

# 13

## Experimental Designs

*What is Experimental Research?*

*Internal Invalidity*

*External Invalidity*

*Types of Designs*

---

We've previously discussed aspects of three dissertations which embraced an experimental design. My Southwestern dissertation compared three approaches to teaching adults in a local Southern Baptist church: Skinnerian behaviorism, Brunerian cognitivism, and an eclectic approach of the two in 1978. **Dr. Stephen Tam** compared three approaches to teaching with Chinese students in Hong Kong seminary: interactivity, gaming, and lecture in 1989. **Dr. Mark Cook** studied the role of active participation in adult learning in a local church in 1994.<sup>1</sup>

### What Is Experimental Research?

The research methods we have examined in the past few chapters are generally considered **descriptive studies**. A descriptive study analyzes a present condition in order to describe it completely. It answers the question "What is?"

Experimental research, on the other hand, answers the question "What if?" The researcher *manipulates independent variables* (e.g., type of treatment, teaching method, communication strategy) and *measures dependent variables* (anxiety level, Bible comprehension, marital satisfaction) in order to establish cause-and-effect relationships between them. Notice, the **independent variable** is controlled or set by the researcher. The **dependent variable** is measured by the researcher. An "experiment" is a prescribed set of conditions which permit measurement of the effects of a particular treatment.<sup>2</sup>

In our varied curricula — education, administration, age-group ministry, counseling and social work — there is a need to discover the "If p then q" links in the world of local church ministry. In this chapter we will explain threats to internal and external experimental validity as well as illustrate both true- and quasi- experimental research designs.

There are numerous hindrances to planning a good experiment. A "good" experi-

---

<sup>1</sup>See Yount "A Critical Analysis..."; Tam, "A Comparative Study..."; and Cook, "A Study of..."

<sup>2</sup>See Babbie, "Chapter 8: Experiments," pp. 186-207; Borg and Gall, "Chapters 15-16: Experimental Designs, Parts I and II," pp. 631-731; Clifford J. Drew and Michael L. Hardman, "Chapter 5: Designing Experimental Research," *Designing and Conducting Behavioral Research* (New York: Pergamon Press, 1985), pp. 77-105; Sax, "Chapter 6: Research Design: Factors Affecting Internal and External Validity," pp. 116-151 and "Chapter 7: Research Design, Types of Experimental Designs," pp. 152-178; and True, "Chapter 8: Experiments and Quasi-experiments," pp. 233-258.

ment is one that confines the variation of measurement scores to variation caused by the treatment itself. The hindrances to good research design are called **sources of experimental invalidity**. These sources fall under two major subdivisions: internal invalidity and external invalidity. Let's define further these sources of experimental invalidity.<sup>2</sup>

## Internal Invalidity

History

Maturation

Testing

Instrumentation

Regression

Selection

Mortality

Interaction

John Henry

Diffusion

Internal invalidity asks the question, *“Are the measurements I make on my dependent (i.e., the variable I measure) variable influenced only by the treatment, or are there other influences which change it?”* An experimental design suffers from internal invalidity when the other influences, called extraneous sources of variation, have not been controlled by the researcher. When extraneous variables have been *controlled*, researchers can be reasonably sure that post-treatment measurements are influenced by the experimental treatment, and not by extraneous variables.

Donald Campbell and Julian Stanley wrote a chapter of a text on research designs that has become a classic in the field.<sup>3</sup> In this chapter they list eight extraneous variables: *history, maturation, testing, instrumentation, statistical regression, differential selection, experimental mortality, and selection-maturation interaction*. Borg and Gall list two more: *the John Henry effect and experimental treatment diffusion*.<sup>4</sup>

### History

*History* refers to events other than the treatment that occur *during the course of an experiment* which may influence the post-treatment measure of treatment effect. If the explosion of the nuclear reactor in Chernobyl, Ukraine had occurred in the middle of a six-month treatment to help people reduce their “anxiety of nuclear power,” it is likely that post-test anxiety scores would be higher than they would have been without the disaster. *History does not refer to the background of the subject*. Since *history* is an internal source of invalidity, its influence must occur during the experiment.

If you study two groups, one which receives the treatment and a similar one which does not, you “control” for history (which is why this second group is called a “control group”) since both groups are statistically<sup>5</sup> affected the same way by events outside the experiment. Any differences between the two groups at the end of the experiment could reasonably be linked to the treatment.

