Experimental Design in the Complex World

Gerald J. Hahn

Corporate Research and Development
General Electric Company
Schenectady, NY 12301

The design of an experiment involves much more than deciding on a matrix of experimental points. This is demonstrated by six recent experiments, dealing with such diverse subjects as evaluating the reaction of birds to different noises, comparing computer programming methods, and optimizing a chemical reaction. These experiences and others are generalized in some basic guidelines for designing experiments. Some technical challenges and educational programs are also discussed.

KEY WORDS: Fractional factorials; Randomization; Statistical consulting.

1. INTRODUCTION

Many books and articles provide examples of designed experiments that usually are neatly packaged, contain well-defined program objectives and experimental variables, and have few practical constraints. All that is required of the statistician is to propose a matrix of experimental points and analyze the results. In most applications, however, selection of the formal test plan represents only the tip of the proverbial iceberg. This article is concerned principally with these other considerations. We describe recent experiences in planning experimental programs (Sec. 2), generalize these and others to provide some basic guidelines (Sec. 3), suggest some technical challenges (Sec. 4), and comment on educational programs (Sec. 5). Many of our observations, or similar ones, have been made by others. For example, see the books by Cox (1958), Fisher (1935), and Youden (1951) and the papers by Bishop et al. (1982), Box (1954), Box and Wilson (1951), Box and Youle (1955), Marquardt (1979), Hooke (1980), Hunter (1981), Joiner (1977, 1981) and Price (1982).

2. EXAMPLES OF DESIGNING REAL-WORLD EXPERIMENTS

The first three examples illustrate the statistician's role in designing an experiment. The last three deal, more briefly, with experiments that were different in one way or another.

2.1 An Experiment That Showed Statistics Is for the Birds

A report dealing with airplane accidents resulting from the ingestion of birds in jet engines suggested that a contributing factor to such accidents might have been that the noise of the engine sounded like a distress call to some birds, thus attracting these birds to airplanes. To assess this claim a four-week experiment was to be conducted in a meadow populated by a "representative group" of birds. Recordings of engine noises and distress calls were to be played to the birds and the birds' reactions noted. A team consisting of an acoustical engineer, an ornithologist (or "birdman"), and the author met to develop an experimental plan. At the team's initial meeting, I raised the following questions:

- Are the birds that live near the selected meadow a random sample from the population of those that live near airports? The ensuing discussion revealed that this had been a major consideration in selecting the test site.
- Will the results be biased by the fact that airport birds have adapted to engine noise, whereas the birds in the experiment are hearing such noise for the first time? It was felt that the four-week test period would provide a sufficiently long "learning time," and that the data could be analyzed to assess time trends.
- Will the recordings of jet engine noise and of bird distress calls be a "random sample" of the noises about which conclusions are to be drawn? Recordings of four different engines and four different types of bird calls were carefully selected for this experiment. In addition, a "white noise" recording was to be used as a control.
- Is there a seasonal effect? Does the fact that this experiment would run for only a one-month period (during the early summer) significantly limit the generality of the findings? In the ornithologist's opinion, the time period chosen was the one in which the birds would most likely be attracted to the engines. There-

fore, the assumption of nonexistence of a seasonal effect was conservative.

- Does the fact that the experiment was limited to simulating sound sensations, ignoring sight and smell, seriously detract from the applicability of the findings? This would have to be one of the underlying assumptions of the experiment, but was not expected to be an important limitation.

Each of these questions had been considered previously. However, raising them at the meeting focused added attention on them. Moreover, the discussion led to an improved recognition of the assumptions to which the findings would be subject, irrespective of how good the experimental plan and analysis might be.

The team then discussed questions related to the protocol of the experiment, including

- How can one simulate different distances of the birds' locations from the noise source? This would be accomplished by a "step-stress" approach—the recordings were to be played on each test run at increasing noise levels during a two-minute period.
- How can one obtain a realistic and consistent quantitative assessment of the birds' "reactions"? Provisions were made to record, as accurately as possible, data on (a) the number of birds on site before, during, and at the conclusion of each run; and (b) the proportion of birds hovering over the noise source. In addition, movies of the birds' reactions to the recordings were to be taken for later review, if needed.
- How much time should elapse between experimental runs to allow the birds to return to a stable situation? It was decided that it would be conservative to run tests every hour. This would permit a series of nine runs for each nine-hour period. Test periods were to be started at different times of the day and night.

Eventually, a randomized block design, with each nine-hour period representing a block, was proposed. The nine treatments within each block consisted of the nine different recordings. The sequencing of the runs, that is, of playing the recordings within each block, was randomized.

The experiment was to be conducted in 26 blocks, preceded by some pilot runs. Various changes (e.g., the placement of the loudspeakers) were planned over the course of the experiment. Some other requirements, such as that there be a minimum number of birds on site for each run, were also included in the experimental protocol. Provisions were also made to record data on potentially significant covariates.

The data were reviewed each day. A formal analysis was conducted at the conclusion of the experiment. An effective noise figure was calculated by adjusting the actual noise by a factor related to the average distance of the birds from the noise source. Logistic curves were fitted to the results of each run to obtain estimates of the noise levels at which 10% and 50% of the birds reacted to the noise by flying away or hovering above. These estimates were used to compare the birds' reactions to the different noises. Analyses with covariates, such as the sequence of testing, the test day, and the wind speed, were also conducted.

The statistical analysis confirmed what was evident from an inspection and some simple plots of the data. The birds reacted differently to the different noises. They were attracted by the distress calls and, when sufficiently loud, scared away by the engine noise.

The major contribution of the statistical plan was to add discipline to the experiment and to help ensure that it would result in as valid conclusions as possible, subject to the constraints imposed by the testing situation.

