Early Experimentation

The earliest assumptions of experimental research were based on what was known as the law of the single variable. In 1873 John Stuart Mill provided a definition for this principle. He stated five rules or canons that he believed would include all types of logical procedure required to establish order among controlled events.

One of his canons, known as the method of difference, states:

*If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur have every circumstance in common save one, that one occurring only in the former, the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause of the phenomenon.* (Mill, 1873, p. 222)

In simpler language, if two situations are alike in every respect, and one element is added to one but not the other, any difference that develops is the effect of the added element; or, if two situations are alike in every respect, and one element is removed from one but not from the other, any difference that develops may be attributed to the subtracted element.

The law of the single variable provided the basis for much early laboratory experimentation. In 1662 Robert Boyle, an Irish physician, used this method in arriving at a principle on which he formulated his law of gases: When temperature is held constant, the volume of an ideal gas is inversely proportional to the pressure exerted on it. In other words, when pressure is raised, volume decreases; when pressure is lowered, volume increases. In Boyle’s law, pressure is the single variable:

\[
\frac{V_1}{V_2} = \frac{P_2}{P_1}
\]

A little more than a century later, Jacques A. C. Charles, a French physicist, discovered a companion principle, now known as Charles’s law. He observed that when the pressure was held constant, the volume of an ideal gas was directly proportional to the temperature. When temperature is raised, volume increases; when temperature is lowered, volume decreases. In Charles’s law, temperature is the single variable:

\[
\frac{V_1}{V_2} = \frac{T_1}{T_2}
\]

Although the concept of the single variable proved useful in some areas of the physical sciences, it failed to provide a sound approach to experimentation in the behavioral sciences. Despite its appealing simplicity and apparent logic, it did not provide an adequate method for studying complex problems. It assumed a highly artificial and restricted relationship between single variables. Rarely, if ever, are
human events the result of single causes. They are usually the result of the interaction of many variables, and an attempt to limit variables so that one can be isolated and observed proves impossible.

The contributions of R. A. Fisher, first applied in agricultural experimentation, have provided a much more effective way of conducting realistic experimentation in the behavioral sciences. His concept of achieving pre-experimental equating of conditions through random selection of subjects and random assignment of treatments, and his concepts of analysis of variance and analysis of covariance made possible the study of complex interactions through factorial designs in which the influence of more than one independent variable on more than one dependent variable could be observed. Current uses of this type of design will be discussed more fully later in this chapter.

Experimental and Control Groups

An experiment involves the comparison of the effects of a particular treatment with that of a different treatment or of no treatment. In a simple conventional experiment, reference is usually made to an experimental group and to a control group.

These groups are equated as nearly as possible. The experimental group is exposed to the influence of the factor under consideration; the control group is not. Observations are then made to determine whether the difference appears or what change or modification occurs in the experimental as contrasted with the control group.

Sometimes it is also necessary to control for the effect of actually participating in an experiment. Medical researchers have long recognized that patients who receive any medication, regardless of its real efficacy, tend to feel better or perform more effectively. In medical experiments a harmless or inert substitute is administered to the control group to offset the psychological effect of medication. These substitutes, or placebos, are indistinguishable from the real medication under investigation, and neither experimental nor control subjects know whether they are receiving the medication or the placebo. The effectiveness of the true medication is the difference between the effect of the medication and that of the placebo.

What seems to be a similar psychological effect was recognized in a series of experiments at the Hawthorne Plant of the Western Electric Company and originally published in 1933 (Mayo, 1960). The studies concerned the relationships between certain working conditions and worker output efficiency. Illumination was one of these manipulated experimental variables. The researchers found that as light intensity was increased, worker output increased. After a certain peak was apparently reached, it was decided to see what effect the reduction of intensity of illumination would have. To the surprise of the researchers, as intensity was decreased by stages, output continued to increase. The researchers concluded that the attention given the workers and their awareness of participation in the experiment apparently were important motivating factors. From these studies the term Hawthorne Effect was introduced into the psychological literature.

Variables

Independent and Dependent Variables

Variables are the conditions or characteristics that the experimenter manipulates, controls, or observes. The independent variables are the conditions or characteristics
that the experimenter manipulates or controls in his or her attempt to ascertain the relationship to observed phenomena. The dependent variables are the characteristics that appear, disappear, or change as the experimenter introduces, removes, or changes independent variables.

In educational research an independent variable may be a particular teaching method, a type of teaching material, a reward, a period of exposure to a particular condition, or an attribute such as sex or level of intelligence. The dependent variable may be a test score, the number of errors, or measured speed in performing a task. Thus, the dependent variables are the measured changes in pupil performance attributable to the influence of the independent variables.

There are two types of independent variables: treatment and organismic or attribute variables. Treatment variables are those factors that the experimenter manipulates and to which he or she assigns subjects. Attribute variables are the characteristics that cannot be altered by the experimenter. Such independent variables as age, sex, race, and intelligence level have already been determined, but the experimenter can decide to include them or remove them as variables to be studied. The question of whether 5-year-old girls show greater reading achievement than 8-year-old boys is an example of the use of an organismic variable. The teaching procedure is the same for both groups so there is no treatment independent variable.

Confounding Variables
Confounding variables are those aspects of a study or sample that might influence the dependent variable (outcome measure) and whose effect may be confused with the effects of the independent variable. Confounding variables are of two types: intervening and extraneous variables.

Intervening variables are hard, if not impossible, to observe because they usually have to do with an individual's feelings (e.g., boredom, fatigue, excitement). Extraneous variables, on the other hand, are more readily observed or measured and thus more easily controlled. Both of these confounding variables should be thought of as potential confounds since they are only confounding variables if they are systematically related to the independent variable and have an effect on the dependent variable. Because it is impossible to know prior to the study which of these conditions may exist, all potential confounds need to be controlled as much as possible.

Intervening Variables
In many types of behavioral research the relationship between the independent and dependent variables is not a simple one of stimulus to response. Certain variables cannot be controlled or measured directly may have an important effect on the outcome. These modifying variables intervene between the cause and the effect.

In a classroom language experiment a researcher is interested in determining the effect of immediate reinforcement on learning the parts of speech. He or she suspects that certain factors or variables other than the one being studied, immediate reinforcement, may be influencing the results, even though they cannot be observed directly. These factors—fatigue, motivation, for example—may be intervening variables. They are difficult to define in operational, observable, terms, but they cannot be ignored. Rather, they must be controlled as much as is feasible through the use of appropriate design.

Extraneous Variables
Extraneous variables are those uncontrolled variables (i.e., variables not manipulated by the experimenter) that may have a significant influence on the results of a study. Many research conclusions are questionable because of the influence of these extraneous variables.

In a widely publicized study, the effectiveness of three methods of social studies teaching was compared. Intact classes were used, and the researchers were unable to randomize or control such variables as teacher competence or enthusiasm or the age, socioeconomic level, or academic ability of the student subjects. The criterion of effectiveness was achievement measured by scores on standardized tests. It would seem clear that the many extraneous variables precluded valid conclusions about the relative effectiveness of the independent variables, which were teaching methods. It should be noted that in order for an extraneous variable to confound the results of a study it must be correlated strongly enough with both the independent and dependent variables so that its influence can be mistaken for that of the independent variable.

Although it is impossible to eliminate all extraneous variables, particularly in classroom research, sound experimental design enables the researcher to largely neutralize their influence.

Controlling Extraneous Variables
Variables that are of interest to the researcher can be controlled by building them into the study as independent variables. For instance, a researcher comparing two different reading programs may wish to control for the potentially confounding extraneous variable of sex by making it an independent attribute variable and thereby investigating the effect of sex on the two different reading programs.

Variables not of direct interest to the researcher may be removed or their influence minimized by several methods, which are discussed in the following sections.

Removing the Variable
Variables may be controlled by eliminating them completely. Observer distraction may be removed by separating the observer from both experimental and control groups by a one-way glass partition. Some variables between subjects may be eliminated by selecting cases with uniform characteristics. Using only female subjects removes sex as a variable but thereby reduces the generalization from the study to only females.
Randomization

Randomization involves pure chance selection and assignment of subjects to experimental and control groups for a limited supply of available subjects. Random selection was discussed in Chapter 1. Here we are referring to random assignment, the method by which everyone already selected for the sample has an equal chance of being assigned to the various treatment conditions (e.g., experimental and control). If two groups are involved, randomization can be achieved by tossing a coin, assigning a subject to one group if heads appear, to the other if the toss is tails. With more than two groups, randomization can be achieved by tossing a coin or having a table of random numbers that can be used.