### Maturation

Subjects change over the course of an experiment. These changes can be physical, mental, emotional, or spiritual. Perspective can change. *The natural process of human growth can result in changes in post-test scores quite apart from the treatment. Question: How would a “control group” control this source of internal invalidity?*<sup>6</sup>

---

<sup>2</sup> I use the term “invalidity” to differentiate this concept from “test validity” discussed in Chapter 8. Be careful, however. Many texts use the terms “*experimental validity*” and “*test validity*.”

<sup>3</sup>Donald T. Campbell and Julian C. Stanley, “Experimental and Quasi-experimental Designs for Research on Teaching,” in *Handbook of Research on Teaching*, ed. N. L. Gage (Chicago: Rand McNally, 1963)

<sup>4</sup>Borg and Gall, 635-637

<sup>5</sup>Individuals might be affected, but the *groups* will not significantly differ from each other.

<sup>6</sup>Subjects in both groups will mature, on average, the same.

## Testing

A common research design is to give a group a pre-test, a treatment, and then a post-test (see p. 13-6). If you use the same test both times, *the group may show an improvement simply because of their experience with the test*. This is especially true when the treatment period is short and the tests are given within a short time.

Unless you must specifically measure changes during the experiment -- requiring testing before and after the treatment -- *it is better to only give a post-test*. Randomly assign subjects to groups to render the dependent variable (as well as all others!) statistically equal at the beginning of the study.

## Instrumentation

In the previous section we discussed the problem of using the *same test twice* in pre- and post-measurements. But *if you use different tests for pre- and post-measurements, then the change in pre- and post-scores may be due to differences between the tests rather than the treatment*. The best remedy, as we have already discussed, is to use randomization and a post-test only design. But if you must have pre-test scores — you must use intact groups and need to know if the groups are “equivalent”, or you want to study changes over time — then you must develop “equivalent tests” using the parallel forms techniques discussed in Chapter Eight. *How does use of a control group relate to instrumentation?*<sup>7</sup>

## Statistical regression

Set a glass of cold milk and a hot cup of coffee on a table. Over time, the cold milk will get warmer and the hot coffee colder. They *both regress toward the room temperature*. *Statistical regression refers to the tendency of extreme scores, whether low or high, to move toward the average on a second testing*. Subjects who score very high or very low on one test will probably score less high or low when they take the test again. That is, they *regress toward the mean*.

Let's say you are analyzing how much a particular reading enrichment program enhances the reading skills of 3rd grade children. You give a reading skills test and select for your experiment *every child who scores in the bottom third* of the group. You provide a three-month treatment of reading enrichment, and then measure the reading ability of the group. On the basis of the scores on the children's first and second tests, you find that reading skills improved significantly. *What, in your opinion, is wrong with this study?*<sup>8</sup>

Do not study groups formed from extreme scores. **Study the full range of scores**. The question we need to answer is: Does the reading enrichment program significantly improve reading skills of randomly selected subjects over a control group?

## Differential selection

If we select groups for “treatment” and “control” differently, then the results may be due to the differences between groups before treatment. Say you select high school

---

<sup>7</sup>Even if tests are not “equivalent” both experimental and control groups answer the same test. This controls for the effects of instrumentation on the treatment group. It isolates treatment group changes to the given treatment.

<sup>8</sup>The group would have scored, on average, better on the second testing regardless of the treatment, simply due to statistical regression. In addition, there is no control by which to measure the treatment.

seniors who volunteer for a special Bible study program as your treatment group, and compare their scores with a control group of high school seniors who did not volunteer. Do your post-test scores measure the effect of the Bible study treatment, or the differences between volunteers and non-volunteers? You cannot say. **Randomization solves this problem by statistically equating groups.**

### Experimental mortality

**Experimental mortality, also called "attrition," refers to the loss of subjects from the experiment.** If there is a systematic bias in the subjects who drop out, then post-test scores will be biased. For example, if subjects drop out because they are aware that they're not improving as they should, then the post-test scores of all those who complete the treatment will be positively biased. Your results will appear more favorable than they really are. How does use of a control group solve the problem of attrition?<sup>9</sup>

### Selection-Maturation Interaction of Subjects

**Interaction means the mixing or combining of separate elements.** If you draw a group of subjects from one church to serve as the treatment group, and a second group from a different church to serve as a control, you could well find -- beyond the simple problem of selection differences ("Are the two groups equivalent?") -- **a mixing of selection and maturation factors to compound the extraneous influence on your measurements.** For example, if the two churches differ in the average age of their members, they may well respond to the treatment differently due to inherent maturational factors. Randomly selecting all subjects from a defined population solves this problem.