2.2 An Experiment to Compare Computer Programming Methods

An experiment was recently conducted at the General Electric Research and Development Center to compare the merits of the following approaches to computer programming:

- Conventional one-person manual programming and checkout, to be referred to as the *conventional method.*
- An automated dual programming approach, whereby two programmers independently develop the same program and then check it using an automatic test driver. The automatic test driver randomly generates test cases from the program's input domain and identifies cases in which the two programs give different results. The programmers are informed of all differences between their results, rework their programs, and again use the automatic test driver to compare the results. The process is continued until the same results are obtained by both programs on 10,000 test cases, at which point the programs are submitted. This approach is referred to as *the dual method.*
- An automated one-person checkout, whereby the program is developed by a single programmer who may, however, access the automatic test driver. This approach is referred to as the *solo random method.*

An experiment was to be conducted to compare the results of using these three programming methods. Different programmers would be required to write a variety of programs using one of the three methods for each program. The major criterion for comparing the methods would be a comparison of the number of errors that remain in the final submitted program. These would be determined by comparing the answers obtained using the submitted test program with those obtained using a "correct" version that had previously

been checked over a large number of test cases generated by the automatic test driver. A secondary criterion would be the required number of programming hours. These two criteria were to be combined to create a third one—the number of programming hours per error removed.

Designation of the required matrix of test points was again only a small part of the total experimental design. The experiment had to provide a statistically valid, and reasonably efficient comparison of the three programming methods, in order to withstand attacks against the conclusions concerning the relative merits of the methods—no matter what these conclusions would be. This, in fact, is what led to the inclusion of the solo random method in the test program. Originally, only the dual and conventional methods were to be compared. We pointed out, however, that if the dual method gave better results than the conventional method, it would not be clear to what extent this result is due to (a) the use of two programmers or (b) the use of the automated checkout. The solo random method was, therefore, included in the experiment to resolve this question. (A fourth method, involving two programmers and conventional checkout, was also proposed. This approach had, however, been studied previously by others, and the experimenter felt it was not of sufficient interest to warrant inclusion.)

The following were also considered:

1. *Selection of Programmers.* The experimenter was asked to define the population of programmers for which the results were to apply. For example, was the dual method meant only for relatively inexperienced programmers or for all programmers? (The answer was "for all programmers.") In theory, one would like to select the programmers for this study at random from the defined population. In practice, however, this is not possible. After some discussion, the experimenter recruited a supposedly "representative" group of programmers from among his associates, summer students working on MS degrees in computer science, and a second group of programmers recently hired by a company component and awaiting security clearance. Obviously, this was not a random sample. For example, the sample tended to overrepresent programmers with extensive formal education and to underrepresent those with appreciable work experience. It seemed, however, to be the best sample that could be obtained under the circumstances and was probably a fairly reasonable selection for the desired evaluation.

I urged that a relatively large sample of programmers be selected, subject to the requirement that each use the three programming methods at least once. As a result, 24 programmers were chosen. Background information on each was obtained; this could be used as covariates in the data analysis.

2. *Selection of Computer Programs.* Like programmers, computer programs had to be chosen to provide a realistic (if not random) representation of the population of programs for which the programming methods were to be compared. I again recommended that a relatively large number of programs be used. The development and checkout of programs, however, was complex. Thus, the study was limited to four highly different programs that the experimenter felt were most representative of those encountered in practice (a line editor, parser, network database, and macrodefinition processor).

3. *General Test Protocol.* Numerous further questions involving the implementation of the experiment were raised and resolved. For example,

- Under the dual method, how much communication, if any, would be allowed between the two programmers?
- What instructions should the programmers be given about the importance of accuracy versus speed?
- For the solo random and the dual methods, should the programmers also be encouraged to submit their own test cases for checkout, in addition to those randomly generated?
- What is the best way of measuring the number of independent remaining errors associated with a submitted program?
- How should the number of programming hours be measured and recorded?

In such discussions, the statistician's major functions are to help structure the problem, to identify important issues and practical constraints, and to indicate the effect of various compromises on the inferences that can be validly drawn from the experimental data. At this stage, the statistician serves as a logician and interacts with the experimenter in a manner somewhat similar to that of a psychologist who helps a patient cope more effectively with personal problems.

After the test protocol was developed and documented, a pilot run of the experiment was conducted. A formal test plan, consisting of a series of six modified Latin Squares, was proposed. These called for four programmers (different in each Latin Square) to write (the same) four programs using the conventional method once, the solo random method once, and the dual method twice with different partners each time. (The Latin Square modification was due to each programmer's use of the dual method twice.) The pairing of programmers and the testing sequence were randomized.

The experiment was conducted in two stages, involving 12 programmers and two sets of three modified Latin Squares each. With two minor exceptions

(involving the premature withdrawal of one of the programmers, and the fact that one run, involving the conventional method, could not be scored due to the programmer's misunderstanding of the program's basic objectives), everything went as planned in the first stage. The results were overwhelmingly in favor of the dual method. In particular, the total number of errors in the submitted programs was (a) 65 in 11 programs using the conventional method, (b) 48 in 12 programs using the solo random method, and (c) 2 in 11 programs using the dual method. No sophisticated statistical analysis was required to demonstrate the superiority of the dual method in removing errors! A comprehensive tabulation and histogram of the data sufficed. However, to be responsive to a request to formally establish "the statistical significance" of the results, we noted that for each of the 10 cases where direct comparison was possible, fewer errors were committed for the dual method than for the conventional method. The probability of this occurring purely by chance, if the conventional method were as good as or better than the dual method, was one in $(\frac{1}{2})^{10}$ or one in 1,024. Thus, one can reject the null hypothesis that the conventional method is no worse than the dual method at the .1% significance level. A similar result was obtained in comparing the dual method with the solo random method. The statistical analysis, however, provided only the proverbial "icing on the cake" by fine-tuning conclusions, and giving more quantitative estimates. More generally, I have found that a formal statistical analysis of the data is often superfluous for a well-designed and well-executed experiment.

In the second stage of the experiment, control was poorer and for several reasons the original experimental plan was not followed closely. Although the results appeared to be similar to those from the first stage, further analysis seemed prudent. Thus, a number of regression runs on the combined data from the two stages were conducted (using programming method, programs, and programmers as dummy variables). These analyses, thus, verified the previous findings. (See Panzl 1981 for further discussion of this experiment and its results.)

2.3 An Experiment to Improve a Chemical Reaction

This experiment involved a laboratory scale-down of the manufacture of a plastic material. The objective was to evaluate the effect of certain process variables (operating temperature, pressure, water addition, and amount of catalyst) on process performance, as measured by various yield and product quality characteristics.