Randomization provides the most effective method of eliminating systematic bias and of minimizing the effect of extraneous variables. The principle is based on the assumption that through random assignment differences between groups result only from the operation of probability or chance. These differences are known as sampling error or error variance, and their magnitude can be established by the researcher.

In an experiment differences in the dependent variables that may be attributed to the effect of the independent variables are known as experimental variance. The significance of an experiment may be tested by comparing experimental variance with error variance. If at the conclusion of the experiment the differences between the experimental and control groups are too great to attribute to error variance, it may be assumed that these differences are attributable to experimental variance. This process is described in detail in Chapter 11.

Matching Cases

When randomization is not feasible (e.g., there are too few subjects), selecting pairs or sets of individuals with identical or nearly identical characteristics and assigning one of them to the experimental group and the other to the control group provides another method of control. This method is limited by the difficulty of matching on more than one variable. It is also likely that some individuals will be excluded from the experiment if a matching subject is not available. Matching is not considered satisfactory unless the members of the pairs or sets are then randomly assigned to the treatment groups, a method known as matched randomization.

Balancing Cases or Group Matching

Balancing cases consists of assigning subjects to experimental and control groups in such a way that the means and the variances of the groups are as nearly equal as possible. Because identical balancing of groups is impossible, the researcher must decide how much departure from equality can be tolerated without loss of satisfactory control. This method also presents a similar difficulty noted in the matching method, namely, the difficulty of equating groups on the basis of more than one characteristic or variable.

Analysis of Covariance (ANCOVA)

This method permits the experimenter to eliminate initial differences on several variables, including but not limited to the pretest, between the experimental and control groups by statistical methods. The use of pretest mean scores as covariates is considered preferable to the use of gain scores. Gain scores are problematic for two reasons. First, not all individuals have an equal chance for an equal gain. A person who scores very low on a pretest has a great deal more room for improvement than one who scores high. For instance, does a person who gains 50 points from a pretest of 20 to a posttest of 70 actually show more improvement than a person who gains just 30 points but goes from 70 to 100? ANCOVA takes care of this problem statistically. The second problem with gain scores is their lack of reliability.

Analysis of covariance is a rather complicated statistical procedure, beyond the scope of this elementary treatment. For a complete discussion, readers may wish to consult Glass and Hopkins (1986), Hays (1981), Kerlinger (1986), Kirk (1982), or Shavelson (1986).

Experimental Validity

To make a significant contribution to the development of knowledge, an experiment must be valid. Campbell and Stanley (1966) described two types of experimental validity, internal validity and external validity. Cook and Campbell (1979) further divided experimental validity, adding two other types, statistical validity and construct validity. For purposes of this introductory treatment of the issue, the discussion is confined to the two types of experimental validity described by Campbell and Stanley.

Internal Validity

An experiment has internal validity to the extent that the factors that have been manipulated (independent variables) actually have a genuine effect on the observed consequences (dependent variables) in the experimental setting.

External Validity

The researcher would achieve little of practical value if those observed variable relationships were valid only in the experimental setting and only for those participating. External validity is the extent to which the variable relationships can be generalized to other settings, other treatment variables, other measurement variables, and other populations.

Experimental validity is an ideal to aspire to, for it is unlikely that it can ever be completely achieved. Internal validity is very difficult to achieve in the nonlaboratory setting of the behavioral experiment in which there are so many extraneous variables to attempt to control. When experimental controls are tightened to achieve internal validity, the more artificial, less realistic situation may prevail, reducing the external validity or generalizability of the experiment. Some compromise is inevitable so that a reasonable balance may be established between control and generalizability between internal and external validity.
Threats to Internal Experimental Validity

In educational experiments, or in any behavioral experiments, a number of extraneous variables are present in the situation or are generated by the experimental design and procedures. These variables influence the results of the experiment and are always difficult to evaluate. In a sense, they introduce rival hypotheses that could account for experimental change not attributable to the experimental variables under consideration. Although these extraneous variables usually cannot be completely eliminated, many of them can be identified. It is important that behavioral researchers anticipate them and take all possible precautions to minimize their influence through sound experimental design and execution.

A number of factors jeopardize the power of the experimenter to evaluate the effects of independent variables unambiguously. Campbell and Stanley (1963) have discussed these factors in their excellent definitive treatment. They include the following:

Maturation
Subjects change (biologically and psychologically) in many ways over time, and these changes may be attributed to the effect of the independent variables under consideration. During the course of a study, the subjects might become more tired, wiser, hungrier, older, and so on. They may be influenced by the incidental learning or experiences encountered through normal maturation. This threat is best controlled by randomly assigning subjects to experimental and control groups. Differences between the groups are then considered to be due to the treatment rather than to maturation.

History
Specific external events occurring between the first and second measurements and beyond the control of the researcher may have a stimulating or disturbing effect on the performance of subjects. The effect of a fire drill, the emotional tirade of a teacher, a pop session, the anxiety produced by a pending examination, or a catastrophic event in the community may significantly affect the test performance of a group of students.

In many experiments these external events will have a similar effect on both experimental and control subjects, in which case this threat is controlled. However, because they are specific events, they may affect one group but not the other. The effect of these uncontrolled external events is one of the hazards inherent in experiments carried out outside the laboratory. In laboratory experiments these extraneous variables can be controlled more effectively.

Testing
The process of pretesting at the beginning of an experiment can produce a change in subjects. Pretesting may produce a practice effect making subjects more proficient in subsequent test performance. Testing presents a threat to internal validity that is common to pretest–posttest experiments. Of course, an equivalent control group would be affected by the test in a similar way as the experimental group. Thus, having experimental and control groups controls for this threat in the same way that it does for the threat of maturation. In the next section on threats to external validity, we will discuss another type of threat related to testing that is not controlled by having an equivalent control group, the interaction effect of testing.

Unstable Instrumentation
Unreliable instruments or techniques used to describe and measure aspects of behavior are threats to the validity of an experiment. If tests used as instruments of observation are not accurate or consistent, a serious element of error is introduced. If observers are used to describe behavior changes in subjects, changes in observers or in their standards due to fatigue, increased insight, or skill in judgment over a period of time—all these are likely to introduce error. The instability of measurement deals with the topic of test reliability and the measurement of observer reliability, which are discussed in Chapter 9.

Statistical Regression
Statistical regression, also known as regression to the mean, is a phenomenon that sometimes operates when subjects are selected on the basis of extremely high or extremely low pretest scores and when the measurement device is not totally reliable, a situation which is common. Subjects who score very high, near the ceiling, on a pretest will most likely score lower (nearer the mean) on a subsequent testing. Subjects who score very low, near the floor, on a pretest will most likely score higher (nearer the mean) on a subsequent testing, with or without anything pertinent to their performance (e.g., instruction) occurring in the meantime. The reader should be aware that this phenomenon occurs only when subjects are selected as a group because of their extreme scores and that the regression referred to is for the group as a whole, not for all individuals. Posttest scores for individuals may go in the opposite direction expected by this phenomenon for the group.

The purpose of a study may require the researcher to select subjects based on their extreme scores. A study of the effects of different remedial reading programs assumes that the subjects must need remedial reading instruction and, therefore, have very low reading scores on the pretest. To control for regression to the mean, the researcher would randomly assign his or her sample of poor readers to the experimental and control groups. Because both groups would be expected to improve equally because of regression to the mean, if the experimental group improved significantly more than the control group, the researcher could conclude that this was because of the experimental treatment rather than statistical regression.

Selection Bias
Selection bias is represented by the nonequivalence of experimental and control groups, and its most effective deterrent is the random assignment of subjects to treatments. Selection bias is likely when, on invitation, volunteers are used as members of an experimental group. Although they may appear to be equated to the
nonvolunteers, their characteristics of higher motivation may introduce a bias that would invalidate reasonable comparison. Selection bias may be introduced when intact classes are used as experimental and control groups: Because of scheduling arrangements, an English class meeting during the fourth period may consist of particularly able students who are scheduled at that period because they are all enrolled in an advanced mathematics class. Another example of this effect occurs when students and/or their parents have the opportunity to volunteer for a special program. The volunteers or their parents may be more motivated than the students who did not volunteer. Thus, the only good control group would be a random group of volunteers who were unable to participate because the program had more volunteers than could be enrolled. This control group would usually be on the waiting list and get the special program a later date.