### The John Henry Effect

John Henry, the legendary "steel drivin' man," set himself to prove he could drive railroad spikes faster and better than the newly invented steam-powered machine driver. He exerted himself so much in trying to outdo the "experimental" condition that he died of a ruptured heart. **If subjects in a control group find out they are in competition with those in an experimental treatment, they tend to work harder.** When this occurs, differences between control and treatment groups is decreased, minimizing the perceived treatment effect.

### Treatment diffusion

Similar to the John Henry effect is treatment diffusion. If subjects in the control group perceive the treatment as very desirable, they may try to find out what's being done. For example, a sample of church members are selected to use an innovative program of discipleship training, while the control group uses a traditional approach. Over the course of the experiment, some of the materials of the treatment group may be borrowed by the control group members. Over time, **the treatment "diffuses" to the control group, minimizing the treatment effect.** This often happens when the groups are in close proximity (members of the same church, for example). Both the John Henry Effect and Treatment Diffusion can be controlled if experimental and control groups are isolated.

---

<sup>9</sup> Subjects will tend to drop out of both treatment and control groups equally. Those who remain in both groups provide a better picture of "difference" than before-and-after type designs.

## External Invalidity

External invalidity asks, “*How confidently can I generalize my experimental findings to the world?*” Sources of external invalidity cause changes in the experimental groups so that they no longer reflect the population from which they were drawn. The whole point of inferential research is to secure representative samples to study so that inferences can be made back to the population from which the samples were drawn (Chapter Seven). **External invalidity hinders the ability to infer back.**

Campbell and Stanley list four sources of external invalidity: the reactive effects of testing, the interaction of treatment and subject, the interaction of testing and subject, and multiple treatment interference.

### Reactive effects of testing

Subjects in your samples may respond differently to experimental treatments merely because they are being tested. **Since the population at large is not tested, experimental effects may be due to the testing procedures rather than the treatment itself. This reduces generalizability.**

One type of reactive effect is **pretest sensitization**. Subjects who take a pre-test are sensitized to the treatment which is to follow (educators sometimes use a pre-test as an advanced organizer to prepare students for learning). This preparation **changes the research subjects from the population from which they were drawn**, and therefore reduces the ability to generalize findings back to the (untested) population. **The best experimental designs do not use pretests.**

Another type of reactive effect is **post-test sensitization**. The posttest can be, in itself, a learning experience that helps subjects to “put all the pieces together.” Different results would be obtained if the treatment were given without a posttest. While researchers must make measurements, care must be taken to **measure treatment effect**, not add to it, with a post-test.

### Treatment and Subject Interaction

**Subjects in a sample may react to the experimental treatment in ways that are hard to predict.** This limits the ability of the researcher to generalize findings outside the experiment itself. If there is a systematic bias in a sample, then treatment effects may be different when applied to a different sample.

### Testing and Subject Interaction

**Subjects in a sample may react to the process of testing in ways that are hard to predict.** This limits the ability of the researcher to generalize findings outside the experiment itself. If there is a systematic bias of test anxiety or “test-wiseness” in a sample, then treatment effects will be different when applied to a different sample.

### Multiple Treatment Effect

Normally we find a single treatment in an experiment. If, however, an experiment exposes subjects to, say, three treatments (A, B, and C) and test scores show that treatment C produced the best results, one cannot declare treatment C the best. It may have been the **combination of the treatments that led to the results**. Treatment C, given alone, may produce different results.

## Summary

Designing an experiment that produces reliable, valid, and objective data is not easy. *But experimental research is the only direct way to measure cause and effect relationships among variables.* What a help it would be to Kingdom service if we could develop effective experimental researchers who are also committed ministers of Gospel -- learning from direct research how to teach and counsel and manage and serve in ways that directly enhance our ministry.

