is varied to its "optimal" setting, often by looking at whether the last change in the factor made results better or worse. Once the "best" setting for a factor has been found, it is then held there and the next factor of interest is then similarly examined. Of course, "optimal" and "best" must be qualified by scare quotes, as this strategy ignores both experimental noise and the possibility of interactions among the factors, issues that will be discussed further shortly. The bottom line is that more careful approaches are needed.

In any industrial or business process, there are always potentially many factors that might affect the response(s) of interest. The classic **Ishikawa "fishbone" diagram**, Ishikawa [1], is one well-known quality tool that can be used by those who share process knowledge to identify from dozens to literally hundreds of methodological details, raw material variations, equipment settings, environmental factors, and even maintenance procedures that might impact results.

As a canonical example, consider a measurement process to measure "purity" (of a chemical, food product, or drug, say). Such a process requires a procedure for obtaining and preparing samples to be measured, for storing them over the course of the preparation, to prevent outside contamination or chemical reactions with the environment, and for obtaining a readout of the prepared samples in an instrument. Typically, the protocols defining these procedures contain many steps – how to sample, how to handle the samples, what reagents to use, what times and temperatures the samples can be stored under, how the instrument is to be set up, calibrated, maintained, and so forth. A long list of individual factors that might affect the process would result.

In reality, for a process to work at all, prior knowledge and experience will generally dictate how to control most, if not all, of these many details satisfactorily. When this is not the case, typically a relatively small subset of process variables – less than, say, 15 or so and perhaps only 1 or 2 – can be identified for empirical investigation and adjustment, that is, experimentation. That is the focus of this article.

If "first principles" and prior experience have whittled the number of process variables down to at most one or two, it is generally not difficult to come up with a reasonably efficient experimental strategy that achieves success. Here, "efficient" simply means with an acceptable expenditure of time and effort. It is important to note, however, that "efficiency" can be defined in a statistical way that formalizes this idea of minimizing experimental effort in order to design "optimally" efficient experiments (*see* **Optimal Design**). While this theory turns out not to be quite so "optimal" in practice, it nevertheless yields very useful insights and approaches that can greatly reduce experimental effort. The strategy of sequential experimentation implicitly uses much of this methodology. So while no claims of statistical optimality will be made, a major advantage of using this approach is that it achieves success with less experimental effort.

But what happens if one cannot identify one or two variables that can be manipulated and cannot achieve a satisfactory result in those that are used. This is often the case, for example, when there are several competing criteria that together constitute "satisfactory" performance: it may be easy to optimize the criteria individually but difficult to find a compromise that is satisfactory for all.

However it comes to pass, when there is a need to investigate the settings of more than just one or two process variables, it is no longer so obvious how to proceed. Following Fisher's observation, the experimental strategy should be a frugal one. That is, whenever possible (it isn't always!), one would prefer to gradually accumulate understanding through a sequence of relatively small experiments, rather than gambling everything on a large, one-shot effort. Not only does this permit later experiments to take advantage of what has been learned, it also allows the experimenter to recover from "mistakes" – misjudgments about what variables should be investigated and over what ranges, for example – before experimental resources (or managerial patience!) have been exhausted. No experimenter who is truly on the frontier of knowledge can avoid making such mistakes, so it makes sense to employ a strategy that quickly reveals and recovers from errors so that more fruitful directions can be pursued.

An Outline of the Strategy

There are three principles that underpin the sequential experimentation strategy.

1. Exploit Occam's razor, which is an early version (from the fourteenth century English logician and friar, William of Ockham) of the familiar ***Pareto principle*** that asserts that in any process, about 80% of the process variation is caused by about 20% of the process variables. In this context, it means that one can expect relatively few (sometimes none!) of the experimental variables being considered to meaningfully affect the response. By employing a strategy to efficiently identify and quantify the effects of just these few, one can learn how to achieve the desired performance with great experimental economy. Because time and resources are always limited, this is often crucial.
2. Distinguish signal from noise. To be useful, an experiment must yield results that are replicable. In other words, to infer that observed changes in a response were actually caused by changes in an experimental variable – which can then be used to control the response – those changes must not in fact have been due to uncontrolled and unknown sources of experimental variability. Of course, this is a fundamental precept of science, not merely of statistical experimental design.
3. Allow for interaction (*see* **Interactions**). Factors often affect results interactively. Example: the concentration of glucose needed for best growth of a cell culture depends on the concentration of certain amino acids present in the growth medium. One cannot specify how glucose affects growth without simultaneously specifying how much amino acid is present (and conversely). This kind of behavior is clearly not uncommon, and so it is important that multifactor experiments allow for such possibilities (*see* **Factorial Experiments** for further discussion).

