**An Introduction to Research Design[[1]](#footnote-1)©**

### Bivariate Experimental Research

Let me start by sketching a simple picture of a basic bivariate (focus on two variables) research paradigm.

“IV” stands for “**independent variable**” (also called the “**treatment**”), “DV” for “**dependent variable**,” and “EV” for “**extraneous variable**.” In **experimental research** we manipulate the IV and observe any resulting change in the DV. Because we are manipulating it experimentally, the IV will probably assume only a very few values, maybe as few as two. The DV may be categorical or may be continuous. The EVs are variables other than the IV which may affect the DV. To be able to detect the effect of the IV upon the DV, we must be able to control the EVs.

Consider the following experiment. I go to each of 100 classrooms on campus. At each, I flip a coin to determine whether I will assign the classroom to Group 1 (level 1 of the IV) or to Group 2. The classrooms are my “experimental units” or “subjects.” In psychology, when our subjects are humans, we prefer to refer to them as “participants,” or “respondents,” but in statistics, the use of the word “subjects” is quite common, and I shall use it as a generic term for “experimental units.” For subjects assigned to Group 1, I turn the room’s light switch off. For Group 2 I turn it on. My DV is the brightness of the room, as measured by a photographic light meter. EVs would include factors such as time of day, season of the year, weather outside, condition of the light bulbs in the room, etc.

Think of the effect of the IV on the DV as a signal you wish to detect. EVs can make it difficult to detect the effect of the IV by contributing “**noise**” to the DV – that is, by producing variation in the DV that is not due to the IV. Consider the following experiment. A junior high school science student is conducting research on the effect of the size of a coin (dime versus silver dollar) on the height of the wave produced when the coin is tossed into a pool of water. She goes to a public pool, installs a wave measuring device, and starts tossing coins. In the pool at the time are a dozen rowdy youngsters, jumping in and out and splashing, etc. These youngsters’ activities are EVs, and the noise they produce would make it pretty hard to detect the effect of the size of the coin.

Sometimes an EV is “**confounded**” with the IV. That is, it is entangled with the IV in such a way that you cannot separate the effect of the IV from that of the DV. Consider the pool example again. Suppose that the youngsters notice what the student is doing and conspire to confound her research. Every time she throws the silver dollar in, they stay still. But when she throws the dime in, they all cannonball in at the same time. The student reports back remarkable results: Dimes produce waves much higher than silver dollars.

Here is another example of a confound. When I was a graduate student at ECU, one of my professors was conducting research on a new method of instruction. He assigned one of his learning classes to be taught with method A. This class met at 0800. His other class was taught with method B. This class met at 1000. On examinations, the class taught with method B was superior. Does that mean that method B is better than method A? Perhaps not. Perhaps the difference between the two classes was due to the time the class was taught rather than the method of instruction. Maybe most students just learn better at 10 than at 8 – they certainly attend better at 8 than at 10. Maybe the two groups of students were not equivalent prior to being taught differently. Most students tend to avoid classes at 8. Upperclassmen get to register before underclassmen. Some people who hate classes at 8 are bright enough to learn how to avoid them, others not. Campbell and Stanley (*Experimental and quasi-experimental designs for research*, 1963, Chicago: Rand McNally) wrote about the importance of “achieving pre-experimental equation of groups through randomization.” Note that the students in the research described here were not randomly assigned to the treatments, and thus any post-treatment differences might have been contaminated by pre-treatment differences.

### Nonexperimental Research

Much research in the behavioral sciences is not experimental (no variable is manipulated), but rather “**observational**”. Some use the term “correlational” to describe such a design, but that nomenclature leads to confusion, so I suggest you avoid it. Consider the following research. I recruit participants in downtown Greenville one evening. Each participant is asked whether or not e has been drinking alcohol that evening. I test each participant on a reaction time task. I find that those who report that they have been drinking have longer (slower) reaction times than those who were not drinking. In observational research like this, a nonmanipulated categorical variable, is often referred to as the independent variable, but this can lead to confusion. Better practice is to reserve the word “independent variable” for manipulated variables in experimental research.