The laboratory in which the experiment was to be conducted is about 1,000 miles from my office. This precluded the frequent, relatively short meetings, interspersed with thinking periods, that characterized the planning of the experiment to compare programming methods. Prior to our first meeting, I asked the experimenter to prepare detailed documentation, described in the next section, about the nature and purpose of the experiment. This forewarned him about the type of information needed and set the stage for subsequent discussions. Since the meeting occurred in my office, I, unfortunately, had to forgo the opportunity of seeing the experimental setup.

From the discussions with the experimenter I determined that

- The primary purpose of the experiment was to optimize the reaction within the current operating region and in an expanded region. (The expanded region would require equipment modifications on the production line.) Optimum performance was well defined on an individual variable basis, that is, increasing the yield and raising the quality level. However, the trade-off between quality and yield was not known and, in fact, depended upon production requirements. Thus, it was not clear how to combine the various criteria in seeking an overall optimum, and it was not possible to develop a general desirability function to be optimized, as suggested by Derringer and Suich (1980), or even to pursue a simultaneous optimization approach (See Khuri and Conlon 1981). Instead, it was agreed that the major objective would be the more basic one of estimating the relationships between the process variables and each of the performance variables and drawing curves therefrom. A good knowledge of the relationships, however, was of greatest interest in regions where the process would give promising results. These objectives had to be understood in deciding whether the experimental strategy should concentrate on searching for an optimum or on representing the response surface over a broad region in the experimental space. It appeared that some of each was desired.

- Some preliminary runs (a) verified that the laboratory experiment provided a reasonable simulation of the manufacturing process, and (b) demonstrated acceptable repeatability.

- There were no serious practical impediments to randomization.

- A minimum of one day was required to conduct an experimental run, and it seemed desirable to include an option for evaluating the data after approximately every 10 to 12 runs to ensure that "things were going right."

The following three-stage experimental plan was eventually proposed:

Stage 1. A half replicate of a four factor, two-level factorial plan, supplemented by two repeat runs at the

center condition (See Mendenhall 1968) and a further test at the current production-line operating conditions (total of 11 runs).

Stage 2 (Tentative). Additional runs to convert the fractional factorial plan into a central composite design and further single repeat runs at the center condition and at the current production-line operating condition (total of 10 runs).

Stage 3 (Tentative). Additional runs to convert the half replicate fractional factorial base to a three-quarter replicate (see Diamond 1981) and further single repeat runs at the center condition and at the current production-line operating condition (total of six further runs).

The experimenter reviewed the proposed plan with his technician and suggested some minor changes in the conditions of some of the variables. The experimental design was revised accordingly.

We proposed at least an informal analysis of the results after the first stage of the experiment to ensure that the proposed subsequent stages still made sense. This seemed particularly appropriate in light of the interest in searching for a generally optimum region, and the fact that the selection of the levels of the variables had been somewhat arbitrary.

The first stage involved a Resolution 4 design, that is, main effects could be estimated independently of two-factor interactions, but two-factor interactions were confounded with one another (see Box and Hunter 1961 and Daniel 1976). Also, curvilinear terms could not be estimated separately for each variable. Moreover, a comparison of the lack-of-fit term from the fitted model with the pure error term (based upon a single degree of freedom resulting from the two repeat runs at the center condition) suggested the model that could be fitted at this point to be inadequate. Nevertheless, the supplemented fractional factorial plan provided a good spanning of the experimental region and, in light of the apparent good repeatability, a scanning of the test results would, it was hoped, identify promising subregions. Indeed, a number of conditions were found to give better results than those at the current plant operating condition. Thus, inspection of the data was more informative, at this point, than the formal statistical analyses. In fact, it would seem that the information that can be gained from scanning the data is often a major justification for using saturated fractional factorial and Plackett and Burman (1946) designs.

Inspection of the data from the first stage of the experiment, as well as the approximate model fit, indicated that parts of the experimental region did not merit further study. Thus, we decided against implementing the previously developed second and third stages and developed a revised second stage. In the revision it was decided to (a) change the range of variation for two of the process variables (temperature and percentage catalyst) to concentrate testing on what appeared to be the most desirable subregion of the experimental space; and (b) hold one of the process variables (water addition) constant for most of the runs at the condition that seemed to provide optimum results within the current operating region.

A three-variable central composite design using a factorial base was proposed for the revised second stage. This was supplemented by two additional runs involving perturbation of the variable held constant for the other runs (water addition). These two runs were conducted at the center condition of the other variables. (This permitted some further assessment of water addition without making it a full-fledged variable.) Single repeat runs from the previous stage of the experiment were also performed at the current plant operating conditions and at the center point of the Stage 1 design.

It turned out that a reaction could not be obtained at two of the conditions. Thus, these tests were rerun at nearby feasible conditions. At the conclusion of this stage, a second-order regression model and various reduced models were fitted based upon the data from the second stage only and the combined data from both stages. Since readings were obtained on each run at various times, separate analyses were performed using the results at selected single times only and all the data. Care had to be exercised in drawing statistical inferences in the second case because the readings on the same reaction at different times are not independent of one another.

The analyses indicated

- Contrary to expectations, there was a significant difference in results between the two stages. This was indicated by (a) the statistical significance of a dummy variable associated with stages in a regression analysis of the entire data and (b) a comparison of the results from the two pairs of runs that were conducted at the same condition in both stages. The experimenter ascribed the difference to the use of two different batches of material, although it could, of course, have been due to some other reason. In any case, sufficient precautions had apparently not been taken in planning the experiment to eliminate between-stage differences.

- A specific condition at the boundary of the current plant operating region appeared to maximize yield within the currently feasible plant operating conditions, without a loss in quality. An inspection of the individual data points, similar to that conducted after the first stage, provided further confirmation. The overall optimum, however, appeared to be beyond the experimental region and current plant operating conditions.

This experiment, like many others, began as a nonstatistical "vary one-factor-at-a-time" program. The experimenter contacted me at the recommendation of one of his associates. Recently, I asked him whether he felt a statistically planned experiment had been worthwhile. (Since I was asking the questions, a biased response was likely.) He responded that the statistical approach was valuable because it allowed him to progress more rapidly than would otherwise have been possible. It also added discipline and quantification to the program. On the other hand, he indicated that this approach was "not as much fun" for him as a direct hands-on approach, where he could make all the decisions.