Interaction of Selection and Maturation
This type of threat to the internal validity of a study is not the same as selection bias. The interaction of selection and maturation may occur whenever the subjects can select which treatment (e.g., which instructional method) they will receive. Even though the groups may be equivalent on the pretest and on other criteria such as measures, the reason some people choose one treatment over another may be related to the outcome to be measured (dependent variable). Thus, if more motivated students chose method A for learning calculus over method B because method A appears harder and requires greater academic motivation, that differential motivation might be confused for the effects of the experimental variable.

Experimental Mortality
Mortality, or loss of subjects, particularly likely in a long-term experiment, introduces a potentially confounding element. Although experimental and control groups are randomly assigned, the survivors might represent groups that are quite different from the unbiased groups that began the experiment. Those who survive a period of experimentation are likely to be healthier, more capable, or more highly motivated than those who are absent frequently or who drop out of school and do not remain for the duration of the experiment. The major concern here is whether the groups experienced different loss rates or reasons for dropouts that might confound the results. Usually a comparison of the pretest scores of those remaining in the study and those who dropped out will help determine if the dropout of the experimental and control groups was from different reasons and has resulted in significantly different groups than began the study.

Experimenter Bias
This type of bias is introduced when the researcher has some previous knowledge about the subjects in an experiment. This knowledge of subject status may cause the researcher to convey some clue that affects the subjects' reaction or may affect the objectivity of his or her judgment.

In medical research it is common practice to conceal from the subject the knowledge of who is receiving the placebo and who the experimental medication. This is known as a blind. Having someone other than the experimenter administer the treatments and record which subjects are receiving the medication and which the placebo provides an additional safeguard. This practice, known as a double blind, helps to minimize contamination.

Beginners in educational research have been known to contaminate a study by classifying student performance when they know the nature of the variable to be correlated with that performance. In a simple ex post facto study a student proposed to determine the relationship between academic achievement and citizenship grades in her class. Because she proposed to assign the citizenship grades herself, it would seem apparent that an element of contamination would result. Her knowledge of the student's previous academic achievement would tend to precondition her judgment in assigning citizenship grades.

In educational studies of this type, researchers would minimize contamination if outside observers rated the subjects without any knowledge of their academic status.

Threats to External Experimental Validity
Laboratory research has the virtue of permitting the experimenter to carefully avoid threats to internal validity. However, the artificial nature of such a setting greatly reduces the generalizability of the findings from such research. Because educational researchers are primarily concerned with the practical uses of their findings, they frequently conduct their studies in real classroom situations. Although these real-life situations present opportunities for greater generalization, they do not automatically result in externally valid research. Campbell and Stanley (1966) also discussed the factors that may lead to reduced generalizability of research to other settings, persons, variables, and measurement instruments. The factors they discussed include the following:

Interference of Prior Treatment
In some types of experiments the effect of one treatment may carry over to subsequent treatments. In an educational experiment learning produced by the first treatment is not completely erased, and its influence may accrue to the advantage, or disadvantage, of the second treatment. This is one of the major limitations of the single-group, equated-materials experimental design in which the same subjects serve as members of both control and experimental groups. If an equated-materials design is necessary, a counterbalanced design will generally control for this threat.

The Artificiality of the Experimental Setting
In an effort to control extraneous variables the researcher imposes careful controls that may introduce a sterile or artificial atmosphere not at all like the real life situation to which generalization is desired. The reactive effect of the experimental process is a constant threat. Even in a classroom setting, if the study brings increased resources that result in a better student/teacher ratio than is typical, this threat is present.
Interaction Effect of Testing
The use of a pretest at the beginning of a study may sensitize individuals by making them more aware of concealed purposes of the researcher and may serve as a stimulus to change. This is a different potential problem than that of testing, discussed earlier as a threat to internal validity.

With testing, the threat is that the pretest will affect the subjects' performance on the posttest in a direct fashion. That is easily controlled by having a control group. In the case of the interaction effect of testing, there is a more difficult problem. Here the pretest may alert the experimental group to some aspect of the interventions that is not present for the control group. That is, the pretest may interact differently with the experimental intervention than it does with the control or placebo conditions. To avoid this threat requires random assignment and either a pretest or the Solomon four-group design discussed in the next section, Experimental Design.

Interaction of Selection and Treatment
Educational researchers are rarely, if ever, able to randomly select samples from the wide population of interest or randomly assign to groups; consequently, generalization from samples to populations is hazardous. Samples used in multiple classroom experiments are usually composed of intact groups, not of randomly selected individuals. They are based on an accepted invitation to participate; others refuse. One cannot assume that samples taken from cooperating schools are necessarily representative of the target population, which includes schools that would not cooperate. Such schools are usually characterized by faculties that have higher morale, less insecurity, greater willingness to try a new approach, and a greater desire to improve their performance.

The Extent of Treatment Verification
Because of the potential threat of experimenter bias, most researchers have research assistants or others who are not directly involved in the formulation of the research hypotheses deliver the treatment. This leads to a potential threat to external validity. Was the treatment administered as intended and described by the researcher? The researcher must have a verification procedure (e.g., direct observation, video tape) to make sure that the treatment was properly administered.

After reading about these threats to experimental validity, the beginner is probably ready to conclude that behavioral research is too hazardous to attempt. Particularly outside of the laboratory, ideal experimental conditions and controls are never likely to prevail. However, an understanding of these threats is important so that the researcher can make every effort to remove or minimize their influence. If one were to wait for a research setting free from all threats, no research would ever be carried out. Knowing the limitations and doing the best that he or she can under the circumstances, the researcher may conduct experiments, reach valid conclusions, provide answers to important questions, and solve significant problems.

Experimental Design
Experimental design is the blueprint of the procedures that enable the researcher to test hypotheses by reaching valid conclusions about relationships between independent and dependent variables. Selection of a particular design is based on the purposes of the experiment, the type of variables to be manipulated, and the conditions or limiting factors under which it is conducted. The design deals with such practical problems as how subjects are to be assigned to experimental and control groups, the way variables are to be manipulated and controlled, the way extraneous variables are to be controlled, how observations are to be made, and the type of statistical analysis to be employed in interpreting data relationships.

The adequacy of experimental designs is judged by the degree to which they eliminate or minimize threats to experimental validity. Three categories are presented here:

1. Pre-experimental design is the least effective, for it provides either no control group or no way of equating the groups that are used.
2. True experimental design employs randomization to provide for control of the equivalence of groups and exposure to treatment.
3. Quasi-experimental design provides a less satisfactory degree of control, used only when randomization is not feasible.

A complete discussion of experimental design would be too lengthy and complex for this introductory treatment. Therefore, only a relatively few designs are described. Readers may wish to refer to Campbell and Stanley's (1966) and Cook and Campbell's (1979) excellent treatments of the subject, in which many more designs are described.

In discussing experimental designs, we have followed Campbell and Stanley's symbol system:

- \( R \) random assignment of subjects to groups or treatments
- \( X \) exposure of a group to an experimental (treatment) variable
- \( O \) observation or test administered
- \( C \) exposure of a group to the control or placebo condition

Pre-Experimental Designs
The least adequate of designs is characterized by (1) the lack of a control group, or (2) a failure to provide for the equivalence of a control group.

The One-Shot Case Study

\( X \)
Carefully studied results of a treatment are compared with a general expectation of what would have happened if the treatment had not been applied. This difference provides the weakest basis for generalization.

Mr. Jones used a 25-minute film on racial integration in his junior high school history class. In a test administered after the showing of the film, the mean score was 86 (a high score indicated a favorable attitude toward acceptance of all racial groups). Mr. Jones believes that the mean score was higher than it would have been had the film not been viewed and, as he recalls, higher than the mean score of a test that he had administered to a similar class several years before. He concludes that the film has been effective in reducing racial prejudice.

However, Mr. Jones has come to this conclusion on the basis of inadequate data. The reader has no way of knowing if a change has occurred because of the lack of a pretest or if a similar group who had not seen the film (a control group) would have scored differently from the group viewing the film. This design is the poorest available and should not be used.

The One-Group, Pretest-Posttest Design

\[ O_1 \times O_2 \]

\[ O_1 = \text{pretest} \quad O_2 = \text{posttest} \]

This design provides some improvement over the first, for the effects of the treatment are judged by the difference between the pretest and the posttest scores. However, no comparison with a control group is provided.