In experimental research one typically has only a few levels of the manipulated variable and most often has a continuous dependent variable. The most commonly used statistical procedures for such research are *t* tests and ANOVA, and the (manipulated) categorical variable is appropriately referred to as the “independent variable.” These procedures may also be used in nonexperimental research when one is testing the relationship between a nonmanipulated categorical variable and a continuous variable. I have speculated that the bad practice of referring to nonmanipulated categorical variables as “independent variables” stems from the association between “categorical variable” and “ independent variable” in experimental research. Referring to a nonmanipulated categorical variable as an independent variable even leads many researchers into believing that one can make causal inferences from nonexperimental research as long as a procedure like *t* or ANOVA is used to do the analysis, but not if correlation/regression analysis is employed. That is absolute BS! In fact, *t* and ANOVA can be shown to be mathematically identical to a correlation/regression analysis.

Better practice is to referring to the categorical variable in nonexperimental research as a “**grouping variable**” and the continuous variable as the “**criterion variable**” or “**test variable**.”

A variable that is measured earlier in time is more likely to be called an independent variable (even in nonexperimental research) than is one measured later in time, since causes precede effects. For example, when developing a regression to predict future Y from past X, X is commonly referred to as the “independent variable” and Y as the “dependent variable,” Better practice would be to refer to X as the “predictor variable” and Y as the “criterion variable.”

It is important that you recognize that drinking and reaction time research described above is observational, not experimental. With observational research like this, the results may suggest a causal relationship, but there are always alternative explanations. For example, there may be a “**third variable**” involved here. Maybe some people are, for whatever reason, mentally dull, while other people are bright. Maybe mental dullness tends to cause people to consume alcohol, and, independently of such consumption, to have slow reaction times. If that were the case, the observed relationship between drinking status and reaction time would be explained by the relationship between the third variable and the other variables, without any direct casual relationship between drinking alcohol and reaction time.

For my drinking research, I could do the statistical analysis with a method often thought of as being associated with experimental research, like a *t* test or an ANOVA, or with a method thought of as being associated with observational research, a correlation analysis. With the former analysis, I would compute *t* or *F*, test the null hypothesis that the two populations (drinkers and nondrinkers) have identical mean reaction times, and obtain a *p*, which, if low enough, would cause me to conclude that those two populations have different reaction times. With the latter analysis I would compute Pearson *r* (which is called a point biserial *r* when computed between a dichotomous variable and a continuous variable). To test the null hypothesis that there is, in the population, zero correlation between drinking status and reaction time, I would convert that *r* to a *t* and then to a *p*. If the *p* were sufficiently low, I would conclude that there is an association between drinking and reaction time. The value of *t* and of *p* would be exactly the same for these two analyses, because *t* tests and ANOVA are, quite simply, just special cases of correlation or multiple correlation analysis. Whether you can make a causal attribution or not depends not on the type of analysis done, but on how the data were collected (experimentally with adequate EV control or not). Some psychologists mistakenly think that one can never make firm causal inferences on the basis of a correlation analysis but that one always can on the basis of a *t* test or an ANOVA. These researchers have confused the “correlational” (better called “observational”) research design with the correlation analysis. This is why I discourage the use of the term “correlational” when referring to a research design and the use of the terms “independent variable” and ‘dependent variable” when referring to nonexperimental research.

The demonstration of a correlation between variables X and Y is necessary, but not sufficient, to establish a causal relationship between X and Y. To establish the causal relationship, you have to rule out alternative explanations for the observed correlation. That is, to establish that X causes Y you must show the following:

* X precedes Y.
* X and Y are correlated.
* Noncausal explanations of the correlation are ruled out.

**Design Notation**

To describe research designs, I shall use notation similar to that employed by Donald T. Campbell and Julian C. Stanley. For example, consider the design illustrated at the right. Each row represents one group. Time is represented by the horizontal dimension, so that elements in a single column occur at the same time. At the far left of each row is a letter that indicates how subjects were assigned to groups, R for randomly or N for not random (nonequivalent). X stands for a treatment. When the treatment is not experimentally manipulated, I shall enclose the X in parentheses. O stands for an observation. If the O is subscripted, the numbers indicate different variables. The design illustrated above has two nonequivalent groups, one which receives a treatment and one which does not. After the time of the treatment both groups are observed on two variables.

