Experiments

Planning, Analysis, and Parameter Design Optimization

C. F. Jeff Wu

Michael Hamada

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

Contents

CHAPTER 1

Basic Principles and Experiments with a Single Factor

Some basic concepts and principles in experimental design are introduced in this chapter. The simplest class of experiments, that with a single factor, is then considered. Analysis techniques like regression analysis, variable selection, analysis of variance (ANOVA), multiple comparisons, and residual analysis are presented. These techniques are applicable to more complex experiments to be considered in later chapters.

1.1 INTRODUCTION AND HISTORICAL PERSPECTIVE

Experimentation is one of the most common activities that people engage in. It covers a wide range of applications from household work like food preparation to technological innovation in material science, semiconductors, robotics, life science, etc. It allows an investigator to find out what happens to the output or response when the settings of the input variables in a system are purposely changed. Statistical or common-sense analysis can then be used to study the relationship between the input and output values. A better understanding of how the input variables affect the performance of a system can thereby be achieved. This provides a basis for selecting optimum input settings. Experimental design is a body of knowledge and techniques that enables an investigator to conduct better experiments, analyze data efficiently, and make the connections between the conclusions from the analysis and the original objectives of the investigation.

Experimentation is used to understand and/or improve a system. A system can be a product or process. A product can be one developed in engineering, biology, or the physical sciences. A process can be a manufacturing process, a process that describes a physical phenomenon, or a non-physical process such as those found in service or administration. Although most examples in the book are from engineering or the physical and biological sciences, the methods can also be applied to other disciplines, such as business, medicine, and psychology. For example, in studying the efficiency and cost of a payroll operation, the entire payroll operation can be viewed as a process with key input variables such as the number of supervisors, the number of clerks, method of bank deposit, level of automation, administrative structure, etc. A computer simulation model can then be used to study the effects of changing these input variables on cost and efficiency.

Modern experimental design dates back to the pioneering work of the great statistician R. A. Fisher in the 1930s at the Rothamsted Agricultural Experimental Station in the United Kingdom. Fisher's work and the notable contributions by F. Yates and D. J. Finney were motivated by problems in agriculture and biology. Because of the nature of agricultural experiments, they tend to be large in scale, take a long time to complete, and must cope with variations in the field. Such considerations led to the development of blocking, randomization, replication, orthogonality, and the use of analysis of variance and fractional factorial designs. The theory of combinatorial designs, to which R. C. Bose has made fundamental contributions, was also stimulated by problems in block designs and fractional factorial designs. The work in this era also found applications in social science research and in the textile and woolen industries.

The next era of rapid development came soon after World War II. In attempting to apply previous techniques to solve problems in the chemical industries, G. E. P. Box and co-workers at Imperial Chemical Industries discovered that new techniques and concepts had to be developed to cope with the unique features in process industries. The new techniques focused on process modeling and optimization rather than on treatment comparisons, which was the primary objective in agricultural experiments. The experiments in process industries tend to take less time but put a premium on run size economy because of the cost of experimentation. These time and cost factors naturally favor sequential experimentation. The same considerations led to the development of new techniques for experimental planning, notably central composite designs and optimal designs. Their analysis relies more heavily on regression modeling and graphical analysis. Process optimization based on the fitted model is also emphasized. Because the choice of design is often linked to a particular model (e.g., a second-order central composite design for a second-order regression model) and the experimental region may be irregularly shaped, a flexible strategy for finding designs to suit a particular model and/or experimental region is called for. With the availability of fast computational algorithms, optimal designs (which was pioneered by J. Kiefer) have become an important part of this strategy.

The relatively recent emphasis on variation reduction has provided a new source of inspiration and techniques in experimental design. In manufacturing the ability to make many parts with few defects is a competitive advantage. Therefore variation reduction in the quality characteristics of these parts has become a major focus of quality and productivity improvement. G. Taguchi advocated the use of robust parameter design to improve a system (i.e., a product or process) by making it less sensitive to variation, which is hard to control during normal operating or use conditions of the product or process. The input variables of a system can be divided into two broad types: control factors, whose values remain fixed once they are chosen, and noise factors, which are hard to control during normal conditions. By exploiting the interactions between the control and noise factors, one can achieve robustness by choosing control factor settings that make the system less sensitive to noise variation. The new paradigm is variation modeling and reduction. Traditionally, when the mean and variance are both considered, variance is used to assess the variability of the sample mean as in the t test or of the treatment comparisons as in the analysis of variance. The focus on variation and the division of factors into two types led to the development of new concepts and techniques in the planning and analysis of robust parameter design experiments. The original problem formulation and some basic concepts were developed by G. Taguchi. Other basic concepts and many sound statistical techniques have been developed by statisticians since the mid-1980s.