In many respects, these principles seem almost self-evident; but in fact, OFAT experiments violate all of them! They fail the first, because equal effort must be expended to investigate each experimental factor; there is no "information sharing" that allows, for example, *hidden replication* (*see* **Fractional Factorial Designs**) or elucidation of interactions in the space of *active factors* (various aspects of which are discussed in **Aliasing in Fractional Designs**; **Foldover Designs**; **Fractional Factorial Designs**; **Plackett–Burman Designs**; **Projectivity in Experimental Designs**). They fail the second, because they make poor use of experimental information and therefore are only reliable in sufficiently high signal/noise situations. This can only be achieved by (appropriately!) replicating the individual factor changes a large number of times, thereby incurring excessive experimental effort. They fail the third because interactions cannot be "seen" when only OFAT is changed: by definition, interactions require the simultaneous change of two or more experimental factors to evidence themselves (*see* **Interactions**).

To gain a rough sense of how ineffective OFAT can be compared to the strategy outlined here, suppose that one has 10 experimental factors to study. By 1, it is reasonable to expect that no more than two or three of them will have large enough effects – either individually or jointly (interactions) – to be of concern. With this in mind, one could use a so-called 12-run Plackett–Burman design (named after the two British mathematicians who discovered it in 1946 – *see* **Plackett–Burman Designs**), perhaps with a few added center points (*see* **Center Points**) to determine what the important factors were and how they affect the response of interest. An additional few runs or foldover (*see* **Foldover Designs**) are occasionally required to resolve possible uncertainties that might exist (e.g., due to aliasing, *see* **Aliasing in Fractional Designs**). So perhaps 15–20 experimental runs might be required to distinguish the few most active factors (if any!) from noise (2) and gain a general sense of their individual and joint effects (3).

By contrast, to gain the same information on the individual effects using the OFAT strategy, a minimum of 120 experimental runs (and some center points, perhaps) are required. This is roughly six to eight times the experimental effort! Moreover, even with this large number of runs, there is no information about any possible interactions!

The strategy outlined here was developed over the course of about 70 years (starting about 1920) by Fisher, Frank Yates, Oscar Kempthorne, Gertrude Cox, George Snedecor, George Box, and numerous other statisticians. Interestingly, most of the work prior to 1950 had its roots in agricultural experimentation to improve crop and livestock yields, reduce pests, and so forth. More recent (post-1990) developments by Box, John Tyssedal, Ching-Shui Cheng (*see* **Projectivity in Experimental Designs**), and others make use of the mathematical properties of Hadamard matrices and other two-level orthogonal arrays (*see* **Orthogonal Arrays**) originally discovered by the French mathematician, Jacques Hadamard and further developed by Rao and others in the first half of the twentieth century.

The strategy proceeds in stages as follows:

