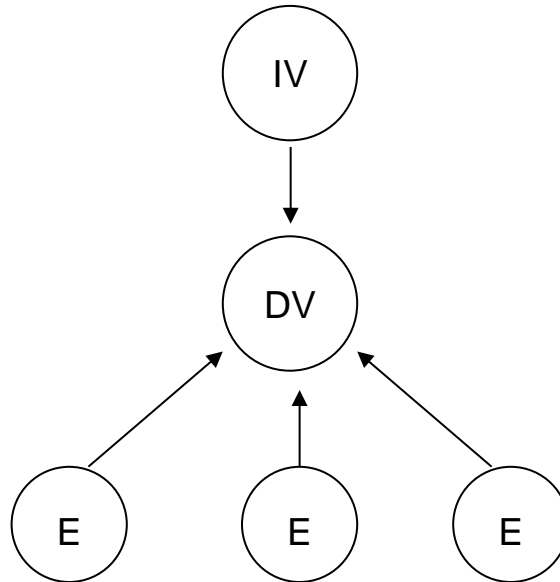


## Bivariate Experimental Research

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 \(1963\)](#) 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. I may refer to the drinking status variable as my IV, but note that it was not manipulated. In observational research like this, the variable that we think of as being a cause rather than an effect, especially if it is a grouping variable (has few values, as is generally case with the IV in experimental research), is often referred to as the IV. Also, a variable that is measured earlier in time is more likely to be called an IV than one measured later in time, since causes precede effects.

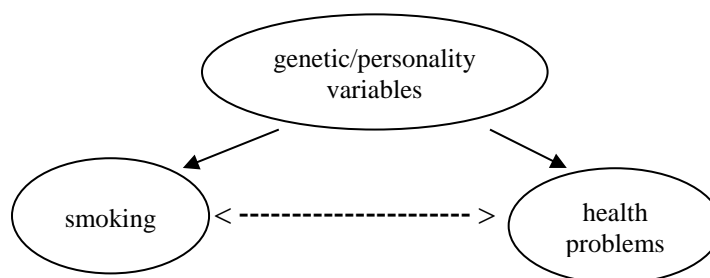
It is important, however, that you recognize that this design is observational, not experimental. With observational research like this, the results may suggest a causal relationship, but there are always alternative explanations. For example, there may be a “**third variable**” involved here. Maybe 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.

Do note that the drinking research could have been done experimentally. We could randomly assign participants to drink or not, administer the treatments, and then test their reaction time. Again, I could do the analysis via a  $t$  test or a Pearson  $r$ , and again the resulting  $p$  value would be identical regardless of statistical method. In this case, if I get a significant correlation between drinking and reaction time, I can conclude that drinking causes altered reaction time. In a nutshell, 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.

Let me give you another example of a third variable problem. Observational research has demonstrated an association between smoking tobacco and developing a variety of health problems. One might argue that this association is due to a third variable rather than any causal relationship between smoking and ill health. Suppose that there is a constellation of third variables, think of them as genetic or personality variables, that cause some people to smoke, and, whether or not they smoke, also cause them to develop health problems. These two effects of the third variable could cause the observed association between smoking and ill health in the absence of any direct causal relationship between smoking and ill health.



How can one rule out such an explanation? It is not feasible to conduct the required experimental research on humans (randomly assigning newborns to be raised as smokers or nonsmokers), but such research has been done on rats. Rats exposed to tobacco smoke develop the same sort of health problems that are associated with smoking in humans. So the tobacco institute has promised not to market tobacco products to rats. By the way, there has been reported an interesting problem with the rats used in such research. When confined to a chamber into which tobacco smoke is pumped, some of them take their fecal boluses and stuff them into the vent from which the smoke is coming.

Side note. Researchers investigating the effects of cigarette smoke on rodents have encountered a problem – many of the subjects respond “by placing feces in the smoke delivery tubing, repeatedly and in quantity. See [Silverman's report](#) on this phenomenon.

