Research Methods

Knowledge Base

William M.K. Trochim

What is the Research Methods Knowledge Base?

The Research Methods Knowledge Base is a comprehensive web-based textbook that addresses all of the topics in a typical introductory undergraduate or graduate course in social research methods. It covers the entire research process including: formulating research questions; sampling (probability and nonprobability); measurement (surveys, scaling, qualitative, unobtrusive); research design (experimental and quasi-experimental); data analysis; and, writing the research paper. It also addresses the major theoretical and philosophical underpinnings of research including: the idea of validity in research; reliability of measures; and ethics. The Knowledge Base was designed to be different from the many typical commercially-available research methods texts. It uses an informal, conversational style to engage both the newcomer and the more experienced student of research. It is a fully hyperlinked text that can be integrated easily into an existing course structure or used as a sourcebook for the experienced researcher who simply wants to browse.

About the Author

William M.K. Trochim is a Professor in the Department of Policy Analysis and Management at Cornell University. He has taught both the undergraduate and graduate required courses in applied social research methods since joining the faculty at Cornell in 1980. He received his Ph.D. in 1980 from the program in Methodology and Evaluation Research of the Department of Psychology at Northwestern University. His research interests include the theory and practice of research, conceptualization methods (including concept mapping and pattern matching), strategic and operational planning methods, performance management and measurement, and change management. He is the developer of The Concept System® and founder of Concept Systems Incorporated. He lives in Ithaca, New York with his wife Mary and daughter Nora.

Acknowledgements

This work, as is true for all significant efforts in life, is a collaborative achievement. I want to thank especially the students and friends who assisted and supported me in various ways over the years. I especially want to thank Dominic Cirillo who has labored tirelessly over several years on both the web and printed versions of the Knowledge Base and without whom I simply would not have survived. There are also the many graduate Teaching Assistants who helped make the transition to a web-based course and have contributed their efforts and insights to this work and the teaching of research methods. And, of course, I want to thank all of the students, both undergraduate and graduate, who participated in my courses over the years and used the Knowledge Base in its various incarnations. You have been both my challenge and inspiration.

Dedication

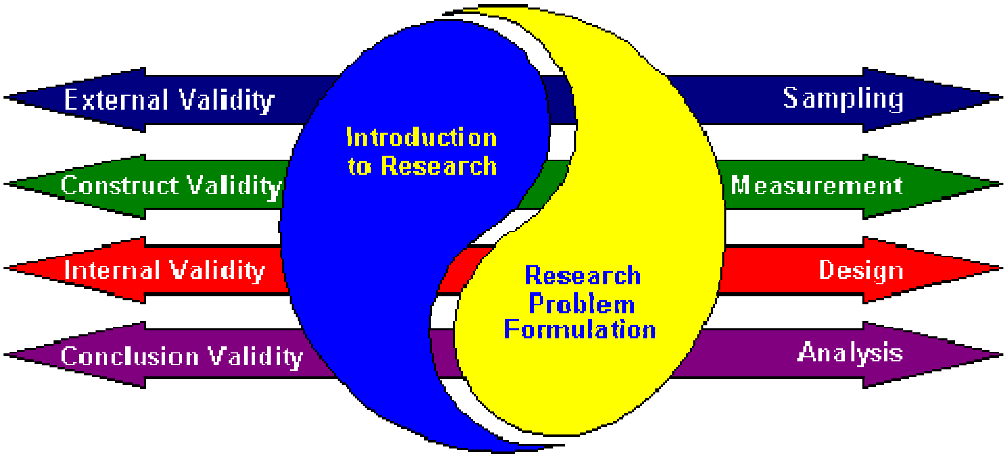
For Mary and Nora, who continue to astonish me with their resilience, patience, and love

Table of Contents

Yin – Yang map

Navigating the Knowledge Base

The Yin-Yang Map



The Yin and the Yang of Research

You can use the figure above to find your way through the material in the Knowledge Base. Click on any part of the figure to move to that topic.

The figure shows one way of structuring the material in the Knowledge Base. The left side of the figure refers to the ***theory*** of research. The right side of the figure refers to the ***practice*** of research.

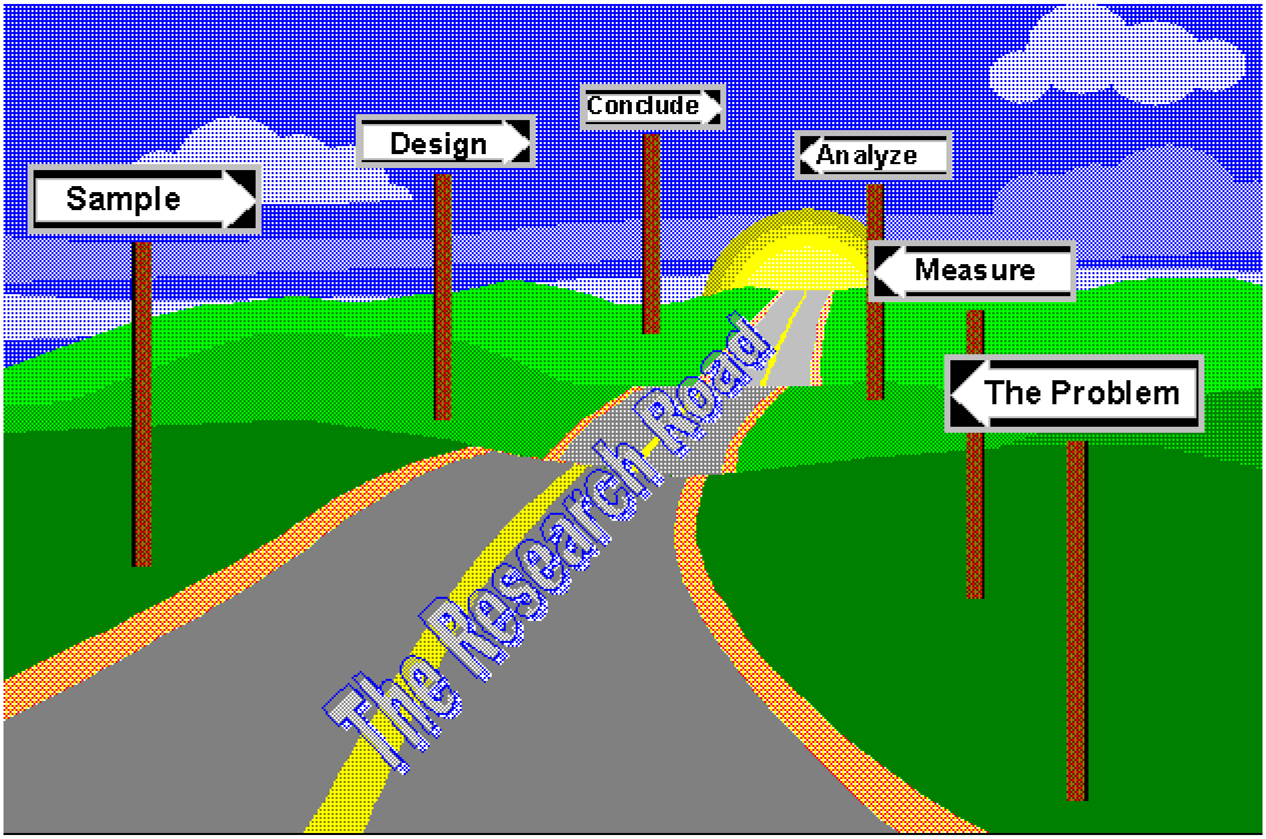
The yin-yang figure in the center links you to a ***theoretical*** introduction to research on the left and to the ***practical*** issue of how we formulate research projects on the right.

The four arrow links on the left describe the four ***types of validity*** in research. The idea of validity provides us with a unifying theory for understanding the criteria for good research. The four arrow links on the right point to the ***research practice areas*** that correspond with each validity type. For instance, external validity is related to the theory of how we generalize research results. Its corresponding practice area is sampling methodology which is concerned with how to draw representative samples so that generalizations are possible.

The Road Map

Navigating the Knowledge Base

The Road Map



The Road to Research

Remember all those Bob Hope and Bing Crosby films? The Road to Singapore? Of course you don't -- you're much too young! Well, I thought it might be useful to visualize the research endeavor sequentially, like taking a trip, like moving down a road -- the Road to Research. The figure above shows a very applied way to view the content of a research methods course that helps you consider the research process *practically*. You might visualize a research project as a journey where you must stop at certain points along your way. Every research project needs to start with a clear problem formulation. As you develop your project, you will find critical junctions where you will make choices about how you will proceed. Consider issues of sampling, measurement, design, and analysis - as well as the theories of validity behind each step. In the end, you will need to think about the whole picture, or "What can we conclude?" Then you might write-up your findings or report your evaluation. You even might find yourself backtracking and evaluating your previous decisions! Don't forget that this is a **two-way** road; planning and evaluation are critical and interdependent. The asphalt of the road is the foundation of research philosophy and practice. Without consideration of the basics in research, you'll find yourself bogged down in the mud!

Foundation

This section provides an overview the major issues in research and in evaluation. This is probably the best place for you to begin learning about research.

We have to begin somewhere. (Although, if you think about it, the whole idea of hyperlinked text sort of runs contrary to the notion that there is a single place to begin -- you can begin anywhere, go anywhere, and leave anytime. Unfortunately, you can only be in one place at a time and, even less fortunately for you, you happen to be right here right now, so we may as well consider this a place to begin.) And what better place to begin than an introduction? Here's where we take care of all the stuff you think you already know, and probably should already know, but most likely don't know as well as you think you do.

The first thing we have to get straight is the ***language of research***. If we don't, we're going to have a hard time discussing research.

With the basic terminology under our belts, we can look a little more deeply at some of the underlying ***philosophical issues*** that drive the research endeavor.

We also need to recognize that social research always occurs in a social context. It is a human endeavor. Therefore, it's important to consider the critical ***ethical issues*** that affect the researcher, research participants, and the research effort generally.

Where do research problems come from? How do we develop a research question? We consider these issues under ***conceptualization***.

Finally, we look at a specific, and very applied, type of social research known as ***evaluation research***.

That ought to be enough to get you started. At least it ought to be enough to get you thoroughly confused. But don't worry, there's stuff that's far more confusing than this yet to come.

Language of research

- **Five Big Words**

Research involves an eclectic blending of an enormous range of skills and activities. To be a good social researcher, you have to be able to work well with a wide variety of people, understand the specific methods used to conduct research, understand the subject that you are studying, be able to convince someone to give you the funds to study it, stay on track and on schedule, speak and write persuasively, and on and on.

Here, I want to introduce you to five terms that I think help to describe some of the key aspects of contemporary social research. (This list is not exhaustive. It's really just the first five terms that came into my mind when I was thinking about this and thinking about how I might be able to impress someone with really big/complex words to describe fairly straightforward concepts).

I present the first two terms -- ***theoretical*** and ***empirical*** -- together because they are often contrasted with each other. Social research is theoretical, meaning that much of it is concerned with developing, exploring or testing the theories or ideas that social researchers have about how the world operates. But it is also empirical, meaning that it is based on observations and measurements of reality -- on what we perceive of the world around us. You can even think of most research as a blending of these two terms -- a comparison of our theories about how the world operates with our observations of its operation.

The next term -- ***nomothetic*** -- comes (I think) from the writings of the psychologist Gordon Allport. Nomothetic refers to laws or rules that pertain to the general case (nomos in Greek) and is contrasted with the term "idiographic" which refers to laws or rules that relate to individuals (idiots in Greek???). In any event, the point here is that most social research is concerned with the nomothetic -- the general case -- rather than the individual. We often study individuals, but usually we are interested in generalizing to more than just the individual.

In our post-positivist view of science, we no longer regard certainty as attainable. Thus, the fourth big word that describes much contemporary social research is ***probabilistic***, or based on probabilities. The inferences that we make in social research have probabilities associated with them -- they are seldom meant to be considered covering laws that pertain to all cases. Part of the reason we have seen statistics become so dominant in social research is that it allows us to estimate probabilities for the situations we study.

The last term I want to introduce is ***causal***. You've got to be very careful with this term. Note that it is spelled *causal* not *casual*. You'll really be embarrassed if you write about the "casual hypothesis" in your study! The term causal means that most social research is interested (at some point) in looking at cause-effect relationships. This doesn't mean that most studies actually study cause-effect relationships. There are some studies that simply observe -- for instance, surveys that seek to describe the percent of people holding a particular opinion. And, there are many studies that explore relationships -- for example, studies that attempt to see whether there is a relationship between gender and salary. Probably the vast majority of applied social research consists of these descriptive and correlational studies. So why am I talking about causal studies? Because for most social sciences, it is important that we go beyond just looking at the world or looking at relationships. We would like to be able to change the world, to improve it and eliminate some of its major problems. If we want to change the world (especially if we want to do this in an organized, scientific way), we are automatically interested in causal relationships -- ones that tell us how our causes (e.g., programs, treatments) affect the outcomes of interest.

Types of Questions

There are three basic types of questions that research projects can address:

1. ***Descriptive*****:** When a study is designed primarily to describe what is going on or what exists. Public opinion polls that seek only to describe the proportion of people who hold various opinions are primarily descriptive in nature. For instance, if we want to know what percent of the population would vote for a Democratic or a Republican in the next presidential election, we are simply interested in describing something.
2. ***Relational*****:** When a study is designed to look at the relationships between two or more variables. A public opinion poll that compares what proportion of males and females say they would vote for a Democratic or a Republican candidate in the next presidential election is essentially studying the relationship between gender and voting preference.
3. ***Causal*****:** When a study is designed to determine whether one or more variables (e.g., a program or treatment variable) causes or affects one or more outcome variables. If we did a public opinion poll to try to determine whether a recent political advertising campaign changed voter preferences, we would essentially be studying whether the campaign (cause) changed the proportion of voters who would vote Democratic or Republican (effect).

The three question types can be viewed as cumulative. That is, a relational study assumes that you can first describe (by measuring or observing) each of the variables you are trying to relate. And, a causal study assumes that you can describe both the cause and effect variables and that you can show that they are related to each other. Causal studies are probably the most demanding of the three.

- **Time in Research**

Time is an important element of any research design, and here I want to introduce one of the most fundamental distinctions in research design nomenclature: ***cross-sectional*** versus ***longitudinal*** studies. A *cross-sectional* study is one that takes place at a single point in time. In effect, we are taking a 'slice' or cross-section of whatever it is we're observing or measuring. A *longitudinal* study is one that takes place over time -- we have at least two (and often more) waves of measurement in a longitudinal design.

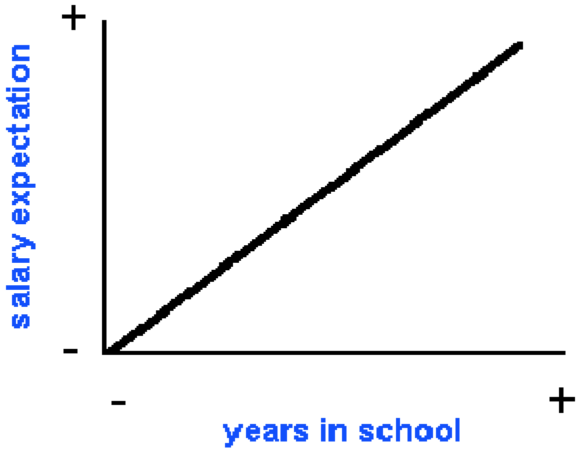
A further distinction is made between two types of longitudinal designs: ***repeated measures*** and ***time series***. There is no universally agreed upon rule for distinguishing these two terms, but in general, if you have two or a few waves of measurement, you are using a *repeated measures* design. If you have many waves of measurement over time, you have a *time series*. How many is 'many'? Usually, we wouldn't use the term time series unless we had at least twenty waves of measurement, and often far more. Sometimes the way we distinguish these is with the analysis methods we would use. Time series analysis requires that you have at least twenty or so observations. Repeated measures analyses (like repeated measures ANOVA) aren't often used with as many as twenty waves of measurement.

- **Types of Relationships**

A relationship refers to the *correspondence* between two variables. When we talk about types of relationships, we can mean that in at least two ways: the *nature* of the relationship or the *pattern* of it.

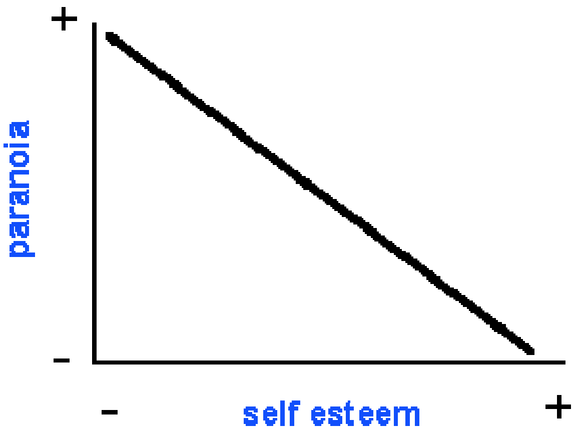
The Nature of a Relationship

While all relationships tell about the correspondence between two variables, there is a special type of relationship that holds that the two variables are not only in correspondence, but that one *causes* the other. This is the key distinction between a simple *correlational relationship* and a *causal relationship*. A correlational relationship simply says that two things perform in a synchronized manner. For instance, we often talk of a correlation between inflation and unemployment. When inflation is high, unemployment also tends to be high. When inflation is low, unemployment also tends to be low. The two variables are correlated. But knowing that two variables are correlated does not tell us whether one *causes* the other. We know, for instance, that there is a correlation between the number of roads built in Europe and the number of children born in the United States. Does that mean that is we want fewer children in the U.S., we should stop building so many roads in Europe? Or, does it mean that if we don't have enough roads in Europe, we should encourage U.S. citizens to have more babies? Of course not. (At least, I hope not). While there is a relationship between the number of roads built and the number of babies, we don't believe that the relationship is a *causal* one. This leads to consideration of what is often termed the *third variable problem*. In this example, it may be that there is a third variable that is causing both the building of roads and the birthrate, that is causing the correlation we observe. For instance, perhaps the general world economy is responsible for both. When the economy is good more roads are built in Europe and more children are born in the U.S. The key lesson here is that you have to be careful when you interpret correlations. If you observe a correlation between the number of hours students use the computer to study and their grade point averages (with high computer users getting higher grades), you *cannot* assume that the relationship is *causal*: that computer use improves grades. In this case, the third variable might be socioeconomic status -- richer students who have greater resources at their disposal tend to both use computers and do better in their grades. It's the resources that drives both use and grades, not computer use that causes the change in the grade point average.



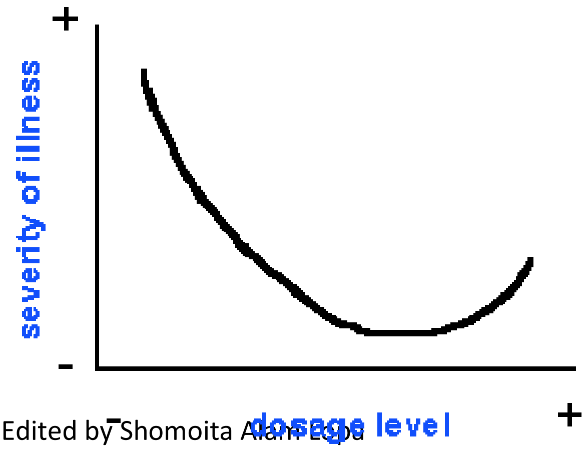
Patterns of Relationships

We have several terms to describe the major different types of patterns one might find in a relationship. First, there is the case of ***no relationship*** at all. If you know the values on one variable, you don't know anything about the values on the other. For instance, I suspect that there is no relationship between the length of the lifeline on your hand and your grade point average. If I know your GPA, I don't have any idea how long your lifeline is.



Then, we have the ***positive relationship***. In a positive relationship, high values on one variable are associated with high values on the other and low values on one are associated with low values on the other. In this example, we assume an idealized positive relationship between years of education and the salary one might expect to be making.

On the other hand a ***negative*** relationship implies that high values on one variable are associated with low values on the other. This is also sometimes termed an ***inverse*** relationship. Here, we show an idealized negative relationship between a measure of self esteem and a measure of paranoia in psychiatric patients.



These are the simplest types of relationships we might typically estimate in research. But the pattern of a relationship can be more complex than this. For instance, the figure on the left shows a relationship that changes over the range of both variables, a curvilinear relationship. In this example, the horizontal axis represents dosage of a drug for an illness and the vertical axis represents a severity of illness measure. As dosage rises, severity of illness goes down. But at some point, the patient begins to experience negative side effects associated with too high a dosage, and the severity of illness begins to increase again.

- **Variables**

You won't be able to do very much in research unless you know how to talk about variables. A ***variable*** is *any entity that can take on different values*. OK, so what does that mean? Anything that can vary can be considered a variable. For instance, *age* can be considered a variable because age can take different values for different people or for the same person at different times. Similarly, *country* can be considered a variable because a person's country can be assigned a value.

Variables aren't always 'quantitative' or numerical. The variable 'gender' consists of two text values: 'male' and 'female'. We can, if it is useful, assign quantitative values instead of (or in place of) the text values, but we don't have to assign numbers in order for something to be a variable. It's also important to realize that variables aren't only things that we measure in the traditional sense. For instance, in much social research and in program evaluation, we consider the treatment or program to be made up of one or more variables (i.e., the 'cause' can be considered a variable). An educational program can have varying amounts of 'time on task', 'classroom settings', 'student-teacher ratios', and so on. So even the program can be considered a variable (which can be made up of a number of sub-variables).

An ***attribute*** is a specific value on a variable. For instance, the variable *sex* or *gender* has two attributes: *male* and *female*. Or, the variable *agreement* might be defined as having five attributes:

- 1 = strongly disagree
- 2 = disagree
- 3 = neutral
- 4 = agree
- 5 = strongly agree

Another important distinction having to do with the term 'variable' is the distinction between an ***independent*** and ***dependent*** variable. This distinction is particularly relevant when you are investigating cause-effect relationships. It took me the longest time to learn this distinction. (Of course, I'm someone who gets confused about the signs for 'arrivals' and 'departures' at airports -- do I go to arrivals because I'm arriving at the airport or does the person I'm picking up go to arrivals because they're arriving on the plane!). I originally thought that an independent variable was one that would be free to vary or respond to some program or treatment, and that a dependent variable must be one that *depends* on my efforts (that is, it's the *treatment*). But this is entirely backwards! In fact ***the independent variable is what you (or nature) manipulates*** -- a treatment or program or cause. The ***dependent variable is what is affected by the independent variable*** -- your effects or outcomes. For example, if you are studying the effects of a new educational program on student achievement, the program is the independent variable and your measures of achievement are the dependent ones.

Finally, there are two traits of variables that should always be achieved. Each variable should be ***exhaustive***, it should include all possible answerable responses. For instance, if the variable is "religion" and the only options are "Protestant", "Jewish", and "Muslim", there are quite a few religions I can think of that haven't been included. The list does not exhaust all possibilities. On the other hand, if you exhaust all the possibilities with some variables -- religion being one of them -- you would simply have too many responses. The way to deal with this is to explicitly list the most common attributes and then use a general category like "Other" to account for all remaining ones. In addition to being exhaustive, the attributes of a variable should be ***mutually exclusive***, no respondent should be able to have two attributes simultaneously. While this might seem obvious, it is often rather tricky in practice. For instance, you might be tempted to represent the variable "Employment Status" with the two attributes "employed" and "unemployed." But these attributes are not necessarily mutually exclusive -- a person who is looking for a second job while employed would be able to check both attributes! But don't we often use questions on surveys that ask the respondent to "check all that apply" and then list a series of categories? Yes, we do, but technically speaking, each of the categories in a question like that is its own variable and is treated dichotomously as either "checked" or "unchecked", attributes that *are* mutually exclusive.

- **Hypotheses**

A hypothesis is a specific statement of prediction. It describes in concrete (rather than theoretical) terms what you expect will happen in your study. Not all studies have hypotheses. Sometimes a study is designed to be exploratory (see inductive research). There is no formal hypothesis, and perhaps the purpose of the study is to explore some area more thoroughly in order to develop some specific hypothesis or prediction that can be tested in future research. A single study may have one or many hypotheses.

Actually, whenever I talk about an hypothesis, I am really thinking simultaneously about *two* hypotheses. Let's say that you predict that there will be a relationship between two variables in your study. The way we would formally set up the hypothesis test is to formulate two hypothesis statements, one that describes your prediction and one that describes all the other possible outcomes with respect to the hypothesized relationship. Your prediction is that variable A and variable B will be related (you don't care whether it's a positive or negative relationship). Then the only other possible outcome would be that variable A and variable B are *not* related. Usually, we call the hypothesis that you support (your prediction) the ***alternative*** hypothesis, and we call the hypothesis that describes the remaining possible outcomes the ***null*** hypothesis. Sometimes we use a notation like H_A or H_1 to represent the alternative hypothesis or your prediction, and H_O

or H_0 to represent the null case. You have to be careful here, though. In some studies, your prediction might very well be that there will be no difference or change. In this case, you are essentially trying to find support for the null hypothesis and you are opposed to the alternative.

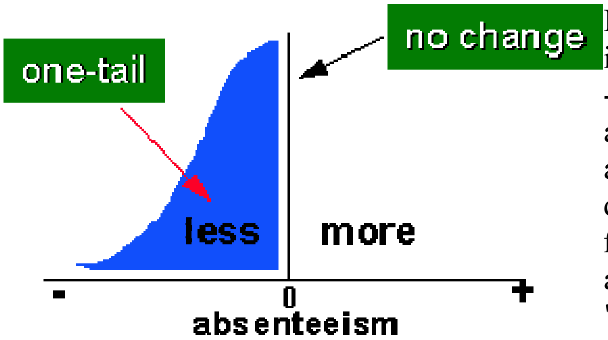
If your prediction specifies a direction, and the null therefore is the no difference prediction and the prediction of the opposite direction, we call this a ***one-tailed hypothesis***. For instance, let's imagine that you are investigating the effects of a new employee training program and that you believe one of the outcomes will be that there will be *less* employee absenteeism. Your two hypotheses might be stated something like this:

The null hypothesis for this study is:

H_O: As a result of the XYZ company employee training program, there will either be no significant difference in employee absenteeism or there will be a significant *increase*.

which is tested against the alternative hypothesis:

H_A: As a result of the XYZ company employee training program, there will be a significant *decrease* in employee absenteeism.



In the figure on the left, we see this situation illustrated graphically. The alternative hypothesis -- your prediction that the program will decrease absenteeism -- is shown there. The null must account for the other two possible conditions: no difference, or an increase in absenteeism. The figure shows a hypothetical distribution of absenteeism differences. We can see that the term "one-tailed" refers to the tail of the distribution on the outcome variable.

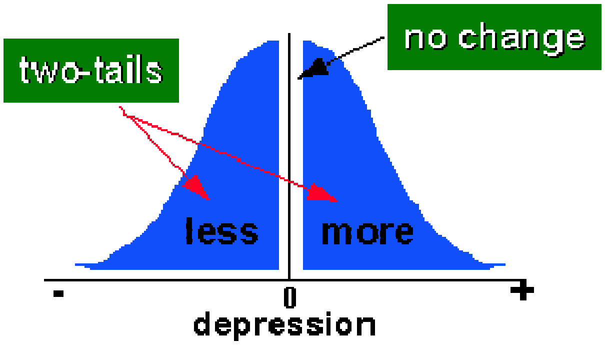
When your prediction does *not* specify a direction, we say you have a ***two-tailed hypothesis***. For instance, let's assume you are studying a new drug treatment for depression. The drug has gone through some initial animal trials, but has not yet been tested on humans. You believe (based on theory and the previous research) that the drug will have an effect, but you are not confident enough to hypothesize a direction and say the drug will reduce depression (after all, you've seen more than enough promising drug treatments come along that eventually were shown to have severe side effects that actually worsened symptoms). In this case, you might state the two hypotheses like this:

The null hypothesis for this study is:

H_O: As a result of 300mg./day of the ABC drug, there will be no significant difference in depression.

Which is tested against the alternative hypothesis:

H_A: As a result of 300mg./day of the ABC drug, there will be a significant difference in depression.



The figure on the right illustrates this two-tailed prediction for this case. Again, notice that the term "two-tailed" refers to the tails of the distribution for your outcome variable.

The important thing to remember about stating hypotheses is that you formulate your prediction (directional or not), and then you formulate a second hypothesis that is mutually exclusive of the first and incorporates all possible alternative outcomes for that case. When your study analysis is completed, the idea is that you will have to choose between the two hypotheses. If your prediction was correct, then you would (usually) reject the null hypothesis and accept the alternative. If your original prediction was not supported in the data, then you will accept the null hypothesis and reject the alternative. The logic of hypothesis testing is based on these two basic principles:

- the formulation of two mutually exclusive hypothesis statements that, together, exhaust all possible outcomes
- the testing of these so that one is necessarily accepted and the other rejected

OK, I know it's a convoluted, awkward and formalistic way to ask research questions. But it encompasses a long tradition in statistics called the ***hypothetical-deductive model***, and sometimes we just have to do things because they're traditions. And anyway, if all of this hypothesis testing was easy enough so anybody could understand it, how do you think statisticians would stay employed?

- **Types of Data**

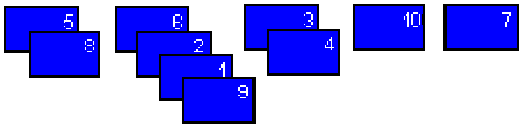
We'll talk about data in lots of places in The Knowledge Base, but here I just want to make a fundamental distinction between two types of data: **qualitative** and **quantitative**. The way we typically define them, we call data 'quantitative' if it is in numerical form and 'qualitative' if it is not. Notice that qualitative data could be much more than just words or text. Photographs, videos, sound recordings and so on, can be considered qualitative data.