1. Use small two-level screening designs (*see* **Screening Designs**) to identify the most important few factors of interest and roughly quantify how they affect the response. Fractional factorials (*see* **Fractional Factorial Designs**) and other designs of projectivity 3 or 4 (*see* **Plackett–Burman Designs**; **Projectivity in Experimental Designs**) can be used. Depending on which, full or partial aliasing may require a small number of carefully chosen follow-up runs to resolve any confounding and clearly identify the influential factors. While there is statistical methodology available to guide this choice, for example, using foldover (*see* **Foldover Designs**) or as in Meyer *et al.* [2] and Box *et al.* [3], in practice, this is rarely necessary. Subject matter expertise and the hierarchy of effects – which asserts that main effects are more likely than two-factor interactions (2 fi's), 2 fi's are more likely than 3 fi's, and so forth – are usually sufficient.
2. Stage 1 identifies the important factors and quantifies their effects. This provides an approximate direction for improvement that can now be rapidly explored by running several experimental runs in this direction. This is essentially an empirical version of the *method of steepest ascent* (*see* **Response Surface Methodology**). For example, suppose two factors of 10, A and B, have been identified, and the projection of the fitted **response surface** (equivalently, the fitted model coefficients) indicates that the response will improve most rapidly if A is increased and B is decreased in approximately a 2:1 ratio in standardized scaling units (i.e., units in which the ranges of all experimental variables are -1 to $+1$). Suppose the low and high levels of A and B in stage 1 in their natural (unstandardized) units are l_a, l_b, h_a, and h_b respectively, and δ_a and δ_b are $h_a - l_a$ and $h_b - l_b$. Then in stage 2, the values of the other eight experimental factors would be held constant (at the centers of their range or similar "standard" values chosen for economic or other reasons), and A and B pairs of the form $m + t\Delta$ would be run, where $m = ((h_a + l_a)/2, (h_b + l_b)/2)$, the center of the stage 1 design in A and B, and $\Delta = (2\delta_a, -\delta_b)$. To control for the possibility of an additive shift in response between the stages, typically one would include one or more replicates of the center point, $t = 0$, in this sequence. Note that this can be considered a one variable design in the *constructed variable*, Δ. However, unlike OFAT, the variable was constructed based on experiment – it reflects what nature has to say, not prior ignorance.
3. As t is increased, the response may improve for a while, but eventually one will go too far, and it will crash or various physical or economic constraints will make further increments in t impossible. At this point, there are several possibilities.
   (a) If results are deemed to be satisfactory, you're done: declare victory and move on.
   (b) A new design region in A and B will have been identified. With A and B at these new levels some of the other eight previously unimportant factors may now have meaningful effects. This is equivalent to the existence of interactions of A and B with the remaining factors over the wider range of A and B from those of stage 1 to the current levels.[a] Hence one might wish to return to stage 1 and repeat a screening design in the original factors with A and B either held constant or varied around their new levels.
4. Stages 1–3 may be repeated until either no further improvement results or other constraints

(technological, economic) prevent further change in the variables. At this point, one will typically have completed a final two-level screening design in the final optimal region and identified a relatively small number of variables that affect the response in this region. At this point, it may be necessary to explore the region of the optimum in greater detail to determine the sensitivity of the response to changes of the critical variables/validate ranges within which to hold these variables to keep the response within desired limits. In currently fashionable language, this would constitute a kind of process *robustness* study (*see* **Robust Design**; **Tolerance Design**). A wide variety of approaches to do this have been proposed (*see* **Performance Measures for Robust Design**; **Product Array Designs**; **Tolerance Design**; **Box–Meyer Method for Dispersion Effects**, for example), but the classical and most widely used approach is to add *axial points* to the existing screening design to create a second-order *central composite design* (*see* **Central Composite Designs**). The fitted response surface can then be both graphically and mathematically explored using, for example, ridge analysis (*see* **Ridge Analysis in Experimental Design**) to provide often revealing insight into the behavior of the system under study.

A great virtue of this approach is *sequential assembly*, Box *et al.* [3] (*see* **Central Composite Designs**): the axial and cube portions of the full second-order design comprise separate *blocks* (*see* **Blocking**), and so any additive shift that may occur is confounded with blocks. Running center points in both the earlier cube portion and the newer axial portion would allow **estimation** of such a shift; but even if this were not the case, both the first- and second-order (including interaction) variable effects would not be corrupted by a systematic shift in response levels. In other words, valid estimates of the effects – and therefore of the response surface – would be obtained even if there were a shift in system behavior between the times the cube portion and axial portion were run. Some of the **robust design** methods – for example, product array designs (*see* **Product Array Designs**) – that have been proposed do not easily fit into this paradigm and require large designs run *de novo*. While there are certainly situations where this is justifiable, it may come at a considerable cost. Sequential assembly attempts to avoid those costs.

Although this overall strategy may seem complex, in concept it is actually rather simple. Here is a little constructed example and picture that shows how it works as compared to standard OFAT experimentation. Suppose there is interest in reducing the amount of a certain contaminant – a reaction by-product – that is formed along with the main reaction product in a biochemical reaction. The contaminant is measured as a percentage of the main peak by ion exchange chromatography, and engineers believe that changes (typically reductions) in reaction time and concentration of a reagent should help reduce it. Figure 1 shows a typical OFAT approach: the initial setting has a time of 9.3 s and a concentration of 11.8. The first set of experiments, shown as white circles, vary the time up and down while holding the concentration fixed. The true, but of course unknown to the experimenters, behavior of the system is shown by the shaded contamination contour plot. If results are very precise, these results would show a slight reduction in contamination by reducing the time to about 8.4 s, as indicated by the X on the horizontal line. Then a second set of experiments would be conducted holding the time fixed at this "optimal" value and varying the concentration as shown by the black circles. The optimal concentration would then be found to be about 10.7, indicated by the second X.