**N  X  O1,2**

**N     O1,2**

**Internal Validity**

Campbell and Stanley used the term “internal validity” to refer to the degree to which the research design allows one to determine whether or not the experimental treatments, as applied in a particular piece of research, with a particular group of subjects, affected the dependent variable, as measured in that research. They listed a dozen types of threats to internal validity. Here I give you a definition and an example for each type of threat.

**Threats to the Internal Validity of the One Group Pretest-Posttest Design**

Campbell and Stanley called this a “pre-experimental” design, but I consider it to be experimental (since the X is experimentally manipulated), but with potentially serious problems. In some circumstances we may be able to eliminate the problems with this design (for example, when our subjects are inanimate objects whose environments we control completely, as we might imagine things are in the physics or chemistry laboratory). Statistically, the comparison between means on O1 and O2 could be made with correlated samples *t* or a nonparametric equivalent.

**O   X   O**

**History**. The problem presents itself when events other than the experimental treatment occur between pretest and posttest. Without a control group, these other events will be confounded with the experimental treatment. You make an observation at time 1, administer a treatment, and then make an observation at time 2. Extraneous events between time 1 and time 2 may confound your comparison. Suppose that your treatment is an educational campaign directed at the residents of some community. It is designed to teach the residents the importance of conserving energy and how to do so. The treatment period lasts three months. You measure your subjects’ energy consumption for a one month period before the treatment and a one month period after the treatment. Although their energy consumption goes way down after the treatment, you are confounded, because international events that took place shortly after the pre-testing caused the price of energy to go up 50%. Is the reduction in energy consumption due to your treatment or to the increased price of energy?

**Maturation**. This threat involves processes that cause your subjects to change across time, independent of the existence of any special events (including your experimental treatment). In the one-group pretest-posttest design, these changes may be mistaken for the effect of the treatment. For example, suppose that you wish to evaluate the effect of a new program on employees’ morale. You measure the morale of a group of newly hired employees, administer the treatment across a six month period, and then measure their morale again. To your dismay, you find that their morale has gone down. Was your treatment a failure, or did the drop in morale just reflect a common change that takes place across the first several months in a new job – you know, at first you think this is going to be a great job, and then after a while you find that it just as boring as all those other jobs you have had.

**Testing**. The problem here is that pretesting subjects can change them. Suppose you are still trying to get people to conserve energy and other resources. You give them a pretest which asks them whether or not they practice a number of conservation behaviors (things like using low flow toilets, lowering the thermostat in the water heater, recycling, etc.). The treatment is completion of a two week course module in environmental biology. The module includes information on how our planet is being adversely affected by our modern lifestyle. After the treatment, subjects are asked again about their conservation behaviors. You find that the frequency of conservation behaviors has increased. Did it increase because of your treatment, or just because of the pretest? Perhaps the pretest functioned to inform the subjects of several things they could do to conserve, and, so informed, they would have started doing those things whether or not they were exposed to the treatment.

**Instrumentation**. During the course of an experiment, the instrument used to measure the DV may change, and these changes may be mistaken for a treatment effect. Suppose we are going fishing, and want to see if we get bigger fish in the morning or the afternoon. On the way we stop to get bait, beer, and a scale to weigh the fish. If we buy the expensive scale, we can’t afford the beer, so we get the $1.99 cheapo scale – it has a poorly made spring with a hook on it, and the heavier the fish, the more the spring stretches, pointing to a higher measurement. Each time you stretch this cheap spring, it stretches a bit further, and that makes the apparent weight of the fish we catch in the afternoon larger than those we caught in the morning, due to instrumentation error. Often the “instrument” is a human observer. For example, you have trained computer lab assistants to find and count the number of unauthorized installations of software on lab computers, and then remove them. You establish a treatment that is intended to stop users of the lab from installing unauthorized software. Your dependent variable is the number of unauthorized installations found and the amount of time it takes to repair the damage done by such installations. Both the number and the time go down, but is that due to the treatment, or are your assistants just getting bored with the task and missing many unauthorized installations, or getting better at repairing them and thus taking less time?