Personally, while I find the distinction between qualitative and quantitative data to have some utility, I think most people draw too hard a distinction, and that can lead to all sorts of confusion. In some areas of social research, the qualitative-quantitative distinction has led to protracted arguments with the proponents of each arguing the superiority of their kind of data over the other. The quantitative types argue that their data is 'hard', 'rigorous', 'credible', and 'scientific'. The qualitative proponents counter that their data is 'sensitive', 'nuanced', 'detailed', and 'contextual'.

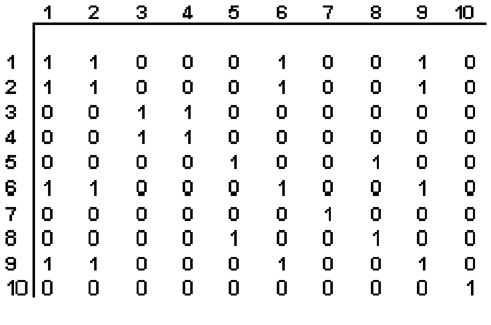
For many of us in social research, this kind of polarized debate has become less than productive. And, it obscures the fact that qualitative and quantitative data are intimately related to each other. ***All quantitative data is based upon qualitative judgments; and all qualitative data can be described and manipulated numerically.*** For instance, think about a very common quantitative measure in social research -- a self esteem scale. The researchers who develop such instruments had to make countless judgments in constructing them: how to define self esteem; how to distinguish it from other related concepts; how to word potential scale items; how to make sure the items would be understandable to the intended respondents; what kinds of contexts it could be used in; what kinds of cultural and language constraints might be present; and on and on. The researcher who decides to use such a scale in their study has to make another set of judgments: how well does the scale measure the intended concept; how reliable or consistent is it; how appropriate is it for the research context and intended respondents; and on and on. Believe it or not, even the respondents make many judgments when filling out such a scale: what is meant by various terms and phrases; why is the researcher giving this scale to them; how much energy and effort do they want to expend to complete it, and so on. Even the consumers and readers of the research will make lots of judgments about the self esteem measure and its appropriateness in that research context. What may look like a simple, straightforward, cut-and-dried quantitative measure is actually based on lots of qualitative judgments made by lots of different people.

On the other hand, all qualitative information can be easily converted into quantitative, and there are many times when doing so would add considerable value to your research. The simplest way to do this is to divide the qualitative information into units and number them! I know that sounds trivial, but even that simple nominal enumeration can enable you to organize and process qualitative information more efficiently. Perhaps more to the point, we might take text information (say, excerpts from transcripts) and pile these excerpts into piles of similar statements. When we do something even as easy as this simple grouping or piling task, we can describe the results quantitatively. For instance, if we had ten statements and we grouped these into five piles (as shown in the figure), we could describe the piles using a 10 x 10 table of **0**'s and **1**'s. If two statements were placed together in the same pile, we would put a **1** in their row-column juncture. If two statements were placed in different piles, we would use a **0**. The resulting matrix or table describes the grouping of the ten statements in terms of their similarity. Even though the data in this example consists of qualitative statements (one per card), the result of our simple qualitative procedure (grouping similar excerpts into the same piles) is *quantitative* in nature. "So what?" you ask. Once we have the data in numerical form, we can manipulate it numerically. For instance, we could have five different judges sort the 10 excerpts and obtain a 0-1 matrix like this for each judge. Then we could average the five matrices into a single one that shows the proportions of judges who grouped each pair together. This proportion could be

Sorting of 10 qualitative items



Binary Square Similarity Matrix for the sort



	1	2	3	4	5	6	7	8	9	10
1	1	1	0	0	0	1	0	0	1	0
2	1	1	0	0	0	1	0	0	1	0
3	0	0	1	1	0	0	0	0	0	0
4	0	0	1	1	0	0	0	0	0	0
5	0	0	0	0	1	0	0	1	0	0
6	1	1	0	0	0	1	0	0	1	0
7	0	0	0	0	0	0	1	0	0	0
8	0	0	0	0	1	0	0	1	0	0
9	1	1	0	0	0	1	0	0	1	0
10	0	0	0	0	0	0	0	0	0	1

considered an estimate of the similarity (across independent judges) of the excerpts. While this might not seem too exciting or useful, it is exactly this kind of procedure that I use as an integral part of the process of developing 'concept maps' of ideas for groups of people (something that *is* useful!).

Unit of Analysis

One of the most important ideas in a research project is the ***unit of analysis***. The unit of analysis is the major entity that you are analyzing in your study. For instance, any of the following could be a unit of analysis in a study:

- individuals
- groups
- artifacts (books, photos, newspapers)
- geographical units (town, census tract, state)
- social interactions (dyadic relations, divorces, arrests)

Why is it called the 'unit of analysis' and not something else (like, the unit of sampling)? Because *it is the analysis you do in your study that determines what the unit is*. For instance, if you are comparing the children in two classrooms on achievement test scores, the unit is the individual child because you have a score for each child. On the other hand, if you are comparing the two classes on classroom climate, your unit of analysis is the group, in this case the classroom, because you only have a classroom climate score for the class as a whole and not for each individual student. For different analyses in the same study you may have different units of analysis. If you decide to base an analysis on student scores, the individual is the unit. But you might decide to compare average classroom performance. In this case, since the data that goes into the analysis is the average itself (and not the individuals' scores) the unit of analysis is actually the group. Even though you had data at the student level, you use aggregates in the analysis. In many areas of social research these hierarchies of analysis units have become particularly important and have spawned a whole area of statistical analysis sometimes referred to as ***hierarchical modeling***. This is true in education, for instance, where we often compare classroom performance but collected achievement data at the individual student level.

Two Research Fallacies

A *fallacy* is an error in reasoning, usually based on mistaken assumptions. Researchers are very familiar with all the ways they could go wrong, with the fallacies they are susceptible to. Here, I discuss two of the most important.

The ***ecological fallacy*** occurs when you make conclusions about individuals based only on analyses of group data. For instance, assume that you measured the math scores of a particular classroom and found that they had the highest average score in the district. Later (probably at the mall) you run into one of the kids from that class and you think to yourself "she must be a math whiz." Aha! Fallacy! Just because she comes from the class with the highest *average* doesn't mean that she is automatically a high-scorer in math. She could be the lowest math scorer in a class that otherwise consists of math geniuses!

An ***exception fallacy*** is sort of the reverse of the ecological fallacy. It occurs when you reach a group conclusion on the basis of exceptional cases. This is the kind of fallacious reasoning that is at the core of a lot of sexism and racism. The stereotype is of the guy who sees a woman make a driving error and concludes that "women are terrible drivers." Wrong! Fallacy!

Both of these fallacies point to some of the traps that exist in both research and everyday reasoning. They also point out how important it is that we do research. We need to determine empirically how individuals perform (not just rely on group averages). Similarly, we need to look at whether there are correlations between certain behaviors and certain groups (you might look at the whole controversy around the book *The Bell Curve* as an attempt to examine whether the supposed relationship between race and IQ is real or a fallacy.

Philosophy of Research

You probably think of research as something very abstract and complicated. It can be, but you'll see (I hope) that if you understand the different parts or phases of a research project and how these fit together, it's not nearly as complicated as it may seem at first glance. A research project has a well-known structure -- a beginning, middle and end. We introduce the basic ***phases*** of a research project in ***The Structure of Research***. In that section, we also introduce some important distinctions in research: the different ***types of questions*** you can ask in a research project; and, the major ***components*** or parts of a research project.

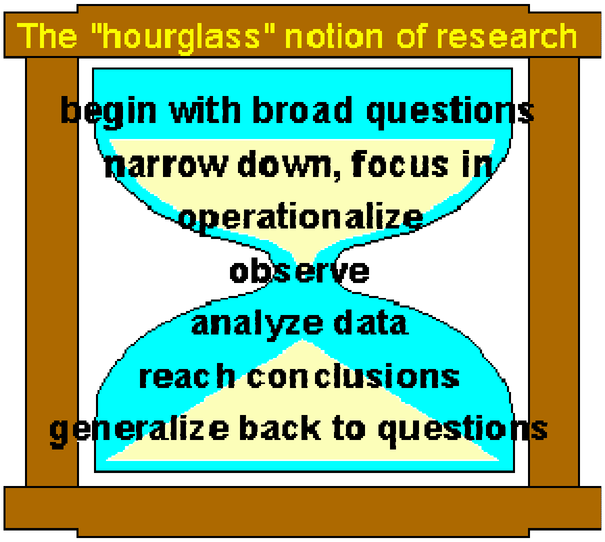
Before the modern idea of research emerged, we had a term for what philosophers used to call research -- logical reasoning. So, it should come as no surprise that some of the basic distinctions in logic have carried over into contemporary research. In ***Systems of Logic*** we discuss how two major logical systems, the inductive and deductive methods of reasoning, are related to modern research.

OK, you knew that no introduction would be complete without considering something having to do with assumptions and philosophy. (I thought I very cleverly snuck in the stuff about logic in the last paragraph). All research is based on assumptions about how the world is perceived and how we can best come to understand it. Of course, nobody really *knows* how we can best understand the world, and philosophers have been arguing about that very question for at least two millennia now, so all we're going to do is look at how most contemporary social scientists approach the question of how we know about the world around us. We consider two major philosophical schools of thought -- ***Positivism and Post-Positivism*** -- that are especially important perspectives for contemporary social research (OK, I'm only considering positivism and post-positivism here because these are the major schools of thought. Forgive me for not considering the hotly debated alternatives like relativism, subjectivism, hermeneutics, deconstructivism, constructivism, feminism, etc. If you really want to cover that stuff, start your own Web site and send me your URL to stick in here).

Quality is one of the most important issues in research. We introduce the idea of ***validity*** to refer to the quality of various conclusions you might reach based on a research project. Here's where I've got to give you the pitch about validity. When I mention validity, most students roll their eyes, curl up into a fetal position or go to sleep. They think validity is just something abstract and philosophical (and I guess it is at some level). But I think if you can understand validity -- the principles that we use to judge the quality of research -- you'll be able to do much more than just complete a research project. You'll be able to be a virtuoso at research, because you'll have an understanding of *why* we need to do certain things in order to assure quality. You won't just be plugging in standard procedures you learned in school -- sampling method X, measurement tool Y -- you'll be able to help create the next generation of research technology. Enough for now -- more on this later.

- **Structure of Research**

Most research projects share the same general structure. You might think of this structure as following the shape of an hourglass. The research process usually starts with a broad area of interest, the initial problem that the researcher wishes to study. For instance, the researcher could be interested in how to use computers to improve the performance of students in mathematics. But this initial interest is far too broad to study in any single research project (it might not even be addressable in a lifetime of research). The researcher has to narrow the question down to one that can reasonably be studied in a research project. This might involve formulating a hypothesis or a focus question. For instance, the researcher might hypothesize that a particular method of computer instruction in math will improve the ability of elementary school students in a specific district. At the narrowest point of the research hourglass, the researcher is engaged in direct measurement or observation of the question of interest.


The "hourglass" notion of research
begin with broad questions
narrow down, focus in
operationalize
observe
analyze data
reach conclusions
generalize back to questions

Once the basic data is collected, the researcher begins to try to understand it, usually by analyzing it in a variety of ways. Even for a single hypothesis there are a number of analyses a researcher might typically conduct. At this point, the researcher begins to formulate some initial conclusions about what happened as a result of the computerized math program. Finally, the researcher often will attempt to address the original broad question of interest by generalizing from the results of this specific study to other related situations. For instance, on the basis of strong results indicating that the math program had a positive effect on student performance, the researcher might conclude that other school districts similar to the one in the study might expect similar results.

Components of a Study

What are the basic components or parts of a research study? Here, we'll describe the basic components involved in a causal study. Because causal studies presuppose descriptive and relational questions, many of the components of causal studies will also be found in those others.

Most social research originates from some general ***problem*** or question. You might, for instance, be interested in what programs enable the unemployed to get jobs. Usually, the problem is broad enough that you could not hope to address it adequately in a single research study. Consequently, we typically narrow the problem down to a more specific **research question** that we can hope to address. The research question is often stated in the context of some theory that has been advanced to address the problem. For instance, we might have the theory that ongoing support services are needed to assure that the newly employed remain employed. The research question is the central issue being addressed in the study and is often phrased in the language of theory. For instance, a research question might be:

Is a program of supported employment more effective (than no program at all) at keeping newly employed persons on the job?

The problem with such a question is that it is still too general to be studied directly. Consequently, in most research we develop an even more specific statement, called an ***hypothesis*** that describes in *operational* terms exactly what we think will happen in the study. For instance, the hypothesis for our employment study might be something like:

The Metropolitan Supported Employment Program will significantly increase rates of employment after six months for persons who are newly employed (after being out of work for at least one year) compared with persons who receive no comparable program.

Notice that this hypothesis is specific enough that a reader can understand quite well what the study is trying to assess.

In causal studies, we have at least two major variables of interest, *the* ***cause*** *and the* ***effect***. Usually the cause is some type of event, program, or treatment. We make a distinction between causes that the researcher can control (such as a program) versus causes that occur naturally or outside the researcher's influence (such as a change in interest rates, or the occurrence of an earthquake). The effect is the outcome that you wish to study. For both the cause and effect we make a distinction between our idea of them (the construct) and how they are actually manifested in reality. For instance, when we think about what a program of support services for the newly employed might be, we are thinking of the "*construct.*" On the other hand, the real world is not always what we think it is. In research, we remind ourselves of this by distinguishing our view of an entity (the construct) from the entity as it exists (the operationalization). Ideally, we would like the two to agree.

Social research is always conducted in a social context. We ask people questions, or observe families interacting, or measure the opinions of people in a city. An important component of a research project is the ***units*** that participate in the project. Units are directly related to the question of sampling. In most projects, we cannot involve all of the people we might like to involve. For instance, in studying a program of support services for the newly employed we can't possibly include in our study everyone in the world, or even in the country, who is newly employed. Instead, we have to try to obtain a representative sample of such people. When ***sampling***, we make a distinction between the theoretical population of interest to our study and the final sample that we actually measure in our study. Usually the term "units" refers to the *people* that we sample and from whom we gather information. But for some projects the units are organizations, groups, or geographical entities like cities or towns. Sometimes our sampling strategy is multi-level: we sample a number of cities and within them sample families.

In causal studies, we are interested in the effects of some cause on one or more **outcomes**. The outcomes are directly related to the research problem -- we are usually most interested in outcomes that are most reflective of the problem. In our hypothetical supported employment study, we would probably be most interested in measures of employment -- is the person currently employed, or, what is their rate of absenteeism.

Finally, in a causal study we usually are comparing the effects of our cause of interest (e.g., the program) relative to other conditions (e.g., another program or no program at all). Thus, a key component in a causal study concerns how we decide what units (e.g., people) receive our program and which are placed in an alternative condition. This issue is directly related to the ***research design*** that we use in the study. One of the central questions in research design is determining how people wind up in or are placed in various programs or treatments that we are comparing.

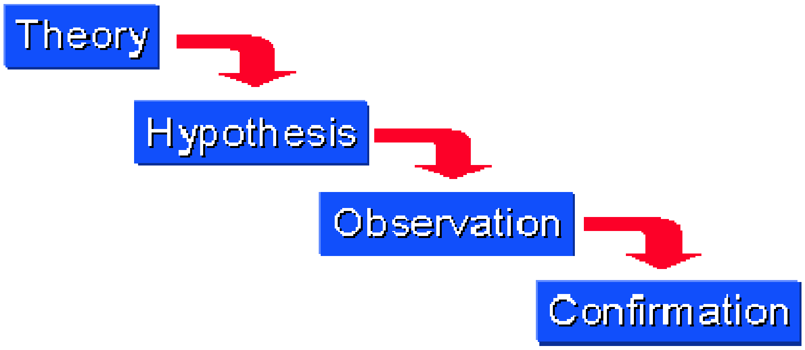
These, then, are the major components in a causal study:

- The Research Problem
- The Research Question
- The Program (Cause)
- The Units
- The Outcomes (Effect)
- The Design

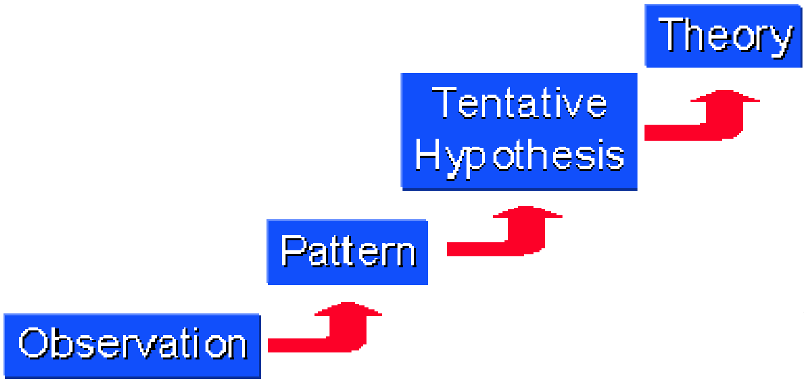
- **Deduction & Induction**

Deductive and Inductive Thinking

In logic, we often refer to the two broad methods of reasoning as the ***deductive*** and ***inductive*** approaches.



Deductive reasoning works from the more general to the more specific. Sometimes this is informally called a "top-down" approach. We might begin with thinking up a *theory* about our topic of interest. We then narrow that down into more specific *hypotheses* that we can test. We narrow down even further when we collect *observations* to address the hypotheses. This ultimately leads us to be able to test the hypotheses with specific data -- a *confirmation* (or not) of our original theories.



Inductive reasoning works the other way, moving from specific observations to broader generalizations and theories. Informally, we sometimes call this a "bottom up" approach (please note that it's "bottom up" and *not* "bottoms up" which is the kind of thing the bartender says to customers when he's trying to close for the night!). In inductive reasoning, we begin with specific observations and measures, begin to detect patterns and regularities, formulate some tentative hypotheses that we can explore, and finally end up developing some general conclusions or theories.

These two methods of reasoning have a very different "feel" to them when you're conducting research. Inductive reasoning, by its very nature, is more open-ended and exploratory, especially at the beginning. Deductive reasoning is more narrow in nature and is concerned with testing or confirming hypotheses. Even though a particular study may look like it's purely deductive (e.g., an experiment designed to test the hypothesized effects of some treatment on some outcome), most social research involves both inductive and deductive reasoning processes at some time in the project. In fact, it doesn't take a rocket scientist to see that we could assemble the two graphs above into a single circular one that continually cycles from theories down to observations and back up again to theories. Even in the most constrained experiment, the researchers may observe patterns in the data that lead them to develop new theories.

- **Positivism & Post-Positivism**

Let's start our very brief discussion of philosophy of science with a simple distinction between ***epistemology*** and ***methodology***. The term epistemology comes from the Greek word epistêmê, their term for knowledge. In simple terms, epistemology is the philosophy of knowledge or of how we come to know. Methodology is also concerned with how we come to know, but is much more practical in nature. Methodology is focused on the specific ways -- the methods -- that we can use to try to understand our world better. Epistemology and methodology are intimately related: the former involves the *philosophy* of how we come to know the world and the latter involves the *practice*.

When most people in our society think about science, they think about some guy in a white lab coat working at a lab bench mixing up chemicals. They think of science as boring, cut-and-dry, and they think of the scientist as narrow-minded and esoteric (the ultimate nerd -- think of the humorous but nonetheless mad scientist in the *Back to the Future* movies, for instance). A lot of our stereotypes about science come from a period where science was dominated by a particular philosophy -- ***positivism*** -- that tended to support some of these views. Here, I want to suggest (no matter what the movie industry may think) that science has moved on in its thinking into an era of ***post-positivism*** where many of those stereotypes of the scientist no longer hold up.

Let's begin by considering what positivism is. In its broadest sense, positivism is a rejection of metaphysics (I leave it you to look up that term if you're not familiar with it). It is a position that holds that the goal of knowledge is simply to describe the phenomena that we experience. The purpose of science is simply to stick to what we can observe and measure. Knowledge of anything beyond that, a positivist would hold, is impossible. When I think of positivism (and the related philosophy of logical positivism) I think of the behaviorists in mid-20th Century psychology. These were the mythical 'rat runners' who believed that psychology could only study what could be directly observed and measured. Since we can't directly observe emotions, thoughts, etc. (although we may be able to measure some of the physical and physiological accompaniments), these were not legitimate topics for a scientific psychology. B.F. Skinner argued that psychology needed to concentrate only on the positive and negative reinforcers of behavior in order to predict how people will behave -- everything else in between (like what the person is thinking) is irrelevant because it can't be measured.

In a positivist view of the world, science was seen as the way to get at truth, to understand the world well enough so that we might predict and control it. The world and the universe were deterministic -- they operated by laws of cause and effect that we could discern if we applied the unique approach of the scientific method. Science was largely a mechanistic or mechanical affair. We use deductive reasoning to postulate theories that we can test. Based on the results of our studies, we may learn that our theory doesn't fit the facts well and so we need to revise our theory to better predict reality. The positivist believed in *empiricism* -- the idea that observation and measurement was the core of the scientific endeavor. The key approach of the scientific method is the experiment, the attempt to discern natural laws through direct manipulation and observation.

OK, I am exaggerating the positivist position (although you may be amazed at how close to this some of them actually came) in order to make a point. Things have changed in our views of science since the middle part of the 20th century. Probably the most important has been our shift away from positivism into what we term *post-positivism*. By post-positivism, I don't mean a slight adjustment to or revision of the positivist position -- post-positivism is a wholesale rejection of the central tenets of positivism. A post-positivist might begin by recognizing that the way scientists think and work and the way we think in our everyday life are not distinctly different. Scientific reasoning and common sense reasoning are essentially the same process. There is no difference in kind between the two, only a difference in degree. Scientists, for example, follow specific procedures to assure that observations are verifiable, accurate and consistent. In everyday reasoning, we don't always proceed so carefully (although, if you think about it, when the stakes are high, even in everyday life we become much more cautious about measurement. Think of the way most responsible parents keep continuous watch over their infants, noticing details that non-parents would never detect).

One of the most common forms of post-positivism is a philosophy called ***critical realism***. A critical realist believes that there is a reality independent of our thinking about it that science can study. (This is in contrast with a ***subjectivist*** who would hold that there is no external reality -- we're each making this all up!). Positivists were also realists. The difference is that the post-positivist critical realist recognizes that all observation is fallible and has error and that all theory is revisable. In other words, the critical realist is *critical* of our ability to know reality with certainty. Where the positivist believed that the goal of science was to uncover the truth, the post-positivist critical realist believes that *the goal of science is to hold steadfastly to the goal of getting it right about reality, even though we can never achieve that goal*! Because all measurement is fallible, the post-positivist emphasizes the importance of multiple measures and observations, each of which may possess different types of error, and the need to use ***triangulation*** across these multiple errorful sources to try to get a better bead on what's happening in reality. The post-positivist also believes that all observations are theory-laden and that scientists (and everyone else, for that matter) are inherently biased by their cultural experiences, world views, and so on. This is not cause to give up in despair, however. Just because I have my world view based on my experiences and you have yours doesn't mean that we can't hope to translate from each other's experiences or understand each other. That is, post-positivism rejects the ***relativist*** idea of the ***incommensurability*** of different perspectives, the idea that we can never understand each other because we come from different experiences and cultures. Most post-positivists are ***constructivists*** who believe that we each construct our view of the world based on our perceptions of it. Because perception and observation is fallible, our constructions must be imperfect. So what is meant by ***objectivity*** in a post-positivist world? Positivists believed that objectivity was a characteristic that resided in the individual scientist. Scientists are responsible for putting aside their biases and beliefs and seeing the world as it 'really' is. Post-positivists reject the idea that any individual can see the world perfectly as it really is. We are all biased and all of our observations are affected (theory-laden). Our best hope for achieving objectivity is to triangulate across multiple fallible perspectives! Thus, objectivity is not the characteristic of an individual, it is inherently a social phenomenon. It is what multiple individuals are trying to achieve when they criticize each other's work. We never achieve objectivity perfectly, but we can approach it. The best way for us to improve the objectivity of what we do is to do it within the context of a broader contentious community of truth-seekers (including other scientists) who criticize each other's work. The theories that survive such intense scrutiny are a bit like the species that survive in the evolutionary struggle. (This is sometimes called the ***natural selection theory of knowledge*** and holds that ideas have 'survival value' and that knowledge evolves through a process of variation, selection and retention). They have adaptive value and are probably as close as our species can come to being objective and understanding reality.

Clearly, all of this stuff is not for the faint-of-heart. I've seen many a graduate student get lost in the maze of philosophical assumptions that contemporary philosophers of science argue about. And don't think that I believe this is not important stuff. But, in the end, I tend to turn pragmatist on these matters. Philosophers have been debating these issues for thousands of years and there is every reason to believe that they will continue to debate them for thousands of years more. Those of us who are practicing scientists should check in on this debate from time to time (perhaps every hundred years or so would be about right). We should think about the assumptions we make about the world when we conduct research. But in the meantime, we can't wait for the philosophers to settle the matter. After all, we do have our own work to do!

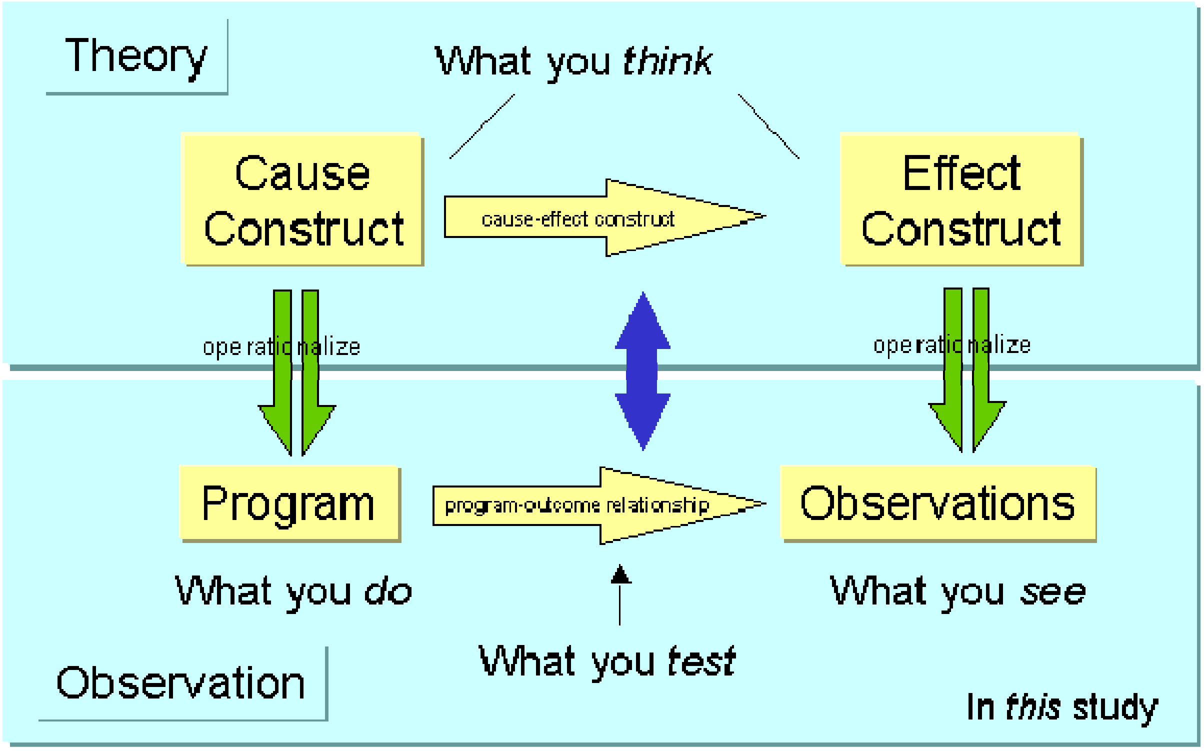
- **Introduction to Validity**

Validity:
the best available approximation to the truth of a given proposition, inference, or conclusion

The first thing we have to ask is: "validity of *what*?" When we think about validity in research, most of us think about research components. We might say that a measure is a valid one, or that a valid sample was drawn, or that the design had strong validity. But all of those statements are technically incorrect. Measures, samples and designs don't 'have' validity -- only propositions can be said to be valid. Technically, we should say that a measure leads to valid conclusions or that a sample enables valid inferences, and so on. It is a proposition, inference or conclusion that can 'have' validity.

We make lots of different inferences or conclusions while conducting research. Many of these are related to the process of doing research and are not the major hypotheses of the study. Nevertheless, like the bricks that go into building a wall, these intermediate process and methodological propositions provide the foundation for the substantive conclusions that we wish to address. For instance, virtually all-social research involves measurement or observation. And, whenever we measure or observe we are concerned with whether we are measuring what we intend to measure or with how our observations are influenced by the circumstances in which they are made. We reach conclusions about the quality of our measures -- conclusions that will play an important role in addressing the broader substantive issues of our study. When we talk about the validity of research, we are often referring to these to the many conclusions we reach about the quality of different parts of our research methodology.

We subdivide validity into four types. Each type addresses a specific methodological question. In order to understand the types of validity, you have to know something about how we investigate a research question. Because all four validity types are really only operative when studying causal questions, we will use a causal study to set the context.