Given this historical background, we now classify experimental problems into five broad categories according to their objectives.

1. *Treatment Comparisons.* The main purpose is to compare several treatments and select the best ones. For example, in the comparison of six barley varieties, are they different in terms of yield and resistance to drought? If they are indeed different, how are they different and which are the best? Examples of treatments include varieties (rice, barley, corn, etc.) in agricultural trials, sitting positions in ergonomic studies, instructional methods, machine types, suppliers, etc.

2. *Variable Screening.* If there is a large number of variables in a system but only a relatively small number of them is important, a screening experiment can be conducted to identify the important variables. Such an experiment tends to be economical in that it has few degrees of freedom left for estimating error variance and higher order terms like quadratic effects or interactions. Once the important variables are identified, a follow-up experiment can be conducted to study their effects more thoroughly. This latter phase of the study falls into the category discussed next.

3. *Response Surface Exploration.* Once a smaller number of variables is identified as important, their effects on the response need to be explored. The relationship between the response and these variables is sometimes referred to as a response surface. Usually the experiment is based on a design that allows the linear and quadratic effects of the variables and some of the interactions between the variables to be estimated. This experiment tends to be larger (relative to the number of variables under study) than the screening experiment. Both parametric and semi-parametric models may be considered. The latter is more computer intensive but also more flexible in model fitting.

4. *System Optimization.* In many investigations, interest lies in the optimization of the system. For example, the throughput of an assembly plant or

the yield of a chemical process is to be maximized; the amount of scrap or number of reworked pieces in a stamping operation is to be minimized; the time required to process a travel claim reimbursement is to be reduced. If a response surface has been identified, it can be used for optimization. For the purpose of finding an optimum, it is, however, not necessary to map out the whole surface as in a response surface exploration. An intelligent sequential strategy can quickly move the experiment to a region containing the optimum settings of the variables. Only within this region is a thorough exploration of the response surface warranted.

5. *System Robustness*. Besides optimizing the response, it is important in quality improvement to make the system robust against noise (i.e., hard-to-control) variation. This is often achieved by choosing control factor settings at which the system is less sensitive to noise variation. Even though the noise variation is hard to control in normal conditions, it needs to be systematically varied during experimentation. The response in the statistical analysis is often the variance (or its transformation) among the noise replicates for a given control factor setting.

1.2 A SYSTEMATIC APPROACH TO THE PLANNING AND IMPLEMENTATION OF EXPERIMENTS

In this section, we provide some guidelines on the planning and implementation of experiments. The following seven-step procedure summarizes the important steps that the experimenter must address.

1. *State Objective*. The objective of the experiment needs to be clearly stated. All stakeholders should provide input. For example, for a manufactured product, the stakeholders may include design engineers who design the product, process engineers who design the manufacturing process, line engineers who run the manufacturing process, suppliers, lineworkers, customers, marketers, and managers.

2. *Choose Response*. The response is the experimental outcome or observation. There may be multiple responses in an experiment. Several issues arise in choosing a response. Responses may be *discrete* or *continuous*. Discrete responses can be counts or categories, e.g., binary (good, bad) or ordinal (easy, normal, hard). Continuous responses are generally preferable. For example, a continuous force measurement for opening a door is better than an ordinal (easy, normal, hard to open) judgment; the recording of a continuous characteristic is preferred to the recording of the percent that the characteristic is within its specifications. Trade-offs may need to be made. For example, an ordinal measurement of force to open a door may be preferable to delaying the experiment until a device to take continuous measurements can be developed. Most importantly, there should be a good