## Types of Designs

The following is a summary of some of the more important designs of Campbell and Stanley. I will briefly describe the design, give an example of how the design would be used in a research study, and indicate possible sources of internal and external invalidity. In the design diagrams which follow, a test is designated by "O," a treatment by "X," and randomization by an "R."

### True Experimental Designs

Experimental designs are considered *true* experiments when they employ randomization in the selection of their samples and control for extraneous influences of variation on the dependent variable. The three designs we will consider in this section are the **best choices for an experimental dissertation**. These are the pretest-posttest control group design, the Posttest Only Control Group design, and the Solomon Four Group design.

#### Pretest-Posttest Control Group

Two randomly selected groups are measured before ( $O_1$  and  $O_3$ ) and after ( $O_2$  and  $O_4$ ) one of the groups receives a treatment (X).

R	$O_1$	X	$O_2$
R	$O_3$		$O_4$

**Example.** Third graders are randomly assigned to two groups and tested for knowledge of Paul. Then one group gets a special Bible study on Paul. Both are then tested again.

**Analysis.** The t-test for independent samples (Chapter 20) can be used to determine if there is a significant difference between the average scores of the groups ( $O_2$  and  $O_4$ ). You can also compute gain scores ( $O_2 - O_1$  and  $O_4 - O_3$ ) and test the significance of the average gain scores with the matched samples t-test.

**Comments.** This design's only weakness is *pre-test sensitization* and the possible interaction between pretest and treatment.

#### Posttest Only Control Group

Subjects are randomly selected and assigned to two groups. Due to randomization, the two groups are statistically equal. *No pretest is given.* One group receives the treatment.

R	X	O <sub>1</sub>
R		O <sub>2</sub>

**Example.** Third graders are randomly assigned to two groups. Then one group receives a special study on the life of Paul (no pre-test). Both are tested on their knowledge of Paul at the conclusion of the study.

**Analysis.** The difference between group means (O<sub>1</sub> and O<sub>2</sub>) can be computed by an **independent groups t-test**.

[Other procedures that can be used include one-way ANOVA (though usually used with three or more groups - see Chapter 24), the ordinal procedures *Wilcoxin Rank Sum test* or *Mann-Whitney U* (see Chapter 21). We'll discuss these later].

### Solomon Four-Group

Subjects are randomly selected and assigned to one of four groups. Group 1 is tested before and after receiving the treatment; Group 2 is tested before and after receiving no treatment; Group 3 is tested only after receiving the treatment; and Group 4 is tested after receiving no treatment.

1	R	O <sub>1</sub>	X	O <sub>2</sub>
2	R	O <sub>3</sub>		O <sub>4</sub>
3	R		X	O <sub>5</sub>
4	R			O <sub>6</sub>

The Solomon design is actually a combination of the Pre-Test Post-Test Design (groups 1 and 2) and the Post-Test Only design (groups 3 and 4). Look!

1	R	O <sub>1</sub>	X	O <sub>2</sub>
2	R	O <sub>3</sub>		O <sub>4</sub>
3	R		X	O <sub>5</sub>
4	R			O <sub>6</sub>

**Example.** Third graders are randomly assigned to 1 of 4 groups. The “knowledge of Paul” is measured in groups 1 and 2. Groups 1 and 3 are given a special study on the life of Paul. When the special study is over, all four groups are tested.

**Analysis.** **One-way ANOVA** can be used to test the differences in the four post-test mean scores (O<sub>2</sub>, O<sub>4</sub>, O<sub>5</sub>, O<sub>6</sub>). The **effects of the pretest** can be analyzed by applying a t-test to the means of O<sub>4</sub> (pretest but no treatment) and O<sub>6</sub> (neither pretest or treatment). The **effects of the treatment** can be analyzed by applying a t-test to the means of O<sub>5</sub> (treatment but no pretest) and O<sub>6</sub> (neither pretest or treatment). **Subject maturation** can be analyzed by comparing the combined means of O<sub>1</sub> and O<sub>3</sub> against O<sub>6</sub>.