**Statistical regression**. If you have scores which contain a “random” error component, and you retest subjects who had very high or very low scores, you expect them to score closer to the mean upon retesting. Such regression towards the mean might be mistaken for a treatment effect if only subjects with very high (or very low) scores on a pretest were given the treatment. Consider this demonstration. You have a class of 50 students. You tell them you are giving an ESP test. You have a ten item True-False quiz in the right hand drawer of your desk, but you are not going to pass it out. Instead, they must try to use special powers to read that quiz. You give them two minutes to record their answers. Then you give them an answer key and they score their quizzes. Clearly this measurement has a high (100%) random error component. In a class of 50, a couple of students will, by chance, have pretty high scores. Identify them and congratulate them on their fantastic ESP. Almost certainly a couple will have very low scores too. Identify them and tell them that you can help them get some ESP power, if only the two high scorers will cooperate. Say that you have the ability to transfer ESP power from one person to another. Put your hands on the heads of the high scorers, quiver a bit and mumble something mysterious, and then do the same on the heads of the low scorers. Now you are ready to give the posttest, but only to those given this special treatment. In all probability, those who had the very high scores will score lower on the posttest (see, you did take some of their ESP ability) while those who had very low scores will show some gain.

Years ago, while in the bookstore at Miami University, I overhead a professor of education explaining to a student how intelligence is not a stable characteristic. He explained how he had chosen a group of students who had tested low on IQ, given them a special educational treatment, and then retested them. They got smarter, as evidenced by increased posttest scores. I bit my tongue. Then he went on to explain that such educational interventions must be tailored to the audience. He said that he had tried the same educational intervention on a group of students who had scored very high on the IQ pretest, and, surprise of surprises, they got less intelligent, as indicated by their lower scores on the posttest. I could not help myself. The phrase “regression to the mean” leaped out of my mouth, to the great displeasure of the professor.

**Mortality.** Here the problem is that some subjects drop out of the research before the posttest -- this may be because they actually die, but more often it is due to other reasons. Suppose I am conducting research on the effect of a new drug which I hope will help persons who have a wasting disease. My dependent variable is body weight. I recruit 20 patients into the study. Their mean weight prior to treatment is 97 lbs. Ten subjects drop out before the end of the study -- some of them because they died, others for reasons unknown. I ignore the mortality problem and compute the mean weight of the 10 patients remaining at the end of the study period. Their mean weight is 125 lb. I conclude that my treatment was effective in producing weight gain.

So, what is wrong with my conclusion? One problem is that the patients who dropped out might have been the sickest of the group, those who weighed the least, so by including them in the pretest mean but excluding them from the posttest mean I am artificially inflating the posttest mean. Of course, you could compare the mean pretest weight of those who dropped out with the mean pretest weight of those who stayed in to see if this was the problem, but there are other possible problems.

One other possible problem is that my experimental treatment might work well for some patients but not well for others. It might make some feel better and gain weight while it has no effect on others or even makes others feel worse and lose weight. Those for whom the treatment is not working well would be more likely to drop out than those for whom the treatment is working well, again biasing the results.

A partial solution to the mortality problem is to compare pretest mean to posttest mean for only those subjects who completed the study. This, however, creates a problem of external validity -- you can now generalize your results only to that type of person who would complete a treatment program like that you offered.