Another example of a third variable problem concerns the air traffic controllers strike that took place when Reagan was president. The controllers contended that the high stress of working in an air traffic control tower caused a variety of health problems known to be associated with stress. It is true that those working in that profession had higher incidences of such problems than did those in most other professions. The strikers wanted improved health benefits and working conditions to help with these stress related problems -- but the government alleged that it was not the job that caused the health problems, it was a constellation of third variables (personality/genetic) that on the one hand caused persons of a certain disposition (Type A folks, perfectionists) to be attracted to the air traffic controllers profession, and that same constellation of third variables caused persons with such a disposition to have these health problems, whether or not they worked in an air traffic control tower. One FAA official went so far as to say that working in an air traffic control tower is no more stressful than driving the beltway around DC. Personally, I find such driving very stressful. The government won, the union was busted. I suppose they could solve the problem by hiring as air traffic controllers only folks with a different disposition (Type B, lay back, what me worry, so what if those two little blips on the screen on headed towards one another).

## Internal Validity

Donald T. Campbell and Julian C. 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.

**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. Suppose that you are using a [one-group pretest-posttest design](#): 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 overheard 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

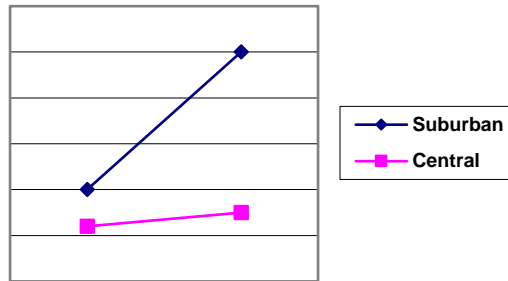
had scored very high on the IQ pretest, and, surprise of surprises, they got less intelligent, as indicated by their lower scores on the posttest. I could not help myself. The phrase “regression to the mean” leaped out of my mouth, to the great displeasure of the professor.

**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 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.

This problem 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.

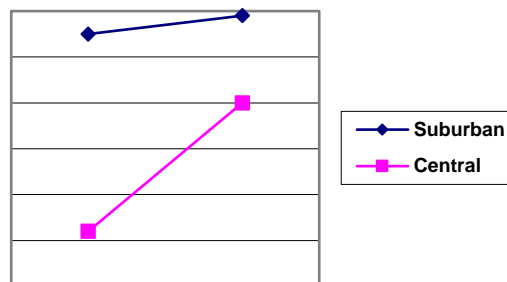
**Differential mortality.** This involves the loss of subjects in a way that results in the subjects remaining in the one group being different than the subjects remaining in the other group. Sometimes the apparent effect of an experimental treatment is simply due to its effectiveness in causing attrition of some types of subjects and not others. For example, suppose you are comparing two programs designed to produce weight loss. Program A subjects come in to a group meeting once a week and stand on a scale. If the subject has not lost at least two pounds since the last week, she is forced to do 50 pushups, while the rest of the class shouts derogatory comments at her. Program B also has a weekly weigh in, but in that program, subjects who are not losing weight are given more positive encouragement to keep at it. Both programs start out with 50 participants. After two months, 10 participants remain in Program A, and their mean weight loss is 21 pounds, while in Program B, 40 participants remain, but mean weight loss is 5 pounds. Is A the more effective program, or was it just more effective at chasing off those who were unable or unwilling to lose weight?

**Selection x (Maturation, History, Instrumentation) interaction.** Here the effect of maturation, history, or instrumentation 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 experimental program is being used at Suburban High, the traditional program at Central High. The DV is scores on an achievement test.



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 the effect of events occurring between pretest and posttest is 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.

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 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.

## 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 DVs, it is probably reasonable to assume generalization to most humans, but for others it may well not be safe to assume such generalization.

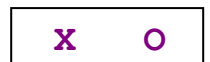
**Reactive effects of experimental arrangements.** It is difficult if not impossible to observe with affecting that which is observed. Can you generalize the results found with observed subjects to subjects who are not being observed?

**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 Research Designs

I shall sketch out common research designs using notation similar, but not identical, to that employed by Campbell and Stanley. **O<sub>i</sub>** will stand for the observation of the criterion (dependent) variable at time *i*. **X** will stand for the presence of the experimental treatment. If we have two groups, the symbols representing the chain of events for the experimental group will be on a line above that for the control group (the group that does not receive the experimental treatment -- of course, you should recognize that our two groups may not include a "control" group but might rather be two groups that have received different experimental treatments -- also, we could have more than two comparison groups). If the lines are separated by + signs, then subjects were assigned to the two comparisons groups randomly (or in another fashion that should result in them being equivalent prior to any special treatment). If the lines are separated by ? marks, then subjects were assigned to comparison groups in a way that could have resulted in the groups being nonequivalent even in the absence of the X treatment.