In the same setting Mr. Jones administered a pretest before showing the film and a posttest after the viewing. He computed the mean difference between the pretest and the posttest scores and found that the mean had increased from 52 to 80, a gain of 28 score points. He also apparently detected some temporary improvement in attitude toward racial integration. He concludes that there has been a significant improvement in attitude as a result of the students' viewing the film. But what about the sensitizing effect of the pretest items that may have made the students aware of issues that they had not even thought of before? What would the gain have been if the pretest and the posttest had been administered to another class that had not viewed the film? Threats to the internal validity that are not controlled include history, maturation, testing, and so forth. External validity is also poor.

The Static-Group Comparison Design

\[ X \quad O \]

\[ C \]

This design compares the status of a group that has received an experimental treatment with one that has not. There is no provision for establishing the equivalence of the experimental and control groups, a very serious limitation.

A beginning researcher administered the 25-minute racial integration film to a group of elementary teachers in one school. He then administered the attitude scale and computed the mean score. At another elementary school he administered the attitude scale to teachers who had not viewed the film. A comparison of mean scores showed that the teachers who had viewed the film had a higher mean score than those who had not. He concluded that the film was an effective device in reducing racial prejudice.

What evidence did he have that the initial attitudes of the groups were equivalent? Without some evidence of equivalence of the control and experimental groups, attributing the difference to the experimental variable is unwarranted.

Campbell and Stanley (1966) provide a table that quickly describes which threats to the internal and external validity of a study are controlled by each of the pre-experimental and true experimental designs. Table 6.1 presents Campbell and Stanley's Table 1. They refer to two threats to external validity differently in this table than is done in the text. Campbell and Stanley's "reactive arrangements" is the artificiality of experimental setting and their "multiple X [X means treatment] interference" is the interference of prior treatments. They also shorten the names of some threats in the list of internal threats.

True Experimental Designs

In a true experiment the equivalence of the experimental and control groups is provided by random assignment of subjects to experimental and control treatments. Although it is difficult to arrange a true experimental design, particularly in school classroom research, it is the strongest type of design and should be used whenever possible. Three experimental designs are discussed in the following sections:

The Posttest-Only, Equivalent-Groups Design

\[ R \times O_1 \]

\[ R \times C \times O_2 \]

This design is one of the most effective in minimizing the threats to experimental validity. It differs from the static group comparison design in that experimental and control groups are equated by random assignment. At the conclusion of the experimental period the difference between the mean posttest scores of the experimental and control groups is subjected to a test of statistical significance, usually a t test or an analysis of variance. The assumption is that the means of randomly assigned experimental and control groups from the same population will differ only to the extent that random sample means from the same population will differ as a result of sampling error. If the difference between the means is too great to attribute to sampling error, the difference may be attributed to the treatment variable effect.

The researcher randomly selects 80 students from a school population of 450 sophomores. The 80 students are randomly assigned to experimental and control treatments, using 40 as the experimental group and 40 as the control group. The experimental group is taught the concepts of congruence of triangles by an experimental procedure method X, and the control group is taught the same set of
TABLE 6.1 Summary Table from Campbell & Stanley (1966)

<table>
<thead>
<tr>
<th>Sources of Invalidity</th>
<th>Internal</th>
<th>External</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>History</td>
<td>Maturation</td>
</tr>
<tr>
<td>Pre-Experimental Designs:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. One-Shot Case Study</td>
<td>X O</td>
<td>- - -</td>
</tr>
<tr>
<td>2. One-Group Pretest-Posttest Design</td>
<td>O X O</td>
<td>- - -</td>
</tr>
<tr>
<td>3. Static-Group Comparison</td>
<td>X O</td>
<td>+ ? + + +</td>
</tr>
<tr>
<td>True Experimental Designs:</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Pretest-Posttest Control Group Design</td>
<td>R O X O</td>
<td>+ + + + + + + + +</td>
</tr>
<tr>
<td>5. Solomon Four-Group Design</td>
<td>R O X O</td>
<td>+ + + + + + + + +</td>
</tr>
<tr>
<td>6. Posttest-Only Control Group Design</td>
<td>R O X O</td>
<td>+ + + + + + + + +</td>
</tr>
</tbody>
</table>

Note: In the table, a minus indicates a definite weakness, a plus indicates that the factor is controlled, a question mark indicates a possible source of concern, and a blank indicates that the factor is not relevant. It is with extreme reluctance that these summary tables are presented because they are apt to be "too helpful," and to be depended upon in place of the more complex and qualified presentation in the text. No + or - indicator should be respected unless the reader comprehends why it is placed there. In particular, it is against the spirit of this presentation to create uncomprehended fears of, or confidence in, specific designs.

Concepts by the usual method, method C. Time of day, treatment length in time, and other factors are equated. At the end of a 3-week period the experimental and control groups are administered a test, and the difference between mean scores is subjected to a test of statistical significance. The difference between mean scores is found to favor the experimental group but not by an amount that is statistically significant. The researcher rightly concludes that the superiority of the X group could well have been the result of sampling error and that there was no evidence of the superiority of the X method. If, on the other hand, the finding was that the difference between the means was sufficient to be statistically significant, the researcher would have rejected the null hypothesis and stated that her research hypothesis was supported.

The Pretest-Posttest Equivalent-Groups Design

\[ R \ O_1 \times \ O_2 \quad O_1 \ O_3 = \text{pretests} \]
\[ R \ O_3 \ C \ O_4 \quad O_2 \ O_4 = \text{posttests} \]

This design is similar to the previously described design, except that pretests are administered before the application of the experimental and control treatments and posttests at the end of the treatment period. Pretest scores can be used in analysis of covariance to statistically control for any differences between the groups at the beginning of the study. This is a strong design, but there may be a possibility of the influence of the interaction effect of testing with the experimental variable.

Laney (1963) conducted a pretest-posttest true experiment to study the relative impact of two interventions on first graders' understanding of economic concepts. He randomly assigned 31 first-grade students to two groups. Both groups set up and operated a market economy in their classrooms, but one group received economic instruction (referred to as "debriefing") after each market day, and the other group did not. The purpose of the study was to determine what misconceptions first-grade students have about economic concepts and to determine if the debriefing sessions when added to the market experience, were superior to the experience alone, as the theory being studied suggested. The results clearly indicated that the experience with instruction was superior to the experience-only condition. Laney added a third group the following year that did not get the market experience but did get the instruction (debriefing). This group was added to study whether the instruction was the sole critical condition for the learning of economics rather than experience being critical, too. This group was found to be superior to the experience-only group, but the experience and instruction group was found to have a more advanced understanding of economics than the instruction-only control group. Thus, economic concepts can be learned best by first graders by combining experience with instruction. The complete text of this study appears at the end of this chapter as the sample article.

Watanabe, Hare, and Lomax (1984) also reported on a study that included a pretest-posttest equivalent-groups design. This study compared a procedure for
teaching eighth-grade students to be better able to predict the content of newspaper stories from their headlines than a control group of eighth-grade students. In a pilot study, reported in their article, indicated that even good middle school readers have difficulty predicting the content of news stories from the headlines. The authors felt that college students have no trouble with this task. Because the eighth graders they surveyed reported reading primarily comics, movie, and sport sections (a fact which might explain their poor prediction of content from headlines) and because most teachers would prefer that their students read more of the newspaper, the authors felt that it would be useful to determine if a training program could teach eighth graders how to better understand headlines.

Watanabe et al. randomly assigned 46 eighth graders to either headline reading instruction (experimental group) or regular reading instruction (control group). The 46 students were asked to read 20 headlines and predict story content prior to, and after, a 3-week period of instruction. The authors scored each attempt to predict story content on a scale of 0 to 4, with 0 indicating that the students’ response explained nothing and 4 indicating an “on-target potential prediction” (pp. 439–440). The each student could receive a score from 0 to 80 on each of the testings.

At the end of the 3 weeks of instruction, the authors compared the two groups using analysis of covariance (ANCOVA) and found that the experimental group was better able to predict story content from headlines after training than the control group. ANCOVA was used because even with random assignment the groups were not exactly equal. ANCOVA permitted the authors to statistically control for differences on the pretest so that posttest differences would not be due to initial differences before training.