While significant improvement results – the off-peak percentage is reduced from around 2.5 to around 1% – one could still clearly do better. Figure 2 shows how. The 2^2 design indicated by the four white circles is run instead. Analysis of these results would show that that the best direction for improvement is that indicated by the arrow. A series of subsequent trials in this direction would clearly lead to the optimal region at the lower left. Incidentally, the key feature of this process is that the time and concentration have an interactive effect on the percentage contamination. The 2^2 design can "see" this interaction, but the OFAT experiment cannot.

Of course, there are both variations and enhancements that may be required to handle special situations, some of which will be mentioned in the next section. But these are technical details. One should not lose sight of the fact that the underlying strategy is a straightforward implementation of the three basic principles that are essential to sound and efficient

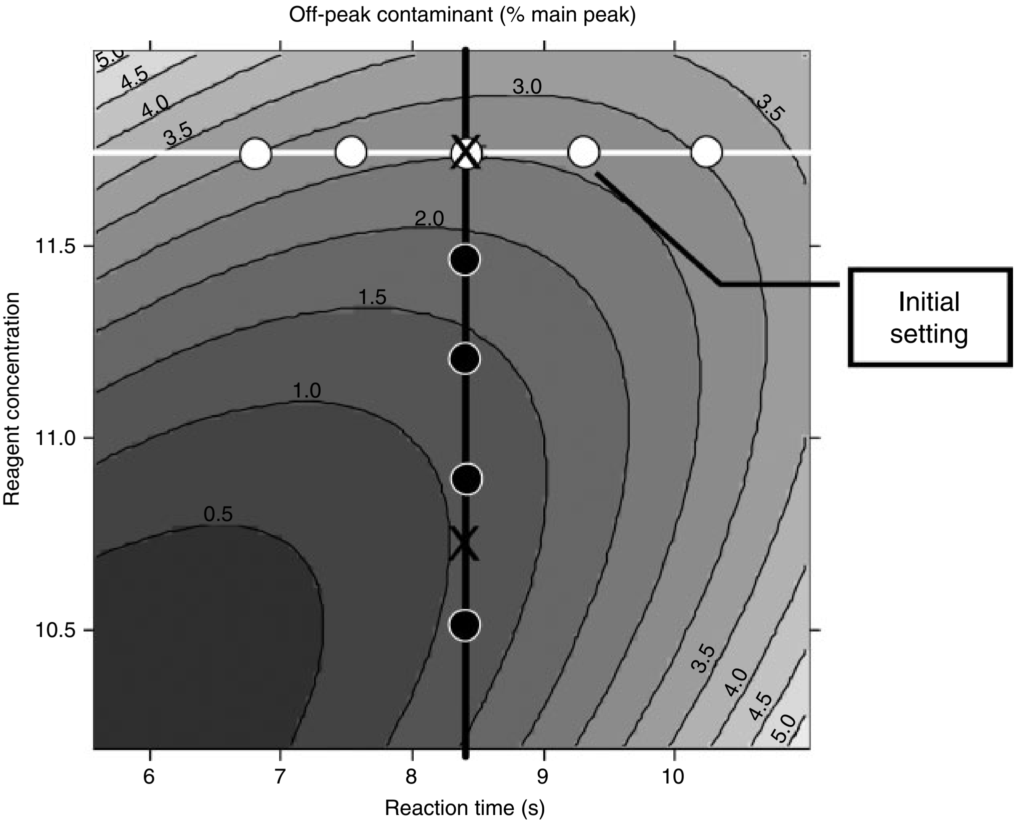

Figure 1 Graphical representation of a one-factor-at-a-time experiment. First, time is varied holding concentration constant; then concentration is varied holding time at the previously found best setting. This leads to the false optimum indicated by the X on the vertical line

experimentation. This is why the strategy has such wide potential application and value.