**Threats to the Internal Validity of Designs with Two or More Groups**

**Selection**. The problem here is that comparison groups are selected, or subjects are selected into comparison groups, in such a way that they might have been different on the criterion variable (dependent variable) prior to the one group having received some special treatment that the other group did not receive. Campbell and Stanley discussed this threat with respect to what they called the **static-group comparison design**, in which the researcher finds two existing groups, one which has experienced some special treatment and another which has not. The two existing groups are then measured on some characteristic and if they are found to differ on that characteristic then it is inferred that the special treatment in the one group caused the observed difference in the measured characteristic. Since no pretest is given, we have no way of knowing whether or not the two groups were equivalent prior to the one group having experienced the special treatment. Campbell and Stanley classified this design as “pre-experimental” in that the researcher does not manipulate the X, but rather simply finds one group which has already experienced the X and compares that group to another group that has not experienced the X. Independent samples *t* or a nonparametric equivalent could be employed to compare the two groups’ means.

**N  (X)  O**

**N       O**

The selection threat may exist with any design where subjects are selected into the comparison groups in such a way that they are already different before the treatment is applied to the experimental group -- in which case any difference between the groups after the treatment may be due to that initial difference between the groups rather than due to the treatment. For example, you wish to evaluate the effectiveness of a tutorial program. You announce that it is available on the computers in the lab and that it covers the material on which the students will be tested on the next exam. You note who uses the tutorial and who does not. After the next exam, you compare these two groups’ performance on that exam. If the tutorial group does not do as well as the control group, does that mean the tutorial is just a failure, or might it be that students who were having difficulty selected themselves into the tutorial program, and did do better than they would have otherwise, but still not as well as those who had no need to do the tutorial and skipped it. If the tutorial group does better than the controls, does that mean the tutorial was effective, or might it be that only the highly motivated students bothered with the tutorial, and they would have done better than the unmotivated students tutorial or no tutorial.

**Selection x (Maturation, History, Instrumentation, Mortality, Testing, or Regression) Interaction**. Here the effect of maturation, history, instrumentation, mortality, testing, or regression is not the same in the one comparison group as in the other. Suppose that you are comparing the effectiveness of one educational program with another. The one program is being used at Suburban High, the other program at Central High. The dependent variable is scores on an achievement test. The design here is **Pretest-Posttest Nonequivalent Groups Design.** It is considered a “quasi-experimental” design because it has some of the characteristics of experiments (manipulation of the treatment) but the comparison groups are not equated prior to treatment.

**N  O  X  O**

**N  O     O**

It is clear that the pre to post gain at Suburban is greater than at Central, but is that because of the special program at Suburban, or might it be due to a **Selection x Maturation interaction**. That is, might the students at Suburban be maturing (intellectually) at a rate faster than those at Central, in which case they would have made greater gains at Suburban than at Central regardless of any special treatment?

Alternatively, our results might be due to a **Selection x History interaction**, where extraneous events occurring between pretest and posttest are different for the one group than for the other group. For example, there might have been a teacher’s strike and a student riot at Central, while Suburban had a quiet year.

Also possible is a **Selection x Testing interaction**, where the effect of the pretest is different for the one group than for the other group -- perhaps the staff at Suburban High stressed to the students the importance of learning the concepts that were on the pretest, but no such comments were made to the students at Central High.

We should also consider the possibility of a **Selection x Mortality interaction**, that is, differential attrition in the two groups. For example, suppose that the program at Suburban High is run by a retired maniac drill sergeant, Sergeant Stedanko.. Every time a student makes a mistake, Stedanko slaps em in the head and demands 100 push ups (or other such abuse). Although 80 students are enrolled in the program, 60 of them drop out before the end of the semester. The means are computed using the 20 students who completed the program. At Central High things are quite different -- no abuse of the students and little pressure to perform -- and only 10 of 80 enrolled students fail to complete the program. Given these circumstances, we need to ask ourselves who is that drops out of a program like that at Suburban High. My suspicion would be that anybody who was not very motivated and capable to do well in the program would drop out at Suburban High, but that there would be no such pressure to drop out at Central High. In other words, I am suggesting that the apparent effectiveness of the program at Suburban High was mostly due to Sergeant Stedanko’s ability to chase off the weaker students.