The reader may wonder why these authors did not just subtract the pretest scores from the posttest scores and compare these gain scores in a simpler statistical analysis. There are two reasons why good research does not use gain scores. First, an analysis comparing gain scores is not as reliable as an analysis comparing posttest scores. Second, and perhaps more apparent, is that not everyone has an equal chance to show gains on the posttest. For instance, a person who scores very low on the pretest has a great deal more room to improve on the posttest than a person who scores near the ceiling on the pretest (e.g., 95 out of a possible 100). It is much harder to gain 15 points from 40 to 55 or to go from 95 to 100 (a perfect score).

The Solomon Four-Group Design

| R | O₁ \( X \) O₂ |
| R | O₃ \( C \) O₄ |

In this design

1. Subjects are randomly assigned to four groups.
2. Two groups receive the experimental treatment (X).
3. One experimental group receives a pretest (O₁).
4. Two groups (control) do not receive treatment (O₂).
5. One control group receives a pretest (O₃).
6. All four groups receive posttests (O₄, O₅, O₆, O₇).

The design is really a combination of the two two-group designs previously described, the posttest only and the pretest–posttest. The Solomon Four-Group Design permits the evaluation of the effects of testing, history, and maturation. Analysis of variance is used to compare the four posttest scores; analysis of covariance may be used to compare changes in O₂ and O₄.

Because this design provides for two simultaneous experiments, the advantages of a replication are incorporated. A major difficulty is finding enough subjects to assign randomly to four equivalent groups.

Quasi-Experimental Designs

These designs provide control of when and to whom the measurement is applied, but because random assignment to experimental and control treatments has not been applied, the equivalence of the groups is not assured. Of the many quasi-experimental designs, only five are described. See Cook and Campbell (1979) for a comprehensive review of quasi-experimental designs. Campbell and Stanley (1966) again provide a table that quickly describes which threats to the internal and external validity of a study are controlled by each of the quasi-experimental designs described. Table 6.2 presents Campbell and Stanley’s Table 2.

The Pretest–Posttest Nonequivalent-Groups Design

\[
O₁ \times O₂ \quad O₃ \times O₄
\]

This design is often used in classroom experiments when experimental and control groups are not naturally assembled groups as intact classes, which may be similar. As in the pretest–posttest equivalent group design, analysis of covariance may be used with the pretest as the covariate. Because this design may be the only feasible one, the comparison is justifiable, but, as in all quasi-experimental studies, the results should be interpreted cautiously.

Two first-grade classes in a school were selected for an experiment. One group was taught by the initial teaching alphabet (ITA) approach to reading, and the other was taught by the traditional alphabet approach. Prior to the introduction of the two reading methods and again at the end of the school year, both groups were administered a standardized reading test, and the mean gain scores of the two groups were compared. The ITA group showed a significant superiority in test scores over the conventional alphabet group. However, without some evidence of the equivalence of the groups in intelligence, maturity, readiness, and other factors at the beginning of the experimental period, conclusions should be cautiously interpreted.
TABLE 6.2 Summary Table from Campbell & Stanley (1966)

<table>
<thead>
<tr>
<th>Sources of Invalidity for Quasi-Experimental Designs 7–12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Internal</td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td>History</td>
</tr>
<tr>
<td>Quasi-Experimental Designs</td>
</tr>
<tr>
<td>7. Time Series</td>
</tr>
<tr>
<td>O O O X O O O</td>
</tr>
<tr>
<td>8. Equivalent Time</td>
</tr>
<tr>
<td>Samples Design</td>
</tr>
<tr>
<td>X O X O X O X O, etc.</td>
</tr>
<tr>
<td>9. Equivalent Materials</td>
</tr>
<tr>
<td>Samples Design</td>
</tr>
<tr>
<td>M X O M X O M X O, etc.</td>
</tr>
<tr>
<td>11. Counterbalanced-</td>
</tr>
<tr>
<td>Design</td>
</tr>
<tr>
<td>X O X O X O X O</td>
</tr>
<tr>
<td>X O X O X O X O</td>
</tr>
<tr>
<td>X O X O X O X O</td>
</tr>
</tbody>
</table>

The Follow-Through Planned Variation Study
An interesting example of the pretest-posttest nonequivalent groups design was the Follow-Through Planned Variation Study (Abi Associates, 1977), conceived in the late 1960s and initiated and funded by the United States Office of Education. The purpose of the program was to implement and evaluate a variety of compensatory programs extending the services of Project Head Start for disadvantaged children into the primary grades. Head Start was a large-scale enterprise including many innovative instructional models and involving the expenditure of more than a half billion dollars. The study extended over a period of more than 9 years, with more than 75,000 first-, second-, and third-grade children participating. Of the 20 different instructional models and 170 projects, 17 models and 70 projects were selected for evaluation. Approximately 2% of the total number of children in the program were included in the study.

Participation by school districts was voluntary, with each district selecting the particular model that it wished to implement and helping to choose the groups that were to be used as controls. Treatments were not randomly assigned, nor control groups randomly selected.

The unit of analysis was pupil mean gain for groups K–3 and 1–3 growth scores, statistically compared by instructional model and by project, using variants of linear regression and analysis of covariance. Outcome measures were derived from gain scores on the following measuring instruments:

1. The Metropolitan Achievement Test Battery covering such basic skills as reading comprehension, spelling, word usage and analysis, and mathematical computation, concepts, and problem solving.
2. The Raven's Coloured Matrices Test, a nonverbal test of problem-solving ability, requiring the manipulation of geometric patterns, essentially a measure of intelligence rather than a measure of learning outcomes.
3. The Coopersmith Self-Image Inventory, a measure of self-esteem but questioned on the grounds that it required a maturity of judgment beyond the competence of primary-age children.
4. The Intellectual Achievement Responsibility Scale, which attempts to assess the child's experience of success or failure, indicating the degree to which the child attributes success to internal or external causes. This instrument was also judged to require insights beyond the maturity level of small children.

There have been many analyses and evaluations of the program by official and independent agencies funded by the United States Office of Education and by private philanthropic foundations. Among the evaluating agencies were the Office of Education; Abi Associates, Inc.; The Stanford Research Institute; The Huron Institute; and The Center for Research and Curriculum Evaluation of the University of Illinois.

It is unlikely that any large-scale study has been scrutinized so extensively concerning research design, procedures employed, and interpretation of the data. There have been critiques of the evaluations and critiques of the critiques, with sharp disagreement on most aspects of the study (Anderson, St. Pierre, Proper, & Stebbins, 1978; House, Glass, McLean, & Walker, 1978; Wisler, Burns, & Iwamoto, 1978).

However, the consensus is that the findings were disappointing because most of the experimental effects were negligible. Only a few of the treatment effects produced as much as a one-quarter standard deviation change. (This concept is discussed in Chapter 10.) Of those that met this criterion, two instructional models with at least one positive effect were structured approaches. Three models with at least one negative effect were unstructured approaches. Few of either cognitive, structured approaches or child-centered, unstructured approaches yielded significant effects.

Much of the disagreement centered around the reasons the study was ineffective. Several explanations have been suggested.
1. The research was deficient in design, implementation, statistical analysis, and interpretation. Because experimental treatments were not randomly selected, and control groups were not randomly assigned, mismatching resulted, and comparisons were really made between different populations.

2. There was great intra-site difference in effectiveness within a given instruction model. Most of the within-model differences were greater than the between-models difference. There may have been serious deficiencies in the competence of those who implemented the innovative procedures or in the actual method implemented, even though the teachers and their teacher aides were specially trained and their activities monitored by the project sponsors.

3. The measuring instruments may have been incompatible with the goals of the project because of inadequate identification and definition of appropriate outcome variables. The more effective instruments seemed to focus on basic skill or traditional educational goals rather than on goals ordinarily associated with unstructured approaches to education. Some measured intellectual status rather than achievable learning goals. Others appeared to require a majority of response too complex for primary-age children.

Not all reactions to the study were negative. Hodges, a member of the Follow Through Task Force, listed a number of reasons for viewing the program as significant and worthwhile. "Just because Follow Through has not proved to be an easy, workable, inexpensive solution for all the educational problems of poor children does not mean it should be dismissed as just another failure in compensatory education" (Hodges, 1978, p. 191).