The figure shows that there are really two realms that are involved in research. The first, on the top, is the land of theory. It is what goes on inside our heads as researchers. It is were we keep our theories about how the world operates. The second, on the bottom, is the land of observations. It is the real world into which we translate our ideas -- our programs, treatments, measures and observations. When we conduct research, we are continually flitting back and forth between these two realms, between what we think about the world and what is going on in it. When we are investigating a cause-effect relationship, we have a theory (implicit or otherwise) of what the cause is (***the cause construct***). For instance, if we are testing a new educational program, we have an idea of what it would look like ideally. Similarly, on the effect side, we have an idea of what we are ideally trying to affect and measure (***the effect construct***). But each of these, the cause and the effect, has to be translated into real things, into a program or treatment and a measure or observational method. We use the term ***operationalization*** to describe the act of translating a construct into its manifestation. In effect, we take our idea and describe it as a series of operations or procedures. Now, instead of it only being an idea in our minds, it becomes a public entity that anyone can look at and examine for themselves. It is one thing, for instance, for you to say that you would like to measure self-esteem (a construct). But when you show a ten-item paper-and-pencil self-esteem measure that you developed for that purpose, others can look at it and understand more clearly what you intend by the term self-esteem.

Now, back to explaining the four validity types. They build on one another, with two of them (conclusion and internal) referring to the land of observation on the bottom of the figure, one of them (construct) emphasizing the linkages between the bottom and the top, and the last (external) being primarily concerned about the range of our theory on the top. Imagine that we wish to examine whether use of a World Wide Web (WWW) Virtual Classroom improves student understanding of course material. Assume that we took these two constructs, the cause construct (the WWW site) and the effect (understanding), and operationalized them -- turned them into realities by constructing the WWW site and a measure of knowledge of the course material. Here are the four validity types and the question each addresses:

Conclusion Validity: In this study, is there a ***relationship*** between the two variables?

In the context of the example we're considering, the question might be worded: in this study, is there a relationship between the WWW site and knowledge of course material? There are several conclusions or inferences we might draw to answer such a question. We could, for example, conclude that there is a relationship. We might conclude that there is a positive relationship. We might infer that there is no relationship. We can assess the conclusion validity of each of these conclusions or inferences.

Internal Validity: *Assuming that there is a relationship in this study,* is the relationship a ***causal*** one?

Just because we find that use of the WWW site and knowledge are correlated, we can't necessarily assume that WWW site use *causes* the knowledge. Both could, for example, be caused by the same factor. For instance, it may be that wealthier students who have greater resources would be more likely to use have access to a WWW site and would excel on objective tests. When we want to make a claim that our program or treatment caused the outcomes in our study, we can consider the internal validity of our causal claim.

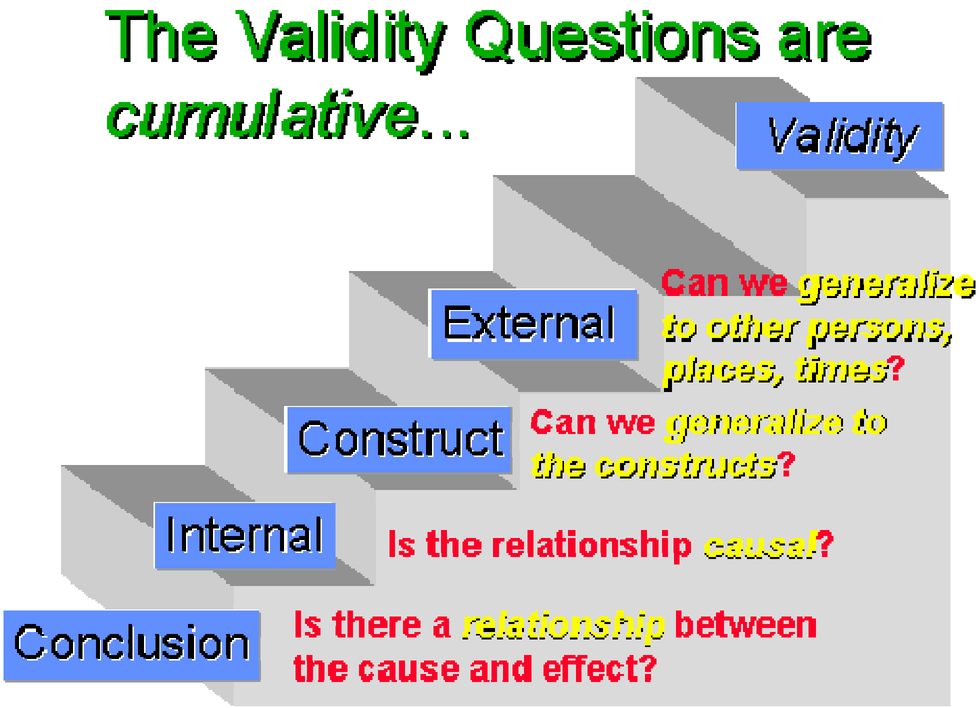
Construct Validity: *Assuming that there is a causal relationship in this study*, can we claim that the program reflected well our ***construct*** of the program and that our measure reflected well our idea of the ***construct*** of the measure?

In simpler terms, did we implement the program we intended to implement and did we measure the outcome we wanted to measure? In yet other terms, did we operationalize well the ideas of the cause and the effect? When our research is over, we would like to be able to conclude that we did a credible job of operationalizing our constructs -- we can assess the construct validity of this conclusion.

External Validity: *Assuming that there is a causal relationship in this study between the constructs of the cause and the effect*, can we ***generalize*** this effect to other persons, places or times?

We are likely to make some claims that our research findings have implications for other groups and individuals in other settings and at other times. When we do, we can examine the external validity of these claims.

Notice how the question that each validity type addresses presupposes an affirmative answer to the previous one. This is what we mean when we say that the validity types build on one another. The figure shows the idea of cumulativeness as a staircase, along with the key question for each validity type.


The Validity Questions are cumulative...
Validity
External
Can we generalize to other persons, places, times?
Construct
Can we generalize to the constructs?
Internal
Is the relationship causal?
Conclusion
Is there a relationship between the cause and effect?

For any inference or conclusion, there are always possible ***threats to validity*** -- reasons the conclusion or inference might be wrong. Ideally, one tries to reduce the plausibility of the most likely threats to validity, thereby leaving as most plausible the conclusion reached in the study. For instance, imagine a study examining whether there is a relationship between the amount of training in a specific technology and subsequent rates of use of that technology. Because the interest is in a relationship, it is considered an issue of conclusion validity. Assume that the study is completed and no significant correlation between amount of training and adoption rates is found. On this basis it is *concluded* that there is no relationship between the two. How could this conclusion be wrong -- that is, what are the "threats to validity"? For one, it's possible that there isn't sufficient statistical power to detect a relationship even if it exists. Perhaps the sample size is too small or the measure of amount of training is unreliable. Or maybe assumptions of the correlational test are violated given the variables used. Perhaps there were random irrelevancies in the study setting or random heterogeneity in the respondents that increased the variability in the data and made it harder to see the relationship of interest. The inference that there is no relationship will be stronger -- have greater conclusion validity -- if one can show that these alternative explanations are not credible. The distributions might be examined to see if they conform with assumptions of the statistical test, or analyses conducted to determine whether there is sufficient statistical power.

The theory of validity, and the many lists of specific threats, provide a useful scheme for assessing the quality of research conclusions. The theory is general in scope and applicability, well-articulated in its philosophical suppositions, and virtually impossible to explain adequately in a few minutes. As a framework for judging the quality of evaluations it is indispensable and well worth understanding.

Ethics in Research

We are going through a time of profound change in our understanding of the ethics of applied social research. From the time immediately after World War II until the early 1990s, there was a gradually developing consensus about the key ethical principles that should underlie the research endeavor. Two marker events stand out (among many others) as symbolic of this consensus. The Nuremberg War Crimes Trial following World War II brought to public view the ways German scientists had used captive human subjects as subjects in oftentimes gruesome experiments. In the 1950s and 1960s, the Tuskegee Syphilis Study involved the withholding of known effective treatment for syphilis from African-American participants who were infected. Events like these forced the reexamination of ethical standards and the gradual development of a consensus that potential human subjects needed to be protected from being used as 'guinea pigs' in scientific research.

By the 1990s, the dynamics of the situation changed. Cancer patients and persons with AIDS fought publicly with the medical research establishment about the long time needed to get approval for and complete research into potential cures for fatal diseases. In many cases, it is the ethical assumptions of the previous thirty years that drive this 'go-slow' mentality. After all, we would rather risk denying treatment for a while until we achieve enough confidence in a treatment, rather than run the risk of harming innocent people (as in the Nuremberg and Tuskegee events). But now, those who were threatened with fatal illness were saying to the research establishment that they *wanted* to be test subjects, even under experimental conditions of considerable risk. You had several very vocal and articulate patient groups who wanted to be experimented on coming up against an ethical review system that was designed to protect them from being experimented on.

Although the last few years in the ethics of research have been tumultuous ones, it is beginning to appear that a new consensus is evolving that involves the stakeholder groups most affected by a problem participating more actively in the formulation of guidelines for research. While it's not entirely clear, at present, what the new consensus will be, it is almost certain that it will not fall at either extreme: protecting against human experimentation at all costs **vs.** allowing anyone who is willing to be experimented on.

Ethical Issues

There are a number of key phrases that describe the system of ethical protections that the contemporary social and medical research establishment have created to try to protect better the rights of their research participants. The principle of ***voluntary participation*** requires that people not be coerced into participating in research. This is especially relevant where researchers had previously relied on 'captive audiences' for their subjects -- prisons, universities, and places like that. Closely related to the notion of voluntary participation is the requirement of ***informed consent***. Essentially, this means that prospective research participants must be fully informed about the procedures and risks involved in research and must give their consent to participate.

Ethical standards also require that researchers not put participants in a situation where they might be at ***risk of harm*** as a result of their participation. Harm can be defined as both physical and psychological. There are two standards that are applied in order to help protect the privacy of research participants. Almost all research guarantees the participants ***confidentiality*** -- they are assured that identifying information will not be made available to anyone who is not directly involved in the study. The stricter standard is the principle of ***anonymity*** which essentially means that the participant will remain anonymous throughout the study -- even to the researchers themselves. Clearly, the anonymity standard is a stronger guarantee of privacy, but it is sometimes difficult to accomplish, especially in situations where participants have to be measured at multiple time points (e.g., a pre-post study). Increasingly, researchers have had to deal with the ethical issue of a person's ***right to service***. Good research practice often requires the use of a no-treatment control group -- a group of participants who do *not* get the treatment or program that is being studied. But when that treatment or program may have beneficial effects, persons assigned to the no-treatment control may feel their rights to equal access to services are being curtailed.

Even when clear ethical standards and principles exist, there will be times when the need to do accurate research runs up against the rights of potential participants. No set of standards can possibly anticipate every ethical circumstance. Furthermore, there needs to be a procedure that assures that researchers will consider all relevant ethical issues in formulating research plans. To address such needs most institutions and organizations have formulated an ***Institutional Review Board (IRB)***, a panel of persons who reviews grant proposals with respect to ethical implications and decides whether additional actions need to be taken to assure the safety and rights of participants. By reviewing proposals for research, IRBs also help to protect both the organization and the researcher against potential legal implications of neglecting to address important ethical issues of participants.

Conceptualizing

One of the most difficult aspects of research -- and one of the least discussed -- is how to develop the idea for the research project in the first place. In training students, most faculty just assume that if you read enough of the research in an area of interest, you will somehow magically be able to produce sensible ideas for further research. Now, that may be true. And heaven knows that's the way we've been doing this higher education thing for some time now. But it troubles me that we haven't been able to do a better job of helping our students learn ***how*** to formulate good research problems. One thing we can do (and some texts at least cover this at a surface level) is to give students a better idea of how professional researchers typically generate research ideas. Some of this is introduced in the discussion of ***problem formulation in applied social research.***

But maybe we can do even better than that. Why can't we turn some of our expertise in developing methods into methods that students and researchers can use to help them formulate ideas for research. I've been working on that area pretty intensively for over a decade now -- I came up with a structured approach that groups can use to map out their ideas on any topic. This approach, called ***concept mapping*** can be used by research teams to help them clarify and map out the key research issues in an area, to help them operationalize the programs or interventions or the outcome measures for their study. The concept mapping method isn't the only method around that might help researchers formulate good research problems and projects. Virtually any method that's used to help individuals and groups to think more effectively would probably be useful in research formulation. Some of the methods that might be included in our toolkit for research formulation might be: brainstorming, brainwriting, nominal group technique, focus groups, Delphi methods, and facet theory. And then, of course, there are all of the methods for identifying relevant literature and previous research work. If you know of any techniques or methods that you think might be useful when formulating the research problem, please feel free to add a notation -- if there's a relevant Website, please point to it in the notation.

- **Problem Formulation**

"Well begun is half done" --Aristotle, quoting an old proverb

Where do research topics come from?

So how do researchers come up with the idea for a research project? Probably one of the most common sources of research ideas is the experience of **practical problems in the field**. Many researchers are directly engaged in social, health or human service program implementation and come up with their ideas based on what they see happening around them. Others aren't directly involved in service contexts, but work with (or survey) people who are in order to learn what needs to be better understood. Many of the ideas would strike the outsider as silly or worse. For instance, in health services areas, there is great interest in the problem of back injuries among nursing staff. It's not necessarily the thing that comes first to mind when we think about the health care field. But if you reflect on it for a minute longer, it should be obvious that nurses and nursing staff do an awful lot of lifting in performing their jobs. They lift and push heavy equipment, and they lift and push oftentimes heavy patients! If 5 or 10 out of every hundred nursing staff were to strain their backs on average over the period of one year, the costs would be enormous -- and that's pretty much what's happening. Even minor injuries can result in increased absenteeism. Major ones can result in lost jobs and expensive medical bills. The nursing industry figures that this is a problem that costs tens of millions of dollars annually in increased health care. And, the health care industry has developed a number of approaches, many of them educational, to try to reduce the scope and cost of the problem. So, even though it might seem silly at first, many of these practical problems that arise in practice can lead to extensive research efforts.

Another source for research ideas is the **literature in your specific field**. Certainly, many researchers get ideas for research by reading the literature and thinking of ways to extend or refine previous research. Another type of literature that acts as a source of good research ideas is the **Requests For Proposals** (**RFP**s) that are published by government agencies and some companies. These RFPs describe some problem that the agency would like researchers to address -- they are virtually handing the researcher an idea! Typically, the RFP describes the problem that needs addressing, the contexts in which it operates, the approach they would like you to take to investigate to address the problem, and the amount they would be willing to pay for such research. Clearly, there's nothing like potential research funding to get researchers to focus on a particular research topic.

And let's not forget the fact that many researchers simply **think up their research** topic on their own. Of course, no one lives in a vacuum, so we would expect that the ideas you come up with on your own are influenced by your background, culture, education and experiences.

Is the study feasible?

Very soon after you get an idea for a study reality begins to kick in and you begin to think about whether the study is feasible at all. There are several major considerations that come into play. Many of these involve making **tradeoffs between rigor and practicality**. To do a study well from a scientific point of view may force you to do things you wouldn't do normally. You may have to control the implementation of your program more carefully than you otherwise might. Or, you may have to ask program participants lots of questions that you usually wouldn't if you weren't doing research. If you had unlimited resources and unbridled control over the circumstances, you would always be able to do the best quality research. But those ideal circumstances seldom exist, and researchers are almost always forced to look for the best tradeoffs they can find in order to get the rigor they desire.

There are several practical considerations that almost always need to be considered when deciding on the ***feasibility*** of a research project. First, you have to think about **how long the research will take** to accomplish. Second, you have to question whether there are important **ethical constraints** that need consideration. Third, can you achieve the **needed cooperation** to take the project to its successful conclusion. And fourth, how significant are the **costs** of conducting the research. Failure to consider any of these factors can mean disaster later.

The Literature Review

One of the most important early steps in a research project is the conducting of the literature review. This is also one of the most humbling experiences you're likely to have. Why? Because you're likely to find out that just about any worthwhile idea you will have has been thought of before, at least to some degree. Every time I teach a research methods course, I have at least one student come to me complaining that they couldn't find anything in the literature that was related to their topic. And virtually every time they have said that, I was able to show them that was only true because they only looked for articles that were *exactly* the same as their research topic. A literature review is designed to identify related research, to set the current research project within a conceptual and theoretical context. When looked at that way, there is almost no topic that is so new or unique that we can't locate relevant and informative related research.

Some tips about conducting the literature review. First, **concentrate your efforts on the *scientific* literature**. Try to determine what the most credible research journals are in your topical area and start with those. Put the greatest emphasis on research journals that use a blind review system. In a blind review, authors submit potential articles to a journal editor who solicits several reviewers who agree to give a critical review of the paper. The paper is sent to these reviewers with no identification of the author so that there will be no personal bias (either for or against the author). Based on the reviewers' recommendations, the editor can accept the article, reject it, or recommend that the author revise and resubmit it. Articles in journals with blind review processes can be expected to have a fairly high level of credibility. Second, **do the review early** in the research process. You are likely to learn a lot in the literature review that will help you in making the tradeoffs you'll need to face. After all, previous researchers also had to face tradeoff decisions.

What should you look for in the literature review? First, you might be able to find a study that is quite similar to the one you are thinking of doing. Since all credible research studies have to review the literature themselves, you can check their literature review to get a quick-start on your own. Second, prior research will help assure that you include all of the major relevant constructs in your study. You may find that other similar studies routinely look at an outcome that you might not have included. If you did your study without that construct, it would not be judged credible if it ignored a major construct. Third, the literature review will help you to find and select appropriate measurement instruments. You will readily see what measurement instruments researchers use themselves in contexts similar to yours. Finally, the literature review will help you to anticipate common problems in your research context. You can use the prior experiences of other to avoid common traps and pitfalls.

- **Concept Mapping**

Social scientists have developed a number of methods and processes that might be useful in helping you to formulate a research project. I would include among these at least the following -- brainstorming, brainwriting, nominal group techniques, focus groups, affinity mapping, Delphi techniques, facet theory, and qualitative text analysis. Here, I'll show you a method that I have developed, called concept mapping, which is especially useful for research problem formulation.

Concept mapping is a general method that can be used to help any individual or group to describe their ideas about some topic in a pictorial form. There are several different types of methods that all currently go by names like "concept mapping", "mental mapping" or "concept webbing." All of them are similar in that they result in a picture of someone's ideas. But the kind of concept mapping I want to describe here is different in a number of important ways. First, it is primarily a group process and so it is especially well-suited for situations where teams or groups of stakeholders have to work together. The other methods work primarily with individuals. Second, it uses a very structured facilitated approach. There are specific steps that are followed by a trained facilitator in helping a group to articulate its ideas and understand them more clearly. Third, the core of concept mapping consists of several state-of-the-art multivariate statistical methods that analyze the input from all of the individuals and yields an aggregate group product. And fourth, the method requires the use of specialized computer programs that can handle the data from this type of process and accomplish the correct analysis and mapping procedures.

Although concept mapping is a general method, it is particularly useful for helping social researchers and research teams develop and detail ideas for research. And, it is especially valuable when researchers want to involve relevant stakeholder groups in the act of creating the research project. Although concept mapping is used for many purposes -- strategic planning, product development, market analysis, decision making, measurement development -- we concentrate here on it's potential for helping researchers formulate their projects.

So what is concept mapping? Essentially, ***concept mapping is a structured process, focused on a topic or construct of interest, involving input from one or more participants, that produces an interpretable pictorial view (concept map) of their ideas and concepts and how these are interrelated***. Concept mapping helps people to think more effectively as a group without losing their individuality. It helps groups to manage the complexity of their ideas without trivializing them or losing detail.

A concept mapping process involves six steps that can take place in a single day or can be spread out over weeks or months depending on the situation. The first step is the **Preparation Step**. There are three things done here. The facilitator of the mapping process works with the initiator(s) (i.e., whoever requests the process initially) to identify who the participants will be. A mapping process can have hundreds or even thousands of stakeholders participating, although we

usually have a relatively small group of between 10 and 20 stakeholders involved. Second, the initiator works with the stakeholders to develop the focus for the project. For instance, the group might decide to focus on defining a program or treatment. Or, they might choose to map all of the outcomes they might expect to see as a result. Finally, the group decides on an appropriate schedule for the mapping. In the **Generation Step** the stakeholders develop a large set of statements that address the focus. For instance, they might generate statements that describe all of the specific activities that will constitute a specific social program. Or, they might generate statements describing specific outcomes that might occur as a result of participating in a program. A wide variety of methods can be used to accomplish this including traditional brainstorming, brainwriting, nominal group techniques, focus groups, qualitative text analysis, and so on. The group can generate up to 200 statements in a concept mapping project. In the **Structuring Step** the participants do two things. First, each participant sorts the statements into piles of similar ones. Most times they do this by sorting a deck of cards that has one statement on each card. But they can also do this directly on a computer by dragging the statements into piles that they create. They can have as few or as many piles as they want. Each participant names each pile with a short descriptive label. Second, each participant rates each of the statements on some scale. Usually the statements are rated on a 1-to-5 scale for their relative importance, where a 1 means the statement is relatively unimportant compared to all the rest, a 3 means that it is moderately important, and a 5 means that it is extremely important. The **Representation Step** is where the analysis is done -- this is the process of taking the sort and rating input and "representing" it in map form. There are two major statistical analyses that are used. The first -- multidimensional scaling -- takes the sort data across all participants and develops the basic map where each statement is a point on the map and statements that were piled together by more people are closer to each other on the map. The second analysis -- cluster analysis -- takes the output of the multidimensional scaling (the point map) and partitions the map into groups of statements or ideas, into clusters. If the statements describe activities of a program, the clusters show how these can be grouped into logical groups of activities. If the statements are specific outcomes, the clusters might be viewed as outcome constructs or concepts. In the fifth step -- the **Interpretation Step** -- the facilitator works with the stakeholder group to help them develop their own labels and interpretations for the various maps. Finally, the **Utilization Step** involves using the maps to help address the original focus. On the program side, the maps can be used as a visual framework for operationalizing the program. on the outcome side, they can be used as the basis for developing measures and displaying results.

This is only a very basic introduction to concept mapping and its uses. If you want to find out more about this method, you might look at some of the articles I've written about concept mapping, including ***An Introduction to Concept Mapping, Concept Mapping: Soft Science or Hard Art?***,or the article entitled ***Using Concept Mapping to Develop a Conceptual Framework of Staff's Views of a Supported Employment Program for Persons with Severe Mental Illness.***

Evaluation Research

One specific form of social research -- evaluation research -- is of particular interest here. The ***Introduction to Evaluation Research*** presents an overview of what evaluation is and how it differs from social research generally. We also introduce several evaluation models to give you some perspective on the evaluation endeavor. Evaluation should not be considered in a vacuum. Here, we consider evaluation as embedded within a larger ***Planning-Evaluation Cycle.***

Evaluation can be a threatening activity. Many groups and organizations struggle with how to build a good evaluation capability into their everyday activities and procedures. This is essentially an organizational culture issue. Here we consider some of the issues a group or organization needs to address in order to ***develop an evaluation culture*** that works in their context.

- **Introduction to Evaluation**

Evaluation is a methodological area that is closely related to, but distinguishable from more traditional social research. Evaluation utilizes many of the same methodologies used in traditional social research, but because evaluation takes place within a political and organizational context, it requires group skills, management ability, political dexterity, sensitivity to multiple stakeholders and other skills that social research in general does not rely on as much. Here we introduce the idea of evaluation and some of the major terms and issues in the field.

Definitions of Evaluation

Probably the most frequently given definition is:

Evaluation is the systematic assessment of the worth or merit of some object

This definition is hardly perfect. There are many types of evaluations that do not *necessarily* result in an assessment of worth or merit -- descriptive studies, implementation analyses, and formative evaluations, to name a few. Better perhaps is a definition that emphasizes the information-processing and feedback functions of evaluation. For instance, one might say:

Evaluation is the systematic acquisition and assessment of information to provide useful feedback about some object

Both definitions agree that evaluation is a *systematic* endeavor and both use the deliberately ambiguous term 'object' which could refer to a program, policy, technology, person, need, activity, and so on. The latter definition emphasizes *acquiring and assessing information* rather than *assessing worth or merit* because all evaluation work involves collecting and sifting through data, making judgements about the validity of the information and of inferences we derive from it, whether or not an assessment of worth or merit results.

The Goals of Evaluation

The generic goal of most evaluations is to provide "useful feedback" to a variety of audiences including sponsors, donors, client-groups, administrators, staff, and other relevant constituencies. Most often, feedback is perceived as "useful" if it aids in decision-making. But the relationship between an evaluation and its impact is not a simple one -- studies that seem critical sometimes fail to influence short-term decisions, and studies that initially seem to have no influence can have a delayed impact when more congenial conditions arise. Despite this, there is broad consensus that the major goal of evaluation should be to influence decision-making or policy formulation through the provision of empirically-driven feedback.

Evaluation Strategies

'Evaluation strategies' means broad, overarching perspectives on evaluation. They encompass the most general groups or "camps" of evaluators; although, at its best, evaluation work borrows eclectically from the perspectives of all these camps. Four major groups of evaluation strategies are discussed here.

Scientific-experimental models are probably the most historically dominant evaluation strategies. Taking their values and methods from the sciences -- especially the social sciences -- they prioritize on the desirability of impartiality, accuracy, objectivity and the validity of the information generated. Included under scientific-experimental models would be: the tradition of experimental and quasi-experimental designs; objectives-based research that comes from education; econometrically-oriented perspectives including cost-effectiveness and cost-benefit analysis; and the recent articulation of theory-driven evaluation.

The second class of strategies are ***management-oriented systems models***. Two of the most common of these are **PERT**, the **P**rogram **E**valuation and **R**eview **T**echnique, and **CPM**, the **C**ritical **P**ath **M**ethod. Both have been widely used in business and government in this country. It would also be legitimate to include the Logical Framework or "Logframe" model developed at U.S. Agency for International Development and general systems theory and operations research approaches in this category. Two management-oriented systems models were originated by evaluators: the **UTOS** model where **U** stands for Units, **T** for Treatments, **O** for Observing Observations and **S** for Settings; and the **CIPP** model where the **C** stands for Context, the **I** for Input, the first **P** for Process and the second **P** for Product. These management-oriented systems models emphasize comprehensiveness in evaluation, placing evaluation within a larger framework of organizational activities.

The third class of strategies are the ***qualitative/anthropological models***. They emphasize the importance of observation, the need to retain the phenomenological quality of the evaluation context, and the value of subjective human interpretation in the evaluation process. Included in this category are the approaches known in evaluation as naturalistic or 'Fourth Generation' evaluation; the various qualitative schools; critical theory and art criticism approaches; and, the 'grounded theory' approach of Glaser and Strauss among others.

Finally, a fourth class of strategies is termed ***participant-oriented models***. As the term suggests, they emphasize the central importance of the evaluation participants, especially clients and users of the program or technology. Client-centered and stakeholder approaches are examples of participant-oriented models, as are consumer-oriented evaluation systems.

With all of these strategies to choose from, how to decide? Debates that rage within the evaluation profession -- and they do rage -- are generally battles between these different strategists, with each claiming the superiority of their position. In reality, most good evaluators are familiar with all four categories and borrow from each as the need arises. There is no inherent incompatibility between these broad strategies -- each of them brings something valuable to the evaluation table. In fact, in recent years attention has increasingly turned to how one might integrate results from evaluations that use different strategies, carried out from different perspectives, and using different methods. Clearly, there are no simple answers here. The problems are complex and the methodologies needed will and should be varied.

Types of Evaluation

There are many different types of evaluations depending on the object being evaluated and the purpose of the evaluation. Perhaps the most important basic distinction in evaluation types is that between ***formative*** and ***summative*** evaluation. Formative evaluations strengthen or improve the object being evaluated -- they help form it by examining the delivery of the program or technology, the quality of its implementation, and the assessment of the organizational context, personnel, procedures, inputs, and so on. Summative evaluations, in contrast, examine the effects or outcomes of some object -- they summarize it by describing what happens subsequent to delivery of the program or technology; assessing whether the object can be said to have caused the outcome; determining the overall impact of the causal factor beyond only the immediate target outcomes; and, estimating the relative costs associated with the object.

Formative evaluation includes several evaluation types:

- ***needs assessment*** determines who needs the program, how great the need is, and what might work to meet the need
- ***evaluability assessment*** determines whether an evaluation is feasible and how stakeholders can help shape its usefulness
- ***structured conceptualization*** helps stakeholders define the program or technology, the target population, and the possible outcomes
- ***implementation evaluation*** monitors the fidelity of the program or technology delivery
- ***process evaluation*** investigates the process of delivering the program or technology, including alternative delivery procedures

Summative evaluation can also be subdivided:

- ***outcome evaluations*** investigate whether the program or technology caused demonstrable effects on specifically defined target outcomes
- ***impact evaluation*** is broader and assesses the overall or net effects -- intended or unintended -- of the program or technology as a whole
- ***cost-effectiveness and cost-benefit analysis*** address questions of efficiency by standardizing outcomes in terms of their dollar costs and values
- ***secondary analysis*** reexamines existing data to address new questions or use methods not previously employed
- ***meta-analysis*** integrates the outcome estimates from multiple studies to arrive at an overall or summary judgement on an evaluation question

Evaluation Questions and Methods

Evaluators ask many different kinds of questions and use a variety of methods to address them. These are considered within the framework of formative and summative evaluation as presented above.

In formative research the major questions and methodologies are:

What is the definition and scope of the problem or issue, or what's the question?

Formulating and conceptualizing methods might be used including brainstorming, focus groups, nominal group techniques, Delphi methods, brainwriting, stakeholder analysis, synectics, lateral thinking, input-output analysis, and concept mapping.