Suppose the results came out differently, as plotted below:

Here it appears that the students at Central made greater gains than those at Suburban -- but this apparent result might be due to a **Selection x Instrumentation interaction**, in which the characteristics of the instrument are different for the one group than for the other group. In this case, it appears that the achievement test is not adequate for testing the students at Suburban. These students are already making close to the maximum score on the test (60) at the beginning of the school year. On that test, there is no room for improvement. They may well have learned a lot during the school year, but the test did not detect it. This is called a **ceiling effect**.

Now consider yet another pattern of results.

Again, it appears that the students at Central made greater gains than those at Suburban, but let us assume that a ceiling effect is not to blame. There still could a different type of **Selection x Instrumentation interaction**. For example, suppose that the examiners at Central High misunderstood the directions for giving the pretest and gave the students only half the time they were supposed to get to take the examination, but they got things right on the posttest, while the examiners at Suburban High had no problems following the instruction either time

Another possibility is a **Selection x Regression interaction**, in which a regression artifact is greater in one group than in the other. For example, suppose that the staff at Central High made sure that all of their weaker students were enrolled in the experimental program being studied (they wanted to help these students get caught up), but no such enrollment bias existed at Suburban High. Regression up towards the mean would be expected to be greater at Central High.

**Social Threats to Internal Validity**

Here we shall talk about threats that stem from the social nature of research. Research participants typically know that they are involved in research and they may know that there is another group of participants who are being treated differently. If they (or their parents, teachers, spouses, etc.) think this is not fair, internal validity may be threatened, as explained below.

**Diffusion/Imitation of Treatment.** Suppose that you randomly assign some students at the middle school to take a special mathematics class. Students not chosen for the special class are enrolled in the regular class. You use some new ideas in the special class, designed to facilitate the students’ learning of mathematical concepts. The students in the special class share these new ideas with their friends in the traditional class, diffusing the treatment to the control group and contaminating the comparisons between groups. Alternatively, the teacher in the regular class may pay attention to what the special teacher is doing and then imitate the special treatment in her class. Again, this would reduce the differences in treatment between the two groups and contaminate the research.

**Compensatory Rivalry.** Suppose that we manage to avoid the diffusion/imitation of treatment problem, but that the teacher in the regular class feels threatened by the special class. The teacher of the regular class may be motivated to do even more than usual to make her students perform well, even encouraging them to compete with “that other class.” Again, this may reduce the difference in the two groups.

**Resentful Demoralization.** The students in the regular class may not believe that assignment was really random, they may conclude that they were not given the special class because the staff think they are not worthy of special treatment. This can lead to demoralization, depressing the performance of students in the regular class, again contaminating the comparisons between the two groups.

**Compensatory Equalization of Treatment.** Suppose that you learn that your child has been assigned to the regular class -- that is, she has been denied the opportunity to benefit from the superior techniques being used in the special class. What are you going to do about that? You may call up the principal and demand that your child be moved to the special class. Some other parents may do the same, and the principal may grant your wishes or may provide the regular class with some special treatment designed to compensate them for not being treated specially. Now you are happy, but the researcher is upset that her random assignment to groups has been destroyed and she now probably has groups that are nonequivalent even before the treatment (there probably is a difference between children whose parents will call up the principal to get the special class and children whose parents do not call the principal), or the difference in treatment between the experimental group and the control group has been diminished or confounded.

**External Validity**

Campbell and Stanley used the term “external validity” when referring to the extent to which the results generalize beyond the specifics of the experimental situation. Would you get the same results if you used different subjects, if you manipulated the IV in a different way, if you measured the DV in a different way, if the setting were different, etc.? Campbell and Stanley discussed four threats to external validity.

**Testing x Treatment interaction**. This is a problem in the **pretest-posttest control group design**. This design simply adds to the one group pretest-posttest design a control group which does not receive the experimental treatment between pretest and posttest. Ideally, subjects are randomly assigned to the two comparison groups so that we do not have to worry about selection. Adding the control group eliminates all threats to internal validity, but we are left wondering whether or not the effect of the treatment would be the same in subjects who were not pretested. If this is of concern to you, you could use the Solomon four group design (to be discussed later), or you could just get rid of the pretest and do a posttest only control group design (which might reduce your statistical power, that is, the probability that you will be able to detect an effect of the experimental treatment).

