

# THE RATIONALE OF SCIENTIFIC EXPERIMENTATION

## JOHN S. DOWD

### INTRODUCTION

Throughout the years I have been working as a statistical quality consultant, I have frequently heard "we don't have the time to do statistical experiments" or various versions of the same theme. The aim of this report is to put forth some thoughts with regard to the advisability and utility of designed experiments.

It is beside the point to discuss the need for or lack of need for 'designed experiments'. All studies that collect data for the purpose of decision making are 'designed'. Some are well designed and some are not. The terms 'planned experiment' or 'planned study' are synonymous with 'designed experiment'.

The use of designed experiments is more than the application of a collection of statistical techniques. It includes a collection of research methodologies that are used to do scientific studies. The aim is to give rise to usable, interpretable results that can be communicated to others. Over time experimental methodologies have been developed that give the experimenter these results in an efficient manner.

Failure to provide usable, interpretable results will usually require additional data and delay or, worse yet, will often lead to mistaken conclusions that can cost companies millions of dollars. In one case a long series of poorly designed experiments led to incorrect conclusions that essentially ended up closing the company and costing hundreds of people their jobs.

There is no such thing as a "correct" design for a given study. A given hypothesis can be studied by different methods using different designs. In fact, the formulation or selection of a suitable experimental plan (design) is usually complicated by numerous practical limitations, not the least important of which are time and cost required to complete the study.

### EXPERIMENTAL STRATEGY

Industrial experiments are complicated by the fact that the experimenter is usually faced with a large number of factors that can affect the quality characteristic of interest. Often these factors interact in complex ways. Simply figuring out which factors have an important effect on the response and which do not can be a major accomplishment.

Thus, in industry we talk about an experimental strategy which usually involves numerous experiments progressing from exploratory work to process optimization.

Don Wheeler in his book "Industrial Experimentation" suggests the following progression:

**Identify Potential Factors**

Initially include all factors that *may* have an effect

**Screen Out Inert Factors**

Delete factors from consideration that do not have a pronounced effect upon the response variable

**Study Active Factors**

Find the best levels for the those factors which do have a pronounced effect upon the response variable

Shewhart outlined the cyclical nature of industrial experimentation in his book *Statistical Method from the Viewpoint of Quality Control*.

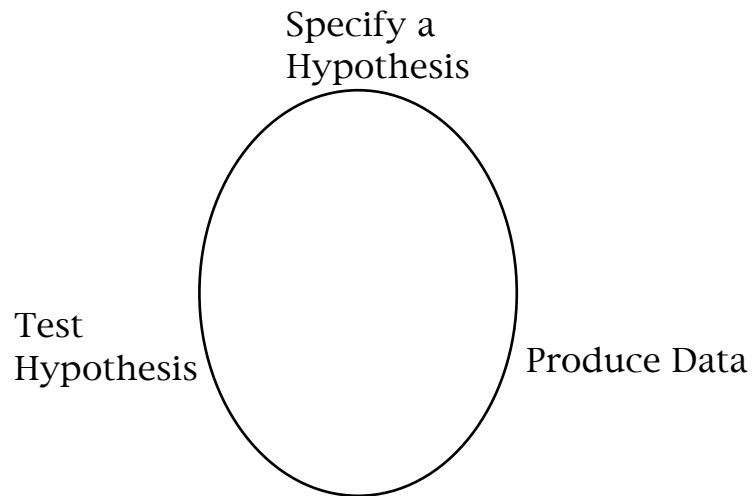


Figure 1. The Shewhart Cycle

Deming later revised this to the now well-known PDSA cycle.