Where is the problem and how big or serious is it?

The most common method used here is "needs assessment" which can include: analysis of existing data sources, and the use of sample surveys, interviews of constituent populations, qualitative research, expert testimony, and focus groups.

How should the program or technology be delivered to address the problem?

Some of the methods already listed apply here, as do detailing methodologies like simulation techniques, or multivariate methods like multiattribute utility theory or exploratory causal modeling; decision-making methods; and project planning and implementation methods like flow charting, PERT/CPM, and project scheduling.

How well is the program or technology delivered?

Qualitative and quantitative monitoring techniques, the use of management information systems, and implementation assessment would be appropriate methodologies here.

The questions and methods addressed under summative evaluation include:

What type of evaluation is feasible?

Evaluability assessment can be used here, as well as standard approaches for selecting an appropriate evaluation design.

What was the effectiveness of the program or technology?

One would choose from observational and correlational methods for demonstrating whether desired effects occurred, and quasi-experimental and experimental designs for determining whether observed effects can reasonably be attributed to the intervention and not to other sources.

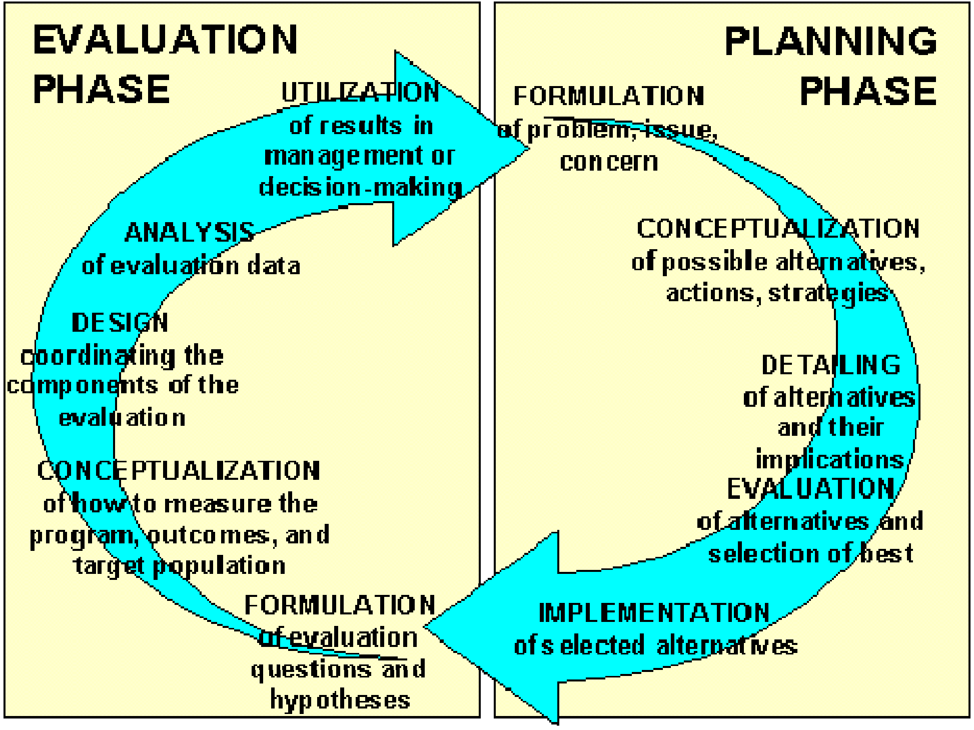
What is the net impact of the program?

Econometric methods for assessing cost effectiveness and cost/benefits would apply here, along with qualitative methods that enable us to summarize the full range of intended and unintended impacts.

Clearly, this introduction is not meant to be exhaustive. Each of these methods, and the many not mentioned, are supported by an extensive methodological research literature. This is a formidable set of tools. But the need to improve, update and adapt these methods to changing circumstances means that methodological research and development needs to have a major place in evaluation work.

- **The Planning-Evaluation Cycle**

Often, ***evaluation*** is construed as part of a larger managerial or administrative process. Sometimes this is referred to as the ***planning-evaluation cycle***. The distinctions between planning and evaluation are not always clear; this cycle is described in many different ways with various phases claimed by both planners and evaluators. Usually, the first stage of such a cycle -- the planning phase -- is designed to elaborate a set of potential actions, programs, or technologies, and select the best for implementation. Depending on the organization and the problem being addressed, a planning process could involve any or all of these stages: the ***formulation*** of the problem, issue, or concern; the broad ***conceptualization*** of the major alternatives that might be considered; the ***detailing*** of these alternatives and their potential implications; the ***evaluation*** of the alternatives and the selection of the best one; and the ***implementation*** of the selected alternative. Although these stages are traditionally considered planning, there is a lot of evaluation work involved. Evaluators are trained in needs assessment, they use methodologies -- like the ***concept mapping*** one presented later -- that help in conceptualization and detailing, and they have the skills to help assess alternatives and make a choice of the best one.


EVALUATION PHASE
PLANNING PHASE
UTILIZATION of results in management or decision-making
ANALYSIS of evaluation data
DESIGN coordinating the components of the evaluation
CONCEPTUALIZATION of how to measure the program, outcomes, and target population
FORMULATION of evaluation questions and hypotheses
FORMULATION of problem, issue, concern
CONCEPTUALIZATION of possible alternatives, actions, strategies
DETAILING of alternatives and their implications
EVALUATION of alternatives and selection of best
IMPLEMENTATION of selected alternatives

The evaluation phase also involves a sequence of stages that typically includes: the ***formulation*** of the major objectives, goals, and hypotheses of the program or technology; the ***conceptualization*** and operationalization of the major components of the evaluation -- the program, participants, setting, and measures; the ***design*** of the evaluation, ***detailing*** how these components will be coordinated; the ***analysis*** of the information, both qualitative and quantitative; and the ***utilization*** of the evaluation results.

- **An Evaluation Culture**

I took the idea of an evaluation culture from a wonderful paper written by Donald Campbell in 1969 entitled 'Methods for an Experimenting Society.' Following in the footsteps of that paper, this one is considerably more naive and utopian. And, I have changed the name of this idealized society to reflect terminology that is perhaps more amenable to the climate of the 1990s. For the term *experimenting*, I have substituted the softer and broader term *evaluating*. And for the term *society*, I have substituted the more internationally-flavored term *culture*. With these shifts in emphasis duly noted, I want you to know that I see the evaluation culture as one that a member of the experimenting society would feel comfortable visiting, and perhaps even thinking of taking as a permanent residence.

What would an evaluation culture look like? What should its values be? You should know at the outset that I fully hope that some version of this fantasy will become an integral part of twenty-first century thought. There is no particular order of importance to the way these ideas are presented -- I'll leave that ordering to subsequent efforts.

First, our evaluation culture will embrace an ***action-oriented*** perspective that actively seeks solutions to problems, trying out tentative ones, weighing the results and consequences of actions, all within an endless cycle of supposition-action-evidence-revision that characterizes good science and good management. In this activist evaluation culture, we will encourage innovative approaches at all levels. But well-intentioned activism by itself is not enough, and may at times be risky, dangerous, and lead to detrimental consequences. In an evaluation culture, we won't act for action's sake -- we'll always attempt to assess the effects of our actions.

This evaluation culture will be an accessible, ***teaching-oriented*** one that emphasizes the unity of formal evaluation and everyday thought. Most of our evaluations will be simple, informal, efficient, practical, low-cost and easily carried out and understood by nontechnicians. Evaluations won't just be delegated to one person or department -- we will encourage everyone in our organizations to become involved in evaluating what they and their organizations do. Where technical expertise is needed we will encourage the experts to also educate us about the technical side of what they do, demanding that they try to find ways to explain their techniques and methods adequately for nontechnicians. We will devote considerable resources to teaching others about evaluation principles.

Our evaluation culture will be ***diverse, inclusive, participatory, responsive and fundamentally non-hierarchical***. World problems cannot be solved by simple "silver bullet" solutions. There is growing recognition in many arenas that our most fundamental problems are systemic, interconnected, and inextricably linked to social and economic issues and factors. Solutions will involve husbanding the resources, talents and insights of a wide range of people. The formulation of problems and potential solutions needs to involve a broad range of constituencies. More than just "research" skills will be needed. Especially important will be skills in negotiation and consensus-building processes. Evaluators are familiar with arguments for greater diversity and inclusiveness -- we've been talking about stakeholder, participative, multiple-constituency research for nearly two decades. No one that I know is seriously debating anymore whether we should move to more inclusive participatory approaches. The real question seems to be how such work might best be accomplished, and despite all the rhetoric about the importance of participatory methods, we have a long way to go in learning how to do them effectively.

Our evaluation culture will be a ***humble, self-critical*** one. We will openly acknowledge our limitations and recognize that what we learn from a single evaluation study, however well designed, will almost always be equivocal and tentative. In this regard, I believe we too often undervalue cowardice in research. I find it wholly appropriate that evaluators resist being drawn into making decisions for others, although certainly the results of our work should help inform the decision makers. A cowardly approach saves the evaluator from being drawn into the political context, helping assure the impartiality needed for objective assessment, and it protects the evaluator from taking responsibility for making decisions that should be left to those who have been duly-authorized -- and who have to live with the consequences. Most program decisions, especially decisions about whether to continue a program or close it down, must include more input than an evaluation alone can ever provide. While evaluators can help to elucidate what has happened in the past or might happen under certain circumstances, it is the responsibility of the organization and society as a whole to determine what ought to happen. The debate about the appropriate role of an evaluator in the decision-making process is an extremely intense one right now in evaluation circles, and my position advocating a cowardly reluctance of the evaluator to undertake a decision-making role may very well be in the minority. We will need to debate this issue vigorously, especially for politically-complex, international-evaluation contexts.

Our evaluation culture will need to be an ***interdisciplinary*** one, doing more than just grafting one discipline onto another through constructing multi-discipline research teams. We'll need such teams, of course, but I mean to imply something deeper, more personally internalized -- we need to move toward being nondisciplinary, consciously putting aside the blinders of our respective specialties in an attempt to foster a more whole view of the phenomena we study. As we consider the programs we are evaluating, we each should be able to speculate about a broad range of implementation factors or potential consequences. We should be able to anticipate some of the organizational and systems-related features of these programs, the economic factors that might enhance or reduce implementation, their social and psychological dimensions, and especially whether the ultimate utilizers can understand or know how to utilize and be willing to utilize the the results of our evaluation work. We should also be able to anticipate a broad spectrum of potential consequences -- system-related, production-related, economic, nutritional, social, environmental.

This evaluation culture will also be an honest, ***truth-seeking*** one that stresses accountability and scientific credibility. In many quarters in contemporary society, it appears that many people have given up on the ideas of truth and validity. Our evaluation culture needs to hold to the goal of getting at the truth while at the same time honestly acknowledging the revisability of all scientific knowledge. We need to be critical of those who have given up on the goal of "getting it right" about reality, especially those among the humanities and social sciences who argue that truth is entirely relative to the knower, objectivity an impossibility, and reality nothing more than a construction or illusion that cannot be examined publicly. For them, the goal of seeking the truth is inappropriate and unacceptable, and science a tool of oppression rather than a road to greater enlightenment. Philosophers have, of course, debated such issues for thousands of years and will undoubtedly do so for thousands more. We in the evaluation culture need to check in on their thinking from time to time, but until they settle these debates, we need to hold steadfastly to the goal of getting at the truth -- the goal of getting it right about reality.

Our evaluation culture will be prospective and ***forward-looking***, anticipating where evaluation feedback will be needed rather than just reacting to situations as they arise. We will construct simple, low-cost evaluation and monitoring information systems when we first initiate a new program or technology -- we cannot wait until a program is complete or a technology is in the field before we turn our attention to its evaluation.

Finally, the evaluation culture I envision is one that will emphasize fair, open, ***ethical and democratic*** processes. We should move away from private ownership of and exclusive access to data. The data from all of our evaluations needs to be accessible to all interested groups allowing more extensive independent secondary analyses and opportunities for replication or refutation of original results. We should encourage open commentary and debate regarding the results of specific evaluations. Especially when there are multiple parties who have a stake in such results, it is important for our reporting procedures to include formal opportunities for competitive review and response. Our evaluation culture must continually strive for greater understanding of the ethical dilemmas posed by our research. Our desire for valid, scientific inference will at times put us in conflict with ethical principles. The situation is likely to be especially complex in international-evaluation contexts where we will often be dealing with multiple cultures and countries that are at different stages of economic development and have different value systems and morals. We need to be ready to deal with potential ethical and political issues posed by our methodologies in an open, direct, and democratic manner.

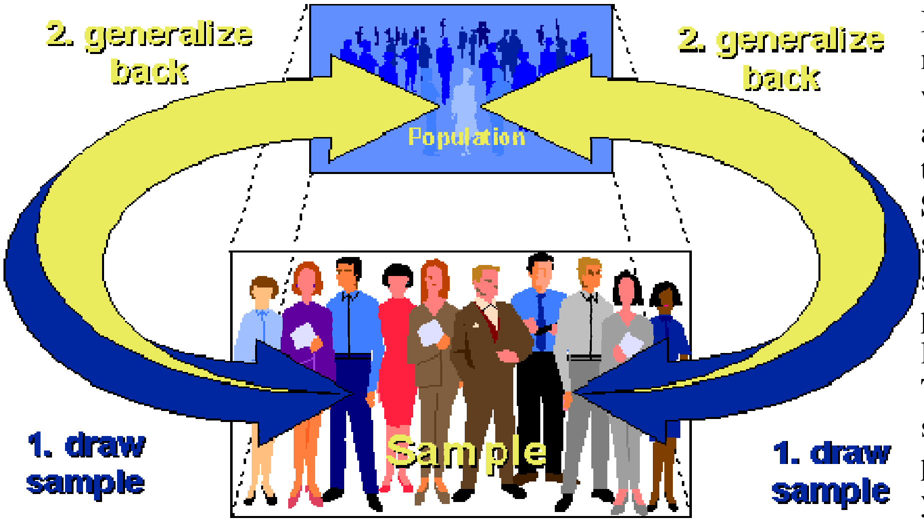
Do you agree with the values I'm describing here? What other characteristics might this evaluation culture have? You tell me. There are many more values and characteristics that ought to be considered. For now, the ones mentioned above, and others in the literature, provide us with a starting point at which we can all join the discussion. I hope you will add to the list, and I encourage each of you to criticize these tentative statements I've offered about the extraordinary potential of the evaluation culture that we are all in the process of creating today.

Sampling

Sampling is the process of selecting units (e.g., people, organizations) from a population of interest so that by studying the sample we may fairly generalize our results back to the population from which they were chosen. Let's begin by covering some of the key terms in sampling like "population" and "sampling frame." Then, because some types of sampling rely upon quantitative models, we'll talk about some of the statistical terms used in sampling. Finally, we'll discuss the major distinction between probability and Nonprobability sampling methods and work through the major types in each.

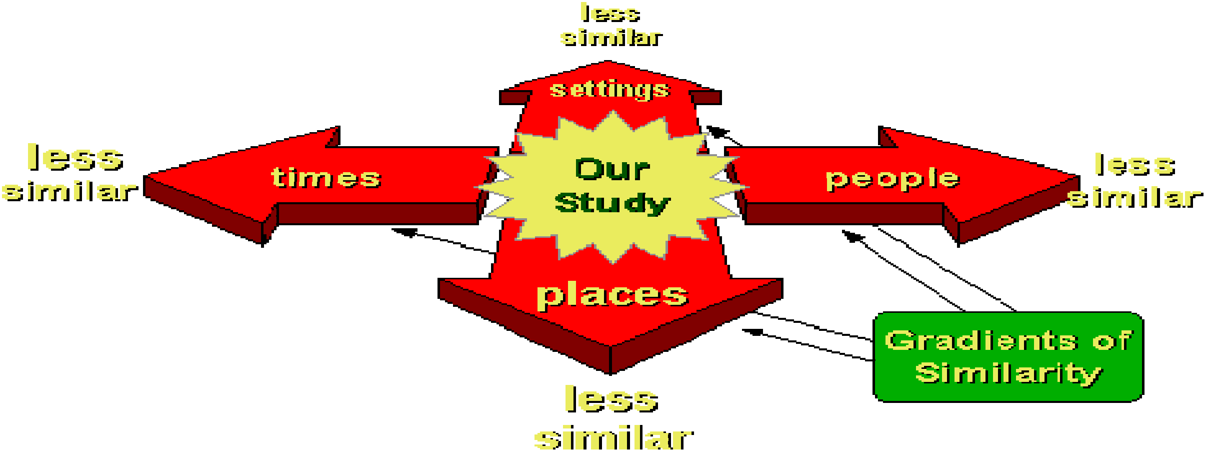
External Validity

External validity is related to generalizing. That's the major thing you need to keep in mind. Recall that validity refers to the approximate truth of propositions, inferences, or conclusions. So, *external* validity refers to the approximate truth of conclusions the involve generalizations. Put in more pedestrian terms, external validity is the degree to which the conclusions in your study would hold for other persons in other places and at other times.



In science there are two major approaches to how we provide evidence for a generalization. I'll call the first approach the **Sampling Model**. In the sampling model, you start by identifying the population you would like to generalize to. Then, you draw a fair sample from that population and conduct your research with the sample. Finally, because the sample is representative of the population, you can automatically generalize your results back to the population. There are several problems with this approach. First, perhaps you don't know at the time of your study who you might ultimately like to generalize to. Second, you may not be easily able to draw a fair or representative sample. Third, it's impossible to sample across all times that you might like to generalize to (like next year).

I'll call the second approach to generalizing the **Proximal Similarity Model**. 'Proximal' means 'nearby' and 'similarity' means... well, it means 'similarity'. The term *proximal similarity* was suggested by Donald T. Campbell as an appropriate relabeling of the term *external validity* (although he was the first to admit that it probably wouldn't catch on!). Under this model, we begin by thinking about different generalizability contexts and developing a theory about which contexts are more like our study and which are less so. For instance, we might imagine several settings that have people who are more similar to the people in our study or people who are less similar. This also holds for times and places. When we place different contexts in terms of their relative similarities, we can call this implicit theoretical a ***gradient of similarity***. Once we have developed this proximal similarity framework, we are able to generalize. How? We conclude that we can generalize the results of our study to other persons, places or times that are more like (that is, more proximally similar) to our study. Notice that here, we can never generalize with certainty -- it is always a question of more or less similar.



Threats to External Validity

A threat to external validity is an explanation of how you might be wrong in making a generalization. For instance, you conclude that the results of your study (which was done in a specific place, with certain types of people, and at a specific time) can be generalized to another context (for instance, another place, with slightly different people, at a slightly later time). There are three major threats to external validity because there are three ways you could be wrong -- people, places or times. Your critics could come along, for example, and argue that the results of your study are due to the unusual type of people who were in the study. Or, they could argue that it might only work because of the unusual place you did the study in (perhaps you did your educational study in a college town with lots of high-achieving educationally-oriented kids). Or, they might suggest that you did your study in a peculiar time. For instance, if you did your smoking cessation study the week after the Surgeon General issues the well-publicized results of the latest smoking and cancer studies, you might get different results than if you had done it the week before.

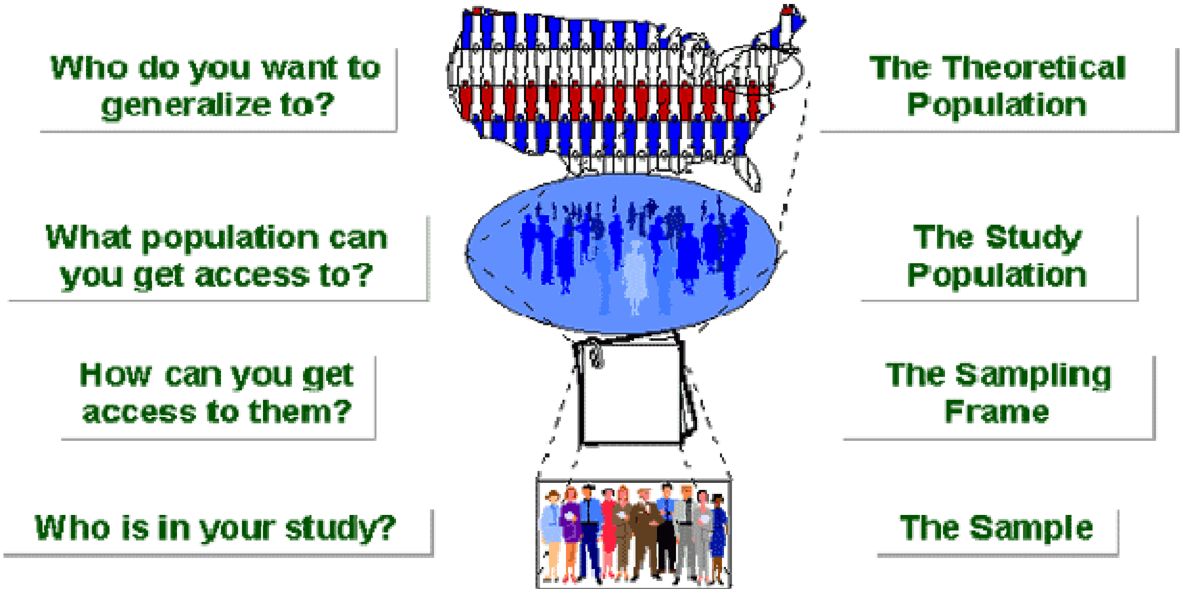
Improving External Validity

How can we improve external validity? One way, based on the sampling model, suggests that you do a good job of drawing a sample from a population. For instance, you should use random selection, if possible, rather than a nonrandom procedure. And, once selected, you should try to assure that the respondents participate in your study and that you keep your dropout rates low. A second approach would be to use the theory of proximal similarity more effectively. How? Perhaps you could do a better job of describing the ways your contexts and others differ, providing lots of data about the degree of similarity between various groups of people, places, and even times. You might even be able to map out the degree of proximal similarity among various contexts with a methodology like concept mapping. Perhaps the best approach to criticisms of generalizations is simply to show them that they're wrong -- do your study in a variety of places, with different people and at different times. That is, your external validity (ability to generalize) will be stronger the more you **replicate** your study.

Sampling Terminology

As with anything else in life you have to learn the language of an area if you're going to ever hope to use it. Here, I want to introduce several different terms for the major groups that are involved in a sampling process and the role that each group plays in the logic of sampling.

The major question that motivates sampling in the first place is: "Who do you want to generalize to?" Or should it be: "To whom do you want to generalize?" In most social research we are interested in more than just the people who directly participate in our study. We would like to be able to talk in general terms and not be confined only to the people who are in our study. Now, there are times when we aren't very concerned about generalizing. Maybe we're just evaluating a program in a local agency and we don't care whether the program would work with other people in other places and at other times. In that case, sampling and generalizing might not be of interest. In other cases, we would really like to be able to generalize almost universally. When psychologists do research, they are often interested in developing theories that would hold for all humans. But in most applied social research, we are interested in generalizing to specific groups. The group you wish to generalize to is often called the **population** in your study. This is the group you would like to sample from because this is the group you are interested in generalizing to. Let's imagine that you wish to generalize to urban homeless males between the ages of 30 and 50 in the United States. If that is the population of interest, you are likely to have a very hard time developing a reasonable sampling plan. You are probably not going to find an accurate listing of this population, and even if you did, you would almost certainly not be able to mount a national sample across hundreds of urban areas. So we probably should make a distinction between the population you would like to generalize to, and the population that will be accessible to you. We'll call the former the **theoretical population** and the latter the **accessible population**. In this example, the accessible population might be homeless males between the ages of 30 and 50 in six selected urban areas across the U.S.


Who do you want to generalize to?
The Theoretical Population
What population can you get access to?
The Study Population
How can you get access to them?
The Sampling Frame
Who is in your study?
The Sample

Once you've identified the theoretical and accessible populations, you have to do one more thing before you can actually draw a sample -- you have to get a list of the members of the accessible population. (Or, you have to spell out in detail how you will contact them to assure representativeness). The listing of the accessible population from which you'll draw your sample is called the **sampling frame**. If you were doing a phone survey and selecting names from the telephone book, the book would be your sampling frame. That wouldn't be a great way to sample because significant subportions of the population either don't have a phone or have moved in or out of the area since the last book was printed. Notice that in this case, you might identify the area code and all three-digit prefixes within that area code and draw a sample simply by randomly dialing numbers (cleverly known as *random-digit-dialing*). In this case, the sampling frame is not a list *per se*, but is rather a procedure that you follow as the actual basis for sampling. Finally, you actually draw your sample (using one of the many sampling procedures). The **sample** is the group of people who you select to be in your study. Notice that I didn't say that the sample was the group of people who are actually *in* your study. You may not be able to contact or recruit all of the people you actually sample, or some could drop out over the course of the study. The group that actually completes your study is a subsample of the sample -- it doesn't include nonrespondents or dropouts. The problem of nonresponse and its effects on a study will be addressed elsewhere.

People often confuse what is meant by random selection with the idea of random assignment. You should make sure that you understand the distinction between random selection and random assignment.

At this point, you should appreciate that sampling is a difficult multi-step process and that there are lots of places you can go wrong. In fact, as we move from each step to the next in identifying a sample, there is the possibility of introducing systematic error or **bias**. For instance, even if you are able to identify perfectly the population of interest, you may not have access to all of them. And even if you do, you may not have a complete and accurate enumeration or sampling frame from which to select. And, even if you do, you may not draw the sample correctly or accurately. And, even if you do, they may not all come and they may not all stay. Depressed yet? This is a very difficult business indeed. At times like this I'm reminded of what Donald Campbell used to say (I'll paraphrase here): "Cousins to the amoeba, it's amazing that we know anything at all!"

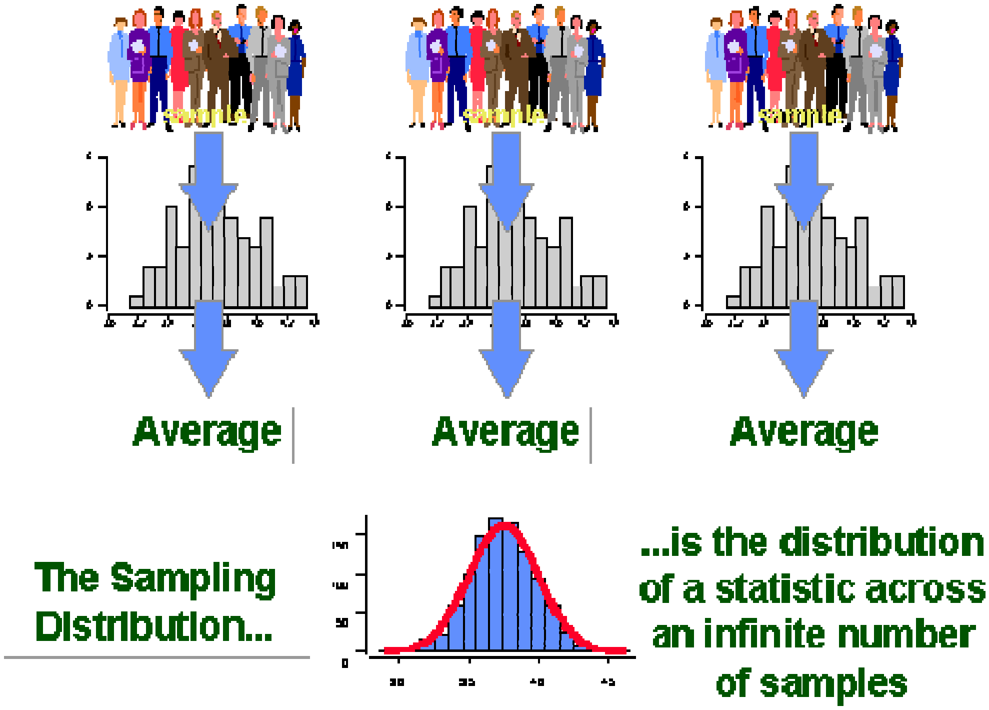
Statistical Terms in Sampling

Let's begin by defining some very simple terms that are relevant here. First, let's look at the results of our sampling efforts. When we sample, the units that we sample -- usually people -- supply us with one or more responses. In this sense, a **response** is a specific measurement value that a sampling unit supplies. In the figure, the person is responding to a survey instrument and gives a response of '4'. When we look across the responses that we get for our entire sample, we use a **statistic**. There are a wide variety of statistics we can use -- mean, median, mode, and so on. In this example, we see that the mean or average for the sample is 3.75. But the reason we sample is so that we might get an estimate for the population we sampled from. If we could, we would much prefer to measure the entire population. If you measure the entire population and calculate a value like a mean or average, we don't refer to this as a statistic, we call it a **parameter** of the population.

The Sampling Distribution

So how do we get from our sample statistic to an estimate of the population parameter? A crucial midway concept you need to understand is the **sampling distribution**. In order to understand it, you have to be able and willing to do a thought experiment. Imagine that instead of just taking a single sample like we do in a typical study, you took three independent samples of the same population. And furthermore, imagine that for each of your three samples, you collected a single response and computed a single statistic, say, the mean of the response. Even though all three samples came from the same population, you wouldn't expect to get the exact same statistic from each. They would differ slightly just due to the random "luck of the draw" or to the natural fluctuations or vagaries of drawing a sample. But you would expect that all three samples would yield a similar statistical estimate because they were drawn from the same population. Now, for the leap of imagination! Imagine that you did an *infinite* number of samples from the same population and computed the average for each one. If you plotted them on a histogram or bar graph you should find that most of them converge on the same central value and that you get fewer and fewer samples that have averages farther away up or down from that central value. In other words, the bar graph would be well described by the *bell curve* shape that is an indication of a "normal" distribution in statistics. The distribution of an infinite number of samples of the same size as the sample in your study is known as the **sampling distribution**. We don't ever actually construct a sampling distribution. Why not? You're not paying attention!