Comments and Caveats

Given the development and claims of the preceding section, one may well ask: "If the sequential experimental strategy is so effective, why is it so little used in the world of science and technology (as any cursory review of the literature demonstrates)?" While a full discussion of this point could probably fill several volumes,[b] it is perhaps more reasonable to frame the issue in terms of the practical problems that one encounters when trying to execute such a strategy. Here are several that come to mind.

1. Screening designs with many factors are complex and difficult to do. They require that the levels of many variables be simultaneously changed, which can be difficult or even impossible in practice. For example, in an experiment to investigate the effect of time after addition of a reagent and the temperature at which the next step of a chemical process takes place, an exothermic reaction may make it difficult to cool the solution down to a desired low temperature if the time after the reagent addition is too short. The logistics of executing simultaneous changes in many variables may also be difficult in some situations, especially when programming automation software is involved.
2. Two-level screening designs and steepest ascent cannot be used when there are categorical factors with more than two categories that need to be tested. While there are sometimes screening-type designs that can accommodate this situation

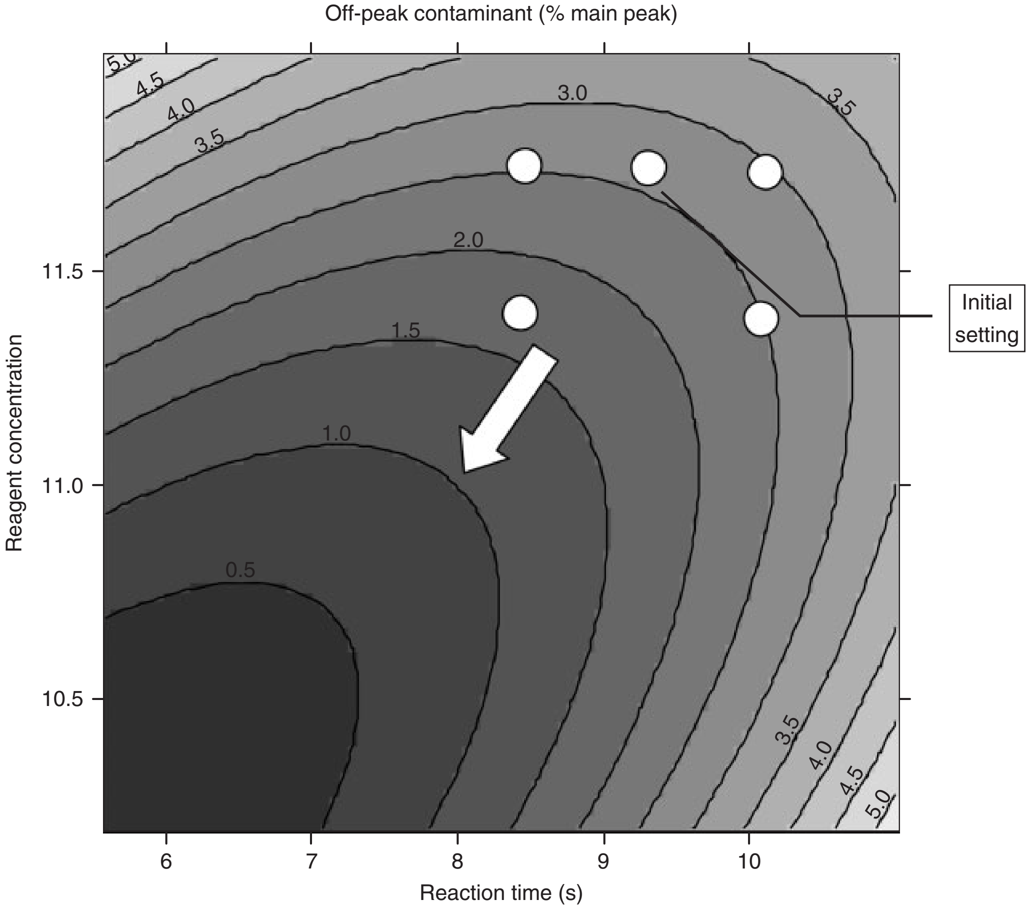

Figure 2 Graphical representation of the sequential experimental strategy. An initial factorial experiment, indicated by the four white circles around the initial setting, shows that the optimal direction of improvement indicated by the arrow. Subsequent experiments in this direction would lead to the true optimum

(*see* **Orthogonal Arrays**), it is the nature of categorical variables that to understand what happens at any combination of categories, you have to test all possible combinations. For example, to test the possible effects and interactions of fixture type, product type, and bearing material on the performance of an assembly, you have to test all possible combinations of fixture, product type, and bearing material. When there are many such factors with many categories each, factorial experiments (*see* **Factorial Experiments**) may become large and unwieldy.