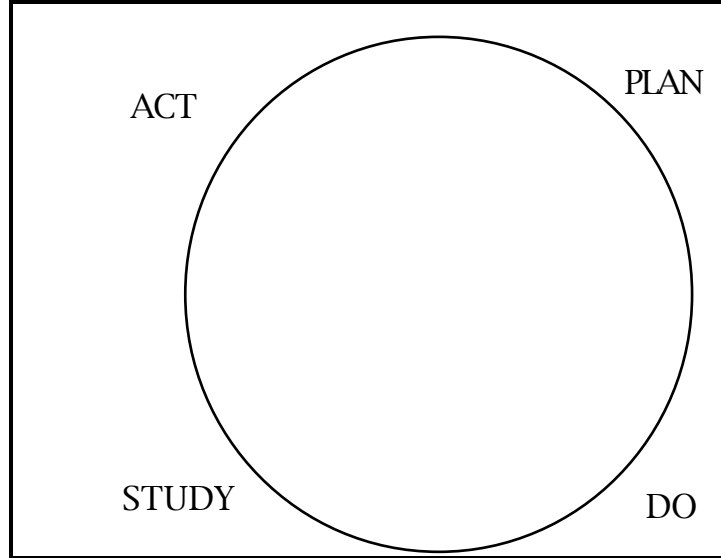


Figure 2. The Deming Cycle

#### **ANALYTIC & ENUMERATIVE STUDIES**

Deming in 1950 in his book *On Some Theory of Sampling* differentiated between two types of studies. He labeled them Enumerative and Analytical. An enumerative study is one that is done for the purpose of taking action on some or all of the items that were included or could have been included in the study. This is called 'full coverage of the sampling frame' (The frame is defined as all of those items that could be included in the sample). Two examples of enumerative studies are the decennial census conducted in the U. S. for the purpose of re-apportioning congressional seats and a study of the defect rate of a lot of incoming material for the purpose of setting a price for it.

Analytic studies are conducted for the purpose of taking action to affect future items of interest, items which could not have been selected to be the subject of the original study. Thus, the concept of a "population" or "representative sample" can not apply since the samples that we are the most interested in haven't happened yet and their characteristics are unknown. An example would be a comparative analysis of two suppliers done for the purpose of choosing one with which to work in the coming years.

It is important to understand this distinction because in the analytic situation, the major source of uncertainty is extrapolation and is, therefore, unquantifiable. The major source of uncertainty in an enumerative study is sampling error and it can be estimated within calculable probability limits. In the world in which we do studies we are almost always doing analytic studies. Thus we are attempting to predict.

For example when an experiment is conducted on a sample of disks to observe the effect(s) of some factor(s), we make a prediction that the results will apply to all disks in the future to which we will apply the factor in that same way. It is critical to understand this. That's why experiments with samples of size one are frequently disastrous. With a sample of one the experimenter has no knowledge of the kind of variability to expect in future samples and the effect observed may be completely erroneous.

## IMPROVING PREDICTIONS

There are two major ways to affect the accuracy of predictions when doing analytic studies. The first is the theoretical and empirical knowledge that can be brought to bear on the study at hand. The second is the logical methodology that can be employed when designing and carrying out the study. **Both are critical.** The most elegantly constructed experiment will be worse than useless if it is based on unsound theoretical considerations. The best theoretical understanding of a subject must always be adapted to see how it applies to the experimental circumstances at hand.

Validity is a general term that applies to experimental results. It is used in two ways. One, called *internal validity* has to do with how the experiment itself is carried out. The higher the internal validity, the more sure we will be that the results were reflective of what actually took place. The other kind of validity is *external validity*. This has to do with the extent to which the experimental results allow valid generalization to the universe of items to which the predictions will be applied.

The issue of validity is of particular importance to us because we are operating in the area of applied science. As Shewhart pointed out over fifty years ago, "Applied science is more exacting than pure science." This is because the applied scientist has to take action based on his results (prediction) and such action will reveal his mistakes, and, as he said, "What is more important, he knows that such mistakes may carry with them serious consequences."

In the experimental situation there exist threats to both internal and external validity. For example problems with measurement can render the experimental results invalid.

Are the measurements being taken on the experimental items obtained from a measurement process that is in statistical control? If the answer to this question is no, experimental results during the experiment (internal validity) may show what appear to be important differences in outcome that are caused by measurement problems and not the result of the experimental manipulation at all. Conversely a special cause shift in the measurement process may mask important differences in the experimental outcome.

Moreover, once action is taken on the experimental results and the future unfolds (external validity), what appeared to be remarkable effects observed in the experimental situation may disappear entirely when subsequent measurements

are taken on different material, in a different setting, using different equipment, run by different people under different circumstances.

The first step in avoiding the threats to internal and external validity is to be aware of them. Unless the experimenter analyzes every proposed experiment in this light, the chances are good that predictions based on the experimental results will be not be as accurate as they could be.

## REPLICATION AND CONFIRMATION

One of the consequences of the analytic situation is that we cannot use probability estimates to confirm the results. The most powerful confirmatory evidence will come from the replication of the results by independent studies. This has strong implications for how the experiments are performed and especially in how the data are analyzed. For example, in the analytic situation it is often more desirable to conduct a series of smaller scale studies over time than to have one large study done only once.