**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?

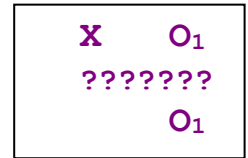


**One Group Pretest-Posttest Design.** Campbell and Stanley called this a "pre-experimental" design, but I consider it to be experimental (since the X is experimentally manipulated), but with potentially serious problems

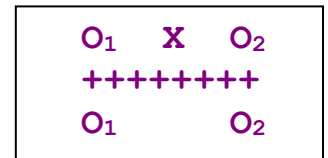


which we have already discussed: History, maturation, testing, instrumentation, and possibly regression. If we have contrived ways to control these threats (which might be possible under if our subjects are inanimate objects whose environments we control completely, as we might imagine things are in the physics or chemistry laboratory), then this design could be OK. Statistically, the comparison between means on  $O_1$  and  $O_2$  could be made with correlated samples  $t$  or a nonparametric equivalent.

**Static Group Comparison.** We discussed this design earlier. As noted by Campbell and Stanley, it is “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.



**Pretest-Posttest Control Group Design.** Here we have added a control group to the one-group pretest-posttest design. 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  $O_1$  to  $O_2$  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 the two groups in a way (such as random assignment) that makes 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.



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  $t$ .

One interesting modification of the Pretest-Posttest Control Group Design is the **Wait-List Control Group Design**. This is especially useful when the experimental treatment should be of value to all subjects.

Kate Cutitta proposed research evaluating the effectiveness of an intervention designed to reduce anxiety and depression and increase physical activity in children who have received implantable cardioverter defibrillators. Her basic research design is a modification of the basic Pretest-Posttest Control Group Design.

Intervention (X)	Pre (no X)	Post (X)	Followup (X)	Followup 2	Followup 3
Waitlist Control	Time 1 (no X)	Time 2 (no X)	Time 3 (no X)	Post X	Followup

All subjects get the treatment, but a randomly selected half (the wait-list control group) get it after the other group. This provides good control for the major threats to the internal validity of the one-group pretest-posttest design: History, maturation, testing, and instrumentation. Campbell and Stanley considered the one-group pretest-posttest to be of such little value that they classified it as "pre-experimental."

After her proposal had been tied up in the IRB approval process for seven months, Kate received an email from "<Alfred E. Neuman>, MPH, IRB Administrator. In this letter Mr. <Neuman> expressed reservations about the use of the wait-list control group, which he suggested was a not necessary part of the research design -- "I feel like the Committee will ask what the purpose is of randomizing and delaying the intervention for half of the participants. It is not necessarily a research ethics question, more of a study design one."

Apparently getting a master's degree in public health does not involve any instruction in basic research methods.

[Donald T. Campbell and Julian C. Stanley](#) are turning over in their graves.

**Posttest Only Control Group Design.** Here we simply assign subjects to groups in a way that should assure pretreatment equivalence, 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

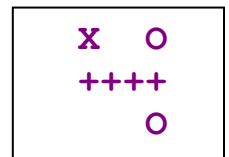


Diagram illustrating the addition of two 10-bit numbers, showing the carry propagation:

```

      O1      X      O2
    ++++++
      O1      O2
    ++++++
           X      O2
    ++++++
                O2
  
```

### Treatment effect but no testing or Testing x Treatment interaction

	Treatment		
Pre-tested	Control	X	marginal
no	10	30	20
yes	10 (10)	30 (10)	20
marginal	10	30	20

Look at the marginal means for Treatment, 10 and 30. This shows a main effect of 20 points for the treatment. Now look at the marginal means for Pre-testing, 20 and 20. Since these marginal means are identical, the overall effect of pretesting is nil. Now look at the cell means. In the subjects not pretested the treated subjects scored 20 points higher than the control subjects. That is, the simple main effect (aka conditional main effect) of the treatment is 20 points for subjects not pretested. Now look at the cell means for those who were pretested. The treatment group scored 20 points higher than did the control group. Since the effect of the treatment is identical for the two pretesting groups, there is no Testing x Treatment interaction.