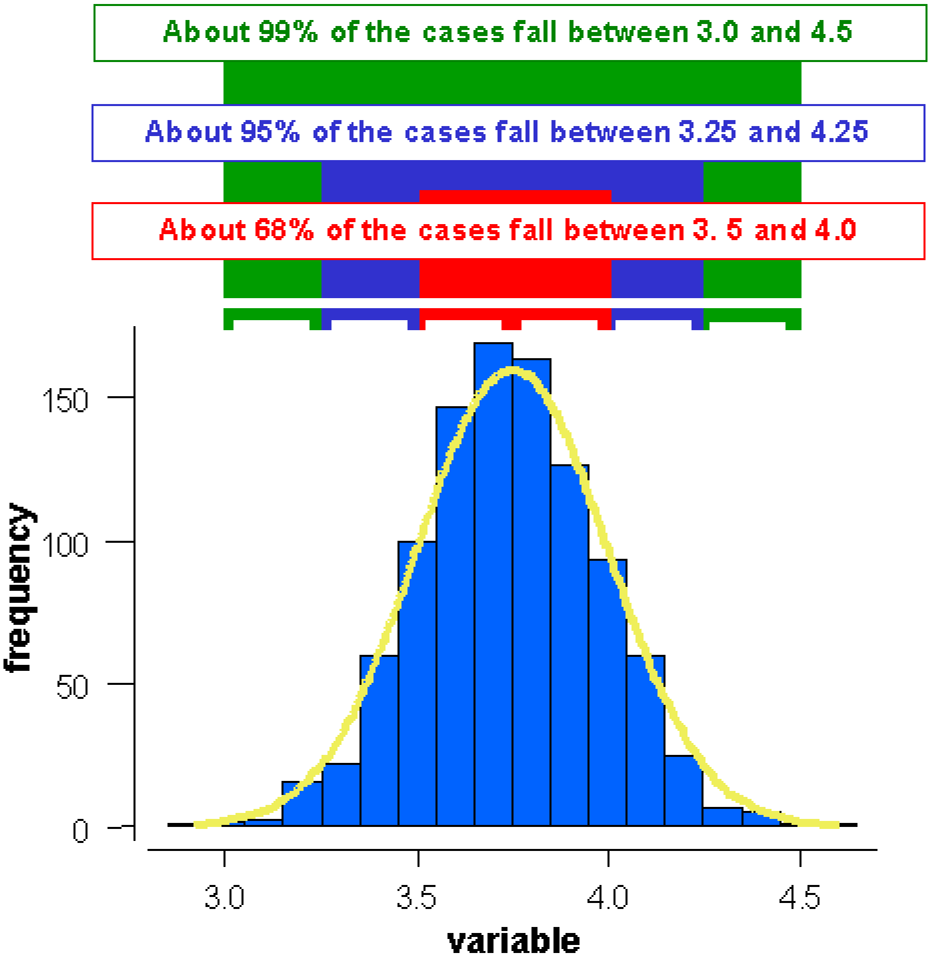
Because to construct it we would have to take an *infinite* number of samples and at least the last time I checked, on this planet infinite is not a number we know how to reach. So why do we even talk about a sampling distribution? Now that's a good question! Because we need to realize that our sample is just one of a potentially infinite number of samples that we could have taken. When we keep the sampling distribution in mind, we realize that while the statistic we got from our sample is probably near the center of the sampling distribution (because most of the samples would be there) we could have gotten one of the extreme samples just by the luck of the draw. If we take the average of the sampling distribution -- the average of the averages of an infinite number of samples -- we would be much closer to the true population average -- the parameter of interest. So the average of the sampling distribution is essentially equivalent to the parameter. But what is the standard deviation of the sampling distribution (OK, never had statistics? There are any number of places on the web where you can learn about them or even just brush up if you've gotten rusty. This isn't one of them. I'm going to assume that you at least know what a standard deviation is, or that you're capable of finding out relatively quickly). The standard deviation of the sampling distribution tells us something about how different samples would be distributed. In statistics it is referred to as the **standard error** (so we can keep it separate in our minds from standard deviations. Getting confused? Go get a cup of coffee and come back in ten minutes...OK, let's try once more... A standard deviation is the spread of the scores around the average in a single sample. The standard error is the spread of the averages around the average of averages in a sampling distribution. Got it?)


sample
sample
sample
Average
Average
Average
The Sampling Distribution...
...is the distribution of a statistic across an infinite number of samples

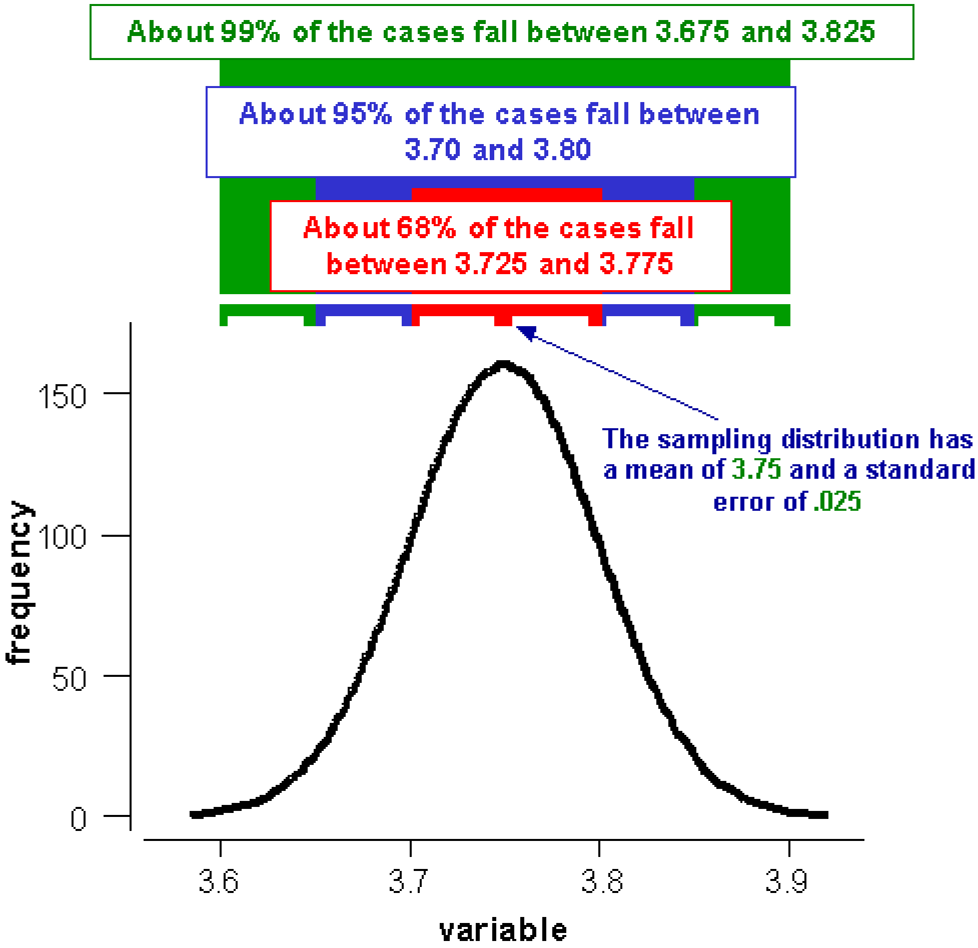
Sampling Error

In sampling contexts, the standard error is called **sampling error**. Sampling error gives us some idea of the precision of our statistical estimate. A low sampling error means that we had relatively less variability or range in the sampling distribution. But here we go again -- we never actually see the sampling distribution! So how do we calculate sampling error? We base our calculation *on the standard deviation of our sample*. The greater the sample standard deviation, the greater the standard error (and the sampling error). The standard error is also related to the sample size. The greater your sample size, the smaller the standard error. Why? Because the greater the sample size, the closer your sample is to the actual population itself. If you take a sample that consists of the entire population you actually have no sampling error because you don't have a sample, you have the entire population. In that case, the mean you estimate is the parameter.

The 68, 95, 99 Percent Rule



You've probably heard this one before, but it's so important that it's always worth repeating... There is a general rule that applies whenever we have a normal or bell-shaped distribution. Start with the average -- the center of the distribution. If you go up and down (i.e., left and right) one standard unit, you will include approximately 68% of the cases in the distribution (i.e., 68% of the area under the curve). If you go up and down two standard units, you will include approximately 95% of the cases. And if you go plus-and-minus three standard units, you will include about 99% of the cases. Notice that I didn't specify in the previous few sentences whether I was talking about standard deviation units or standard error units. That's because the same rule holds for both types of distributions (i.e., the raw data and sampling distributions). For instance, in the figure, the mean of the distribution is 3.75 and the standard unit is .25 (If this was a distribution of raw data, we would be talking in standard deviation units. If it's a sampling distribution, we'd be talking in standard error units). If we go up and down one standard unit from the mean, we would be going up and down .25 from the mean of 3.75. Within this range -- 3.5 to 4.0 -- we would expect to see approximately 68% of the cases. This section is marked in red on the figure. I leave to you to figure out the other ranges. But what does this all mean you ask? If we are dealing with raw data and we know the mean and standard deviation of a sample, we can *predict* the intervals within which 68, 95 and 99% of our cases would be expected to fall. We call these intervals the -- guess what -- 68, 95 and 99% confidence intervals.



Now, here's where everything should come together in one great aha! experience if you've been following along. If we had a *sampling distribution*, we would be able to predict the 68, 95 and 99% confidence intervals for where the population parameter should be! And isn't that why we sampled in the first place? So that we could predict where the population is on that variable? There's only one hitch. We don't actually have the sampling distribution (now this is the third time I've said this in this essay)! But we do have the distribution for the sample itself. And we can from that distribution estimate the standard error (the sampling error) because it is based on the standard deviation and we have that. And, of course, we don't actually know the population parameter value -- we're trying to find that out -- but we can use our best estimate for that -- the sample statistic. Now, if we have the mean of the sampling distribution (or set it to the mean from our sample) and we have an estimate of the standard error (we calculate that from our sample) then we have the two key ingredients that we need for our sampling distribution in order to estimate confidence intervals for the population parameter.

Perhaps an example will help. Let's assume we did a study and drew a single sample from the population. Furthermore, let's assume that the average for the sample was 3.75 and the standard deviation was .25. This is the raw data distribution depicted above. now, what would the sampling distribution be in this case? Well, we don't actually construct it (because we would need to take an infinite number of samples) but we *can* estimate it. For starters, we assume that the mean of the sampling distribution is the mean of the sample, which is 3.75. Then, we calculate the standard error. To do this, we use the standard deviation for our sample and the sample size (in this case N=100) and we come up with a standard error of .025 (just trust me on this). Now we have everything we need to estimate a confidence interval for the population parameter. We would estimate that the probability is 68% that the true parameter value falls between 3.725 and 3.775 (i.e., 3.75 plus and minus .025); that the 95% confidence interval is 3.700 to 3.800; and that we can say with 99% confidence that the population value is between 3.675 and 3.825. The real value (in this fictitious example) was 3.72 and so we have correctly estimated that value with our sample.

Probability Sampling

A **probability sampling** method is any method of sampling that utilizes some form of *random selection*. In order to have a random selection method, you must set up some process or procedure that assures that the different units in your population have equal probabilities of being chosen. Humans have long practiced various forms of random selection, such as picking a name out of a hat, or choosing the short straw. These days, we tend to use computers as the mechanism for generating random numbers as the basis for random selection.

Some Definitions

Before I can explain the various probability methods we have to define some basic terms. These are:

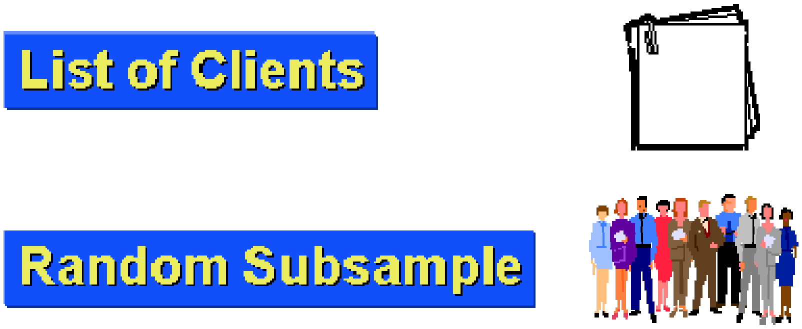
- **N** = the number of cases in the sampling frame
- **n** = the number of cases in the sample
- $_N C_n$ = the number of combinations (subsets) of n from N
- **f** = n/N = the sampling fraction

That's it. With those terms defined we can begin to define the different probability sampling methods.

Simple Random Sampling

The simplest form of random sampling is called **simple random sampling**. Pretty tricky, huh? Here's the quick description of simple random sampling:

- **Objective**: To select *n* units out of *N* such that each $_N C_n$ has an equal chance of being selected.
- **Procedure**: Use a table of random numbers, a computer random number generator, or a mechanical device to select the sample.



A somewhat stilted, if accurate, definition. Let's see if we can make it a little more real. How do we select a simple random sample? Let's assume that we are doing some research with a small service agency that wishes to assess client's views of quality of service over the past year. First, we have to get the sampling frame organized. To accomplish this, we'll go through agency records to identify every client over the past 12 months. If we're lucky, the agency has good accurate computerized records and can quickly produce such a list. Then, we have to actually draw the sample. Decide on the number of clients you would like to have in the final sample. For the sake of the example, let's say you want to select 100 clients to survey and that there were 1000 clients over the past 12 months. Then, the sampling fraction is f = n/N = 100/1000 = .10 or 10%. Now, to actually draw the sample, you have several options. You could print off the list of 1000 clients, tear then into separate strips, put the strips in a hat, mix them up real good, close your eyes and pull out the first 100. But this mechanical procedure would be tedious and the quality of the sample would depend on how thoroughly you mixed them up and how randomly you reached in. Perhaps a better procedure would be to use the kind of ball machine that is popular with many of the state lotteries. You would need three sets of balls numbered 0 to 9, one set for each of the digits from 000 to 999 (if we select 000 we'll call that 1000). Number the list of names from 1 to 1000 and then use the ball machine to select the three digits that selects each person. The obvious disadvantage here is that you need to get the ball machines. (Where do they make those things, anyway? Is there a ball machine industry?).

Neither of these mechanical procedures is very feasible and, with the development of inexpensive computers there is a much easier way. Here's a simple procedure that's especially useful if you have the names of the clients already on the computer. Many computer programs can generate a series of random numbers. Let's assume you can copy and paste the list of client names into a column in an EXCEL spreadsheet. Then, in the column right next to it paste the function =RAND() which is EXCEL's way of putting a random number between 0 and 1 in the cells. Then, sort both columns -- the list of names and the random number -- by the random numbers. This rearranges the list in random order from the lowest to the highest random number. Then, all you have to do is take the first hundred names in this sorted list. pretty simple. You could probably accomplish the whole thing in under a minute.

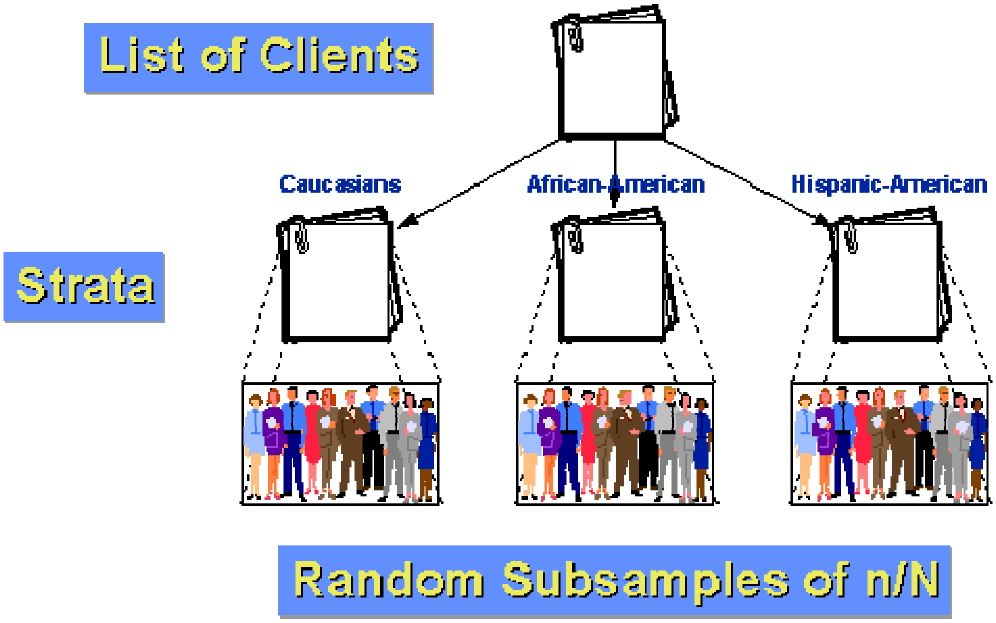
Simple random sampling is simple to accomplish and is easy to explain to others. Because simple random sampling is a fair way to select a sample, it is reasonable to generalize the results from the sample back to the population. Simple random sampling is not the most statistically efficient method of sampling and you may, just because of the luck of the draw, not get good representation of subgroups in a population. To deal with these issues, we have to turn to other sampling methods.

Stratified Random Sampling

Stratified Random Sampling, also sometimes called *proportional* or *quota* random sampling, involves dividing your population into homogeneous subgroups and then taking a simple random sample in each subgroup. In more formal terms:

Objective: Divide the population into non-overlapping groups (i.e., *strata*) N_1, N_2, N_3, ... N_i, such that $N_1 + N_2 + N_3 + ... + N_i = N$. Then do a simple random sample of f = n/N in each strata.

There are several major reasons why you might prefer stratified sampling over simple random sampling. First, it assures that you will be able to represent not only the overall population, but also key subgroups of the population, especially small minority groups. If you want to be able to talk about subgroups, this may be the only way to effectively assure you'll be able to. If the subgroup is extremely small, you can use different sampling fractions (f) within the different strata to randomly over-sample the small group (although you'll then have to weight the within-group estimates using the sampling fraction whenever you want overall population estimates). When we use the same sampling fraction within strata we are conducting *proportionate* stratified random sampling. When we use different sampling fractions in the strata, we call this *disproportionate* stratified random sampling. Second, stratified random sampling will generally have more statistical precision than simple random sampling. This will only be true if the strata or groups are homogeneous. If they are, we expect that the variability within-groups is lower than the variability for the population as a whole. Stratified sampling capitalizes on that fact.

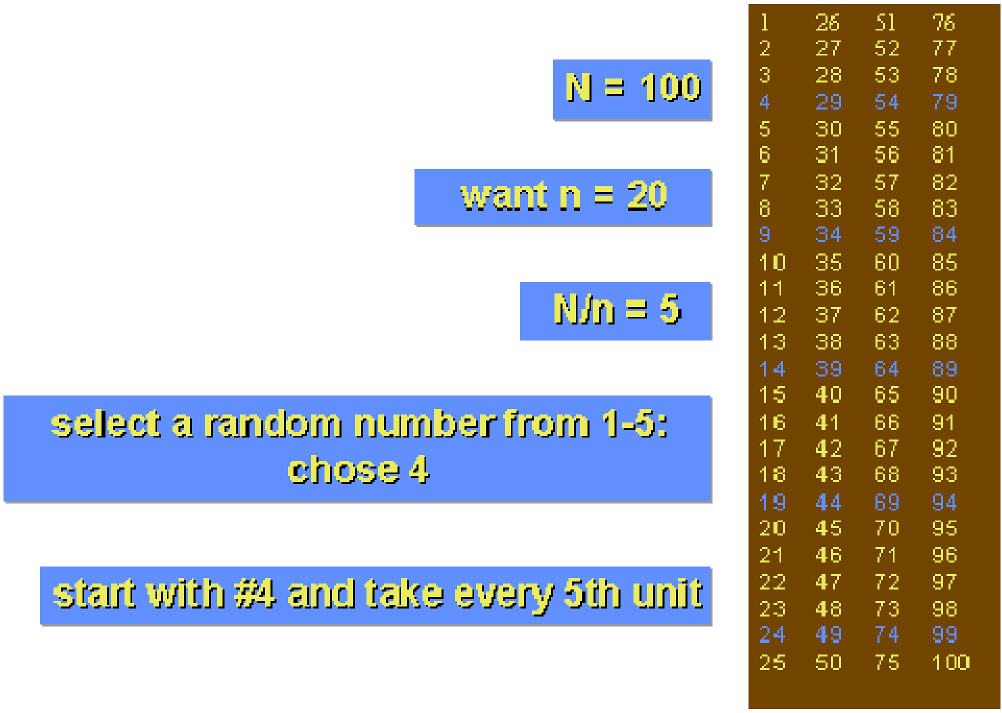


For example, let's say that the population of clients for our agency can be divided into three groups: Caucasian, African-American and Hispanic-American. Furthermore, let's assume that both the African-Americans and Hispanic-Americans are relatively small minorities of the clientele (10% and 5% respectively). If we just did a simple random sample of n=100 with a sampling fraction of 10%, we would expect by chance alone that we would only get 10 and 5 persons from each of our two smaller groups. And, by chance, we could get fewer than that! If we stratify, we can do better. First, let's determine how many people we want to have in each group. Let's say we still want to take a sample of 100 from the population of 1000 clients over the past year. But we think that in order to say anything about subgroups we will need at least 25 cases in each group. So, let's sample 50 Caucasians, 25 African-Americans, and 25 Hispanic-Americans. We know that 10% of the population, or 100 clients, are African-American. If we randomly sample 25 of these, we have a within-stratum sampling fraction of 25/100 = 25%. Similarly, we know that 5% or 50 clients are Hispanic-American. So our within-stratum sampling fraction will be 25/50 = 50%. Finally, by subtraction we know that there are 850 Caucasian clients. Our within-stratum sampling fraction for them is 50/850 = about 5.88%. Because the groups are more homogeneous within-group than across the population as a whole, we can expect greater statistical precision (less variance). And, because we stratified, we know we will have enough cases from each group to make meaningful subgroup inferences.

Systematic Random Sampling

Here are the steps you need to follow in order to achieve a **systematic random sample**:

- number the units in the population from 1 to N
- decide on the n (sample size) that you want or need
- k = N/n = the interval size
- randomly select an integer between 1 to k
- then take every kth unit



N = 100

want n = 20

N/n = 5

select a random number from 1-5: chose 4

start with #4 and take every 5th unit

1	26	51	76
2	27	52	77
3	28	53	78
4	29	54	79
5	30	55	80
6	31	56	81
7	32	57	82
8	33	58	83
9	34	59	84
10	35	60	85
11	36	61	86
12	37	62	87
13	38	63	88
14	39	64	89
15	40	65	90
16	41	66	91
17	42	67	92
18	43	68	93
19	44	69	94
20	45	70	95
21	46	71	96
22	47	72	97
23	48	73	98
24	49	74	99
25	50	75	100

All of this will be much clearer with an example. Let's assume that we have a population that only has N=100 people in it and that you want to take a sample of n=20. To use systematic sampling, the population must be listed in a random order. The sampling fraction would be f = 20/100 = 20%. in this case, the interval size, k, is equal to N/n = 100/20 = 5. Now, select a random integer from 1 to 5. In our example, imagine that you chose 4. Now, to select the sample, start with the 4th unit in the list and take every k-th unit (every 5th, because k=5). You would be sampling units 4, 9, 14, 19, and so on to 100 and you would wind up with 20 units in your sample.

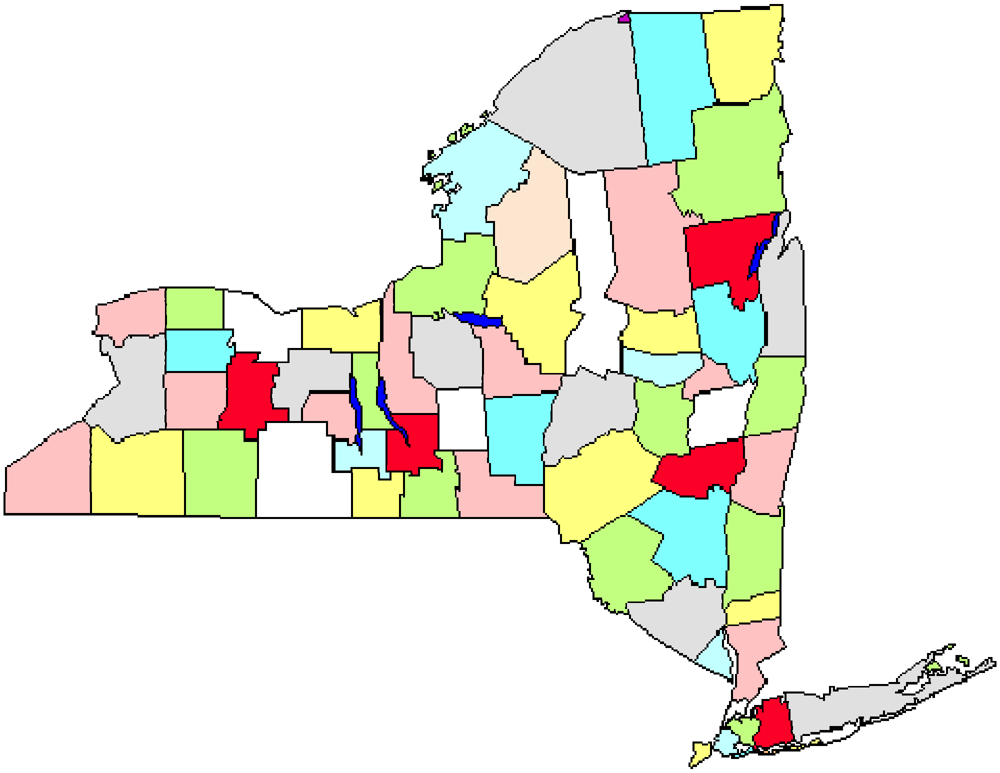
For this to work, it is essential that the units in the population are randomly ordered, at least with respect to the characteristics you are measuring. Why would you ever want to use systematic random sampling? For one thing, it is fairly easy to do. You only have to select a single random number to start things off. It may also be more precise than simple random sampling. Finally, in some situations there is simply no easier way to do random sampling. For instance, I once had to do a study that involved sampling from all the books in a library. Once selected, I would have to go to the shelf, locate the book, and record when it last circulated. I knew that I had a fairly good sampling frame in the form of the shelf list (which is a card catalog where the entries are arranged in the order they occur on the shelf). To do a simple random sample, I could have estimated the total number of books and generated random numbers to draw the sample; but how would I find book #74,329 easily if that is the number I selected? I couldn't very well count the cards until I came to 74,329! Stratifying wouldn't solve that problem either. For instance, I could have stratified by card catalog drawer and drawn a simple random sample within each drawer. But I'd still be stuck counting cards. Instead, I did a systematic random sample. I estimated the number of books in the entire collection. Let's imagine it was 100,000. I decided that I wanted to take a sample of 1000 for a sampling fraction of 1000/100,000 = 1%. To get the sampling interval k, I divided N/n = 100,000/1000 = 100. Then I selected a random integer between 1 and 100. Let's say I got 57. Next I did a little side study to determine how thick a thousand cards are in the card catalog (taking into account the varying ages of the cards). Let's say that on average I found that two cards that were separated by 100 cards were about .75 inches apart in the catalog drawer. That information gave me everything I needed to draw the sample. I counted to the 57th by hand and recorded the book information. Then, I took a compass. (Remember those from your high-school math class? They're the funny little metal instruments with a sharp pin on one end and a pencil on the other that you used to draw circles in geometry class.) Then I set the compass at .75", stuck the pin end in at the 57th card and pointed with the pencil end to the next card (approximately 100 books away). In this way, I approximated selecting the 157th, 257th, 357th, and so on. I was able to accomplish the entire selection procedure in very little time using this systematic random sampling approach. I'd probably still be there counting cards if I'd tried another random sampling method. (Okay, so I have no life. I got compensated nicely, I don't mind saying, for coming up with this scheme.)

Cluster (Area) Random Sampling

The problem with random sampling methods when we have to sample a population that's disbursed across a wide geographic region is that you will have to cover a lot of ground geographically in order to get to each of the units you sampled. Imagine taking a simple random sample of all the residents of New York State in order to conduct personal interviews. By the luck of the draw you will wind up with respondents who come from all over the state. Your interviewers are going to have a lot of traveling to do. It is for precisely this problem that **cluster or area random sampling** was invented.

In cluster sampling, we follow these steps:

- divide population into clusters (usually along geographic boundaries)
- randomly sample clusters
- measure all units within sampled clusters



For instance, in the figure we see a map of the counties in New York State. Let's say that we have to do a survey of town governments that will require us going to the towns personally. If we do a simple random sample state-wide we'll have to cover the entire state geographically. Instead, we decide to do a cluster sampling of five counties (marked in red in the figure). Once these are selected, we go to *every* town government in the five areas. Clearly this strategy will help us to economize on our mileage. Cluster or area sampling, then, is useful in situations like this, and is done primarily for efficiency of administration. Note also, that we probably don't have to worry about using this approach if we are conducting a mail or telephone survey because it doesn't matter as much (or cost more or raise inefficiency) where we call or send letters to.