A more recent study examined the relative effects of age and schooling on the rapid developmental changes that generally occur between 5 and 7 years of age (Morrison, Smith, & Dow-Elversberger, 1995). This phenomenon of a major shift in the way children think, cognitively, socially, and even morally, is generally accepted and usually has been thought to be due to maturation. This view is based on Piaget's theory of cognitive development. However, in recent years, research has suggested that the experiences children have in school may play a major role in these changes. Because children are arbitrarily assigned to start school in a given year depending on their birth date (e.g., must be 6 by March 1, December 1, or September 1, depending on the district/state, to enter first grade), it is possible to keep age relatively stable and study only the schooling effect by using children who just made the cutoff date to enter school or who just missed the date. This is what Morrison, Smith, and Dow-Elversberger did. The cutoff date in the community they studied was January 1, or March 1, to enter first grade. Thus, they compared children who turned 6 in January or February and entered 1st grade (young first graders) with those who turned 6 in March or April and entered kindergarten instead (old kindergarten students). The first-grade group of 10 children were, on average, 41 days older than the 10 children in the kindergarten group. The children were tested three times on measures of short-term memory and phonological skill at the beginning of the study in grade 1 or kindergarten (depending on the age group they were in), at the beginning of first or second grade, and at the beginning of second or third grade, again depending on their group. The findings of this study indicated a strong schooling effect in particular for short-term memory. It seems that grade 1 experiences enhanced short-term memory regardless of age in first grade. The phonological measure was affected by both schooling and age. Thus, this design permitted the researchers to support maturational theories (e.g., Piaget) but also demonstrate that experiences in school play an important role in the changes that seem to occur between the ages of 5 and 7.

Most important for readers of this text, this study had to be quasi-experimental. Although a better design would have taken children born between January 1 and April 30 and randomly assigned them to kindergarten or first grade, this was not possible for two reasons. First, the school district would not go along with such random assignment because it would change a basic policy and might have led to parents being very upset and possibly suing the district. Second, random assignment would be unethical because of the unknown effects of having older children in kindergarten and younger children in first grade than would be typical for this district.

In behavioral research the random selection and assignment of subjects to experimental and control groups may be impracticable. Because of administrative difficulties in arranging school experiments, it may be necessary to use the same group as both the experimental and control group. These designs have two apparently attractive features. Three of these designs (time-series, equivalent time-samples, and equivalent materials) are described below. They can be carried out with one intact group without a noticeable reorganization of the classroom schedule. The changes in procedures and testing can be concealed within ordinary classroom routines. Artificiality can be minimized, for the procedures can be introduced without the subjects' awareness of participation in an experiment.

The Time-Series Design

At periodic intervals observations (measurements) are applied to individuals or a group. An experimental variable (X) is introduced, and its effect may be judged by the change or gain from the measurement immediately before to the one just after its introduction. The purpose of the series of measurements before and after the intervention or treatment is to demonstrate little or no change except from immediately before to just after the intervention.

In the time-series experimental design, a measured change or gain from observation 4 (O₄) to observation 5 (O₅) would indicate that the treatment had an effect. This design is particularly sensitive to the failure to control the extraneous variable, history, for it is possible that some distracting, simultaneous event at the time of the intervention would provide a rival hypothesis for the change.

\[ O₁, O₂, O₃, O₄, X, O₅, O₆, O₇, O₈ \]

The diagram showing one X and several Os does not necessarily represent the relative number of sessions for each. It may be that each O represents one measurement, and the single X represents an intervention of several weeks. Although it is better to have several observations, as shown, it is not always possible to have this many. For instance, a time-series experiment by a student of
the second author used only two pre-intervention and two post-intervention measures. Because this study was measuring the effect of a program to reduce the number of criminal victimizations of students with disabilities, it was necessary to have a 2-month period between measurements in order to have a sufficient number of victimizations for each period measured. That is, O₁ in November measured September and October crimes against the subjects; O₂ in January measured crimes in November and December, and so on for O₃ and O₄. In this study, the intervention/instruction occurred in January, followed two and four months later with measures of victimizations. The findings showed no changes between O₁ and O₂ or between O₃ and O₄ but a significant reduction in victimizations from O₂ to O₃. Thus, the intervention/instruction appeared successful in reducing crimes committed against persons with disabilities.

**The Equivalent Time-Samples Design**

Instead of using equivalent samples of persons, it may be necessary to use one group as the experimental and control group. In this design the experimental condition X is present between some observations and not (C) between others. This may be diagrammed as shown below, although the number of observations and interventions vary and the alternation of the experimental condition with the control condition is normally random rather than systematic as shown here. The design is the group design that is analogous to the A-B-A-B design for single-subject research discussed in Chapter 7.

\[
O₁ \rightarrow X₁ \rightarrow O₂ \rightarrow C \rightarrow O₃ \rightarrow X₂ \rightarrow O₄ \rightarrow C \rightarrow O₅
\]

A study by Hall et al. (1973) provides an excellent illustration of the equivalent time-samples design. Five subjects, identified as the most violently aggressive, were selected from a group of 46 boys with mental retardation living in an institution dormitory. Their ages ranged from 12 to 16 (mean = 13.8), and their IQs from 40 to 62 (mean = 50). Each individual was observed for 10 weeks in 10 randomly selected 3-minute periods during which time acts of aggressive behavior were recorded. Acts were classified as motor aggressive (throwing objects, kicking, fighting, scratching) and nonmotor aggressive (verbal abuse, screaming or shouting, insubordination).

The observations were scheduled in four periods:

1. Observation (baseline) - sessions 1
2. On-reinforcement - sessions 2, 3, 4, 5
3. Off-reinforcement - sessions 6, 7
4. On-reinforcement - sessions 8, 9, 10

Positive reinforcement as a reward for nonaggressive behavior consisted of candy, praise, or trips to the canteen. Punishment following aggressive acts consisted of ostracizing from group activities, taking away a favorite toy, or reprimanding verbally. Two observers were employed, one observing motor aggressive acts, the other, nonmotor aggressive acts.

The researchers concluded that reinforcement affected the amount of aggressive output. Motor aggressive behavior was reduced more effectively than nonmotor aggressive behavior (see Figure 6.1). To assess the permanence of behavior change after the conclusion of the experiment, a phase-out period of 89 days of observation was scheduled. The only reinforcement used was the posting of stars for nonaggressive behavior. Observations during the phase-out period indicated much more acceptable dormitory behavior.

Designs of this type have a number of limitations. Although they may minimize the effect of history, it is possible that they may increase the influence of maturation, unstable instrumentation, testing, and experimental mortality.

![FIGURE 6.1 Number of Motor Aggressive, Nonmotor Aggressive, and Total Aggressive Acts during On-Reinforcement and Off-Reinforcement Experimental Conditions](image)
The Equivalent Materials, Pretest, Posttest Design

\[ O_1 \ X_{MA} \ O_2 \ O_3 \ X_{MB} \ O_4 \]

\[ X_{MA} = \text{teaching method A} \quad X_{MB} = \text{teaching method B} \]

\[ O_2 \ and \ O_3 \ are \ pretests \quad O_3 \ and \ O_4 \ are \ posttests \]

Another experimental design, using the same group or class for both experimental and control groups, involves two or more cycles. The class may be used as a control group in the first cycle and as an experimental group in the second. The order of exposure to experimental and control can be reversed—experimental first and control following.

Essential to this design is the selection of learning materials different but nearly equal as possible in interest to the students and in difficulty of comprehension. An example may help to clarify the procedure.

Ms. Smith hypothesized that the students in her class who were used to background music while doing their homework would learn to spell more efficiently in the classroom if music were provided. Because she was unable to arrange a parallel group experiment, she decided to use her class as both an experimental and a control group.

To equate the words to be learned, she randomly selected 200 words from an appropriate graded word list and randomly assigned 100 words each to list A and list B. For cycle I, the control cycle, she pretested the class on word list A. Then for 20 minutes each day the students studied the words, using drill and the usual spelling rules. At the end of 2 weeks she retested the class and computed the mean gain score in correct spelling.

For cycle II, the experimental cycle, she pretested the class on word list B. Then for 20 minutes each day, with soft, continuous music in the background (the experimental condition), the students studied their word list, using the same drill and spelling rules. At the end of the second 2-week period she retested the class and computed the mean gain score in correct spelling.

The mean gain score for the experimental cycle was significantly greater than the mean gain score for the control cycle. She concluded that the introduction of the experimental variable had indeed improved the effectiveness of the learning experience.

The apparent simplicity and logic of this design are somewhat misleading and when examined in light of the threats of experimental validity, the design's weaknesses become apparent:

1. It is often difficult to select equated materials to be learned. For types of learning other than spelling, finding learning materials that are equally interesting, difficult, and unfamiliar would be a serious problem.
2. As the students enter the second cycle, they are older and more mature (if each instructional method is brief this may not be a problem). They also have more experience.