This last comment reflects the importance of involving the experimenter in the planning of the program and the analysis of the results. In this case, we were able to interest him in expanding his formal background in statistics and the design of experiments and he became the local "statistical expert."

2.4 An Experiment That Succeeded by Failing

Some time ago, we designed a multivariable experiment to identify manufacturing conditions that would be both operationally feasible and economically attractive for a proposed new product. Again, the experimental design had been a central composite plan with a fractional factorial base. We had conducted some regression analyses of the resulting data, discussed their results with the experimenter, and proposed some added tests, which were conducted. Then there was silence—nothing more was heard from the research scientist. Recently, I contacted him for an update.

My call verified that the program had been abandoned. The analysis of the data from the first stage and an inspection of the second-stage results clearly identified many UNECONOMICAL ways of making the product. Unfortunately, no practical conditions that would result in a desirable product were suggested. This was as expected from previous nonstatistical testing. The statistical experiment had, in fact, been proposed as a last-ditch effort to obtain a useful product. Its inability to do so convinced all concerned that the current process would not give the desired results.

The experimenter had earlier raised the question of whether a statistically designed experiment should have been used in the basic research as well as in the subsequent general development phase of the study. I responded that the fundamentals of sound experimental design (e.g., formal definition of the scope of the experiment, proper selection of material batches, definition of what is to be measured, focus on achieving good repeatability, etc.) apply to all phases of experimentation, including basic research. Formal statistical designs, however, are often premature for exploratory work, since the experimenter is frequently searching for "ball-park" assessments rather than subtle evaluations. Broad decisions are often made after almost every run and the program is redirected accordingly. The researcher may not have the resources or the patience to pursue a disciplined, statistically based program. The results of such exploratory testing can, however, lay the groundwork for a later statistical design.

2.5 A Big Bucks Experiment

Sometimes, we are not asked to design an experiment because the out-of-pocket costs for our services represent too large a part of the total project budget. In other cases, however, the costs of designing the experiment represent only a proverbial "drop in the bucket." This is often the case in studies involving heavy equipment, such as a jet engine or a turbine. Unfortunately, such experiments also are often characterized by severe practical constraints and, therefore, do not lend themselves to traditional statistical designs.

Recently, we were asked to develop a design for an experiment to be conducted in a combustion chamber which would compare the performance of different gas turbine fuels for various design configurations under different operating conditions. Because of the high cost of fuel, each run cost more than $10,000. In light of the high cost, such programs are frequently run piecemeal, using only a handful of tests to evaluate a particular design/fuel-type combination. In this case, however, a more comprehensive test program was planned.

Before a meaningful experimental plan could be developed, a significant learning effort was required. This included (a) A review of the purposes of the experiment and the previously prepared "proposed test point schedule"; (b) A visit to the combustion chamber while a similar test was in progress; (c) A review of previous data and the resulting fitted curves; and (d) Identification of the constraints associated with the program. In particular,

- Change in the combustion system required extensive reassembly; change in fuel type necessitated cleaning out the chamber; and temperature could be readily increased, but not decreased. Thus, complete randomization would be impractical.
- Only one batch of each fuel type could be obtained.
- Certain combinations of test conditions were mandated.

I will not describe the proposed experimental plan here. It resembled the engineer's initially proposed test-point schedule more than it did any known statistical design and, at best, can be described as pseudo-

statistical. I did, however, propose (a) the addition of repeat runs to check for trends; (b) randomization at the lowest level of the testing hierarchy; (c) increased emphasis on the consequences of using only one batch for each fuel type, the assumed underlying model, and the methods to be used for analyzing the data.

Jurisdiction over the program was moved before any testing could begin. My recommendations were used as a general guide but were not followed in detail. In fact, I have yet to be involved in any investigation of this type where my recommendations have been implemented. This kind of "messy experimentation" nevertheless deserves the attention of statisticians because of the special challenges it poses and the high potential payoffs resulting from even relatively small gains.

2.6 An On-Line Process Comparison

This program, to be conducted directly on an ongoing manufacturing line, was to evaluate the effect of a process change on product quality. The change was to be introduced into production for a trial period, and the estimated rate of defective units after the change was to be compared with the rate prior to the change. I was called in to answer the question, How long need the new process be run to obtain "statistically significant" results? Inspection of the past data suggested trends in quality over time.

Instead of answering the original question directly, I urged a radical change in the manner in which the investigation was to be conducted. I recommended that production be alternated between running the process with the change and without the change. Since the manufacturing line was shut down each night, it was convenient to make such changes daily. Each pair of days should include one randomly selected day using the old process and one using the new process. Recommendations about sample size requirements were made in this context, based upon an analysis of the past data (see Hahn 1982 for further details).

3. SOME BASIC GUIDELINES

Guidelines that I have found useful in designing experiments are described in this section. Many, but not all, of these concepts were illustrated by one or more of the examples.

3.1 Advise the Experimenter Initially of the Needed Information and Urge Documentation

Some experimenters think that the only information that a statistician needs to design an experiment is the number of experimental variables and the number of levels of each (e.g., please design an experiment for one variable at four levels, three variables at three levels, and two variables at two levels). This viewpoint has, perhaps, been encouraged by the availability of easy-to-use catalogs of experimental designs (see Hahn and Shapiro 1966) and of user-oriented computer programs to design experiments. Olsson (1982) summarizes the problem by stating, "When consulting with engineers, it always seems that I want to learn about the engineering of the problem (or more about what's going on with a proposed experiment), while they're always asking questions about the statistics of the problem."

As suggested by each of the examples in the previous section, to properly design an experiment, the statistician must also know

- The objectives of the experiment.
- The details of the physical set-up.
- The variables to be held constant and how this will be accomplished (as well as those that are to be varied).
- The uncontrolled variables—what they are and which ones are measurable.
- The response variables and how they will be measured.
- The procedures for running a test, including the ease with which each of the variables can be changed from one run to the next.
- Past test data and, especially, any information about different types of repeatability.
- Conditions within the experimental region where the expected outcome is known; the anticipated performance is expected to be inferior, especially for programs where an optimum is sought; and experimentation is impossible or unsafe.
- The budgeted size of the experiment and the deadlines that must be met.
- The desirability and opportunities for running the experiment in stages.
- The anticipated complexity of the relationship between the experimental variables and the response variables and any anticipated interactions.
- Other special considerations.