measurement system for measuring the response. In fact, an experiment called a *gauge repeatability and reproducibility* (R&R) *study* can be performed to assess a continuous measurement system (AIAG, 1990). When there is a single measuring device, the variation due to the measurement system can be divided into two types: variation between the operators and variation within the operators. Ideally, there should be no between-operator variation and small within-operator variation. The gauge R&R study provides estimates for these two components of measurement system variation. Finally, the response should be chosen to increase understanding of mechanisms and physical laws involved in the problem. For example, in a process that is producing underweight soap bars, soap bar weight is the obvious choice for the response in an experiment to improve the underweight problem. By examining the process more closely, there are two subprocesses that have a direct bearing on soap bar weight: the mixing process that affects the soap bar density and the forming process that impacts the dimensions of the soap bars. In order to better understand the mechanism that causes the underweight problem, soap bar density and soap bar dimensions are chosen as the responses. Even though soap bar weight is not used as a response, it can be easily determined from its density and dimensions. Therefore, no information is lost in studying the density and dimensions. Such a study may reveal new information about the mixing and forming subprocesses, which can in turn lead to a better understanding of the underweight problem. Further discussions on and other examples of the choice of responses can be found in Phadke (1989) and León, Shoemaker and Tsui (1993).

The chosen responses can be classified according to the stated objective. Three broad categories will be considered in this book: **nominal-the-best**, **larger-the-better**, and **smaller-the-better**. The first one will be addressed in Section 3.2 and the last two in Section 5.2.

3. *Choose Factors and Levels.* A **factor** is a variable that is studied in the experiment. In order to study the effect of a factor on the response, two or more values of the factor are used. These values are referred to as **levels** or settings. A **treatment** is a combination of factor levels. When there is a single factor, its levels are the treatments. For the success of the experiment, it is crucial that potentially important factors be identified at the planning stage. There are two graphical methods for identifying potential factors. First, a **flow chart** of the process or system is helpful to see where the factors arise in a multi-stage process. In Figure 1.1, a rough sketch of a paper pulp manufacturing process is given which involves raw materials from suppliers, a chemical process to make a slurry which is passed through a mechanical process to produce the pulp. Involving all the stakeholders is invaluable in capturing an accurate description of the process or system. Second, a **cause-and-effect diagram** can be used to list and organize the potential factors that may impact the response. In Figure 1.2, a cause-and-effect diagram is given which lists the factors thought to affect the product quality of an injection molding process. Traditionally, the factors are organized under the headings:

Man, Machine, Measurement, Material, Method, and Environment (Mother Nature for those who like M's). Because of their appearance, cause-and-effect diagrams are also called *fishbone diagrams*. Different characteristics of the factors need to be recognized because they can affect the choice of the experimental design. For example, a factor such as furnace temperature is *hard to change*. That is, after changing the temperature setting, it may take a considerable amount of time before the temperature stabilizes at the new setting. A factor may also be *hard to set* so that the actual level used in the experiment may be different than the intended level. For example, the actual impact force of a pellet projected at an automobile windshield can only be set within 3 psi of the intended impact force. Other factors that may be hard or impossible to control are referred to as *noise* factors. Examples of noise factors include environmental and customer use conditions. (An in-depth discussion of noise factors will be given in Section 10.3.)

Factors may be *quantitative* and *qualitative*. Quantitative factors like temperature, time, and pressure take values over a continuous range. Qualitative factors take on a discrete number of values. Examples of qualitative factors include operation mode, supplier, position, line, etc. Of the two types of factors, there is more freedom in choosing the levels of quantitative factors. For example, if temperature (in °C) is in the range 100–200°C, one could choose 130 and 160°C for two levels or 125, 150, and 175°C for three levels. If only a linear effect is expected, two levels should suffice. If curvature is expected, then three or more levels are required. In general, the levels of quantitative factors must be chosen far enough apart so that an effect can be detected but not too far so that different physical mechanisms are involved (which would make it difficult to do statistical modeling and prediction). There is less flexibility in choosing the levels of qualitative factors. Suppose there are three testing methods under comparison. All three must be included as three levels of the factor "testing method," unless the investigator is willing to postpone the study of one method so that only two methods are compared in a two-level experiment.