**Selection x Treatment interaction**. Does the effect of your treatment interact with characteristics of the experimental units? That is, do your results generalize from the type of subjects you used in your research to other types of subjects? Very often subjects are college students. For some treatments and some dependent variables, it is probably reasonable to assume generalization to most humans, but for others it may well not be safe to assume such generalization.

**Reactive effects of experimental arrangements**. It is difficult if not impossible to observe with affecting that which is observed. Can you generalize the results found with observed subjects to subjects who are not being observed?

**Multiple treatment interference**. In some types of research, subjects are used in many different experiments. For example, monkeys used in medical research are not just discarded after completing one experiment, they are used for additional experiments later. Also, in the psychology lab, if the human subject has to undergo extensive training before serving as a subject in the type of research done in that lab, there are economic benefits to recruiting that subject to serve in additional studies for which that same training would be needed. The question is, do the results found with subjects who have been used in all these different experiments generalize to individuals that do not have such multiple treatment experience.

**Common, Simple Research Designs**

I have already discussed the one-group pretest-posttest design and the static group comparison design , but there are several other common designs that I would like to present at this point.

**One-Shot Case Study**. Campbell and Stanley classified this design as “pre-experimental.” No variable is manipulated. The researchers simply find some group of subjects who have experienced event X and then measure them on some criterion variable. The researcher then tries to related X to O. My kids were in the public schools here when a terrible tornado ripped through the county just south of our house. After the tornado left, psycho-researchers descended on the schools, conducting research to determine the effects of the tornado on the children’s mental health. Of course, they had no pretest data on these children. Without a comparison group, observations like this are of little value. One might suppose that there is an implicit comparison group, such as that provided by “norms” on the measuring instrument, but how do we know whether or not our subjects already differed from the “norms” prior to experiencing the X? Martin discusses this design in Chapter 10 of his text. He calls it the “One-Group Posttest-Only Design.”

**(X)  O**

**Randomized Pretest-Posttest Control Group Design**. Here we have added a control group to the one-group pretest-posttest design and we have randomly assigned subjects to groups. If we can assume that both groups experienced the same history between observations (that is, there is no selection by history interaction), then history is controlled in the sense that it should affect the O1 to O2 difference identically in the two groups. Likewise, maturation, testing, instrumentation, and regression are controlled in the sense of having the same effects in both groups. Selection and selection by maturation interaction are controlled by assigning subjects to groups randomly, making us confident that they were equivalent prior to experimental treatment (and will mature at equivalent rates). Unless we are foolish enough to employ different measuring instruments for the two groups, selection by instrumentation interaction should not be a problem. Of course, testing by treatment interaction is a threat to the external validity of this design.

**R  O  X  O**

**R  O     O**

Statistically, one can compare the two groups’ pretest means (independent *t* or nonparametric equivalent) to reassure oneself (hopefully) that the assignment technique did produce equivalent groups -- sometimes one gets an unpleasant surprise here. For example, when I took experimental psychology at Elmira College, our professor divided us (randomly, he thought) by the first letter of our last name, putting those with letters in the first half of the alphabet into one group, the others in the other group. Each subject was given a pretest of knowledge of ANOVA. Then all were given a lesson on ANOVA. Those in the one group were taught with one method, those in the other group by a different method. Then we were tested again on ANOVA. The professor was showing us how to analyze these data with a factorial ANOVA when I, to his great dismay, demonstrated to him that the two groups differed significantly on the pretest scores. Why? We can only speculate, but during class discussion we discovered that most of those in the one group had taken statistics more recently than those in the other group -- apparently at Elmira course registration requests were processed in alphabetical order, so those with names in the first half of the alphabet got to take the stats course earlier, while those who have suffered alphabetical discrimination all of their lives were closed out of it and had to wait until the next semester to take the stats course -- but having just finished it prior to starting the experimental class (which was taught only once a year), ANOVA was fresh in the minds of those of us at the end of the alphabet.