Multi-Stage Sampling

The four methods we've covered so far -- simple, stratified, systematic and cluster -- are the simplest random sampling strategies. In most real applied social research, we would use sampling methods that are considerably more complex than these simple variations. The most important principle here is that we can combine the simple methods described earlier in a variety of useful ways that help us address our sampling needs in the most efficient and effective manner possible. When we combine sampling methods, we call this **multi-stage sampling**.

For example, consider the idea of sampling New York State residents for face-to-face interviews. Clearly we would want to do some type of cluster sampling as the first stage of the process. We might sample townships or census tracts throughout the state. But in cluster sampling we would then go on to measure everyone in the clusters we select. Even if we are sampling census tracts we may not be able to measure *everyone* who is in the census tract. So, we might set up a stratified sampling process within the clusters. In this case, we would have a two-stage sampling process with stratified samples within cluster samples. Or, consider the problem of sampling students in grade schools. We might begin with a national sample of school districts stratified by economics and educational level. Within selected districts, we might do a simple random sample of schools. Within schools, we might do a simple random sample of classes or grades. And, within classes, we might even do a simple random sample of students. In this case, we have three or four stages in the sampling process and we use both stratified and simple random sampling. By combining different sampling methods we are able to achieve a rich variety of probabilistic sampling methods that can be used in a wide range of social research contexts.

Nonprobability Sampling

The difference between nonprobability and probability sampling is that nonprobability sampling does not involve *random* selection and probability sampling does. Does that mean that nonprobability samples aren't representative of the population? Not necessarily. But it does mean that nonprobability samples cannot depend upon the rationale of probability theory. At least with a probabilistic sample, we know the odds or probability that we have represented the population well. We are able to estimate confidence intervals for the statistic. With nonprobability samples, we may or may not represent the population well, and it will often be hard for us to know how well we've done so. In general, researchers prefer probabilistic or random sampling methods over nonprobabilistic ones, and consider them to be more accurate and rigorous. However, in applied social research there may be circumstances where it is not feasible, practical or theoretically sensible to do random sampling. Here, we consider a wide range of nonprobabilistic alternatives.

We can divide nonprobability sampling methods into two broad types: *accidental* or *purposive*. Most sampling methods are purposive in nature because we usually approach the sampling problem with a specific plan in mind. The most important distinctions among these types of sampling methods are the ones between the different types of purposive sampling approaches.

Accidental, Haphazard or Convenience Sampling

One of the most common methods of sampling goes under the various titles listed here. I would include in this category the traditional "man on the street" (of course, now it's probably the "person on the street") interviews conducted frequently by television news programs to get a quick (although nonrepresentative) reading of public opinion. I would also argue that the typical use of college students in much psychological research is primarily a matter of convenience. (You don't really believe that psychologists use college students because they believe they're representative of the population at large, do you?). In clinical practice,we might use clients who are available to us as our sample. In many research contexts, we sample simply by asking for volunteers. Clearly, the problem with all of these types of samples is that we have no evidence that they are representative of the populations we're interested in generalizing to -- and in many cases we would clearly suspect that they are not.

Purposive Sampling

In purposive sampling, we sample with a *purpose* in mind. We usually would have one or more specific predefined groups we are seeking. For instance, have you ever run into people in a mall or on the street who are carrying a clipboard and who are stopping various people and asking if they could interview them? Most likely they are conducting a purposive sample (and most likely they are engaged in market research). They might be looking for Caucasian females between 30-40 years old. They size up the people passing by and anyone who looks to be in that category they stop to ask if they will participate. One of the first things they're likely to do is verify that the respondent does in fact meet the criteria for being in the sample. Purposive sampling can be very useful for situations where you need to reach a targeted sample quickly and where sampling for proportionality is not the primary concern. With a purposive sample, you are likely to get the opinions of your target population, but you are also likely to overweight subgroups in your population that are more readily accessible.

All of the methods that follow can be considered subcategories of purposive sampling methods. We might sample for specific groups or types of people as in modal instance, expert, or quota sampling. We might sample for diversity as in heterogeneity sampling. Or, we might capitalize on informal social networks to identify specific respondents who are hard to locate otherwise, as in snowball sampling. In all of these methods we know what we want -- we are sampling with a purpose.

- **Modal Instance Sampling**

In statistics, the *mode* is the most frequently occurring value in a distribution. In sampling, when we do a modal instance sample, we are sampling the most frequent case, or the "typical" case. In a lot of informal public opinion polls, for instance, they interview a "typical" voter. There are a number of problems with this sampling approach. First, how do we know what the "typical" or "modal" case is? We could say that the modal voter is a person who is of average age, educational level, and income in the population. But, it's not clear that using the averages of these is the fairest (consider the skewed distribution of income, for instance). And, how do you know that those three variables -- age, education, income -- are the only or event the most relevant for classifying the typical voter? What if religion or ethnicity is an important discriminator? Clearly, modal instance sampling is only sensible for informal sampling contexts.

- **Expert Sampling**

Expert sampling involves the assembling of a sample of persons with known or demonstrable experience and expertise in some area. Often, we convene such a sample under the auspices of a "panel of experts." There are actually two reasons you might do expert sampling. First, because it would be the best way to elicit the views of persons who have specific expertise. In this case, expert sampling is essentially just a specific subcase of purposive sampling. But the other reason you might use expert sampling is to provide evidence for the validity of another sampling approach you've chosen. For instance, let's say you do modal instance sampling and are concerned that the criteria you used for defining the modal instance are subject to criticism. You might convene an expert panel consisting of persons with acknowledged experience and insight into that field or topic and ask them to examine your modal definitions and comment on their appropriateness and validity. The advantage of doing this is that you aren't out on your own trying to defend your decisions -- you have some acknowledged experts to back you. The disadvantage is that even the experts can be, and often are, wrong.

- **Quota Sampling**

In quota sampling, you select people nonrandomly according to some fixed quota. There are two types of quota sampling: *proportional* and *non proportional*. In **proportional quota sampling** you want to represent the major characteristics of the population by sampling a proportional amount of each. For instance, if you know the population has 40% women and 60% men, and that you want a total sample size of 100, you will continue sampling until you get those percentages and then you will stop. So, if you've already got the 40 women for your sample, but not the sixty men, you will continue to sample men but even if legitimate women respondents come along, you will not sample them because you have already "met your quota." The problem here (as in much purposive sampling) is that you have to decide the specific characteristics on which you will base the quota. Will it be by gender, age, education race, religion, etc.?

Nonproportional quota sampling is a bit less restrictive. In this method, you specify the minimum number of sampled units you want in each category. here, you're not concerned with having numbers that match the proportions in the population. Instead, you simply want to have enough to assure that you will be able to talk about even small groups in the population. This method is the nonprobabilistic analogue of stratified random sampling in that it is typically used to assure that smaller groups are adequately represented in your sample.

- **Heterogeneity Sampling**

We sample for heterogeneity when we want to include all opinions or views, and we aren't concerned about representing these views proportionately. Another term for this is sampling for *diversity*. In many brainstorming or nominal group processes (including concept mapping), we would use some form of heterogeneity sampling because our primary interest is in getting broad spectrum of ideas, not identifying the "average" or "modal instance" ones. In effect, what we would like to be sampling is not people, but ideas. We imagine that there is a universe of all possible ideas relevant to some topic and that we want to sample this population, not the population of people who have the ideas. Clearly, in order to get all of the ideas, and especially the "outlier" or unusual ones, we have to include a broad and diverse range of participants. Heterogeneity sampling is, in this sense, almost the opposite of modal instance sampling.

- **Snowball Sampling**

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find. For instance, if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However, if you go to that area and identify one or two, you may find that they know very well who the other homeless people in their vicinity are and how you can find them.

Measurement

Measurement is the process observing and recording the observations that are collected as part of a research effort. There are two major issues that will be considered here.

First, you have to understand the **fundamental ideas** involved in measuring. Here we consider two of major measurement concepts. In Levels of Measurement, I explain the meaning of the four major levels of measurement: nominal, ordinal, interval and ratio. Then we move on to the reliability of measurement, including consideration of true score theory and a variety of reliability estimators.

Second, you have to understand the different **types of measures** that you might use in social research. We consider four broad categories of measurements. Survey research includes the design and implementation of interviews and questionnaires. Scaling involves consideration of the major methods of developing and implementing a scale. Qualitative research provides an overview of the broad range of non-numerical measurement approaches. And unobtrusive measures presents a variety of measurement methods that don't intrude on or interfere with the context of the research.

Construct Validity

Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations were based. Like external validity, construct validity is related to generalizing. But, where external validity involves generalizing from your study context to other people, places or times, construct validity involves generalizing from your program or measures to the *concept* of your program or measures. You might think of construct validity as a "labeling" issue. When you implement a program that you call a "Head Start" program, is your label an accurate one? When you measure what you term "self esteem" is that what you were really measuring?

I would like to tell two major stories here. The first is the more straightforward one. I'll discuss several ways of thinking about the idea of construct validity, several metaphors that might provide you with a foundation in the richness of this idea. Then, I'll discuss the major construct validity threats, the kinds of arguments your critics are likely to raise when you make a claim that your program or measure is valid. In most research methods texts, construct validity is presented in the section on measurement. And, it is typically presented as one of many different types of validity (e.g., face validity, predictive validity, concurrent validity) that you might want to be sure your measures have. I don't see it that way at all. I see construct validity as the overarching quality with all of the other measurement validity labels falling beneath it. And, I don't see construct validity as limited only to measurement. As I've already implied, I think it is as much a part of the independent variable -- the program or treatment -- as it is the dependent variable. So, I'll try to make some sense of the various measurement validity types and try to move you to think instead of the validity of *any* operationalization as falling within the general category of construct validity, with a variety of subcategories and subtypes.

The second story I want to tell is more historical in nature. During World War II, the U.S. government involved hundreds (and perhaps thousands) of psychologists and psychology graduate students in the development of a wide array of measures that were relevant to the war effort. They needed personality screening tests for prospective fighter pilots, personnel measures that would enable sensible assignment of people to job skills, psychophysical measures to test reaction times, and so on. After the war, these psychologists needed to find gainful employment outside of the military context, and it's not surprising that many of them moved into testing and measurement in a civilian context. During the early 1950s, the American Psychological Association began to become increasingly concerned with the quality or validity of all of the new measures that were being generated and decided to convene an effort to set standards for psychological measures. The first formal articulation of the idea of construct validity came from this effort and was couched under the somewhat grandiose idea of the nomological network. The nomological network provided a theoretical basis for the idea of construct validity, but it didn't provide practicing researchers with a way to actually establish whether their measures had construct validity. In 1959, an attempt was made to develop a method for assessing construct validity using what is called a multitrait-multimethod matrix, or MTMM for short. In order to argue that your measures had construct validity under the MTMM approach, you had to demonstrate that there was *both convergent* and *discriminant* validity in your measures. You demonstrated convergent validity when you showed that measures that are theoretically supposed to be highly interrelated are, in practice, highly interrelated. And, you showed discriminant validity when you demonstrated that measures that shouldn't be related to each other in fact were not. While the MTMM did provide a methodology for assessing construct validity, it was a difficult one to implement well, especially in applied social research contexts and, in fact, has seldom been formally attempted. When we examine carefully the thinking about construct validity that underlies both the nomological network and the MTMM, one of the key themes we can identify in both is the idea of "pattern." When we claim that our programs or measures have construct validity, we are essentially claiming that we as researchers understand how our constructs or theories of the programs and measures operate in theory and we claim that we can provide evidence that they behave in practice the way we think they should. The researcher essentially has a theory of how the programs and measures related to each other (and other theoretical terms), a *theoretical pattern* if you will. And, the researcher provides evidence through observation that the programs or measures actually behave that way in reality, an *observed pattern*. When we claim construct validity, we're essentially claiming that our observed pattern -- how things operate in reality -- corresponds with our theoretical pattern -- how we think the world works. I call this process pattern matching, and I believe that it is the heart of construct validity. It is clearly an underlying theme in both the nomological network and the MTMM ideas. And, I think that we can develop concrete and feasible methods that enable practicing researchers to assess pattern matches -- to assess the construct validity of their research. The section on pattern matching lays out my idea of how we might use this approach to assess construct validity.

- **Measurement Validity Types**

There's an awful lot of confusion in the methodological literature that stems from the wide variety of labels that are used to describe the validity of measures. I want to make two cases here. First, it's dumb to limit our scope only to the validity of measures. We really want to talk about the validity of any operationalization. That is, any time you translate a concept or construct into a functioning and operating reality (**the operationalization**), you need to be concerned about how well you did the translation. This issue is as relevant when we are talking about treatments or programs as it is when we are talking about measures. (In fact, come to think of it, we could also think of sampling in this way. The population of interest in your study is the "construct" and the sample is your operationalization. If we think of it this way, we are essentially talking about the construct validity of the sampling!). Second, I want to use the term construct validity to refer to the general case of translating any construct into an operationalization. Let's use all of the other validity terms to reflect different ways you can demonstrate different aspects of construct validity.

With all that in mind, here's a list of the validity types that are typically mentioned in texts and research papers when talking about the quality of measurement:

Construct validity

- **Translation validity**
  - Face validity
  - Content validity
- **Criterion-related validity**
  - Predictive validity
  - Concurrent validity
  - Convergent validity
  - Discriminant validity

I have to warn you here that I made this list up. I've never heard of "translation" validity before, but I needed a good name to summarize what both face and content validity are getting at, and that one seemed sensible. All of the other labels are commonly known, but the way I've organized them is different than I've seen elsewhere.

Let's see if we can make some sense out of this list. First, as mentioned above, I would like to use the term construct validity to be the overarching category. **Construct validity** is the approximate truth of the conclusion that your operationalization accurately reflects its construct. All of the other terms address this general issue in different ways. Second, I make a distinction between two broad types: translation validity and criterion-related validity. That's because I think these correspond to the two major ways you can assure/assess the validity of an operationalization. In **translation validity**, you focus on whether the operationalization is a good reflection of the construct. This approach is definitional in nature -- it assumes you have a good detailed definition of the construct and that you can check the operationalization against it. In **criterion-related validity**, you examine whether the operationalization behaves the way it should given your theory of the construct. This is a more relational approach to construct validity. it assumes that your operationalization should function in predictable ways in relation to other operationalizations based upon your theory of the construct. (If all this seems a bit dense, hang in there until you've gone through the discussion below -- then come back and re-read this paragraph). Let's go through the specific validity types.

Translation Validity

I just made this one up today! (See how easy it is to be a methodologist?) I needed a term that described what both face and content validity are getting at. In essence, both of those validity types are attempting to assess the degree to which you accurately *translated* your construct into the operationalization, and hence the choice of name. Let's look at the two types of translation validity.

Face Validity

In **face validity**, you look at the operationalization and see whether "on its face" it seems like a good translation of the construct. This is probably the weakest way to try to demonstrate construct validity. For instance, you might look at a measure of math ability, read through the questions, and decide that yep, it seems like this is a good measure of math ability (i.e., the label "math ability" seems appropriate for this measure). Or, you might observe a teenage pregnancy prevention program and conclude that, "Yep, this is indeed a teenage pregnancy prevention program." Of course, if this is all you do to assess face validity, it would clearly be weak evidence because it is essentially a subjective judgment call. (Note that just because it is weak evidence doesn't mean that it is wrong. We need to rely on our subjective judgment throughout the research process. It's just that this form of judgment won't be very convincing to others.) We can improve the quality of face validity assessment considerably by making it more systematic. For instance, if you are trying to assess the face validity of a math ability measure, it would be more convincing if you sent the test to a carefully selected sample of experts on math ability testing and they all reported back with the judgment that your measure appears to be a good measure of math ability.

Content Validity

In **content validity**, you essentially check the operationalization against the relevant content domain for the construct. This approach assumes that you have a good detailed description of the content domain, something that's not always true. For instance, we might lay out all of the criteria that should be met in a program that claims to be a "teenage pregnancy prevention program." We would probably include in this domain specification the definition of the target group, criteria for deciding whether the program is preventive in nature (as opposed to treatment-oriented), and lots of criteria that spell out the content that should be included like basic information on pregnancy, the use of abstinence, birth control methods, and so on. Then, armed with these criteria, we could use them as a type of checklist when examining our program. Only programs that meet the criteria can legitimately be defined as "teenage pregnancy prevention programs." This all sounds fairly straightforward, and for many operationalizations it will be. But for other constructs (e.g., self-esteem, intelligence), it will not be easy to decide on the criteria that constitute the content domain.

Criterion-Related Validity

In **criteria-related validity**, you check the performance of your operationalization against some criterion. How is this different from content validity? In content validity, the criteria are the construct definition itself -- it is a direct comparison. In criterion-related validity, we usually make a prediction about how the operationalization will *perform* based on our theory of the construct. The differences among the different criterion-related validity types is in the criteria they use as the standard for judgment.

Predictive Validity

In **predictive validity**, we assess the operationalization's *ability to predict something it should theoretically be able to predict*. For instance, we might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession. We could give our measure to experienced engineers and see if there is a high correlation between scores on the measure and their salaries as engineers. A high correlation would provide evidence for predictive validity -- it would show that our measure can correctly predict something that we theoretically think it should be able to predict.

Concurrent Validity

In **concurrent validity**, we assess the operationalization's *ability to distinguish between groups that it should theoretically be able to distinguish between*. For example, if we come up with a way of assessing manic-depression, our measure should be able to distinguish between people who are diagnosed manic-depression and those diagnosed paranoid schizophrenic. If we want to assess the concurrent validity of a new measure of empowerment, we might give the measure to both migrant farm workers and to the farm owners, theorizing that our measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if you are able to show that you can discriminate between two groups that are very similar.

Convergent Validity

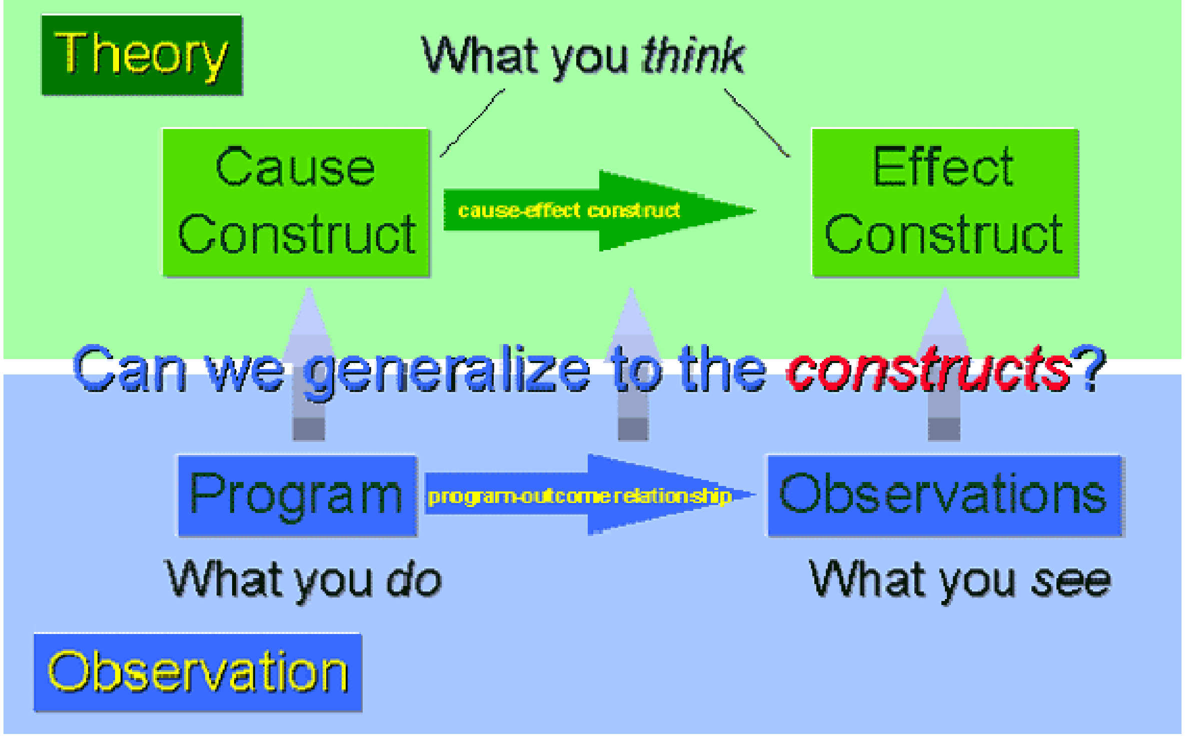
In **convergent validity**, we examine the degree to which the operationalization is similar to (converges on) other operationalizations that it theoretically should be similar to. For instance, to show the convergent validity of a Head Start program, we might gather evidence that shows that the program is similar to other Head Start programs. Or, to show the convergent validity of a test of arithmetic skills, we might correlate the scores on our test with scores on other tests that purport to measure basic math ability, where high correlations would be evidence of convergent validity.

Discriminant Validity

In **discriminant validity**, we examine the degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should be not be similar to. For instance, to show the discriminant validity of a Head Start program, we might gather evidence that shows that the program is *not* similar to other early childhood programs that don't label themselves as Head Start programs. Or, to show the discriminant validity of a test of arithmetic skills, we might correlate the scores on our test with scores on tests that of verbal ability, where *low* correlations would be evidence of discriminant validity.

- **Idea of Construct Validity**

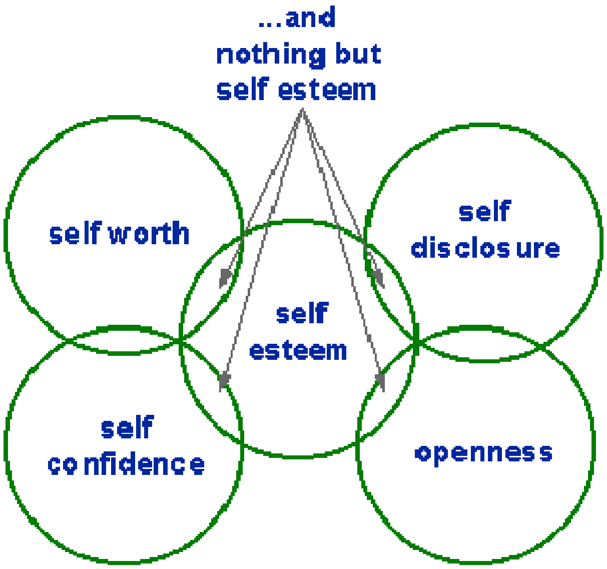
Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations were based. I find that it helps me to divide the issues into two broad territories that I call the "land of theory" and the "land of observation." The land of theory is what goes on inside your mind, and your attempt to explain or articulate this to others. It is all of the ideas, theories, hunches and hypotheses that you have about the world. In the land of theory you will find your idea of the program or treatment as it should be. You will find the idea or construct of the outcomes or measures that you believe you are trying to affect. The land of observation consists of what you see happening in the world around you and the public manifestations of that world. In the land of observation you will find your actual program or treatment, and your actual measures or observational procedures. Presumably, you have constructed the land of observation based on your theories. You developed the program to reflect the kind of program you had in mind. You created the measures to get at what you wanted to get at.



Construct validity is an assessment of how well you translated your ideas or theories into actual programs or measures. Why is this important? Because when you think about the world or talk about it with others (land of theory) you are using words that represent concepts. If you tell someone that a special type of math tutoring will help their child do better in math, you are communicating at the level of concepts or constructs. You aren't describing in operational detail the specific things that the tutor will do with their child. You aren't describing the specific questions that will be on the math test that their child will do better on. You are talking in general terms, using constructs. If you based your recommendation on research that showed that the special type of tutoring improved children' math scores, you would want to be sure that the type of tutoring you are referring to is the same as what that study implemented and that the type of outcome you're saying should occur was the type they measured in their study. Otherwise, you would be mislabeling or misrepresenting the research. In this sense, construct validity can be viewed as a "truth in labeling" kind of issue.

There really are two broad ways of looking at the idea of construct validity. I'll call the first the "definitionalist" perspective because it essentially holds that the way to assure construct validity is to define the construct so precisely that you can operationalize it in a straightforward manner. In a definitionalist view, you have either operationalized the construct correctly or you haven't -- it's an either/or type of thinking. Either this program is a "Type A Tutoring Program" or it isn't. Either you're measuring self esteem or you aren't.

The other perspective I'd call "relationalist." To a relationalist, things are not either/or or black-and-white -- concepts are more or less related to each other. The meaning of terms or constructs differs relatively, not absolutely. The program in your study might be a "Type A Tutoring Program" in some ways, while in others it is not. It might be more that type of program than another program. Your measure might be capturing a lot of the construct of self esteem, but it may not capture all of it. There may be another measure that is closer to the construct of self esteem than yours is. Relationalism suggests that meaning changes gradually. It rejects the idea that we can rely on operational definitions as the basis for construct definition.



To get a clearer idea of this distinction, you might think about how the law approaches the construct of "truth." Most of you have heard the standard oath that a witness in a U.S. court is expected to swear. They are to tell "*the truth, the whole truth and nothing but the truth*." What does this mean? If we only had them swear to tell the truth, they might choose to interpret that as "make sure that what you say is true." But that wouldn't guarantee that they would tell everything they knew to be true. They might leave some important things out. They would still be telling the truth. They just wouldn't be telling everything. On the other hand, they are asked to tell "nothing but the truth." This suggests that we can say simply that Statement X is true and Statement Y is not true.

Now, let's see how this oath translates into a measurement and construct validity context. For instance, we might want our measure to reflect "*the construct, the whole construct, and nothing but the construct*." What does this mean? Let's assume that we have five distinct concepts that are all conceptually related to each other -- self esteem, self worth, self disclosure, self confidence, and openness. Most people would say that these concepts are similar, although they can be distinguished from each other. If we were trying to develop a measure of self esteem, what would it mean to measure "*self esteem, all of self esteem, and nothing but self esteem*?" If the concept of self esteem overlaps with the others, how could we possibly measure all of it (that would presumably include the part that overlaps with others) *and* nothing but it? We couldn't! If you believe that meaning is relational in nature -- that some concepts are "closer" in meaning than others -- then the legal model discussed here does not work well as a model for construct validity.

In fact, we will see that most social research methodologists have (whether they've thought about it or not!) rejected the definitionalist perspective in favor of a relationalist one. In order to establish construct validity you have to meet the following conditions:

- You have to set the construct you want to operationalize (e.g., self esteem) within a **semantic net** (or "net of meaning"). This means that you have to tell us what your construct is more or less similar to in meaning.
- You need to be able to provide direct evidence that you **control** the operationalization of the construct -- that your operationalizations look like what they should theoretically look like. If you are trying to measure self esteem, you have to be able to explain why you operationalized the questions the way you did. If all of your questions are addition

problems, how can you argue that your measure reflects self esteem and not adding ability?

- You have to provide evidence that your **data support your theoretical view** of the relations among constructs. If you believe that self esteem is closer in meaning to self worth than it is to anxiety, you should be able to show that measures of self esteem are more highly correlated with measures of self worth than with ones of anxiety.

- **Convergent & Discriminant Validity**

Convergent and discriminant validity are both considered subcategories or subtypes of construct validity. The important thing to recognize is that they work together -- if you can demonstrate that you have evidence for both convergent and discriminant validity, then you've by definition demonstrated that you have evidence for construct validity. But, neither one alone is sufficient for establishing construct validity.

I find it easiest to think about convergent and discriminant validity as two inter-locking propositions. In simple words I would describe what they are doing as follows:

measures of constructs that theoretically *should* be related to each other are, in fact, observed to be related to each other (that is, you should be able to show a correspondence or *convergence* between similar constructs)

and

measures of constructs that theoretically should *not* be related to each other are, in fact, observed to not be related to each other (that is, you should be able to *discriminate* between dissimilar constructs)

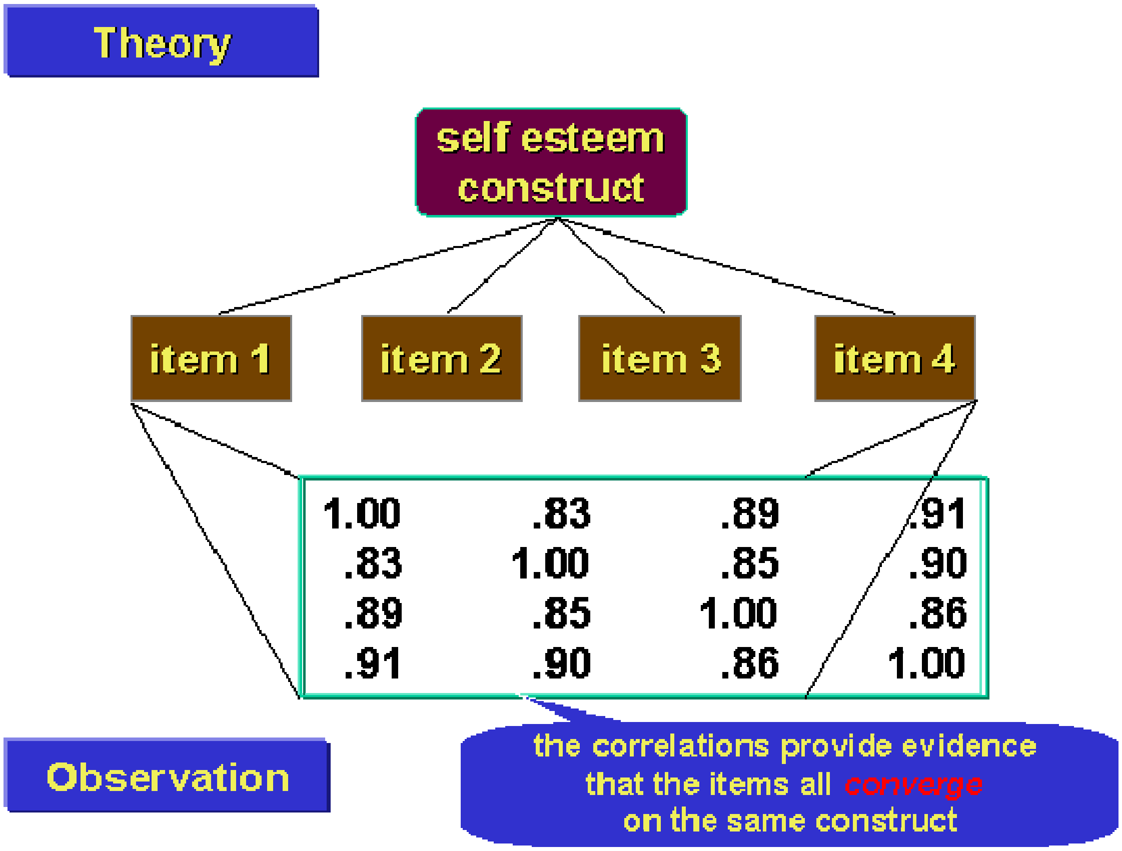
To estimate the degree to which any two measures are related to each other we typically use the correlation coefficient. That is, we look at the patterns of intercorrelations among our measures. Correlations between theoretically similar measures should be "high" while correlations between theoretically dissimilar measures should be "low".