The major mode of communication between the experimenter and the statistician should be face-to-face discussion. The experimenter should, however, also be encouraged to document as much of the above information as possible ahead of time. I advise my clients that such documentation is likely to reduce the amount of time I will be charging against the project and suggest a recent article (Hahn 1977a) as a guide. Documentation forces the experimenter to address fundamental questions early. It can also trigger communication between the experimenter and his or her colleagues and management. Hunter (1981) describes a study in which his question "What is the objective of this investigation?" triggered a lively discussion be-

tween the two principal project investigators. Documentation also provides the statistician with a good overview of the problem and a reference point to return to as the discussion progresses.

Not all experimenters are willing to prepare initial documentation, and time constraints sometimes make this impractical. In any case, such documentation represents only a "first cut" that will be modified and expanded upon in the later discussions.

Finally, the experimenter should be asked to provide other pertinent, readily available information, such as past reports and correspondence, sales blurbs, and relevant articles from the literature. These often give useful background, even if not read in their entirety.

3.2 Understand the True Model and Variables

In some experiments the relationships are based upon known physical theory. This theory should be appreciated by the statistician and considered in the experimental design and analysis. If the relationship is known to have a particular nonlinear form, the design and analysis should accommodate that form and not try to approximate it by a polynomial (see, e.g., Box and Lucas 1959 and Hill and Hunter 1974).

Even if the form of the physical model is not known, the experimental variables should be expressed in a way that makes the most sense physically. In a particular situation, one might have to decide whether current density and time should be the designated experimental variables or whether it is more reasonable, on physical grounds, to consider time and the product of current density with time as the variables that most directly impact performance. Judicious expression of the variables often reduces or eliminates interactions between variables, resulting in a simpler model.

In one recent chemical engineering experiment, I found it useful to ask the experimenters to identify

1. The "true variables" that impact performance—even though they may not be controllable and/or measurable. These included the unknown temperatures within two reactors.
2. The variables that can be directly controlled. These included the inlet temperatures to the two reactors.
3. The functions of the controllable variables that best reflect the effect of the true variables in the model. These included the inlet temperature to the first reactor and the difference in inlet temperatures between the first and second reactor. These were the variables that were eventually specified in the experimental design.

Also, provisions should be made for collecting data on other factors that might prove important. Wood (1982) gives as an example

> the measurement of electrical line voltage in a pilot plant study. When this variable was included in our analysis of experimental data, we were surprised to find it was influential. On checking, we found that every morning, coincident with the startup of other units in the pilot plant, there was a line voltage drop. This reduced the flow of our constant speed pump, reducing the space velocity and providing better yield of product. Had this not been found, erroneous conclusions would have been drawn.

In the "birds" experiment, data were maintained on weather conditions, and this was used in some of the subsequent analyses.

3.3 See the Physical Setup and Become Actively Involved

It is helpful to meet at the experimental site to review the physical setup, especially if the equipment is operational. Such a review is most beneficial near the outset of the discussions—after obtaining a sufficiently detailed explanation of the experiment to appreciate significant points. Such visits were not feasible in the "birds" and chemical reaction experiments, but they were conducted in three of the other programs. It is also useful to return on site to observe the actual implementation of the experiment. Often additional potential sources of variability, not noted by the experimenter, may be observed by the statistician and removed or factored into the analysis.

The statistician should request and review past data, even though such data may not warrant extensive analysis. Such data can often provide a feel for experimental error and some useful insights. For example, the experimenter may suggest that some of the data be ignored on account of some special happening(s). This raises the question of what can be done to minimize the likelihood of such happenings occurring again and spoiling future experiments. In the on-line process comparison, review of the past data made it clear that the approach suggested by the experimenter would not likely yield unambiguous results.

In summary, the statistician should become actively involved in the investigation, obtain a good understanding of the physical setup and constraints, and learn the experimenter's terminology while minimizing the use of statistical jargon. Often the statistician plays an important role by raising fundamental questions and by serving as the devil's advocate. As others have pointed out

- "Clients are well served by statisticians who have a healthy curiosity about underlying mechanisms." (Hunter 1981)
- "Being willing to speak the customer's language as much as possible is one aspect of a good consultant's attitude." (Hooke 1980)
- "The statistician's responsibility in the planning phase ... is to ... insure that the stated objectives are achieved, and that the results are defensible .... We

can serve as catalysts for progress in planning multidisciplinary studies." (Price 1982)

3.4 Obtain Measurements of Real Experimental Error

Before embarking on any major test program, a measure of "the real experimental error" should be obtained. Real experimental error is the variability obtained in the results when the same experimental conditions are repeated independently at different times. (The total experimental error should, if possible, be subdivided into individual components representing repeat measurements, within lot variation, among lots variation, etc.) Repeat experiments were conducted initially in the chemical reaction experiment and were advocated in the "big bucks" experiment.

Experimenters are often overly confident about the magnitude of real experimental error. Satisfactory repeatability should be demonstrated before embarking on any large-scale program. Poor repeatability suggests that the wrong variables are being included in the experiment.

In addition, at least a few (minimum of three) real repeat tests should be included in the experiment, rather than relying on assumed negligible higher-order interactions to estimate experimental error. Sometimes, it is desirable to have all repeat tests at the same condition; at other times, single repeat tests at several different conditions might be preferred. For example, in the chemical reaction experiment, repeat tests were conducted at both the center condition of the experimental design and at current operating conditions.

3.5 Include "Baseline Conditions" in the Experimental Plan

It is often useful to include baseline conditions in the experiment. For example, in the chemical reaction experiment, the current plant operating conditions were selected as the baseline conditions. In other situations, the baseline conditions might be ones the experimenter believes, prior to the experiment, will give the best results.