3. It can be very difficult to determine workable ranges for all the variables in screening experiments with many factors. One reason for this is the so-called curse of dimensionality: the volume of the experimental space increases exponentially with the number of variables. So, for example, if one has only two factors to study, in standardized coordinates the corner points where both factors are at their extremes are only about 40% farther from the origin than individual ranges ($\surd 2$ vs 1); but with 10 factors they are more than three times more distant ($\surd 10$ vs 1)! In any real process in which each of the factors are expected to vary more or less randomly and independently within their individual ranges, it is highly unrealistic to expect more than a few of the factors to be simultaneously at one of their extremes. While effects sparsity (i.e., Occam) may be invoked,

in practice such high-order "cube" points are sometimes obviously unworkable. A statistical restatement of this is that high-order interactions are important in this situation. A simple example is when the amounts of two sugars, glucose and mannose, must be varied to determine how to improve the yield of a cell culture. Individually amounts can be zero, but if both are zero simultaneously – the (–, –) level in a simple 2^2 design – the culture cannot grow.

This reasoning suggests that one might want to make the ranges of experimental factors less extreme when there are more of them than when there are fewer. But how much less extreme? And does this exclude settings in which only one or two factors are extreme that should be investigated? See Voelkel [6] for some related ideas. The problem is that there is a lot of volume in which to experiment, and screening designs are very dependent on Occam in such circumstances just to establish workable initial ranges. One approach is to run preliminary experiments at potentially "dangerous" settings, but this can be time consuming and *ad hoc*.

4. One of the cornerstones of experimental design is randomization (*see* **Randomization in Experimental Designs**). But more often than not it is difficult or impossible to fully randomize. As an example, suppose one is interested in optimizing a plate-based assay for a drug metabolite in blood (these are usually required for **clinical trials** of drugs). Such assays are conducted in 96-well plates, about 10 cm × 15 cm in size, in which the wells are arranged in a 8-row × 12-column matrix. Factors that one would be interested in investigating typically include well-specific components of the biochemistry (pH, buffer type, reagent concentrations) that can be changed well by well within each plate (perhaps with some effort) and plate-specific components like time and temperature of incubation that can only change from one plate to another. It is impossible to fully randomize such experiments, because this would have to occur at the well level, and times and temperatures cannot be changed at that level.

   Such an experiment would have to be conducted as a split-plot design (*see* **Split-Plot Designs**), but with many factors, such designs are not simple to plan or analyze. When experiments are conducted as split-plot designs but analyzed as if they were completely randomized, it is quite possible to reach incorrect conclusions. For example, conducting a two-level unreplicated **fractional factorial** as a split plot would actually mean that there are no **degrees of freedom** for the main effects of the whole plot factors (however unaliased interactions with subplot factors could be estimated). The simple and appealing graphical analyses described elsewhere in this encyclopedia (*see* **Half-Normal Plot**; **Lenth's Method for the Analysis of Unreplicated Experiments**) are completely inappropriate. By contrast, if sufficient replication is done and main effects predominate, OFAT would work.
5. One of the strengths of screening designs is that they are very efficient: each experimental run provides a lot of information. This is just another way of describing why OFAT designs with many factors require so many more runs to gain equivalent information (and even then only on the main effects). But this blessing of high efficiency can also be a curse: if a run is "lost" due to some kind of experimental glitch – a not uncommon occurrence – a lot of information can be lost with it. Perhaps worse, if some kind of experimental problem leads to a severely corrupted result, *but the experimenter is unaware that this has happened,* the single bad value can distort all the estimates in an unreplicated design. In OFAT, losing a run here or there makes little difference, and a corrupted value can distort only the single factor being varied. When there is less information to be gained, there is less information that can be lost.

This litany of problems may seem to be a fatal indictment of the sequential experimentation strategy. Not so. As indicated above, there are practical approaches to deal with all these issues. But it is important that any exposition or advocacy of statistical experimental design methods deal fully and fairly with the real circumstances that one faces in using them. As a rule, the statistical experimental design literature does not concern itself with such vexing realities. Thus, researchers who encounter them either knowingly or unknowingly when attempting to apply the methods often do not achieve the results that were advertised, or may even not achieve useful results at all. Attention to detail is key, but without being told about these details, no attention can be paid.