The main problem that I have with this convergent-discrimination idea has to do with my use of the quotations around the terms "high" and "low" in the sentence above. The question is simple -- how "high" do correlations need to be to provide evidence for convergence and how "low" do they need to be to provide evidence for discrimination? And the answer is -- we don't know! In general we want convergent correlations to be as high as possible and discriminant ones to be as low as possible, but there is no hard and fast rule. Well, let's not let that stop us. One thing that we *can* say is that the convergent correlations should always be *higher* than the discriminant ones. At least that helps a bit.

Before we get too deep into the idea of convergence and discrimination, let's take a look at each one using a simple example.

Convergent Validity

To establish convergent validity, you need to show that measures that should be related are in reality related. In the figure below, we see four measures (each is an item on a scale) that all purport to reflect the construct of self esteem. For instance, Item 1 might be the statement "I feel good about myself" rated using a 1-to-5 Likert-type response format. We theorize that all four items reflect the idea of self esteem (this is why I labeled the top part of the figure *Theory*). On the bottom part of the figure (*Observation*) we see the intercorrelations of the four scale items. This might be based on giving our scale out to a sample of respondents. You should readily see that the item intercorrelations for all item pairings are very high (remember that correlations range from -1.00 to +1.00). This provides evidence that our theory that all four items are related to the same construct is supported.



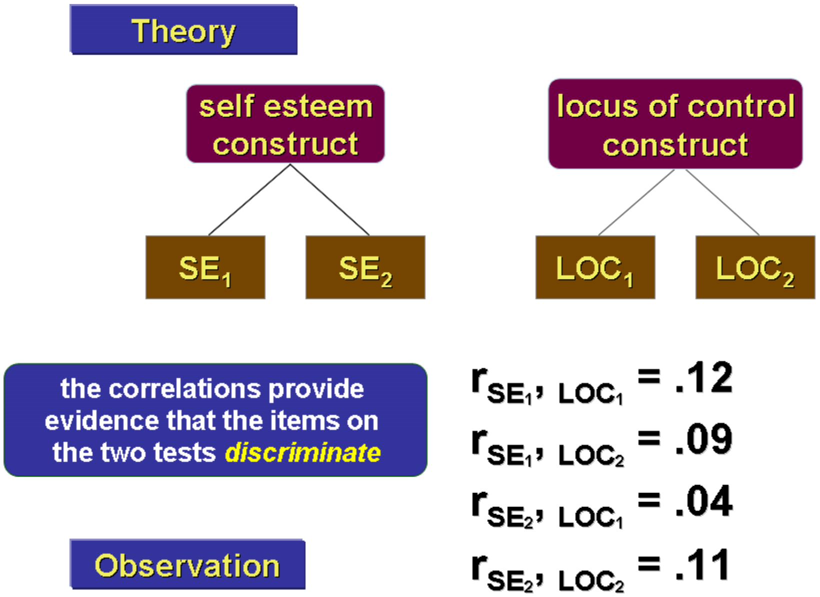
1.00	.83	.89	.91
.83	1.00	.85	.90
.89	.85	1.00	.86
.91	.90	.86	1.00

the correlations provide evidence that the items all converge on the same construct

Notice, however, that while the high intercorrelations demonstrate the the four items are probably related to the *same* construct, that doesn't automatically mean that the construct is *self esteem*. Maybe there's some other construct that all four items are related to (more about this later). But, at the very least, we can assume from the pattern of correlations that the four items are converging on the same thing, whatever we might call it.

Discriminant Validity

To establish discriminant validity, you need to show that measures that should *not* be related are in reality *not* related. In the figure below, we again see four measures (each is an item on a scale). Here, however, two of the items are thought to reflect the construct of self esteem while the other two are thought to reflect locus of control. The top part of the figure shows our theoretically expected relationships among the four items. If we have discriminant validity, the relationship between measures from different constructs should be very low (again, we don't know how low "low" should be, but we'll deal with that later). There are four correlations between measures that reflect different constructs, and these are shown on the bottom of the figure (Observation). You should see immediately that these four cross-construct correlations are very low (i.e., near zero) and certainly much lower than the convergent correlations in the previous figure.



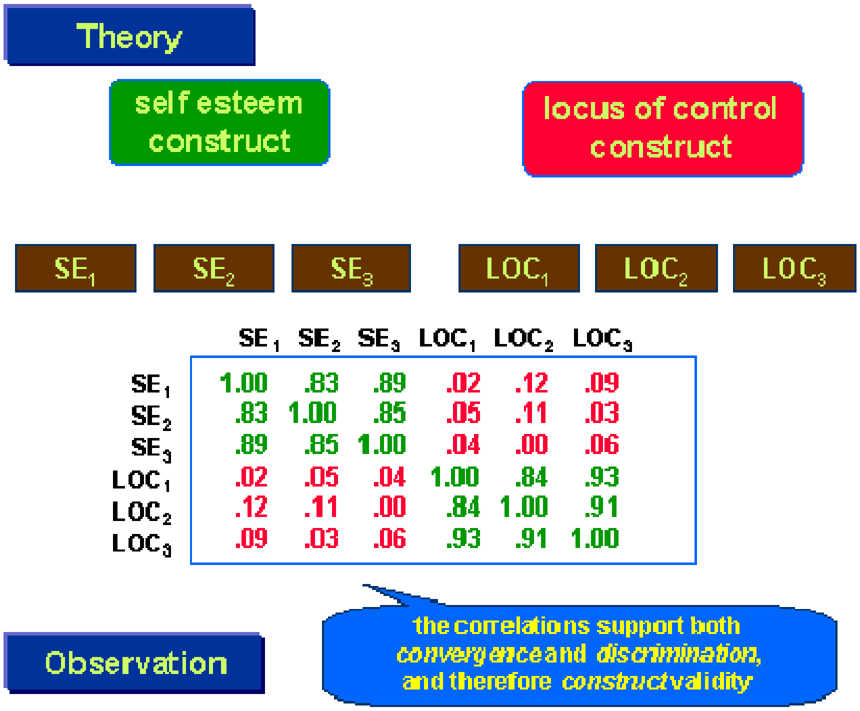
As above, just because we've provided evidence that the two sets of two measures each seem to be related to different constructs (because their intercorrelations are so low) doesn't mean that the constructs they're related to are self esteem and locus of control. But the correlations do provide evidence that the two sets of measures are discriminated from each other.

Putting It All Together

OK, so where does this leave us? I've shown how we go about providing evidence for convergent and discriminant validity separately. But as I said at the outset, in order to argue for construct validity we really need to be able to show that both of these types of validity are supported. Given the above, you should be able to see that we could put both principles together into a single analysis to examine both at the same time. This is illustrated in the figure below.

The figure shows six measures, three that are theoretically related to the construct of self esteem and three that are thought to be related to locus of control. The top part of the figure shows this theoretical arrangement. The bottom of the figure shows what a correlation matrix based on a pilot sample might show. To understand this table, you need to first be able to identify the convergent correlations and the discriminant ones. There are two sets or blocks of convergent coefficients (in green), one 3x3 block for the self esteem intercorrelations and one 3x3 block for the locus of control correlations. There are also two 3x3 blocks of discriminant coefficients (shown in red), although if you're really sharp you'll recognize that they are the same values in mirror image (Do you know why? You might want to read up on correlations to refresh your memory).

How do we make sense of the patterns of correlations? Remember that I said above that we don't have any firm rules for how high or low the correlations need to be to provide evidence for either type of validity. But we *do* know that the convergent correlations should always be higher than the discriminant ones. take a good look at the table and you will see that in this example the convergent correlations are *always* higher than the discriminant ones. I would conclude from this that the correlation matrix provides evidence for both convergent and discriminant validity, all in one analysis!


Theory
self esteem construct
locus of control construct
SE_1
SE_2
SE_3
LOC_1
LOC_2
LOC_3
Observation
the correlations support both *convergence* and *discrimination*, and therefore *construct* validity

	SE_1	SE_2	SE_3	LOC_1	LOC_2	LOC_3
SE_1	1.00	.83	.89	.02	.12	.09
SE_2	.83	1.00	.85	.05	.11	.03
SE_3	.89	.85	1.00	.04	.00	.06
LOC_1	.02	.05	.04	1.00	.84	.93
LOC_2	.12	.11	.00	.84	1.00	.91
LOC_3	.09	.03	.06	.93	.91	1.00

But while the pattern supports discriminant and convergent validity, does it show that the three self esteem measures actually measure self esteem or that the three locus of control measures actually measure locus of control. Of course not. That would be much too easy.

So, what good is this analysis? It does show that, as you predicted, the three self esteem measures seem to reflect the same construct (whatever that might be), the three locus of control measures also seem to reflect the same construct (again, whatever that is) and that the two sets of measures seem to be reflecting two different constructs (whatever they are). That's not bad for one simple analysis.

OK, so how do we get to the really interesting question? How do we show that our measures are actually measuring self esteem or locus of control? I hate to disappoint you, but there is no simple answer to that (I bet you knew that was coming). There's a number of things we can do to address that question. First, we can use other ways to address construct validity to help provide further evidence that we're measuring what we say we're measuring. For instance, we might use a face validity or content validity approach to demonstrate that the measures reflect the constructs we say they are (see the discussion on types of construct validity for more information).

One of the most powerful approaches is to include even more constructs and measures. The more complex our theoretical model (if we find confirmation of the correct pattern in the correlations), the more we are providing evidence that we know what we're talking about (theoretically speaking). Of course, it's also harder to get all the correlations to give you the exact right pattern as you add lots more measures. And, in many studies we simply don't have the luxury to go adding more and more measures because it's too costly or demanding. Despite the impracticality, if we can afford to do it, adding more constructs and measures will enhance our ability to assess construct validity using approaches like the multitrait-multimethod matrix and the nomological network.

Perhaps the most interesting approach to getting at construct validity involves the idea of pattern matching. Instead of viewing convergent and discriminant validity as differences of *kind*, pattern matching views them as differences in *degree*. This seems a more reasonable idea, and helps us avoid the problem of how high or low correlations need to be to say that we've established convergence or discrimination.

• Threats to Construct Validity

Before we launch into a discussion of the most common threats to construct validity, let's recall what a threat to validity is. In a research study you are likely to reach a conclusion that your program was a good operationalization of what you wanted and that your measures reflected what you wanted them to reflect. Would you be correct? How will you be criticized if you make these types of claims? How might you strengthen your claims. The kinds of questions and issues your critics will raise are what I mean by threats to construct validity.

I take the list of threats from the discussion in Cook and Campbell (Cook, T.D. and Campbell, D.T. Quasi-Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston, 1979). While I love their discussion, I do find some of their terminology less than straightforward -- a lot of what I'll do here is try to explain this stuff in terms that the rest of us might hope to understand.

Inadequate Preoperational Explication of Constructs

This one isn't nearly as ponderous as it sounds. Here, **preoperational** means *before translating constructs into measures or treatments*, and **explication** means *explanation* -- in other words, you didn't do a good enough job of *defining* (operationally) what you mean by the construct. How is this a threat? Imagine that your program consisted of a new type of approach to rehabilitation. Your critic comes along and claims that, in fact, your program is neither *new* nor a true *rehabilitation* program. You are being accused of doing a poor job of thinking through your constructs. Some possible solutions:

- think through your concepts better
- use methods (e.g., concept mapping) to articulate your concepts
- get experts to critique your operationalizations

Mono-Operation Bias

Mono-operation bias pertains to the independent variable, cause, program or treatment in your study -- it does not pertain to measures or outcomes (see Mono-method Bias below). If you only use a single version of a program in a single place at a single point in time, you may not be capturing the full breadth of the concept of the program. Every operationalization is flawed relative to the construct on which it is based. If you conclude that your program reflects the construct of the program, your critics are likely to argue that the results of your study only reflect the peculiar version of the program that you implemented, and not the actual construct you had in mind. Solution: try to implement multiple versions of your program.

Mono-Method Bias

Mono-method bias refers to your measures or observations, not to your programs or causes. Otherwise, it's essentially the same issue as mono-operation bias. With only a single version of a self esteem measure, you can't provide much evidence that you're really measuring self esteem. Your critics will suggest that you aren't measuring self esteem -- that you're only measuring part of it, for instance. Solution: try to implement multiple measures of key constructs and try to demonstrate (perhaps through a pilot or side study) that the measures you use behave as you theoretically expect them to.

Interaction of Different Treatments

You give a new program designed to encourage high-risk teenage girls to go to school and not become pregnant. The results of your study show that the girls in your treatment group have higher school attendance and lower birth rates. You're feeling pretty good about your program until your critics point out that the targeted at-risk treatment group in your study is also likely to be involved simultaneously in several other programs designed to have similar effects. Can you really label the program effect as a consequence of your program? The "real" program that the girls received may actually be the *combination* of the separate programs they participated in.

Interaction of Testing and Treatment

Does testing or measurement itself make the groups more sensitive or receptive to the treatment? If it does, then the testing is in effect a part of the treatment, it's inseparable from the effect of the treatment. This is a labeling issue (and, hence, a concern of construct validity) because you want to use the label "program" to refer to the program alone, but in fact it includes the testing.

Restricted Generalizability Across Constructs

This is what I like to refer to as the "unintended consequences" treat to construct validity. You do a study and conclude that Treatment X is effective. In fact, Treatment X does cause a reduction in symptoms, but what you failed to anticipate was the drastic negative consequences of the side effects of the treatment. When you say that Treatment X is effective, you have defined "effective" as only the directly targeted symptom. This threat reminds us that we have to be careful about whether our observed effects (Treatment X is effective) would generalize to other potential outcomes.

Confounding Constructs and Levels of Constructs

Imagine a study to test the effect of a new drug treatment for cancer. A fixed dose of the drug is given to a randomly assigned treatment group and a placebo to the other group. No treatment effects are detected. Perhaps the result that's observed is only true for that dosage level. Slight increases or decreases of the dosage may radically change the results. In this context, it is not "fair" for you to use the label for the drug as a description for your treatment because you only looked at a narrow range of dose. Like the other construct validity threats, this is essentially a labeling issue -- your label is not a good description for what you implemented.

The "Social" Threats to Construct Validity

I've set aside the other major threats to construct validity because they all stem from the social and human nature of the research endeavor.

Hypothesis Guessing

Most people don't just participate passively in a research project. They are trying to figure out what the study is about. They are "guessing" at what the real purpose of the study is. And, they are likely to base their behavior on what they guess, not just on your treatment. In an educational study conducted in a classroom, students might guess that the key dependent variable has to do with class participation levels. If they increase their participation not because of your program but because they think that's what you're studying, then you cannot label the outcome as an effect of the program. It is this labeling issue that makes this a construct validity threat.

Evaluation Apprehension

Many people are anxious about being evaluated. Some are even phobic about testing and measurement situations. If their apprehension makes them perform poorly (and not your program conditions) then you certainly can't label that as a treatment effect. Another form of evaluation apprehension concerns the human tendency to want to "look good" or "look smart" and so on. If, in their desire to look good, participants perform better (and not as a result of your program!) then you would be wrong to label this as a treatment effect. In both cases, the apprehension becomes confounded with the treatment itself and you have to be careful about how you label the outcomes.

Experimenter Expectancies

These days, where we engage in lots of non-laboratory applied social research, we generally don't use the term "experimenter" to describe the person in charge of the research. So, let's relabel this threat "researcher expectancies." The researcher can bias the results of a study in countless ways, both consciously or unconsciously. Sometimes the researcher can communicate what the desired outcome for a study might be (and participant desire to "look good" leads them to react that way). For instance, the researcher might look pleased when participants give a desired answer. If this is what causes the response, it would be wrong to label the response as a treatment effect.

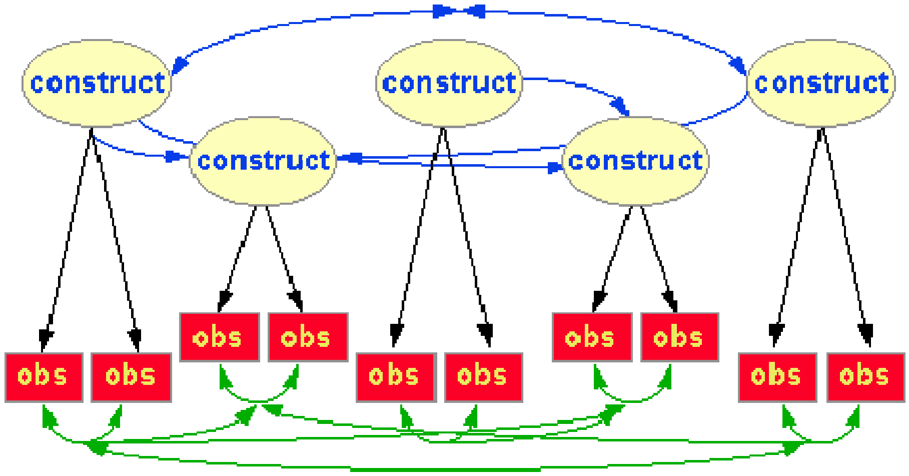
- **The Nomological Network**

What is the Nomological Net?

The **nomological network** is an idea that was developed by Lee Cronbach and Paul Meehl in 1955 (Cronbach, L. and Meehl, P. (1955). Construct validity in psychological tests, *Psychological Bulletin*, 52, 4, 281-302.) as part of the American Psychological Association's efforts to develop standards for psychological testing. The term "nomological" is derived from Greek and means "lawful", so the nomological network can be thought of as the "lawful network." The nomological network was Cronbach and Meehl's view of construct validity. That is, in order to provide evidence that your measure has construct validity, Cronbach and Meehl argued that you had to develop a nomological network for your measure. This network would include the theoretical framework for what you are trying to measure, an empirical framework for how you are going to measure it, and specification of the linkages among and between these two frameworks.

The Nomological Network

a representation of the concepts (constructs) of interest in a study,



...their observable manifestations, *and the interrelationships among and between these*

The nomological network is founded on a number of principles that guide the researcher when trying to establish construct validity. They are:

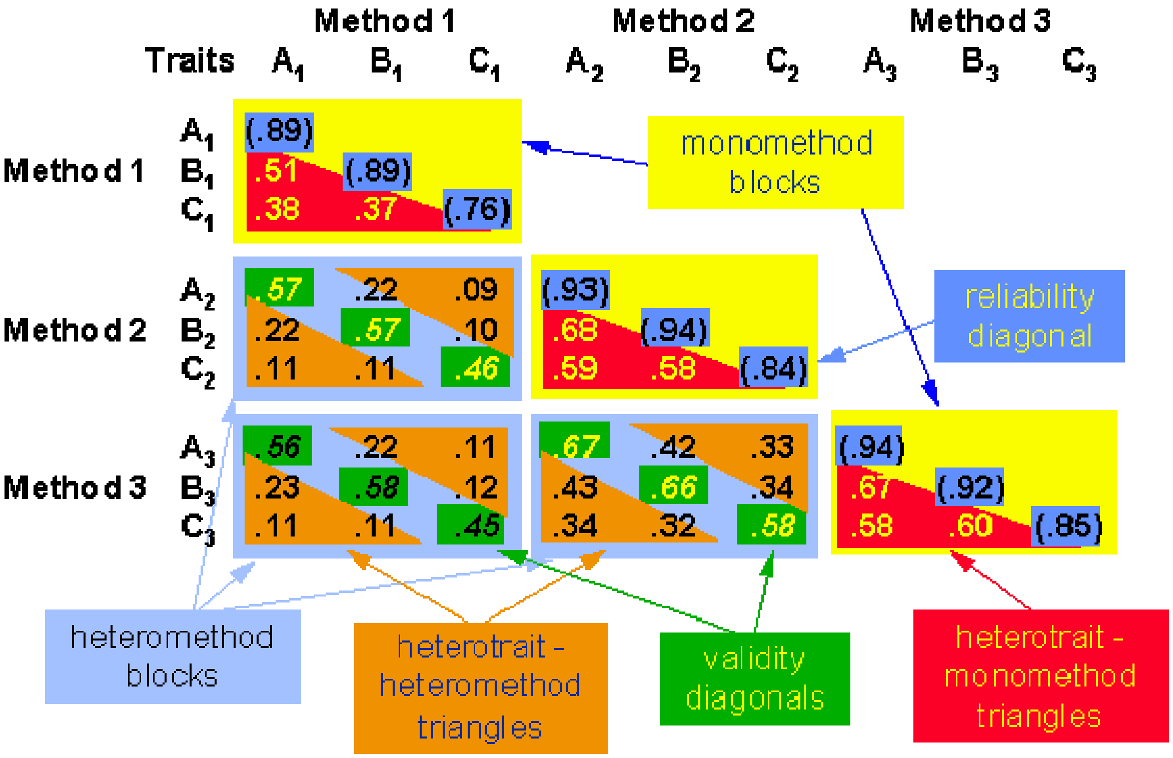
- Scientifically, to make clear what something is or means, so that laws can be set forth in which that something occurs.
- The laws in a nomological network may relate:
    - observable properties or quantities to each other
    - different theoretical constructs to each other
    - theoretical constructs to observables
- At least some of the laws in the network must involve observables.
- "Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs or of increasing the definiteness of its components.
- The basic rule for adding a new construct or relation to a theory is that it must generate laws (nomologicals) confirmed by observation or reduce the number of nomologicals required to predict some observables.
- Operations which are qualitatively different "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable.

What Cronbach and Meehl were trying to do is to link the conceptual/theoretical realm with the observable one, because this is the central concern of construct validity. While the nomological network idea may work as a philosophical foundation for construct validity, it does not provide a practical and usable methodology for actually assessing construct validity. The next phase in the evolution of the idea of construct validity -- the development of the multitrait-multimethod matrix -- moved us a bit further toward a methodological approach to construct validity.

- **The Multitrait-Multimethod Matrix**

What is the Multitrait-Multimethod Matrix?

The Multitrait-Multimethod Matrix (hereafter labeled MTMM) is an approach to assessing the construct validity of a set of measures in a study. It was developed in 1959 by Campbell and Fiske (Campbell, D. and Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. 56, 2, 81-105.) in part as an attempt to provide a practical methodology that researchers could actually use (as opposed to the nomological network idea which was theoretically useful but did not include a methodology). Along with the MTMM, Campbell and Fiske introduced two new types of validity -- convergent and discriminant -- as subcategories of construct validity. **Convergent validity** is the degree to which concepts that should be related theoretically are interrelated in reality. **Discriminant validity** is the degree to which concepts that should *not* be related theoretically are, in fact, *not* interrelated in reality. You can assess both convergent and discriminant validity using the MTMM. In order to be able to claim that your measures have construct validity, you have to demonstrate both convergence and discrimination.



		Method 1			Method 2			Method 3		
	Traits	A_1	B_1	C_1	A_2	B_2	C_2	A_3	B_3	C_3
Method 1	A_1	(.89)								
	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
Method 2	A_2	.57	.22	.09	(.93)					
	B_2	.22	.57	.10	.68	(.94)				
	C_2	.11	.11	.46	.59	.58	(.84)			
Method 3	A_3	.56	.22	.11	.67	.42	.33	(.94)		
	B_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The MTMM is simply a matrix or table of correlations arranged to facilitate the interpretation of the assessment of construct validity. The MTMM assumes that you measure each of several concepts (called *traits* by Campbell and Fiske) by each of several methods (e.g., a paper-and-pencil test, a direct observation, a performance measure). The MTMM is a very restrictive methodology -- ideally you should measure *each* concept by *each* method.

To construct an MTMM, you need to arrange the correlation matrix by concepts within methods. The figure shows an MTMM for three concepts (traits A, B and C) each of which is measured with three different methods (1, 2 and 3) Note that you lay the matrix out in blocks by *method*. Essentially, the MTMM is just a correlation matrix between your measures, with one exception -- instead of 1's along the diagonal (as in the typical correlation matrix) we substitute an estimate of the reliability of each measure as the diagonal.

Before you can interpret an MTMM, you have to understand how to identify the different parts of the matrix. First, you should note that the matrix is consists of nothing but correlations. It is a square, symmetric matrix, so we only need to look at half of it (the figure shows the lower triangle). Second, these correlations can be grouped into three kinds of shapes: diagonals, triangles, and blocks. The specific shapes are:

- **The Reliability Diagonal
(monotrait-monomethod)**

Estimates of the reliability of each measure in the matrix. You can estimate reliabilities a number of different ways (e.g., test-retest, internal consistency). There are as many correlations in the reliability diagonal as there are measures -- in this example there are nine measures and nine reliabilities. The first reliability in the example is the correlation of Trait A, Method 1 with Trait A, Method 1 (hereafter, I'll abbreviate this relationship A1-A1). Notice that this is essentially the correlation of the measure with itself. In fact such a correlation would always be perfect (i.e., r=1.0). Instead, we substitute an estimate of reliability. You could also consider these values to be monotrait-monomethod correlations.

- **The Validity Diagonals
(monotrait-heteromethod)**

Correlations between measures of the same trait measured using different methods. Since the MTMM is organized into method blocks, there is one validity diagonal in each method block. For example, look at the A1-A2 correlation of .57. This is the correlation between two measures of the same trait (A) measured with two different measures (1 and 2). Because the two measures are of the same trait or concept, we would expect them to be strongly correlated. You could also consider these values to be monotrait-heteromethod correlations.

- **The Heterotrait-Monomethod Triangles**

These are the correlations among measures that share the same method of measurement. For instance, A1-B1 = .51 in the upper left heterotrait-monomethod triangle. Note that what these correlations share is method, not trait or concept. If these correlations are high, it is because measuring different things with the same method results in correlated measures. Or, in more straightforward terms, you've got a strong "methods" factor.

- **Heterotrait-Heteromethod Triangles**

These are correlations that differ in both trait and method. For instance, A1-B2 is .22 in the example. Generally, because these correlations share neither trait nor method we expect them to be the lowest in the matrix.

- **The Monomethod Blocks**

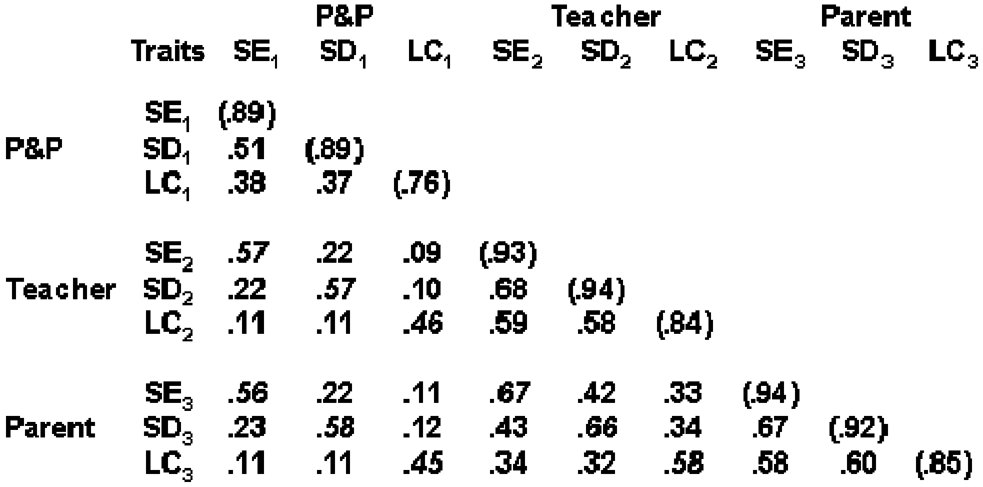
These consist of all of the correlations that share the same method of measurement. There are as many blocks as there are methods of measurement.

- **The Heteromethod Blocks**

These consist of all correlations that do *not* share the same methods. There are (K(K-1))/2 such blocks, where K = the number of methods. In the example, there are 3 methods and so there are (3(3-1))/2 = (3(2))/2 = 6/2 = 3 such blocks.

Principles of Interpretation

Now that you can identify the different parts of the MTMM, you can begin to understand the rules for interpreting it. You should realize that MTMM interpretation requires the researcher to use judgment. Even though some of the principles may be violated in an MTMM, you may still wind up concluding that you have fairly strong construct validity. In other words, you won't necessarily get *perfect* adherence to these principles in applied research settings, even when you do have evidence to support construct validity. To me, interpreting an MTMM is a lot like a physician's reading of an x-ray. A practiced eye can often spot things that the neophyte misses! A researcher who is experienced with MTMM can use it identify weaknesses in measurement as well as for assessing construct validity.