Inclusion of the baseline conditions can

1. Allow a comparison of experimental with expected (or perhaps, desired) results and lead to an assessment of the validity of the test program. In the chemical reaction experiment, the results from the runs at the plant operating conditions supported the validity of the scaled-down laboratory model.
2. Provide a benchmark against which the results at other experimental conditions can be compared.
3. Result in improved precision in estimating the response in an experimental region that is likely to be of high interest.

The baseline condition might be one of the formal experimental points (such as the center point in a central composite plan) or an added point. Repeat tests at the baseline condition are also often helpful.

3.6 Consider a Multistage Plan

Statistically designed experiments have sometimes been criticized on the grounds that they discourage flexibility by requiring the experimenter to follow a rigid test plan. Management might be unhappy at having to wait too long for answers. Moreover, experimentation often involves a learning process, and early results might suggest improvements in the plan for the later tests, as in the chemical reaction experiment.

When feasible, it is often desirable to conduct an experiment in stages, and this was done in three of the experiments described in the preceding section. In the experiment to compare programming methods, the first-stage plan provided encouraging results and was, therefore, followed by a second stage, using a broader sample of programmers. In the chemical reaction experiment, and also in the experiment "that succeeded by failing," the first-stage design involved a highly fractionated factorial plan. Later stages can involve building this into a less fractionated plan, a full factorial, or a central composite design. Often, the early stages identify the important variables and allow one to drop the less important ones. (Differences between stages can be estimated by building up the design in orthogonal blocks.) In addition, in the on-line process comparison, a sequential testing approach was proposed to permit termination of the experiment as soon as definitive results were obtained. Box, Hunter, and Hunter (1978) propose that "As a general rule, not more than one quarter of the experimental effort (budget) should be invested in a first design. ... When the first part of an investigation has been completed, the experimenter will usually know considerably more than when he started and consequently will be able to plan a better second part, which in turn will lead to improved planning of a third part, and so on."

Multistage testing is not feasible when it takes a long time to obtain responses. This is often the case in agricultural experiments, in product life testing, and in dealing with some manufacturing processes such as making integrated circuits where the fabrication cycle may take as long as two months. In the "birds" experiment, the testing had to be conducted within one month and, therefore, it was not practical to plan a multistage program. However, the results were carefully monitored to assure that changes in the experimental plan were not needed.

3.7 Keep It Simple

The experimental plan should be kept as simple as possible and amply justified to the experimenter. The

plan will not be implemented if the experimenter does not agree with it, or if the responsible technician does not understand the instructions. Simplicity is especially important for investigations involving manufacturing processes, such as the on-line comparison.

3.8 Ask the Experimenter to Review the Proposed Experimental Design and to Predict the Expected Outcomes

The experimenter should be asked to review a draft of the proposed experimental plan. This provides another opportunity to identify unfeasible or uninteresting runs and to check that a judicious choice of the ranges for each variable was made. Sometimes it is useful to submit two or more alternative designs, explain the differences between them, and ask the experimenter to pick one. Wood (1982) and others recommend that

> the experimenter be asked to state (preferably in writing) what results are expected. Often the benefit of experimental design is to find the unexpected, that which can not be seen intuitively, either because of interactions, or because of the inability to separate main effects from random noise. Because of faulty memories, there always seems to be someone in management who will say "we knew it all the time". Such records will keep the record straight.

Also, it is worthwhile to consider Feder's (1982) recommendation to submit to the experimenter

> randomly generated sets of data as part of the design process; that is, generate one or more sets of data, incorporating the assumption, variability estimates, and the anticipated form of the response function. This provides an idea of the anticipated expected precision of the resulting estimates and demonstrates how the data will be analyzed.

3.9 Document the Experimental Plan and Protocol

Detailed documentation of the experimental plan avoids misunderstandings. The documentation should include not only the recommended test points (and the proposed sequence for running them), but also the experimental protocol (prepared, perhaps, by the experimenter). Such documentation was prepared for each of the programs described previously.

3.10 Be Prepared for Changes

Things do not always work out as planned. Deadlines change; new information becomes available; conditions expected to be feasible turn out not to be. The statistician should be prepared to modify the experimental design, as required by the changes in circumstances.

Sometimes, a long period elapses between the time the experiment was designed and when it is to be modified or analyzed. In such situations, the statistician will especially appreciate having maintained good formal and informal records.

It is the rule, rather than the exception, that the implemented experiment differs in one way or another from the design. Runs are omitted for various reasons and, sometimes, new ones are added. In the programming experiment, runs were omitted because a few of the programmers had to leave the experiment early. In the chemical reaction experiment, some runs were found not to be feasible and somewhat different ones were run in their place. When the actual experimental conditions differ from the aimed-at ones, the actual conditions should be recorded so they can be used in the statistical analysis. (This, in fact, is one of the reasons that regression analysis is employed more often than the analysis of variance in analyzing the results of industrial experiments.)

Other circumstances also affect the results and need to be carefully noted. For example, sometimes the performance variable is censored, requiring special analyses (see Hahn, Morgan, and Schmee 1981 for an example).

4. SOME TECHNICAL CHALLENGES

4.1 Response Surface Experiments With a Mix of Qualitative and Quantitative Variables

Situations requiring response surface exploration involving qualitative, as well as quantitative, process variables are frequently encountered. For example, in a recent experiment to learn the relationship between three quantitative variables (temperature, application pressure, and material thickness), one two-level qualitative variable (varnish application method), and time to failure for an insulation material, a central composite design with the following modifications was proposed:

- Run the "prong points" for the qualitative variable on the surface of the hypercube; that is, conduct these two tests at the two levels of the qualitative variable and the center condition of the three quantitative variables.
- Run the prong points for each of the quantitative variables of the design twice—once at each of the two levels of the qualitative variable.

The resulting plan involved a total of 30 (instead of the usual 25) points, prior to the inclusion of repeat points. A dummy variable was used for the qualitative variable in the subsequent data analysis.

The preceding scheme can be readily generalized for two or more qualitative variables by running the prong points of the quantitative variables at all combinations of the qualitative variables. However, because this may lead to an unduly large experiment, one might decide to run these prong points only at a fraction of all possible combinations or at the conditions of the qualitative variables of greatest interest.