### Treatment and testing effects but no Testing x Treatment interaction.

Pre-tested	Treatment		marginal
	none	X	
no	10	30	20
yes	20 (10)	40 (10)	30
marginal	15	35	25

Look at the marginal means for Treatment, 15 and 35. This shows a main effect of 20 points for the treatment. Now look at the marginal means for Pre-testing, 20 and 30. This shows a main effect of Pretesting. Subjects who were pretested scored 10 points higher than those who did not. Now look at the cell means. In the subjects not pretested the treated subjects scored 20 points higher than the control subjects. That is, the simple main effect (aka conditional main effect) of the treatment is 20 points for subjects not pretested. Now look at the cell means for those who were pretested. The treatment group scored 20 points higher than did the control group. Since the effect of the treatment is identical for the two pretesting groups, there is no Testing x Treatment interaction.

### Treatment and testing effects and a Testing x Treatment interaction.

Pre-tested	Treatment		marginal
	none	X	
no	10	30	20
yes	20 (10)	60 (10)	40
marginal	15	45	30

The marginal means show a 30 point treatment effect and a 20 point testing effect. Look at the cell means. For subjects not pretested the treatment group scored 20 points higher than the control group. For subjects that were pretested the treatment group scored 40 points higher than the control group. When the effect on Y of variable A on Y differs across levels of variable B, we say that A and B interact (or, that variable B moderates the effect of variable A). There is a Treatment x Pretesting interaction here. The effect of the treatment is not the same for pretested subjects as it is for those not pretested.

### Independent Samples and Correlated Samples Designs

With the **independent samples design** cases are assigned to groups in a way that should not create any correlation between the scores in any one group and the scores in any other group. Such designs are also called **between subjects** designs.

With the **correlated samples design** cases are assigned to groups in a way that should produce a positive correlation between the scores in any one group and the scores in any other group.

One example of a correlated samples design is the **randomized blocks design**. Suppose that you have three experimental treatments to evaluate. You also have one or more blocking variables. A blocking variable is a variable that you have good reason to believe is positively correlated with the dependent variable. You match the subjects up in blocks of three, such that within each block the subjects are nearly identical on the blocking variable(s). Then you randomly assign, within each block, one case to Treatment 1, another to Treatment 2, and another to Treatment 3. If there were only two treatments, then there would be only two cases in each block, and the design could be described as **matched pairs**.

A special case of the randomized blocks design is when you match the subjects up on themselves. That is, each subject serves in each experimental condition. This design is called the **repeated measures** or **within subjects** design. With this design it is important to counterbalance the order in which the treatments are presented to control for order effects.

You may also run across the term “**completely randomized design**.” This refers to a design where you have randomly selected cases from the population of interest and then randomly assigned them to experimental treatments.

	Score	Treatment
1	6	1
2	5	1
3	4	1
4	4	1
5	3	1
6	2	1
7	3	2
8	4	2
9	5	2
10	5	2
11	6	2
12	7	2
13	7	3
14	3	3
15	6	3
16	4	3
17	6	3
18	5	3

If the matching variable(s) are well correlated with the dependent variable, the correlated sample design will have more power than an independent sample design with the same number of scores. Suppose that I have three experimental treatments, A, B, and C. I randomly assign six cases to each experimental treatment. The one-way independent samples ANOVA would be an appropriate analysis, given normality and homogeneity of variance. Here are the contrived data:

Here are the results of a one-way independent samples ANOVA on these data:

### Descriptives

Score

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	6	4.00	1.414	.577	2.52	5.48	2	6
2	6	5.00	1.414	.577	3.52	6.48	3	7
3	6	5.17	1.472	.601	3.62	6.71	3	7
Total	18	4.72	1.447	.341	4.00	5.44	2	7

### ANOVA

Score

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	4.778	2	2.389	1.162	.339
Within Groups	30.833	15	2.056		
Total	35.611	17			

Oh crap, not significant. Suppose we had conducted a randomized blocks design with good correlations between the scores in any one group and those in any other group. Here I have reordered the scores within each condition to produce such correlations.