3. Outside events (history) would be more likely to affect the experience in one cycle than in the other.
4. There would be an influence of prior treatment carrying over from the first cycle to the second.
5. The effects of testing would be more likely to have a greater impact on the measurement of gain in the second cycle.
6. Mortality, or loss of subjects from the experiment, would be more likely in an experimental design spread over a longer period of time.
7. If the experimenter's judgment is a factor in assessment of the students' progress, contamination, the experimenter's knowledge of subject performance in the first cycle could possibly influence evaluation of performance in the second cycle.

Some of the limitations of the equivalent-materials, single-group, pretest-posttest design can be partially minimized by a series of replications in which the order of exposure to experimental and control treatments is reversed. This process, known as rotation, is illustrated by this pattern in a four-cycle experiment.

<table>
<thead>
<tr>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>O_1</td>
<td>X</td>
<td>O_2</td>
<td>C</td>
</tr>
<tr>
<td>O_3</td>
<td>C</td>
<td>O_4</td>
<td>X</td>
</tr>
<tr>
<td>O_5</td>
<td>C</td>
<td>O_6</td>
<td>O_7</td>
</tr>
<tr>
<td>O_8</td>
<td>X</td>
<td>O_9</td>
<td>O_8</td>
</tr>
</tbody>
</table>

If the experimental treatment yielded significantly greater gains regardless of the order of exposure, its effectiveness could be accepted with greater confidence. However, it is apparent that this design is not likely to equate materials, subjects, or experimental conditions.

All single-group experimental designs are sensitive to the influences of many of the threats to validity previously mentioned in this chapter: history, maturation, unstable instrumentation, testing, and experimental mortality. Replication of the studies, using different units as subjects, is an effective way to improve their validity. However, single-group experiments may be performed when randomly equated group designs cannot be arranged.

Counterbalanced Designs

These are designs in which experimental control derives from having all the subjects receive all the treatment conditions. The subjects are placed into, in the case of this example, four groups. Each of the groups then receives all four treatments but in different orders. This may be diagrammed as follows:

<table>
<thead>
<tr>
<th>Replication</th>
<th>O_1X_1</th>
<th>O_2X_2</th>
<th>O_3X_3</th>
<th>O_4X_4</th>
<th>O_5X_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Group A</td>
<td>B</td>
<td>C</td>
<td>D</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Group B</td>
<td>D</td>
<td>A</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Group C</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Group D</td>
<td>C</td>
<td>B</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>
In the first sequence following a pretest (O1), group A receives Treatment 1, group B receives Treatment 2, group C receives Treatment 3, and group D receives Treatment 4. After a second test (O2), each group then receives a second treatment, and so on. Thus, each group receives all treatments, and each treatment is first, second, third, or fourth in the order received by one of the groups.

This design has excellent internal validity because history, maturation, regression, selection, and mortality are all generally well controlled. The major limitation is that an order effect could wipe out any potential differences among the treatments. Four randomly assigned groups in which each group receives a different treatment would therefore be preferable. Thus, this design should be used when random assignment is not possible and when it is expected that different treatments will not interfere too much with each other. This design is particularly useful when the researcher uses preassigned groups (i.e., preexisting classes) and the number of students is a multiplicand of the number of treatments (e.g., treatments with 4, 8, 12, or 16 groups).

**Factorial Designs**

When more than one independent variable is included in a study, whether a true experiment or a quasi-experiment, a factorial design is necessary. Because most real-world outcomes are the result of a number of factors acting in combination, most significant experimentation involves the analysis of the interaction of a number of variable relationships. By using factorial designs, researchers can determine, for example, if the treatment interacts significantly with gender or age. That is, the experimenter can determine if one treatment is more effective with boys and another with girls, or if older girls do better on the treatment than younger girls, whereas older and younger boys do equally well on the treatment.

The simplest case of a factorial design would be to have two independent variables with two conditions of each, known as a 2 x 2 factorial design. This design would be used if a researcher decided to compare a new (experimental) method of teaching reading to reading-disabled children with a commonly used (control) method and also wanted to determine if boys and girls would do differently on the two methods. Such a design would look like Figure 6.2.

With this design we have four cells, each of which represents a subgroup (experimental females, control females, experimental males, and control males). This design permits the researcher to determine if there is a significant overall effect, known as main effect, for treatment and/or gender. It also permits the determination of whether these two variables interact significantly such that boys do best in the experimental condition and girls do best in the control condition. If this were the case, the subjects in Cell 2 would have a higher average score than those in Cell 1, and the subjects in Cell 3 would outperform those in Cell 4.

The study by Morrison/Smith, and Dow-Ehersberger (1995) that we described earlier in the section on pretest-posttest nonequivalent groups/designs included a factorial design. In their study the first independent variable, described earlier, was age, with the conditions being young first graders or old kindergarten students determined by the birth date cutoff for admission to first grade. The other independent variable was experience in school, determined by the grades at which they were tested (e.g., the first test was at the beginning of kindergarten for one group but at the beginning of first grade for the second group). Thus, this study was a 2 x 3 design because there were two groups and three levels of experience, with the children being tested three times at one-year intervals. These authors found one significant interaction for the phonological measure (graphically represented in Figure 6.3). The interaction effect indicated that young first-grade children at the beginning of second grade were superior in phonetic skills to the old kindergarten children at the beginning of first grade (the second test for each group) but that the two groups did not differ at the initial testing (beginning of first grade and beginning of kindergarten, respectively) or at the final testing (at the beginning of second grade and beginning of first grade, respectively). This finding along with the earlier reported finding of enhanced short-term memory from schooling was interpreted as indicating that schooling adds to the general maturational effect that has been found in children going from 5 to 7 years old. Thus, the belief that this change occurs only because of the so-called laws of cognitive development is no longer supported. Schooling appears to add significantly to this effect. In fact, for phonological skills, Figure 6.3 suggests that the experience of first grade is the critical component. As can be seen in Figure 6.3, it was the first-grade experience that made the difference in the growth curve. Both the young first graders and old kindergarten students showed significant growth during first grade (Post 1 for the first graders and Post 2 for the kindergarten students) and limited or no growth at the other times measured, kindergarten for one group and second grade for the other group.

Another factorial study by Nucci and Nucci (1982) examined the responses of children to the social transgressions (such as spitting on the ground) of their peers. They observed boys and girls between 7 and 10 and between 11 and 14 years of age and coded their observations of the responses into one of eight categories. They
found an interaction effect of gender by age for just one of the categories. This interaction effect could be graphically represented as in Figure 6.4. As can be seen, the two lines actually cross, thus clearly indicating that "With increased age the girls provided greater frequencies of ridicule responses to [social transgressions] while the boys responded with approximately the same frequencies as at the younger age" (Nucci & Nucci, 1982, p. 1341). Figure 6.5 shows an example of another type of response, stating the rule being violated, for which Nucci and Nucci found no interaction effect. Here there are two relatively parallel lines.

Of course, factorial designs can have more than two independent variables and more than two conditions of each variable. A study might have three treatment conditions (e.g., three methods of reading instruction), the two genders, three age groups, and three intelligence levels (gifted, average, and mildly retarded) as the independent variables. This would be a $3 \times 2 \times 3 \times 3$ design and would have a total of 54 subgroups or cells. Such designs are too complex for this elementary treatment. We mention such a complex design only to point out that these designs exist and that they are frequently appropriate and necessary. Advanced students may wish to refer to such sources as Glass and Hopkins (1996), Kirk (1995), and Winer (1971) for more detailed information.
This discussion, which has examined the many limitations of the experimental method in behavioral research, may convey a sense of futility. As is true in many other areas of significant human endeavor, researchers do not work under ideal conditions. They must do the best they can under existing circumstances. They will find, however, that in spite of its limitations, the well-designed and well-executed experiment provides a legitimate method for testing hypotheses and making probability decisions about the relationships between variables.

Some variables cannot be manipulated. The ethical problems that would be raised if some others were manipulated indicates a place for such nonexperimental methods as ex post facto research. The researcher starts with the observation of dependent variables and goes back to the observation of independent variables that have previously occurred under uncontrolled conditions. Such studies are not experiments, for the researcher has had no control over the events; they occurred before he or she began the investigation. The description of cigarette-smoking cancer research in Chapter 3 is an example of ex post facto research. Qualitative research methods, discussed in Chapter 8, provide other ways of conducting research into areas that would be difficult or impossible to investigate with experimental methods.