Final Comments

Unfortunately, even this exposition still fails to touch on many important aspects of the sequential design strategy. One of those alluded to in the introduction is that of multiple responses – it is rare that only a single response is of interest; most of the time there are several that "compete" with each other for optimal settings. Often, this is *the* essential problem driving the investigation. Examples: it is easy to increase yield at the **cost of quality** or to improve quality but reduce yield; but it is difficult to do both. It is easy to improve assay precision at the cost of sensitivity or to increase sensitivity at the cost of precision; but it is difficult to do both. One of the real strengths of the sequential design strategy is that it often makes manifest how such trade-offs should be made. Even better, it often provides ways to avoid having to make the trade-offs at all (*see* **Dispersion Effects**; **Performance Measures for Robust Design**; **Product Array Designs**; **Box–Meyer Method for Dispersion Effects** for some aspects of this in robust design).

Another important omitted issue is the use of response transformations to both simplify the statistical model and provide insight into underlying scientific behavior. Box *et al.* [3, 7] provide details and examples.

Another topic that underlies the practical implementation of experimental design is the interplay among empirical experimentation, theory, and prior experience. Statistical experimental design is usually exposited as if all knowledge were in the data, but this is never even close to the case. Experimenters typically work within a firmly established theoretical framework and have a lot of prior experience with the system under study or ones very similar to it. As a consequence, clear results from a well-designed screening experiment often provide sufficient insight to allow the experimenter to cut short further investigation and successfully apply this subject matter knowledge to directly achieve the results that were sought or to recognize inherent constraints that would make further experimentation useless. George Box once described this in an experimental design class as the ability of statistical design "to catalyze the scientific learning process". See Box [8] for further discussion.

Finally, like it or not, the strategy of sequential experimentation is inextricably linked to the philosophy of science: how do theory and experiment interact in the "scientific method"? – what constitutes a "replicated" result? – when is a hypothesis confirmed or contradicted by inevitably noisy data? – when is an experimental result sufficiently unexpected to warrant detailed checking or redoing? Because experimentation is an active process to demonstrate **causality** and not merely to assert passive association for the purpose of prediction, it must deal with such questions. This is perhaps the most interesting and even useful aspect of sequential design, because it focuses on how experimentation alters our understanding of how the world works. Unfortunately, such a discussion is beyond the scope of this article, but perhaps this brief mention will pique the reader's curiosity and stimulate further thought and study.

End Notes

[a.] Another equivalent interpretation is that there is meaningful curvature of the response surface over the wider ranges. Interaction is a second-order effect (*see* **Response Surface Methodology**).

[b.] Indeed, it already has. See, for example, Kahneman *et al.* [4] and Kuhn [5] for two possible aspects of the issue.

References

Note: It should be evident from the text that practically any reference for experimental design is also relevant to sequential experimentation. Hence, the following should be viewed merely as particularly germane to some of the issues discussed, and not a comprehensive or even sufficient compendium.

[1] Ishikawa, K. (1986). *Guide to Quality Control*, Asian Productivity Organization, Tokyo.

[2] Meyer, R., Steinberg, D. & Box, G. (1996). Follow-up designs to resolve confounding in multifactor experiments, *Technometrics* **38**, 303–313.

[3] Box, G. & Draper, N. (2007). *Response Surfaces, Mixtures, and Ridge Analyses*, 2nd ed., John Wiley & Sons, Hoboken.

[4] Kahneman, D., Slovic, P. & Tversky, A. (eds) (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge.

[5] Kuhn, T. (1996). *The Structure of Scientific Revolutions*, 3rd Edition, The University of Chicago Press, Chicago.

[6] Voelkel, J. (2005). The efficiencies of fractional factorial designs, *Technometrics* **47**, 488–494.

[7] Box, G., Hunter, J.S. & Hunter, W.G. (2005). *Statistics for Experimenters: Design, Innovation, and Discovery*, 2nd Edition, John Wiley & Sons, Hoboken.

[8] Box, G. & Friends. (2006). *Improving Almost Anything: Ideas and Essays*, Revised Edition, John Wiley & Sons, Hoboken.

BERT GUNTER