When there is flexibility in choosing the number of levels, the choice may depend on the availability of experimental plans for the given combination of factor levels. In choosing factors and levels, *cost* and *practical constraints* must be considered. If two levels of the factor "material" represent expensive and cheap materials, a negligible effect of material on the response will be welcomed because the cost can be drastically reduced by replacing the expensive material by the cheap alternative. Factor levels must be chosen to meet practical constraints. If a factor combination (e.g., high temperature and long time in an oven) can potentially lead to disastrous results (e.g., burned or overbaked), it should be avoided and a different plan should be chosen.

4. *Choose Experimental Plan.* Use the fundamental principles discussed in Section 1.3 as well as other principles presented throughout the book. The choice of the experimental plan is crucial. A poor design may capture little

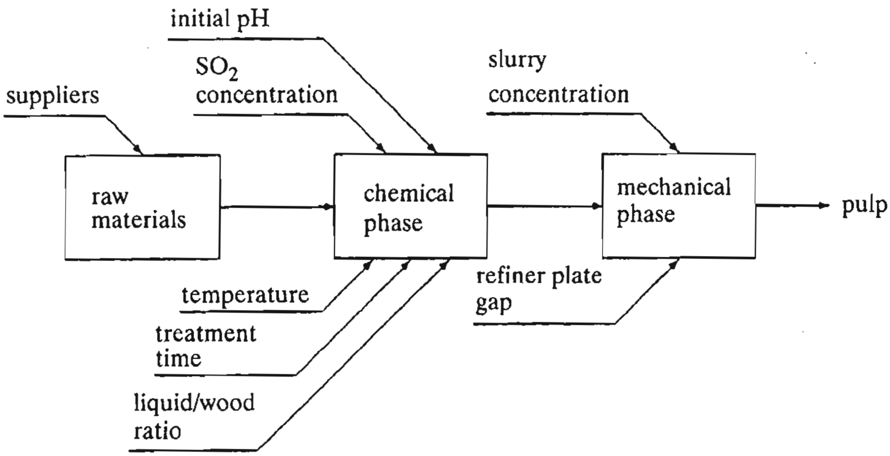

Figure 1.1. Flow Chart, Pulp Manufacturing Process.

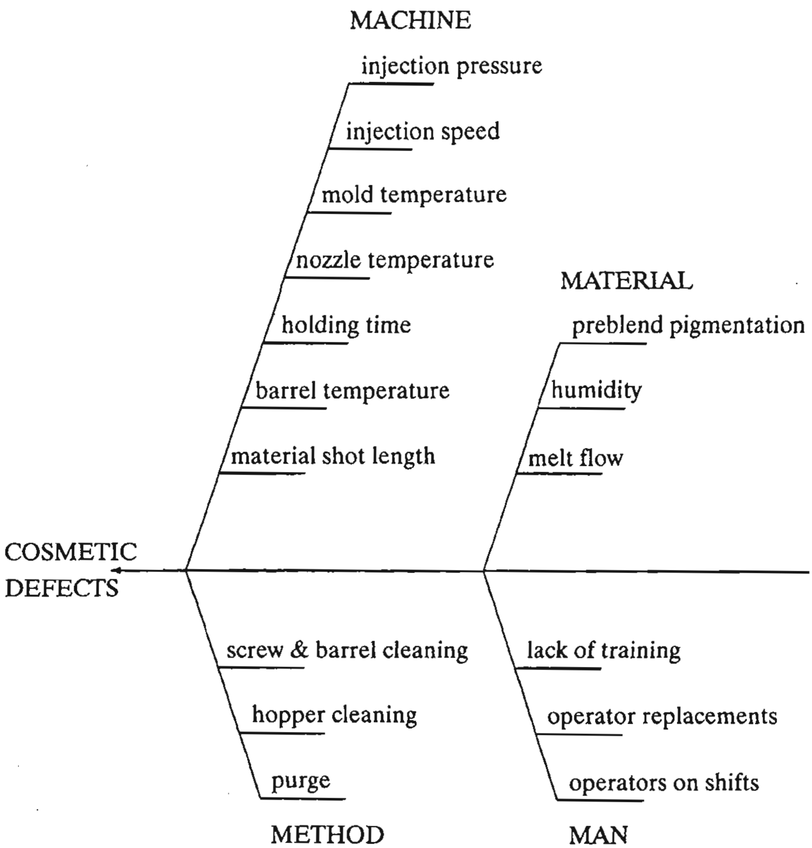

Figure 1.2. Cause-and-Effect Diagram, Injection Molding Experiment.