One can analyze data from this design with a factorial ANOVA (time being a within-subjects factor, group being a between-subjects factor), like my experimental professor did, in which case the primary interest is in the statistical interaction -- did the difference in groups change across time (after the treatment), or, from another perspective, was the change across time different in the two groups. The interaction analysis is absolutely equivalent to the analysis that would be obtained were one simply to compute a difference score for each subject (posttest score minus pretest score) and then use an independent samples *t* to compare the two groups’ means on those difference scores. An alternative analysis is a one-way Analysis of Covariance, employing the pretest scores as a covariate and the posttest scores as the criterion variable -- that is, do the groups differ on the posttest scores after we have removed from them any effect of the pretest scores. All three of these analyses (factorial ANOVA, *t* on difference scores, ANCOV) should be more powerful than simply comparing the posttest means with *t*.

**Randomized Posttest Only Control Group Design**. Here we simply assign subjects to groups randomly, don’t bother with a pretest, administer the treatment to the one group, and then measure the criterion variable. With respect to controlling the previously discussed threats to internal and external validity, this design is the strongest of all I have presented so far. However, this design usually is less powerful than designs that include a pretest-posttest comparison. That is, compared to designs that employ within-subjects comparisons, this design has a higher probability of a Type II error, failing to detect the effect of the treatment variable (failing to reject the null hypothesis of no effect) when that variable really does have an effect. Accordingly, it is appropriate to refer to this threat to internal validity as **statistical conclusion validity**. One can increase the statistical power of this design by converting extraneous variables to covariates or additional factors in a factorial ANOVA, as briefly discussed later in this document (and not-so-briefly discussed later in this course). While it is theoretically possible to make another type of error that would threaten statistical conclusion validity, the Type I error, in which one concludes that the treatment has an effect when in fact it does not (a Type I error), it is my opinion that the Type II error is the error about which we should be more concerned, since it is much more likely to occur than a Type I error, given current conventions associated with conducting statistical analysis.

**R  X  O**

**R     O**

**Extraneous Variable Control**

Controlling extraneous variables is important in terms of eliminating confounds and reducing noise. Here I identify five methods of controlling extraneous variables.

**Constancy**. Here you hold the value of an extraneous variable constant across all subjects. If the EV is not variable, it cannot contribute to the variance in the DV. For example, you could choose to use only female subjects in your research, eliminating any variance in the DV that could be attributable to gender. Do keep in mind that while such noise reduction will increase the statistical “**power**” of your analysis (the ability to detect an effect of the IV, even if that effect is not large), it comes at a potential cost of external validity. If your subjects are all female, you remain uncertain whether or not your results generalize to male individuals.

**Balancing**. Here you assign subjects to treatment groups in such a way that the distribution of the EV is the same in each group. For example, if 60% of the subjects in the experimental group are female, then you make sure that 60% of the subjects in the control group are female. While this will not reduce noise and enhance power, it will prevent the EV from being confounded with the IV.

**Randomization**. If you randomly assign subjects to treatment groups, they should be balanced on subject characteristics (those EVs that subjects bring to the experiment with themselves).

**Matching**. Here we are talking about the research design commonly know as the **randomized blocks** design. On one or more EVs, thought to be well correlated with the DV, we match subjects up in blocks of *k*, where *k* is the number of treatment groups. Within each block, the subjects are identical or nearly identical on the matching variable(s). Within each block, one subject is (randomly) assigned to each treatment group. This will, of course, balance the distribution of the EV across groups, but it will also allow a statistical analysis which removes from the DV the effect of the matching variable, reducing noise and increasing power.

**Statistical control**. One can employ statistical techniques that remove the effect of extraneous variables. I shall discuss this in detail a little later in the semester.

Copyright 2022, Karl L. Wuensch - All rights reserved.

* [Fair Use of this Document](http://core.ecu.edu/psyc/wuenschk/Fair-Use.htm)

1. © Copyright 2022, Karl L. Wuensch - All rights reserved. [↑](#footnote-ref-1)