**Comments.** The Solomon Four Group design provides several ways to analyze data and control sources of extraneous variability. Its major drawback is the large number of subjects required. *Since each group needs to contain at least 30 subjects, one experiment would require 120 subjects.*

## Quasi-experimental Designs

The term *quasi-* (pronounced *kwahz-eye*) means **almost, near, partial, pseudo, or somewhat**. Quasi-experimental designs are used when true experiments cannot be done. A common problem in educational research is the unwillingness of educational administrators to allow the random selection of students out of classes for experimental samples. **Without randomization, there are no true experiments.** So, several designs have been developed for these situations that are “*almost true experiments,*” or **quasi-experimental designs**. We’ll look at three: the time series, the non-equivalent control group design, and the counterbalanced design.

### Time Series

Establish a baseline measure of subjects by administering a series of tests over time ( $O_1$  through  $O_4$  in this case). Expose the group to the treatment and then measure the subjects with another series of tests (e.g.,  $O_5$  through  $O_8$ ).

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

**Example.** A class of third graders is given several tests on Paul before having a special study on him. Several tests are given after the special study is finished.

**Analysis.** I could say something like “data is analyzed by trend analysis for correlated data on  $n$  subjects under  $k$  conditions (linear and polynomial), or the monotonic trend test for correlated samples,” but let me simply say that **data analysis is much more complex with a time series design**. An effective visual analysis can be made by graphing the group’s mean scores on each test over time. Important changes in the group can easily be attributed to the treatment by the shape of the line. One could also average the pre-treatment scores and the post-treatment scores, and apply a t-test for matched samples to the averages!

**Comments.** Since there is **no control group**, one cannot determine the effects of history on the test scores. **Instrumentation** may also be a problem (Are the tests equivalent?) Beyond these internal validity problems, the **reactive effects of repeated testing** of subjects is a source of external invalidity.

### Nonequivalent Control Group Design

Subjects are tested in **existing or “intact” groups** rather than being randomly selected. The dotted line in the diagram represents “non-equivalent” groups. Both groups are measured before and after treatment. Only one group receives the treatment.

$$\begin{array}{ccc} O_1 & X & O_2 \\ \hline O_3 & & O_4 \end{array}$$

**Example.** Two intact third grade classes (no random selection) are tested on their knowledge of Paul before and after one of them receives a special study on the life of Paul.

**Analysis.** One approach to measuring the significance of difference between the

two groups is to **compute gain scores**. This is done by subtracting the pre-test score from the post-test score for each subject. Use gain scores to compute average gain for each group. **Test whether the average gain is significantly different by the t-test for independent samples**. Another approach is to use the **pre-test scores as a covariate** measure to adjust the posttest means. Analysis of covariance (See Chapter 25) is the procedure to use.

**Comments.** This design should be used only when random assignment is impossible. It does not control for **selection-maturation interaction** and may present problems with statistical regression. Beyond these internal sources of invalidity, this design suffers from **pretest sensitization**.

### Counterbalanced Design

Subjects are not randomly selected, but are used in intact groups. Group 1 receives treatment 1 and test 1. Then at a later time, they receive treatment 2 and test 2. Group 2 receives treatment 2 first and then treatment one.

	Time	
	1	2
Group1	$X_1 O$	$X_2 O$
Group2	$X_2 O$	$X_1 O$

**Example.** Two third grade classes receive two special studies on Paul: one in classroom and the other on a computer. Class 1 does the classroom work first, followed by the computer; class 2 does the computer work first. Both groups are tested after both treatments.

**Analysis.** Use the Latin Squares analysis (beyond the scope of this text).

**Comments.** Since randomization is not used in this design, selection-maturation interaction may be a problem. Multiple treatment effect is a possible source of external invalidity.

### Pre-experimental Designs

Pre-experimental designs should not be considered true experiments, and **are not appropriate for formal research**. I include them so that you can contrast them with the better designs. Data collected with these designs is highly suspect. We will consider the One Shot Case Study design, the One Group Pretest Posttest design, and the Static Group comparison design.

#### The One Shot Case Study

A single group is given a treatment and then tested.

$X \quad O$

**Example.** A third grade class is provided a special Bible study course on Paul, after which their knowledge of Paul is tested.

**Analysis.** Very little analysis can be done because there is nothing to compare the posttest against and no basis to determine what influence the treatment had.