	Block	A	B	C	va
1	1	6	7	6	
2	2	4	6	7	
3	3	5	5	6	
4	4	4	5	5	
5	5	3	4	3	
6	6	2	3	4	
7					

Here is the analysis with a correlated samples ANOVA:

		Correlations		
		A	B	C
A	Pearson Correlation	1	.900*	.673
B	Pearson Correlation	.900*	1	.769
C	Pearson Correlation	.673	.769	1

Notice the good correlation between scores in any one group and any other group.

Descriptive Statistics					
	N	Minimum	Maximum	Mean	Std. Deviation
A	6	2	6	4.00	1.414
B	6	3	7	5.00	1.414
C	6	3	7	5.17	1.472

Same means and standard deviations as before.

Tests of Within-Subjects Effects					
Measure: MEASURE_1					
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Treatment	4.778	2	2.389	5.244	.028
Error(Treatment)	4.556	10	.456		

Hot dog, the conditions now differ significantly.

### 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.

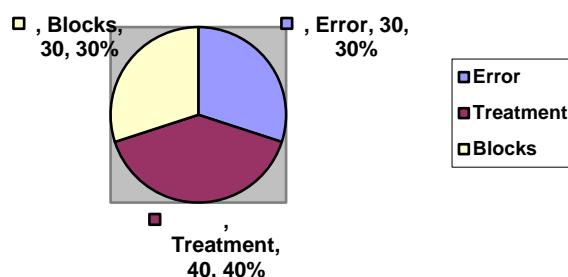
**Randomization.** If you randomly assign subjects to treatment groups, they should be balanced on subject characteristics (those EVs that subjects bring to the experiment with themselves).

**Matching.** Here we are talking about the research design commonly known 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.** Suppose you were going to evaluate the effectiveness of three different methods of teaching young children the alphabet. To enhance power, you wish to use a **randomized blocks design**. You administer to every potential subject a test of readiness to learn the alphabet, and then you match (block) subjects on that variable. Next you randomly assign them (within each block) to groups. In your statistical analysis, the effect of the **matching/blocking variable** is taken out of what would otherwise be “error variance” in your statistical model. Such error variance is generally the denominator of the ratio that you use as the test statistic for a test of statistical significance, and the numerator of that ratio is generally a measure of the apparent magnitude of the treatment effect. Let's look at that ratio.

$$\text{Test statistic} = \frac{\text{treatment effect}}{\text{noise}}, \text{ for example, } t = \frac{\text{difference between means}}{\text{standard error of difference}}, \text{ or}$$

$$F = \frac{\text{among groups variance}}{\text{error variance}}.$$



Look at this pie chart, in which I have partitioned the total variance in the DV into variance due to the treatment, due to the blocking variable, and due to everything else (error). If we had just ignored the blocking variable, rather than controlling it by using the randomized blocks design, the variance identified as due to blocks would be included in the error variance. Look back at the test statistic ratio. Since error variance is in the denominator, removing some of it makes the absolute value of the test statistic greater, giving you more power (a greater probability of obtaining a significant result).

Another statistical way to reduce noise and increase power is to have available for every subject data on one or more **covariate**. Each covariate should be an extraneous variable which is well correlated with the dependent variable. We can then use an **ANCOV (analysis of covariance)** to remove from the error term that variance due to the covariate (just like the randomized blocks

analysis does), but we don't need to do the blocking and random assignment within blocks. This analysis is most straightforward when we are using it along with random assignment of subjects to groups, rather than trying to use ANCOV to "unconfound" a static-group design (more on this later in the semester).

If the EV you wish to control is a categorical variable, one method to remove its effect from the error variance is just to designate the EV as being an IV in a **factorial ANOVA**. More on this later in the semester.

Some of you have already studied "**repeated measures**" or "**within subjects**" designs, where each subject is tested under each treatment condition. This is really just a special case of the randomized blocks design, where subjects are blocked up on all subject variables.

#### Reference

Campbell, D. T., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand-McNally.

- [Mnemonics To Remember Threats to Internal and External Validity](#)
- [Two Case Studies in the Ethics of Scientific Publication](#)
- [Fair Use of this Document](#)

Copyright 2019, Karl L. Wuensch - All rights reserved