We are still faced, however, with the sometimes confusing task of detecting important effects while at the same time not confusing effects with experimental noise. In the face of this challenge it is imperative that the conditions under which the series of experiments are performed arise from a stable state.

Sometimes there is discussion of "holding things constant" or "copying exactly the previous conditions". This, of course cannot be done, but we can assure ourselves, through the use of statistical control charts and other means, that the experimental state shows no instability over time.

As Wheeler points out in his book *Understanding Variation* all data must be interpreted in context. The cyclical nature of industrial experimentation is a part of this context. Except in rare instances it is a mistake to think that one can conduct single so-called "critical" experiments.

## THE ROLE OF THEORY

With regard to the role of theory in experimentation, Wheeler in a paper entitled "Some Differences Between Theory and Practice" suggests the following. Using this model we can see how theory and process knowledge help to guide the empirical process of experimentation.

THEORETICAL	EMPIRICAL
Formulate Hypothesis	
Reason (deductive) From General Hypothesis to Specific Consequences	Plan and Carry Out Experiment
	Determine Methods to Observe Presence or Absence of Specific Consequences
	Plan Data Collection
	Consider Problems that Might Arise with Collection of the Data
	Collect the Data
	Consider Problems that Occurred While Collecting the Data
Expected Outcome	Observed Outcome
Compare Observed and Expected Outcomes	
Make a Prediction Based on Results (Inductive)	

One can have knowledge of the theory but leap to erroneous conclusions or greatly delay a project due to ineffective or inefficient experimental technique.

As previously mentioned, the converse is true as well. One can understand the logic of good experimental design thoroughly, but without process knowledge and knowledge of appropriate theory behind it, that person may well conduct a series of utterly useless experiments.

Knowledge of theory and process is more critical than knowledge of experimental design. After all, the early development of most experimental design techniques was done by practitioners such as botanists, chemists, physicists and so on. There is no need, however, to make this a 'theory vs. design' face-off. Clearly the best case is to have both.

Sometimes an experimenter cannot avail himself of a strong design. For example, a client came to me with some data from the safety group. Over the past several months this person had taken a number of actions intended to decrease the rate of accidents. His experimental hypothesis (theory) is that the actions will reduce accidents. He questions whether or not the actions have had the desired effect.

The design of the study is retrospective. That is, one can only look backward in time at accident rates to see if they have declined. There is no 'control' group. This poses some problems with the interpretation of the results. In this case analysis of the data indicates that the accident rate seems to have declined. Was the decline due to the actions taken or would it have happened anyway? In fact, accidents may have declined more sharply if something different or even nothing else had been done. How would we know?

We'll never know ... there is no way to know because this particular experiment cannot be replicated. If however, we went from organization to organization and carried out similar actions in each of them and in each case there was a subsequent reduction in accidents, our confidence —what Shewhart termed 'degree of belief' — would be higher.

### **ECONOMIC TRADE OFF**

Thus, in industrial experimentation there is an economic question that must be dealt with. How should we trade off the cost (in time and money) of additional replication with the increase in degree of belief confirmatory replication brings?

One way to do this is to have designs that are efficient. One category of such designs is called the factorial design. These designs are constructed in such a way as to give the experimenter a number of independent measures of the effect of the experimental factors while minimizing the number experimental runs that are required.

### **INTERACTIONS**

Another difficulty in industrial experimentation is the existence of interactions. As has been stated, manufacturing processes are complex with many factors involved. In many processes these factors interact. This is particularly so for continuous processes such as plating or sputtering. Saying that the factors interact means more than that they are related to each other. It means that the

effect of one (or more) factors on the response variable(s) changes when one (or more) other factor(s) changes its value.