**Comments.** None of the sources of internal or external invalidity are controlled by this design. It suffers most in the areas of history, maturation, regression, and differential selection. It also suffers from the external source of "treatment and subject." The design is useless for most practical purposes because of numerous uncontrolled sources of difference.

**One-Group Pretest/Posttest**

A single intact group is tested before and after a treatment.

$$O_1 \quad X \quad O_2$$

**Example.** A group of third graders is tested on knowledge of Paul before and after a special study on the life of Paul.

**Analysis.** Test the difference between the pre-test and post-test means using the matched sample t-test (See Chapter 20) or Wilcoxin matched pairs signed rank test (See Chapter 21).

**Comments.** Problems abound with history, maturation, testing, instrumentation, and selection-maturation interaction. The reactive effects of pre-and post- tests and treatment and subject are external sources of invalidity.

**Static-Group comparison**

Two intact groups are tested after one has received the treatment.

$$\begin{array}{r} X \quad O \\ \hline O \end{array}$$

**Example.** Two classes of third graders are tested on their knowledge of Paul after one of them has had the special Bible study.

**Analysis.** Determine whether there is a significant difference between post-test means by using the t-test for independent samples (Chapter 20) or the Mann-Whitney U nonparametric test (Chapter 21). While these statistics will work, *their results are meaningless* since there is no assurance that groups were the same at the beginning of the treatment.

**Comments.** This design suffers most from selection, attrition, and selection-maturation interaction problems. It also fails to control the external invalidity source of treatment and subject.

## Summary

This chapter introduced you to the world of experimental research design. The concepts of **internal and external validity**, **randomization**, and **control** are essential to constructing experiments which provide valid data. **Experimental research is the only type which can establish cause-and-effect relationships between variables.**

## Vocabulary

control group	representative sample which does not receive treatment
differential selection	subjects selected for samples in a non-random manner, i.e., in "different ways"
experimental mortality	loss of subjects from the study
external invalidity	flaw which prevents experimental results from being generalized to the original population
history	events <b>during experiment</b> which influences scores on post test
instrumentation	differences in subject scores due to <b>differences in tests</b> used
interaction of testing/subject	subjects may <b>react to tests unpredictably</b> (generalization?)
interaction of treatment/subject	subjects may <b>react to treatment unpredictably</b> (generalization?)
internal invalidity	condition which alters measurements <b>within the experiment</b>
John Henry effect	Control Group tries harder (distorting the results)
maturation	<b>change in subjects</b> over course of the experiment
posttest sensitization	posttest <b>changes subjects</b> : they 'put it all together' and score higher than they normally would
pretest sensitization	pretest <b>changes subjects</b> : 'advance organizer': prepares subjects for treatment
selection-maturation interaction	samples of subjects may <b>mature differently</b>
statistical regression	top and bottom scoring subjects move toward the average on second test
testing	source of internal invalidity: <b>improvement due to (different) tests</b> , not treatment
treatment diffusion	source of internal invalidity: <b>treatment 'leaked'</b> to Control Group
true experimental research	design which involves <b>random selection</b> and <b>random assignment</b>

## Study Questions

1. Define internal and external invalidity.
2. Explain the ten sources of internal invalidity and four sources of external invalidity.
3. What is required for a research design to be "true experimental"? Why?

### Sample Test Question

Identify each statement on the left as “E”xternal or “I”nternal invalidity by writing an “E” or “I” in the first blank. Then match the type of invalidity on the right with the statements on the right by placing the appropriate letter in the second blank by each statement.

E/I	Ltr		
___	___	1. Subject familiarity with tests	A. History
___	___	2. Systematic differences in drop-out	B. Instrumentation
___	___	3. Exp Groups chosen differently	C. John Henry Effect
___	___	4. Differences between pre/post test	D. Maturation
___	___	5. Control group “tries harder”	E. Mortality
___	___	6. Subjects react differently to experimental treatment	F. Multiple Treatments
___	___	7. Natural changes in subjects	G. Regression
___	___	8. Pre-test sensitization	H. Reactive Effect of Tests
___	___	9. Impact of external events	I. Selection
___	___	10. Treatment “leaked” to Control	J. Testing
___	___	11. “Low” group scores higher on second test	K. Testing & Subject
___	___	12. Subjects react differently to testing procedures	L. Treatment & Subject
			M. Treatment Diffusion