Man, Machine, Measurement, Material, Method, and Environment (Mother Nature for those who like M's). Because of their appearance, cause-and-effect diagrams are also called *fishbone diagrams*. Different characteristics of the factors need to be recognized because they can affect the choice of the experimental design. For example, a factor such as furnace temperature is *hard to change*. That is, after changing the temperature setting, it may take a considerable amount of time before the temperature stabilizes at the new setting. A factor may also be *hard to set* so that the actual level used in the experiment may be different than the intended level. For example, the actual impact force of a pellet projected at an automobile windshield can only be set within 3 psi of the intended impact force. Other factors that may be hard or impossible to control are referred to as *noise* factors. Examples of noise factors include environmental and customer use conditions. (An in-depth discussion of noise factors will be given in Section 10.3.)

Factors may be *quantitative* and *qualitative*. Quantitative factors like temperature, time, and pressure take values over a continuous range. Qualitative factors take on a discrete number of values. Examples of qualitative factors include operation mode, supplier, position, line, etc. Of the two types of factors, there is more freedom in choosing the levels of quantitative factors. For example, if temperature (in °C) is in the range 100–200°C, one could choose 130 and 160°C for two levels or 125, 150, and 175°C for three levels. If only a linear effect is expected, two levels should suffice. If curvature is expected, then three or more levels are required. In general, the levels of quantitative factors must be chosen far enough apart so that an effect can be detected but not too far so that different physical mechanisms are involved (which would make it difficult to do statistical modeling and prediction). There is less flexibility in choosing the levels of qualitative factors. Suppose there are three testing methods under comparison. All three must be included as three levels of the factor "testing method," unless the investigator is willing to postpone the study of one method so that only two methods are compared in a two-level experiment.

When there is flexibility in choosing the number of levels, the choice may depend on the availability of experimental plans for the given combination of factor levels. In choosing factors and levels, *cost* and *practical constraints* must be considered. If two levels of the factor "material" represent expensive and cheap materials, a negligible effect of material on the response will be welcomed because the cost can be drastically reduced by replacing the expensive material by the cheap alternative. Factor levels must be chosen to meet practical constraints. If a factor combination (e.g., high temperature and long time in an oven) can potentially lead to disastrous results (e.g., burned or overbaked), it should be avoided and a different plan should be chosen.

4. *Choose Experimental Plan*. Use the fundamental principles discussed in Section 1.3 as well as other principles presented throughout the book. The choice of the experimental plan is crucial. A poor design may capture little

information which no analysis can rescue. On the other hand, if the experiment is well planned, the results may be obvious so that no sophisticated analysis is needed.

5. *Perform the Experiment.* The use of a *planning matrix* is recommended. This matrix describes the experimental plan in terms of the actual values or settings of the factors. For example, it lists the actual levels such as 50 or 70 psi if the factor is pressure. To avoid confusion and eliminate potential problems of running the wrong combination of factor levels in a multifactor experiment, each of the treatments, such as temperature at 30°C and pressure at 70 psi, should be put on a separate piece of paper and given to the personnel performing the experiment. It is also worthwhile to perform a *trial run* to see if there will be difficulties in running the experiment, namely, if there are problems with setting the factors and measuring the responses. Any deviations from the planned experiment need to be recorded. For example, for hard-to-set factors, the actual values should be recorded.

6. *Analyze the Data.* An analysis appropriate for the design used to collect the data needs to be carried out. This includes model fitting and assessment of the model assumptions through an analysis of residuals. Many analysis methods will be presented throughout the book.

7. *Draw Conclusions and Make Recommendations.* Based on the data analysis, conclusions are presented which include the important factors and a model for the response in terms of the important factors. Recommended settings or levels for the important factors may also be given. The conclusions should refer back to the stated objectives of the experiment. A *confirmation experiment* is worthwhile, for example, to confirm the recommended settings. Recommendations for further experimentation in a *follow-up experiment* may also be given. For example, a follow-up experiment is needed if two models explain the experimental data equally well and one must be chosen for optimization.

For further discussion on the planning of experiments, see Coleman and Montgomery (1993), Knowlton and Keppinger (1993), and Barton (1997).