The statistical consequences, such as the degree of confounding among the resulting estimates of the model parameters, can be assessed by using, for example, the EXPLOR computer program (see Meeker, Hahn, and Feder 1975), and the plan can be compared with alternatives, such as a (fractional) factorial scheme with all quantitative variables at three levels.

4.2 Considerations in Randomization

Randomization is introduced to neutralize the effect of variables that cannot be, or are not, handled otherwise, such as by direct inclusion in the experiment, blocking, holding constant, and so on. Frequently, such potentially contaminating variables are not even specifically identified. Randomization thus provides some insurance against drawing invalid conclusions from the experimental results because of the effect of confounding variables.

Frequently, randomization needs to be considered at various levels, for example, the sequence of preparing the experimental units, the assignment of treatments to the units, the sequence of performing the test runs, and the sequence of taking measurements. This was the case in the chemical reaction experiment.

Unfortunately, all variables are not created equal. Often some can be varied more easily than others. For example, a change in pressure may be implemented by a simple dial adjustment. It might, however, take a long time to reach stability after a change in temperature. Thus, complete randomization is often impractical—as was seen in the "big bucks" experiment. On the other hand, following the experimenter's impulse to perform all the tests at a particular temperature at one time might confound the temperature effect with that of other variables that are changing simultaneously.

The statistician should aim to achieve a practical compromise between the desirability for randomization and the operational constraints of the test program. This requires a good understanding of the mechanics of testing. For example, the experiment might be shut down overnight; at start-up each morning it makes little difference which temperature is used, as long as that temperature is maintained throughout the day. With this type of information the statistician and the experimenter can decide the amount of insurance, in the form of randomization, that is warranted and can be reasonably achieved (see Daniel 1976; Hahn 1977b and 1978; and Joiner and Campbell 1976 for further discussion).

4.3 Many Experimental Programs Have Split-Plot Features

Many experimental programs involve split-plot considerations. In some situations this is due to the nature of the experimental variables. Some time ago we planned an experiment to help Alaskan Indians learn how to best grow vegetables indoors. Some variables, such as type and amount of fertilizer, could be readily varied within the experimental chambers. Others, such as temperature and humidity, had to be varied between chamber runs. The chemical reaction experiment involved "a split-plot in time" as a result of the previously mentioned readings at different times for the same reaction.

Occasionally, split-plotting also is appropriate because it allows one to obtain more precise information about the within-plot variables. In an experiment to evaluate the effect of alloy composition, heat treatment, and varnish coat on tensile strength, master alloys were first prepared using different compositions. Each alloy was split into a number of segments to be subjected to different heat treatments. The treated segments were further divided into subsamples for the application of different coatings. Tensile strength measurements were then obtained on all coated subsamples. The further "down" one went in this hierarchy, the more precise were the comparisons that could be made.

4.4 Test Point(s) That Cannot (or Should Not) Be Run

In some experiments, especially those involving chemical reactions, there are regions in the experimental space where it is not possible to obtain results, or where it might be dangerous to do so. There are other, less extreme, situations where it is known ahead of time that certain conditions would result in poor or, at best, uninteresting results. Such conditions should generally be avoided in the experimental plan—especially if the purpose is to search for an optimum—even though such omission might result in some loss in orthogonality. Sometimes improved definition of the variables or a change in the definition of the experimental region can eliminate this problem.

5. EDUCATIONAL PROGRAMS

The statistical training of engineers and scientists has improved appreciably in recent years, and many have now taken at least one statistics course. Such courses generally discuss elementary techniques and concepts but generally provide little guidance on how to obtain valid data. Some universities offer courses in experimental design as part of their statistics sequences. Unfortunately, many of these require previous training in statistics. Students, however, are often reluctant to take further courses in statistics after having survived the typical elementary course. Moreover, as pointed out by Snee (1982), "many universities do not teach experimental design as a separate course. Some claim it is covered as part of 'linear models' courses. The result is a course on analysis of

experiments, rather than design." As a result, many still think of a statistician as somebody to call in after the data have been obtained.

Fortunately, a number of books (such as those by Box, Hunter, and Hunter 1978; Cox 1958; and Youden 1951) are directed at practitioners with limited or no previous statistical training. Hahn (1980) provides an annotated bibliography of books on experimental design; this includes a summary of the technical level, the applications emphasis, and the subject coverage of available texts. Also, various articles in engineering journals provide an introduction to the concepts of experimental design (see, e.g., Feller 1983; Hahn 1977a; Hendrix 1979; and Mueller and Olsson 1971).

Intensive short courses on experimental design are now being offered in many places, and various taped courses are available to interested groups. Also, many organizations are developing in-house training programs. For example, within my own company, a three-hour "Broadbrush Review of the Design of Experiments" is given for engineers and managers as part of a concentrated course on new technological developments. The course is also offered on site to operating components; the participants are invited to bring their own problems. (Sometimes this is the price of admission.)

6. CONCLUSIONS

Statisticians make their most valuable contributions if they are consulted in the planning stages of an investigation. Proper experimental design is often more important than sophisticated statistical analysis. Results from a well-planned experiment are frequently evident from simple graphical analyses. Indeed, the mark of a well-designed and executed experiment is often that formal statistical analysis is superfluous. The world's best statistical analysis, on the other hand, cannot rescue a poorly planned program. Ginsburg (1982) has put it this way, "When I'm called in after it's all over, I often feel like a coroner. I can sign the death certificate—but do little more."

Designing an experiment is often more an art than an exact science, and there is no single "right way." The statistician should serve as a catalyst to ensure that the objectives, limitations, and assumptions of the investigation are clearly understood, and that the experiment yields the most valid results possible, subject to practical constraints; the specifics clearly depend upon the problem at hand. The resulting test plan must be as robust as possible against potential criticism of the results. Good communication between the experimenter and the statistician is essential. The statistician must be inquisitive, probing, actively involved in the investigation, and ready to raise fundamental questions. The investigator must accept the statistician as a full-fledged team member and be prepared to "tell all."

Specifying the matrix of experimental points is, thus, often only a small part of planning the experiment. The final design must be carefully tailored to fit the real problem—not the other way around.