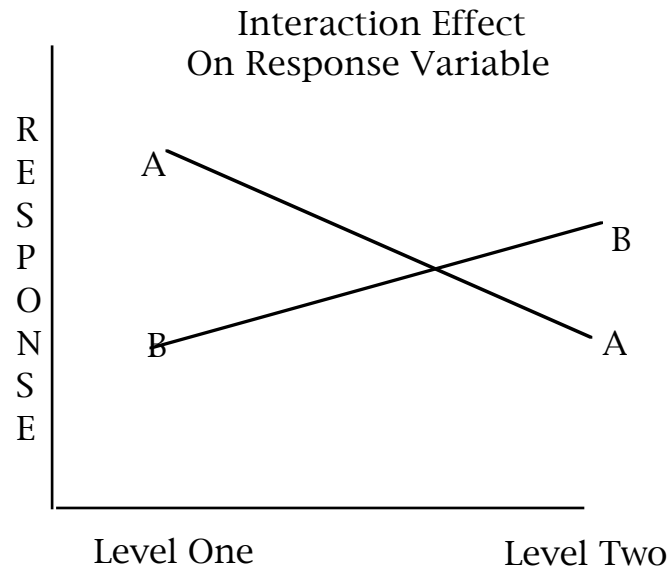


Figure 3. Graphical Depiction of an Interaction

A colleague of mine has developed a training session for design of experiments where the students pop corn. A key quality characteristic is the percentage of kernels of corn that are popped. Thus, the response measure is the number of unpopped kernels. Two factors are studied. One is the type of popper used. In the above figure call this A: two types of poppers are used. The other factor is type of popcorn used. In the above figure call this B. An economy brand and a deluxe brand are tested. The question arises as which is the better popcorn and which is the better popper.

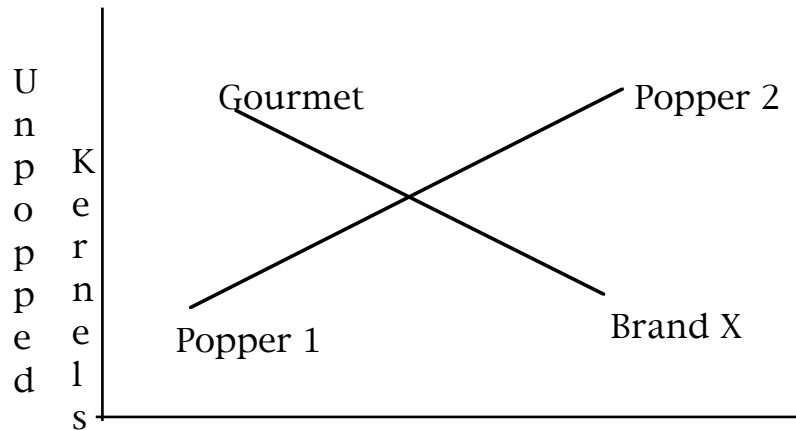


Figure 4. Interaction in Popcorn Experiment

The figure shows that there is an interaction between the two. What kind of popcorn produces the lesser number of unpopped kernels? It depends on which popper is used. What kind of popper produces the lesser number of unpopped kernels? It depends on the type of popcorn used. The Interaction has made it impossible to assess the main effects (type of popper and type of popcorn).

In order to detect interactions and understand the nature of their effects it is necessary to combine the interacting factors into the same experimental runs. The problem is not necessarily knowing in advance if the interactions exist. Sometimes they are predictable with theory. Sometimes they are discovered when the process behaves 'strangely'.

In addition to their efficiency, factorial designs also offer the only method of detecting interactions through experimentation. Because numerous factors can be combined in the same series of experimental runs, the interactions can be detected and the nature of their effects can be evaluated when they are present.

The factorial design strategy, therefore, not only has the advantage of efficiency, but also allows the detection and analysis of interactions which frequently exist in complex processes. It can be seen that one-factor-at-a-time experiments are far less effective. Aside from the risk of making costly mistakes, they use far more time. This loss of time is particularly a problem in companies making products with short life cycles.

There are some variations on the theme of factorial experiments. For example there is a category of experiments called 'fractional factorials' that allows the experimenter to assess the effect of large numbers of factors in an efficient way. The trade off is that some higher order interactions must be assumed to be absent. Fractional factorial designs lie at the core of what has come to be known as Taguchi techniques or, more generically, "robust design".

## SUMMARY

A natural part of the evolution of continuous improvement is the incorporation of the tool of designed experiments into the development and manufacturing processes. There is confusion as to the definition and methodology of designed experiments and as to the advantages offered by employing specific design types. The advantages, however, are well defined and organizations using these techniques will enjoy a competitive advantage over those who don't or do so half-heartedly. Competitiveness is largely driven by increasing the rate at which improvements are made. Fundamental to that increase is the rate at which knowledge is gained and transferred. Designed experiments offer a powerful and efficient means of acquiring and communicating knowledge.