1.3 FUNDAMENTAL PRINCIPLES: REPLICATION, RANDOMIZATION, AND BLOCKING

There are three fundamental principles that need to be considered in the design of an experiment: **replication**, **randomization**, and **blocking**. Other principles will be introduced later in the book as they arise.

An *experimental unit* is a generic term that refers to a basic unit such as material, animal, person, machine, or time period, to which a treatment is applied. By *replication*, we mean that each treatment is applied to experimental units that are representative of the population of units to which the conclusions of the experiment will apply. It enables the estimation of the magnitude of experimental error (i.e., the error variance) against which the differences among treatments are judged. Increasing the number of replications, or *replicates*, decreases the variance of the treatment effect estimates and provides more power for detecting differences in treatments. A distinction needs to be made between replicates and *repetitions*. For example, three readings from the same experimental unit are repetitions while the readings from three separate experimental units are replicates. The error variance from the former is less than that from the latter because repeated readings only measure the variation due to errors in reading while the latter also measures the unit-to-unit variation. Underestimation of the true error variance can result in the false detection of effect significance.

The second principle is that of *randomization*. It should be applied to the allocation of units to treatments, the order in which the treatments are applied in performing the experiment, and the order in which the responses are measured. It provides protection against variables that are unknown to the experimenter but may impact the response. It reduces the unwanted influence of subjective judgment in treatment allocation such as in a physician's assignment of medical treatments to patients. Moreover, randomization ensures validity of the estimate of experimental error and provides a basis for inference in analyzing the experiments. For an in-depth discussion on randomization, see Hinkelmann and Kempthorne (1994).

A group of homogeneous units is referred to as a *block*. Examples of blocks include days, weeks, morning vs. afternoon, batches, lots, sets of twins, and pairs of kidneys. For *blocking* to be effective, the units should be arranged so that the within-block variation is much smaller than the between-block variation. By comparing the treatments within the same block, the block effects are eliminated in the comparison of the treatment effects, thereby making the experiment more efficient. For example, there may be a known day effect on the response so that if all the treatments can be applied within the same day, the day-to-day variation is eliminated.

If blocking is effective, it should be applied to remove the block-to-block variation. Randomization can then be applied to the assignments of treatments to units within the blocks to further reduce the influence of unknown variables. This strategy of **block what you can and randomize what you cannot** is used in randomized block designs, to be discussed in Section 2.2.

These three principles are generally applicable to physical experiments but not to computer experiments because the same input in a computer experiment gives rise to the same output. Computer experiments are not considered in the book.

A simple example will be used to explain these principles. Suppose two keyboards denoted by A and B are being compared in terms of typing efficiency. Six different manuscripts denoted by 1–6 are given to the same typist. First the test is arranged in the following sequence:

$$1.\ A, B,\quad 2.\ A, B,\quad 3.\ A, B,\quad 4.\ A, B,\quad 5.\ A, B,\quad 6.\ A, B.$$

Even though the experiment is replicated six times (with six manuscripts) and blocking is used to compare two keyboards on the same manuscript, the design has a serious flaw. After typing the manuscript on keyboard A, the typist will get more familiar with the content of the manuscript when he or she is typing the same manuscript on keyboard B. This "learning effect" will unfairly help the performance of keyboard B. The observed difference between A and B is the combination of the treatment effects (which measures the intrinsic difference between A and B) and the learning effect. For the given test sequence, it is impossible to disentangle the learning effect from the treatment effect. Randomization would help reduce the unwanted influence of the learning effect, which might not have been known to the investigator who planned the study. By randomizing the typing order for each manuscript, the test sequence may appear as follows:

$$1.\ A,B,\quad 2.\ B,A,\quad 3.\ A,B,\quad 4.\ B,A,\quad 5.\ A,B,\quad 6.\ A,B.$$

With four *AB*'s and two *BA*'s in the sequence, it is a better design than the first one. A further improvement can be made. The design is not balanced because B benefits from the learning effect in four trials while A only benefits from two trials. There is still a residual learning effect not completely eliminated by the second design. The learning effect can be completely eliminated by requiring that half of the trials have the order *AB* and the other half the order *BA*. The actual assignment of *AB* and *BA* to the six manuscripts should be done by randomization. This method is referred to as *balanced randomization.* Balance is a desirable design property, which will be discussed later.