Summary

The experimental method provides a logical, systematic way to answer the question, "If this is done under carefully controlled conditions, what will happen?" To provide a precise answer, experimenters manipulate certain influences, or variables, and observe how the condition or behavior of the subject is affected or changed. Experimenters control or isolate the variables in such a way that they can be reasonably sure that the effects they observe can be attributed to the variables they have manipulated rather than to some other uncontrolled influences. In testing hypotheses or evaluating tentative answers to questions, experimenters make decisions based on probability rather than on certainty. Experimentation, the classic method of the laboratory, is the most powerful method for discovering and developing a body of knowledge about the prediction and control of events. The experimental method has been used with some success in the school classroom, where, to some degree, variables can be controlled.

The early applications of experimental method, based on John Stuart Mill's law of the single variable, have been replaced by the more effective applications of factorial designs made possible by the contributions of R. A. Fisher. His concept of equating groups by random selection of subjects and random assignment of treatments and his development of the analysis of variance and the analysis of covariance have made possible the study of complex multivariate relationships that are basic to the understanding of human behavior.

Experiments must understand and deal with threats to the internal validity of the experiment so that the variable relationships they observe can be interpreted without ambiguity. They must also understand and deal with threats to the external validity of the experiment so that their findings can be extended beyond their experimental subjects and generalized to a wider population of interest.

Experimental design provides a plan, or blueprint, for experimentation. Three pre-experimental, three true-experimental, and five quasi-experimental designs have been presented, and their appropriate use, advantages, and disadvantages have been briefly discussed.

Experimentation is a sophisticated technique for problem solving and may not be an appropriate activity for the beginning researcher. It has been suggested that teachers may make their most effective contribution to educational research by identifying important problems that they encounter in their classrooms and by working cooperatively with research specialists in the conduct and interpretation of classroom experiments.

Exercises

1. Why is it more difficult to control extraneous variables in a classroom experiment than in a pharmaceutical laboratory experiment?
2. What significant element distinguishes a quasi-experiment from a true experiment?
3. Why is an ex post facto study not an experiment?
4. A researcher proposing a research project defines the dependent variable as "achievement in mathematics." What difficulty does this definition present? How would you improve it?
5. How could a double blind be applied in an educational experiment?
6. Under what circumstances could an independent variable in one study be a dependent variable in another study?
7. Why is randomization the best method for dealing with extraneous variables?
8. How could a high degree of experimental mortality seriously affect the validity of an experiment?
9. Read the report of an experiment in an educational research journal.
   a. Was the problem clearly stated?
   b. Were the variables defined in operational terms?
   c. Was the hypothesis clearly stated?
   d. Were the delimitations stated?
   e. Was the design clearly described?
   f. Were extraneous variables recognized? What provisions were made to control them?
   g. Were the population and the sampling methods described?
   h. Were appropriate methods used to analyze the data?
   i. Were the conclusions clearly presented?
   j. Were the conclusions substantiated by the evidence presented?
Exercises in Research Navigator

10. Will computer-based instruction improve learning of critical thinking?

Using text, figures, and other materials on a set of PowerPoint slides, this article reports on two experimental studies that examine the effectiveness of organizations to assess learning of students using web-based audio or video recordings of a lecture. What do you think: the outcomes were? Consider the design in the methods section and consider what confounds might exist and/or how they were controlled. What was the design of the study? To find out, locate the PDF version of this report in the Education database of ContentSelect (item # 6911182) and read the article.

11. Can you identify the independent and dependent variables?

This study used an experimental design to examine the effectiveness of verbal praise on adult learners. What do you think the variables and the design used were? To find out, locate the PDF version of this report in the Education database of ContentSelect (item # 6411043) and read the article.

References


Hancock, D. R. (2002). Influencing graduate students' classroom achievement, homework habits and motivation to learn with verbal praise. Educational Research, 44, 89-100.


Experiential Versus Experience-Based Learning and Instruction*

James D. Laney

Abstract: This study tested claims about the superiority of experience-based over experiential approaches to teaching economic concepts. Students were randomly assigned to three groups—experience-didactic, experience-debriefing, and debriefing-only. At pretest and posttest students were interviewed to probe their understanding of 10 basic economic concepts and to determine their proclivity to use the concept of cost-benefit analysis in a personal decision-making situation. Planned comparisons revealed the following statistically significant differences on the understanding-of-economic-concepts posttest: (a) the combined means of the experience-debriefing and debriefing-only groups were higher than the mean of the experience-didactic group and (b) the mean of the experience-debriefing group was higher than the mean of the debriefing-only group. No significant differences were found between groups on the use-of-cost-benefit analysis measure. Overall, the findings support the superiority of experience-based instruction.

Only a few researchers have explored the economic thinking of preschool and primary-grade children. Of these, most (Ajello, Bomri, & Toccarev, 1978; Armento, 1982; Berti, Bomri, & De Bari, 1986; Burns, 1975; Fox, 1978; Furth, 1980; Schug, 1981, 1983; Schug & Birkey, 1985; and Strauss, 1982) have undertaken interview studies in which they seek developmental patterns in economic reasoning. The findings of these studies suggest that economic thinking develops in an age-related, stage-like sequence. Summarizing across the aforementioned studies, Armento (1982) described children's concept response patterns as progressing "from egocentric to objective; from tautological, literal, and rule-oriented to generalizable; from concrete to abstract, and from inconsistent and narrow to consistent, flexible, and accurate" (p. 89).

According to Schug (1983), many young children exhibit unreflective economic reasoning characterized by (a) a preoccupation with the physical characteristics of the object or process being discussed, (b) egocentric thinking, (c) confusion in identifying causes and effects, and (d) an inclination to treat variables as interchangeable. Research on economic learning has demonstrated that young children tend to have many misconceptions about basic economic concepts indicative of unreflective economic reasoning. Specifically, young children have problems understanding and using such concepts as wants, scarcity, money, monetary value, exchange, change, profits (and what store owners do with the money received from customers), opportunity cost, and cost-benefit analysis.

Most of the research on children's economic reasoning cited above has focused on spontaneous or experiential development that occurs as children experience economies in their daily lives. According to Schug (1983), an important question for future research is to determine whether economic instruction fosters the development of economic reasoning ability. Ajello et al. (1987), Berti et al. (1986), Laney (1989), Armento (1986), and Kourilsky (1983) have provided guidance on how to design economic instruction to achieve this end at the elementary school level.

Berti et al. (1986) and Ajello et al. (1987) discovered that economic training changed third graders' conceptions of (a) profit and (b) work and profit, respectively. In the study by Berti et al., progress toward economic understanding was not dramatic, but it occurred when children (a) were given correct information about economic ideas and (b) found discrepancies between predicted and actual outcomes of economic events. In addition, Berti et al. suggested that having children talk about economic concepts may contribute to their progression in mastering those concepts.

Laney's (1989) findings suggested that real-life experiences are better than vicarious experiences for promoting first graders' learning and retention of the economic concept of opportunity cost. Laney explained his results by suggesting that real-life experiences make economic concepts more meaningful and thus more memorable to students.

According to Armento (1986), children experience economic situations and events on a daily basis. She maintains that (a) play provides the best means for young children to explore their economic world and (b) the economic content emphasized during the early years is best taken from happenings in the children's everyday lives (e.g., buying, selling, making goods and services).

Kourilsky (1983) stressed that it is important for elementary school students to participate in economic experiences that are both personal and active. In defining the role of experience in economic concept acquisition, she distinguishes between experiential learning and experience-based learning. Kourilsky stated that experience with economic concepts (i.e., experiential learning) is not sufficient. She also found that substantive acquisition of economic concepts is dependent on experience-based learning that is, experiences are followed by debriefings and discussions in situations in which situations are analyzed and economic concepts derived. Kourilsky noted that debriefings focus students' attention on relevant ideas. To support her assertion, Kourilsky offered the following analogy: "Most of you played Monopoly as children, but probably few of you learned an extensive amount of economics from participating in the game." (p. 5).

Armento (1986) also downplayed experience, in and of itself, as the cause of a child's direct and predictable learning. She stated that the meanings constructed by a child from his or her economic experiences are most probably attributable to the child's cognitive capabilities at the moment, the value and motivational orientation of the child, the nature of the experience, and the child's prior knowledge.

Kourilsky has authored three experience-based economic education programs for elementary school students—Kinder-Economy (Grades K through