ACKNOWLEDGMENTS

Thanks are due to the many investigators who helped make this article possible by asking that their experiments be designed statistically, subjecting themselves cheerfully to a seemingly never-ending stream of questions, adding much to the final experimental plan, and commenting on a draft of this article. My special appreciation goes to B. Davis, J. Hallgren, D. Panzl, R. Pyles, E. Stone, and R. Wells, all of the General Electric Company. I also wish to thank B. Joiner, N. Bournazian, P. Feder, H. Ginsburg, R. Hooke, D. W. Marquardt, D. Olsson, R. D. Snee, W. T. Tucker, F. S. Wood, and two referees for their many constructive comments that resulted in important improvements in this paper.

[*Received August 1982. Revised September 1983.*]

REFERENCES

BISHOP, T., PETERSEN, B., and TRAYSER, D. (1982), "Another Look at the Statistician's Role in Experimental Planning and Design," *The American Statistician*, 36, 387–389.

BOX, G. E. P. (1954), "Exploration and Exploitation of Response Surfaces," I, *Biometrics*, 10, 16–60.

BOX, G. E. P., HUNTER, W. G., and HUNTER, S. (1978), *Statistics for Experimenters*, New York: John Wiley.

BOX, G. E. P., and HUNTER, J. S. (1961), "The 2^{k-p} Fractional Factorial Designs," I, *Technometrics*, 3, 311–351; II, *Technometrics*, 3, 449–458.

BOX, G. E. P., and LUCAS, H. L. (1959), "Design of Experiments in Nonlinear Situations," *Biometrika*, 46, 77–90.

BOX, G. E. P., and WILSON, K. B. (1951), "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, Ser. B, 13, 1–45.

BOX, G. E. P., and YOULE, P. V. (1955), "Exploration and Exploitation of Response Surfaces," II, *Biometrics*, 11, 287–322.

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

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

DERRINGER, G., and SUICH, R. (1980), "Simultaneous Optimization of Several Response Variables," *Journal of Quality Technology*, 12, 214–219.

DIAMOND, W. J. (1981), *Practical Experiment Designs for Engineers and Scientists*, Belmont, Calif.: Lifetime Learning Publications.

FEDER, P. (1982), Personal communication.

FELLER, J. (1983), "Design Experiments that Control Two or More Variables at Once," *Industrial Research & Development*, 25, 7, 94–95.

FISHER, R. A. (1935), *The Design of Experiments*, New York: John Wiley.

GINSBURG, H. (1982), Personal communication.

HAHN, G. J. (1977a), "Some Things Engineers Should Know

About Experimental Design," *Journal of Quality Technology*, 9, 13–20.

——— (1977b), "Must I Randomize?," *Chemtech*, 7, 10, 630–632.

——— (1978), "More on Randomization," *Chemtech*, 8, 164–168.

——— (1980), "Planning Experiments: An Annotated Bibliography," *Chemtech*, 10, 36–39. Expanded version in "Design of Experiments," *Encyclopedia of Statistical Sciences* (1982), Vol. 2, eds. N. L. Johnson and S. Kotz, New York: John Wiley, 359–366.

——— (1982), "Statistical Assessment of a Process Change," *Journal of Quality Technology*, 14, 1–9.

HAHN, G. J., MORGAN, C. B., and SCHMEE, J. (1981), "The Analysis of a Fractional Factorial Experiment With Censored Data Using Iterative Least Squares," *Technometrics*, 23, 33–36.

HAHN, G. J., and SHAPIRO, S. S. (1966), "A Catalog and Computer Program for Use With Symmetric and Asymmetric Fractional Factorial Experiments," Contributed paper, 1966 Annual Meetings, American Statistical Association.

HENDRIX, C. (1979), "What Every Technologist Should Know About Experimental Design," *Chemtech*, 9, 167–174.

HILL, W. J., and HUNTER, W. G. (1974), "Design of Experiments for Subsets of Parameters," *Technometrics*, 16, 425–434.

HOOKE, R. (1980), "Getting People to Use Statistics Properly," *The American Statistician*, 34, 39–42.

HUNTER, W. G. (1981), "The Practice of Statistics: The Real World Is an Idea Whose Time Has Come," *The American Statistician*, 35, 72–76.

JOINER, B. L. (1977), "Evaluation of Cryogenic Flow Meters: An Example in Nonstandard Experimental Design and Analysis," *Technometrics*, 19, 353–379.

JIONER, B. L. (1981), "Lurking Variables: Some Examples," *The American Statistician*, 35, 227–233.

JOINER, B. L., and CAMPBELL, C. (1976), "Designing Experiments When Run Order Is Important," *Technometrics*, 19, 353–378.

KHURI, A. I., and CONLON, M. (1981), "Simultaneous Optimization of Multiple Responses Represented by Polynomial Regression Functions," *Technometrics*, 23, 363–375.

MARQUARDT, D. W. (1979), "Statistical Consulting in Industry," *The American Statistician*, 33, 102–106.

MEEKER, W., HAHN, G. J., and FEDER, P. I. (1975), "A Computer Program for Evaluating and Comparing Experimental Designs and Some Applications," *The American Statistician*, 29, 60–64.

MENDENHALL, W. (1968), *Introduction to Linear Models and the Design and Analysis of Experiments*, Belmont, Calif.: Wadsworth Publishing Co.

MUELLER, F. X., and OLSSON, D. M. (1971), "Application of Statistical Design for the Solution of Industrial Finishing Problems," *Journal of Paint Technology*, 43, 54–62.

OLSSON, D. (1982), Personal communication.

PANZL, D. J. (1981), "A Method for Evaluating Software Development Techniques," *Journal of Systems and Software*, 2, 133–137.

PLACKETT, R. L., and BURMAN, J. P. (1946), "Design of Optimal Multifactorial Experiments," *Biometrika*, 23, 305–325.

PRICE, B. (1982), "The 'Expectations Gap' in Statistical Studies: Are Statisticians Poor Planners," Invited Presentation 1982 Annual Meetings, American Statistical Association, Cincinnati, Ohio.

SNEE, R. D. (1982), Personal communication.

WOOD, F. S. (1982), Personal communication.

YOUDEN, W. J. (1951), *Statistical Methods for Chemists*, New York: John Wiley.