For simplicity of discussion, we have assumed that only one typist was involved in the experiment. In a practical situation, such an experiment should involve several typists that are representative of the population of typists so that the conclusions made from the study would apply more generally. This and other aspects of the typing experiment will be addressed in the exercises.

With these principles in mind, a useful addition to the cause-and-effect diagram is to indicate how the proposed experimental design addresses each listed factor. The following designations are suggested: **E** for an experimental factor, **B** for a factor handled by blocking, **O** for a factor held constant at one value, and **R** for a factor handled by randomization. This designation clearly indicates how the proposed design deals with each of the potentially important factors. The designation **O**, for "one value," serves to remind the experimenter that the factor is held constant during the current experiment but may be varied in a future experiment. An illustration is given in Figure 1.3 from the injection molding experiment discussed in Section 1.2.

Other designations of factors can be considered. For example, experimental factors can be further divided into two types (control factors and noise factors), as in the discussion on the choice of factors in Section 1.2. For the

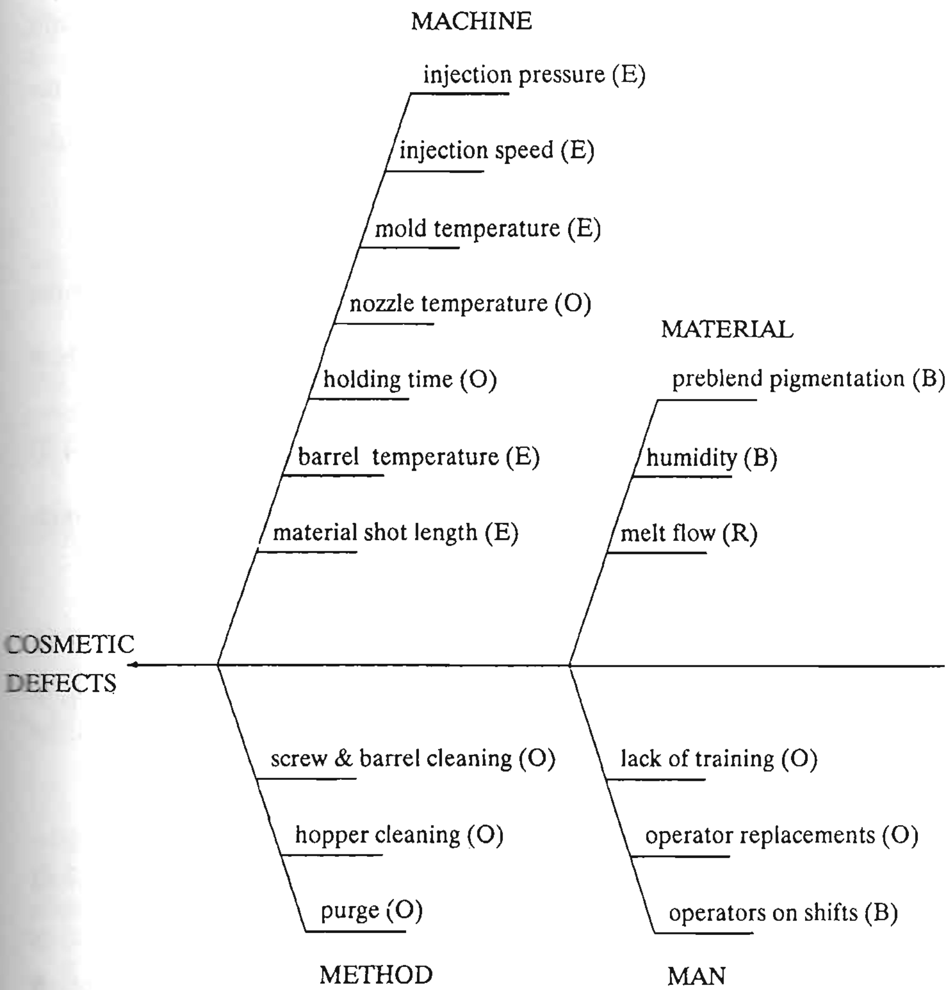

Figure 1.3. Revised Cause-and-Effect Diagram, Injection Molding Experiment.

implementation of experiments, we may also designate an experimental factor as "hard-to-change" or "easy-to-change." These designations will be considered later as they arise.