		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	*.57*	.22	.09	(.93)					
	SD_2	.22	*.57*	.10	.68	(.94)				
	LC_2	.11	.11	*.46*	.59	.58	(.84)			
Parent	SE_3	*.56*	.22	.11	*.67*	.42	.33	(.94)		
	SD_3	.23	*.58*	.12	.43	*.66*	.34	.67	(.92)	
	LC_3	.11	.11	*.45*	.34	.32	*.58*	.58	.60	(.85)

To help make the principles more concrete, let's make the example a bit more realistic. We'll imagine that we are going to conduct a study of sixth grade students and that we want to measure three traits or concepts: Self Esteem (SE), Self Disclosure (SD) and Locus of Control (LC). Furthermore, let's measure each of these three different ways: a Paper-and-Pencil (P&P) measure, a Teacher rating, and a Parent rating. The results are arrayed in the MTMM. As the principles are presented, try to identify the appropriate coefficients in the MTMM and make a judgement yourself about the strength of construct validity claims.

The basic principles or rules for the MTMM are:

- Coefficients in the reliability diagonal should consistently be the highest in the matrix.

That is, a trait should be more highly correlated with itself than with anything else! This is uniformly true in our example.

- Coefficients in the validity diagonals should be significantly different from zero and high enough to warrant further investigation.

This is essentially evidence of convergent validity. All of the correlations in our example meet this criterion.

- A validity coefficient should be higher than values lying in its column and row in the same heteromethod block.

In other words, (SE P&P)-(SE Teacher) should be greater than (SE P&P)-(SD Teacher), (SE P&P)-(LC Teacher), (SE Teacher)-(SD P&P) and (SE Teacher)-(LC P&P). This is true in all cases in our example.

- A validity coefficient should be higher than all coefficients in the heterotrait-monomethod triangles.

This essentially emphasizes that trait factors should be stronger than methods factors. Note that this is *not* true in all cases in our example. For instance, the (LC P&P)-(LC Teacher) correlation of .46 is less than (SE Teacher)-(SD Teacher), (SE Teacher)-(LC Teacher), and (SD Teacher)-(LC Teacher) -- evidence that there might me a methods factor, especially on the Teacher observation method.

- The same *pattern* of trait interrelationship should be seen in all triangles.

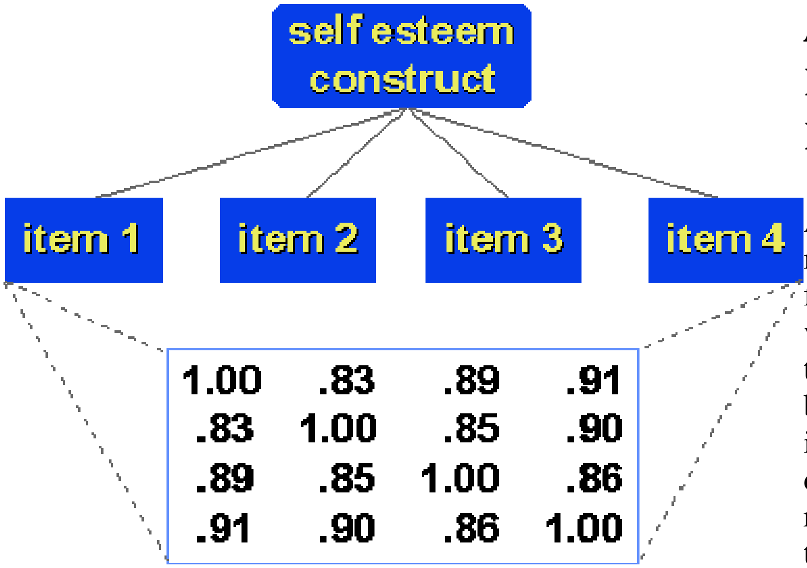
The example clearly meets this criterion. Notice that in all triangles the SE-SD relationship is approximately twice as large as the relationships that involve LC.

Advantages and Disadvantages of MTMM

The MTMM idea provided an operational methodology for assessing construct validity. In the one matrix it was possible to examine both convergent and discriminant validity simultaneously.

By its inclusion of methods on an equal footing with traits, Campbell and Fiske stressed the importance of looking for the effects of how we measure in addition to what we measure. And, MTMM provided a rigorous framework for assessing construct validity.

Despite these advantages, MTMM has received little use since its introduction in 1959. There are several reasons. First, in its purest form, MTMM requires that you have a fully-crossed measurement design -- each of several traits is measured by each of several methods. While Campbell and Fiske explicitly recognized that one could have an incomplete design, they stressed the importance of multiple replication of the same trait across method. In some applied research contexts, it just isn't possible to measure all traits with all desired methods (would you use an "observation" of weight?). In most applied social research, it just wasn't feasible to make methods an explicit part of the research design. Second, the judgmental nature of the MTMM may have worked against its wider adoption (although it should actually be perceived as a strength). many researchers wanted a test for construct validity that would result in a single statistical coefficient that could be tested -- the equivalent of a reliability coefficient. It was impossible with MTMM to quantify the *degree* of construct validity in a study. Finally, the judgmental nature of MTMM meant that different researchers could legitimately arrive at different conclusions.



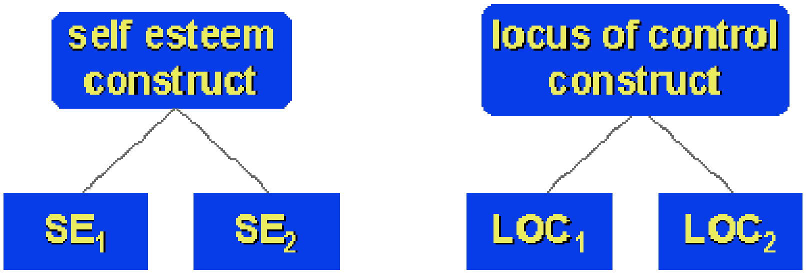
1.00	.83	.89	.91
.83	1.00	.85	.90
.89	.85	1.00	.86
.91	.90	.86	1.00

A Modified MTMM -- Leaving out the Methods Factor

As mentioned above, one of the most difficult aspects of MTMM from an implementation point of view is that it required a design that included all combinations of both traits and methods. But the ideas of convergent and discriminant validity do not require the methods factor. To see this, we have to reconsider what Campbell and Fiske meant by convergent and discriminant validity.

What is convergent validity?

It is the principle that *measures of theoretically similar constructs should be highly intercorrelated*. We can extend this idea further by thinking of a measure that has multiple items, for instance, a four-item scale designed to measure self-esteem. If each of the items actually does reflect the construct of self-esteem, then we would expect the items to be highly intercorrelated as shown in the figure. These strong intercorrelations are evidence in support of convergent validity.



And what is discriminant validity?

It is the principle that *measures of theoretically different constructs should not correlate highly with each other*. We can see that in the example that shows two constructs -- self-esteem and locus of control -- each measured in two instruments. We would expect that, because these are measures of different constructs, the cross-construct correlations would be low, as shown in the figure. These low correlations are evidence for validity. Finally, we can put this all together to see how we can address both convergent and discriminant validity simultaneously. Here, we have two constructs -- self-esteem and locus of control -- each measured with three instruments. The red and green correlations are within-construct ones. They are a reflection of convergent validity and should be strong. The blue correlations are cross-construct and reflect discriminant validity. They should be uniformly lower than the convergent coefficients.

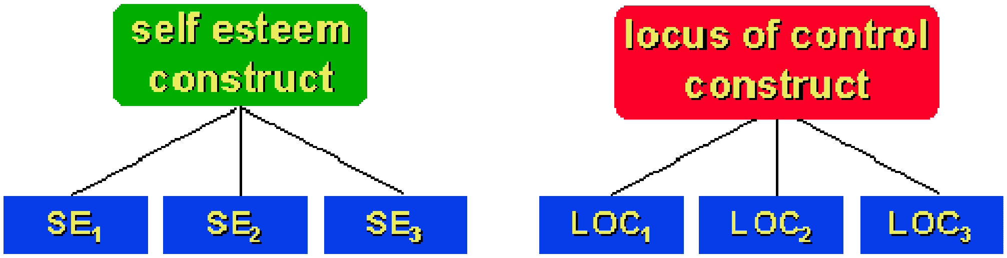
$$r_{SE_1, LOC_1} = .12$$

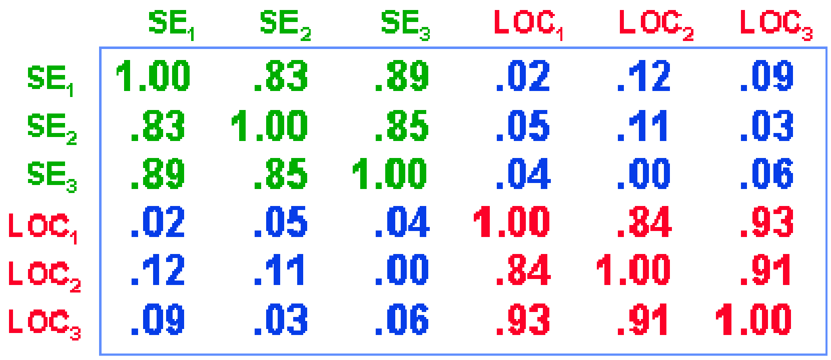
$$r_{SE_1, LOC_2} = .09$$

$$r_{SE_2, LOC_1} = .04$$

$$r_{SE_2, LOC_2} = .11$$

The important thing to notice about this matrix is that *it does not explicitly include a methods factor* as a true MTMM would. The matrix examines both convergent and discriminant validity (like the MTMM) but it only explicitly looks at construct intra- and interrelationships. We can see in this example that the MTMM idea really had two major themes. The first was the idea of looking simultaneously at the pattern of convergence and discrimination. This idea is similar in purpose to the notions implicit in the nomological network -- we are looking at the pattern of interrelationships based upon our theory of the nomological net. The second idea in MTMM was the emphasis on methods as a potential **confounding factor**.





	SE_1	SE_2	SE_3	LOC_1	LOC_2	LOC_3
SE_1	1.00	.83	.89	.02	.12	.09
SE_2	.83	1.00	.85	.05	.11	.03
SE_3	.89	.85	1.00	.04	.00	.06
LOC_1	.02	.05	.04	1.00	.84	.93
LOC_2	.12	.11	.00	.84	1.00	.91
LOC_3	.09	.03	.06	.93	.91	1.00

While methods may confound the results, they won't necessarily do so in any given study. And, while we need to examine our results for the potential for methods factors, it may be that combining this desire to assess the confound with the need to assess construct validity is more than one methodology can feasibly handle. Perhaps if we split the two agendas, we will find that the possibility that we can examine convergent and discriminant validity is greater. But what do we do about methods factors? One way to deal with them is through replication of research projects, rather than trying to incorporate a methods test into a single research study. Thus, if we find a particular outcome in a study using several measures, we might see if that same outcome is obtained when we replicate the study using different measures and methods of measurement for the same constructs. The methods issue is considered more as an issue of generalizability (across measurement methods) rather than one of construct validity.

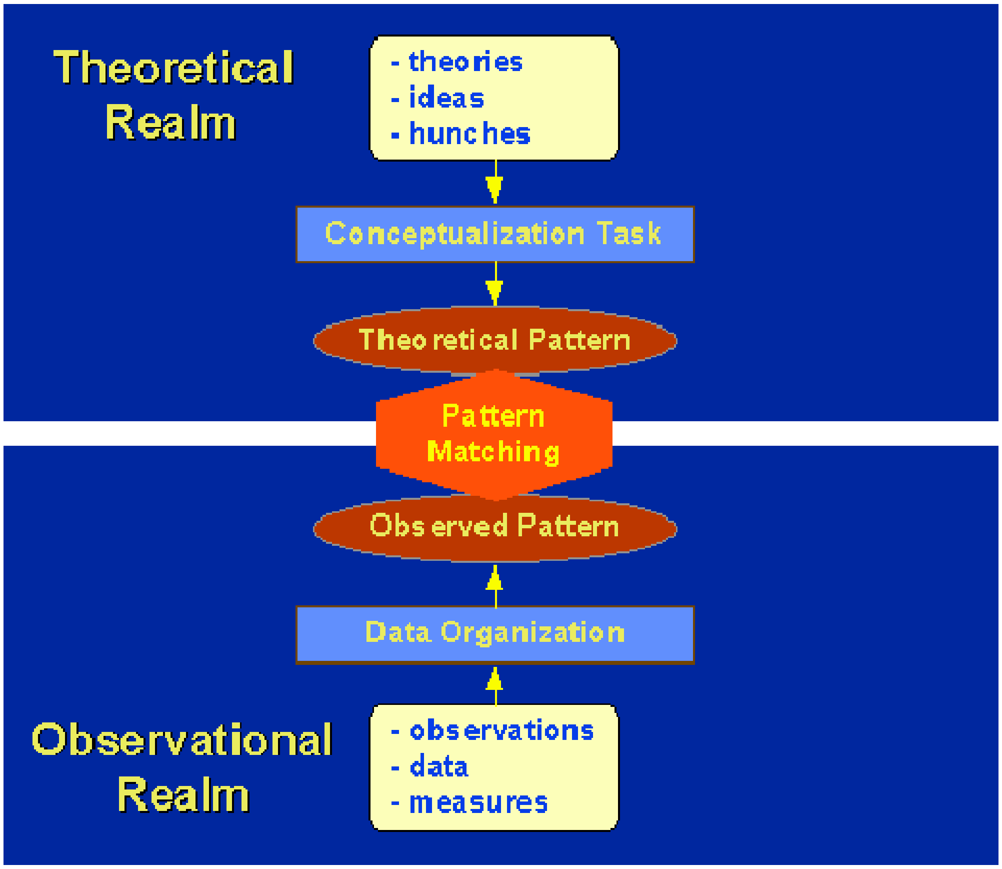
When viewed this way, we have moved from the idea of a MTMM to that of the multitrait matrix that enables us to examine convergent and discriminant validity, and hence construct validity. We will see that when we move away from the explicit consideration of methods and when we begin to see convergence and discrimination as differences of degree, we essentially have the foundation for the pattern matching approach to assessing construct validity.

- **Pattern Matching for Construct Validity**

The idea of using pattern matching as a rubric for assessing construct validity is an area where I have tried to make a contribution (Trochim, W., (1985). Pattern matching, validity, and conceptualization in program evaluation. Evaluation Review, 9, 5, 575-604 and Trochim, W. (1989). Outcome pattern matching and program theory. Evaluation and Program Planning, 12, 355-366.), although my work was very clearly foreshadowed, especially in much of Donald T. Campbell's writings. Here, I'll try to explain what I mean by pattern matching with respect to construct validity.

The Theory of Pattern Matching

A pattern is any arrangement of objects or entities. The term "arrangement" is used here to indicate that a pattern is by definition non-random and at least potentially describable. All theories imply some pattern, but theories and patterns are not the same thing. In general, a theory postulates structural relationships between key constructs. The theory can be used as the basis for generating patterns of predictions. For instance, E=MC2 can be considered a theoretical formulation. A pattern of expectations can be developed from this formula by generating predicted values for one of these variables given fixed values of the others. Not all theories are stated in mathematical form, especially in applied social research, but all theories provide information that enables the generation of patterns of predictions.



Pattern matching always involves an attempt to link two patterns where one is a theoretical pattern and the other is an observed or operational one. The top part of the figure shows the realm of theory. The theory might originate from a formal tradition of theorizing, might be the ideas or "hunches" of the investigator, or might arise from some combination of these. The conceptualization task involves the translation of these ideas into a specifiable theoretical pattern indicated by the top shape in the figure. The bottom part of the figure indicates the realm of observation. This is broadly meant to include direct observation in the form of impressions, field notes, and the like, as well as more formal objective measures. The collection or organization of relevant operationalizations (i.e., relevant to the theoretical pattern) is termed the observational pattern and is indicated by the lower shape in the figure. The inferential task involves the attempt to relate, link or match these two patterns as indicated by the double arrow in the center of the figure. To the extent that the patterns match, one can conclude that the theory and any other theories which might predict the same observed pattern receive support.

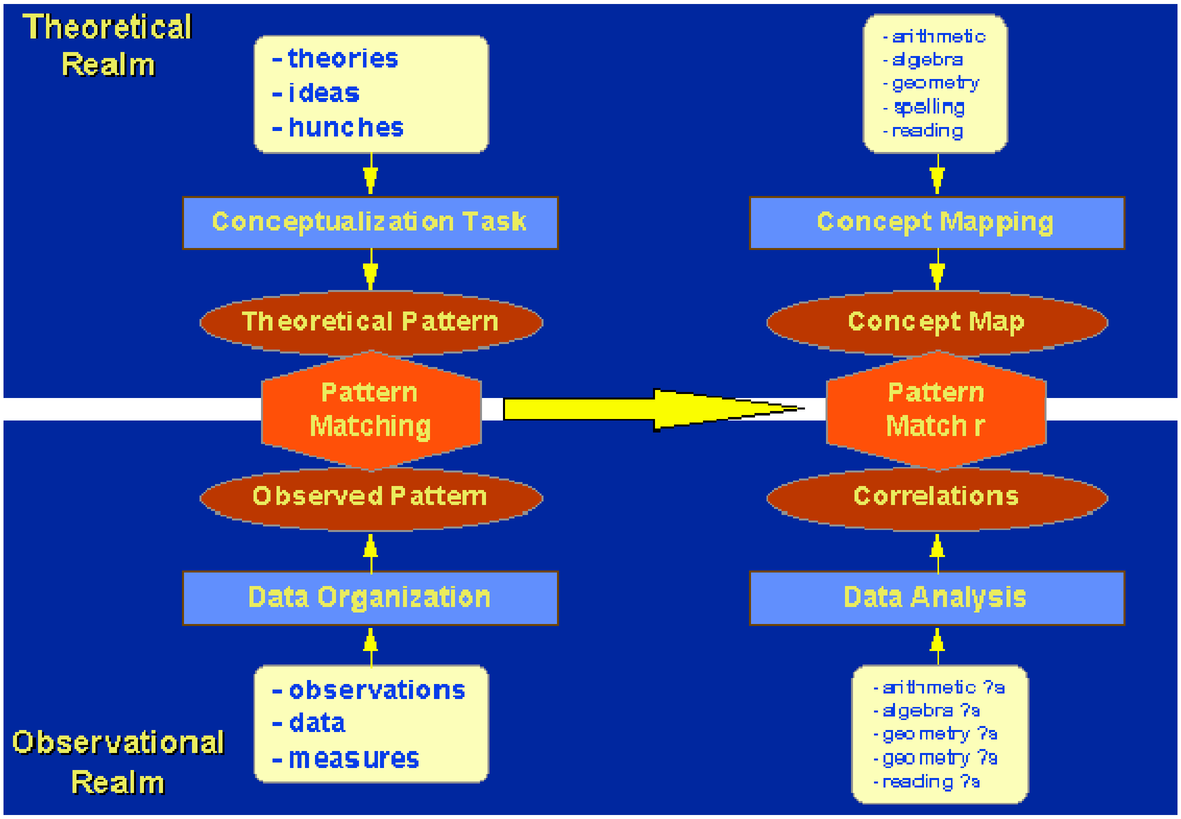
It is important to demonstrate that there are no plausible alternative theories that account for the observed pattern and this task is made much easier when the theoretical pattern of interest is a unique one. In effect, a more complex theoretical pattern is like a unique fingerprint which one is seeking in the observed pattern. With more complex theoretical patterns it is usually more difficult to construe sensible alternative patterns that would also predict the same result. To the extent that theoretical and observed patterns do not match, the theory may be incorrect or poorly formulated, the observations may be inappropriate or inaccurate, or some combination of both states may exist.

All research employs pattern matching principles, although this is seldom done consciously. In the traditional two-group experimental context, for instance, the typical theoretical outcome pattern is the hypothesis that there will be a significant difference between treated and untreated groups. The observed outcome pattern might consist of the averages for the two groups on one or more measures. The pattern match is accomplished by a test of significance such as the t-test or ANOVA. In survey research, pattern matching forms the basis of generalizations across different concepts or population subgroups. In qualitative research pattern matching lies at the heart of any attempt to conduct thematic analyses.

While current research methods can be described in pattern matching terms, the idea of pattern matching implies more, and suggests how one might improve on these current methods. Specifically, pattern matching implies that *more complex patterns, if matched, yield greater validity for the theory*. Pattern matching does not differ fundamentally from traditional hypothesis testing and model building approaches. A theoretical pattern is a hypothesis about what is expected in the data. The observed pattern consists of the data that are used to examine the theoretical model. The major differences between pattern matching and more traditional hypothesis testing approaches are that pattern matching encourages the use of more complex or detailed hypotheses and treats the observations from a multivariate rather than a univariate perspective.

Pattern Matching and Construct Validity

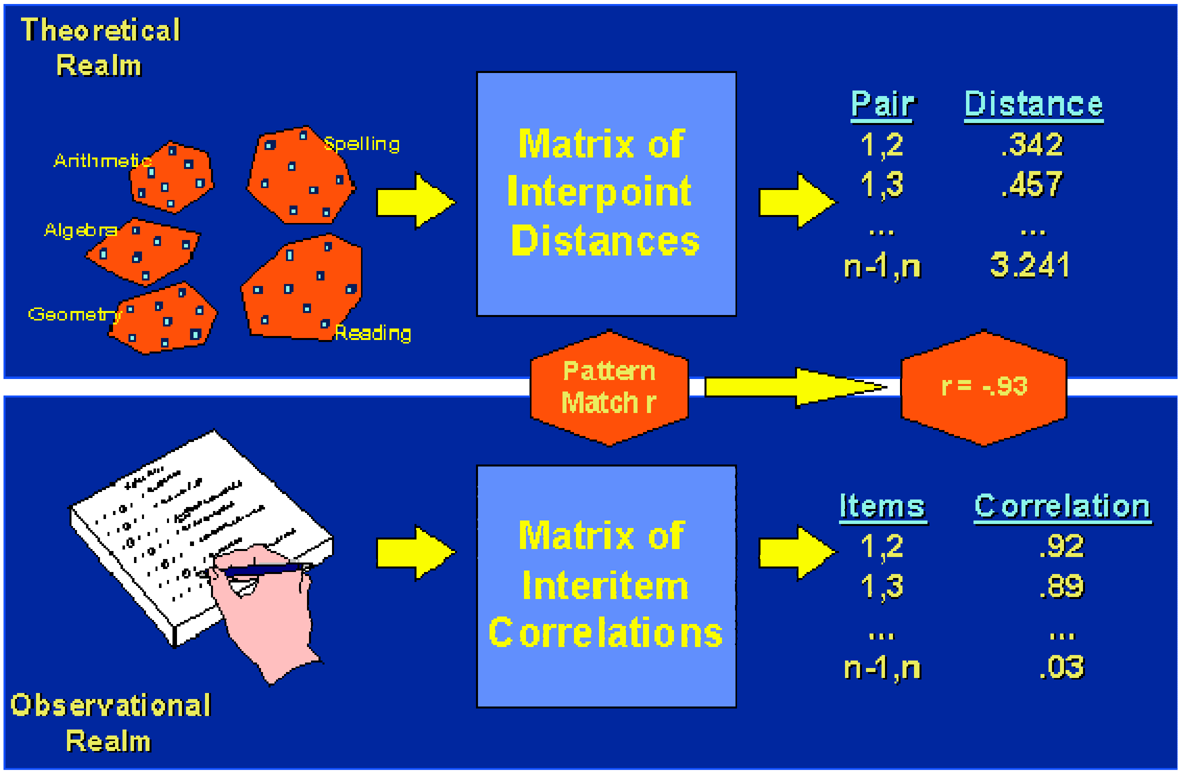
While pattern matching can be used to address a variety of questions in social research, the emphasis here is on its use in assessing construct validity.



The accompanying figure shows the pattern matching structure for an example involving five measurement constructs -- arithmetic, algebra, geometry, spelling, and reading. In this example, we'll use concept mapping to develop the theoretical pattern among these constructs. In the concept mapping we generate a large set of potential arithmetic, algebra, geometry, spelling, and reading questions. We sort them into piles of similar questions and develop a map that shows each question in relation to the others. On the map, questions that are more similar are closer to each other, those less similar are more distant. From the map, we can find the straight-line distances between all pair of points (i.e., all questions). This is the matrix of interpoint distances. We might use the questions from the map in constructing our measurement instrument, or we might sample from these questions. On the observed side, we have one or more test instruments that contain a number of questions about arithmetic, algebra, geometry, spelling, and reading. We analyze the data and construct a matrix of inter-item correlations.

What we want to do is compare the matrix of interpoint distances from our concept map (i.e., the theoretical pattern) with the correlation matrix of the questions (i.e., the observed pattern). How do we achieve this? Let's assume that we had 100 prospective questions on our concept map, 20 for each construct. Correspondingly, we have 100 questions on our measurement instrument, 20 in each area. Thus, both matrices are 100x100 in size. Because both matrices are symmetric, we actually have (N(N-1))/2 = (100(99))/2 = 9900/2 = 4,950 unique pairs (excluding the diagonal). If we "string out" the values in each matrix we can construct a vector or column of 4,950 numbers for each matrix. The first number is the value comparing pair (1,2), the next is (1,3) and so on to (N-1, N) or (99, 100). Now, we can compute the overall correlation between these two columns, which is the correlation between our theoretical and observed patterns, the "pattern matching correlation." In this example, let's assume it is -.93. Why would it be a *negative* correlation? Because we are correlating *distances* on the map with the *similarities* in the correlations and we expect that *greater* distance on the map should be associated with *lower* correlation and *less* distance with *greater* correlation.

The pattern matching correlation is our overall estimate of the degree of construct validity in this example because it estimates the degree to which the operational measures reflect our theoretical expectations.



Advantages and Disadvantages of Pattern Matching

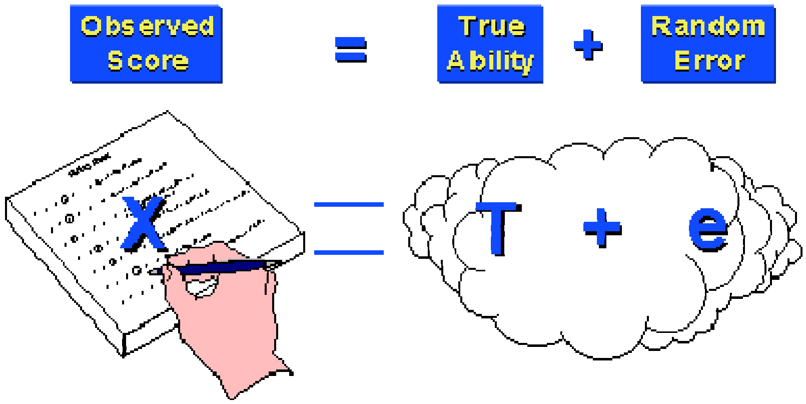
There are several disadvantages of the pattern matching approach to construct validity. The most obvious is that pattern matching requires that you specify your theory of the constructs rather precisely. This is typically not done in applied social research, at least not to the level of specificity implied here. But perhaps it *should* be done. Perhaps the more restrictive assumption is that you are able to structure the theoretical and observed patterns the same way so that you can directly correlate them. We needed to quantify both patterns and, ultimately, describe them in matrices that had the same dimensions. In most research as it is currently done it will be relatively easy to construct a matrix of the inter-item correlations. But we seldom currently use methods like concept mapping that enable us to estimate theoretical patterns that can be linked with observed ones. Again, perhaps we ought to do this more frequently.

There are a number of advantages of the pattern matching approach, especially relative to the multitrait-multimethod matrix (MTMM). First, it is more *general* and *flexible* than MTMM. It does not require that you measure each construct with multiple methods. Second, it treats convergence and discrimination as a *continuum*. Concepts are more or less similar and so their interrelations would be more or less convergent or discriminant. This moves the convergent/discriminant distinction away from the simplistic dichotomous categorical notion to one that is more suitably post-positivist and continuous in nature. Third, the pattern matching approach does make it possible to estimate the overall construct validity for a set of measures in a specific context. Notice that we don't estimate construct validity for a single measure. That's because construct validity, like discrimination, is always a relative metric. Just as we can only ask whether you have distinguished something if there is something to distinguish it from, we can only assess construct validity in terms of a theoretical semantic or nomological net, the conceptual context within which it resides. The pattern matching correlation tells us, for our particular study, whether there is a demonstrable relationship between how we theoretically expect our measures will interrelate and how they do in practice. Finally, because pattern matching requires a more specific theoretical pattern than we typically articulate, it *requires* us to specify what we think about the constructs in our studies. Social research has long been criticized for conceptual sloppiness, for re-packaging old constructs in new terminology and failing to develop an evolution of research around key theoretical constructs. Perhaps the emphasis on theory articulation in pattern matching would encourage us to be more careful about the conceptual underpinnings of our empirical work. And, after all, isn't that what construct validity is all about?

Reliability

Reliability has to do with the quality of measurement. In its everyday sense, reliability is the "consistency" or "repeatability" of your measures. Before we can define reliability precisely we have to lay the groundwork. First, you have to learn about the foundation of reliability, the true score theory of measurement. Along with that, you need to understand the different types of measurement error because errors in measures play a key role in degrading reliability. With this foundation, you can consider the basic theory of reliability, including a precise definition of reliability. There you will find out that we cannot calculate reliability -- we can only estimate it. Because of this, there a variety of different types of reliability that each have multiple ways to estimate reliability for that type. In the end, it's important to integrate the idea of reliability with the other major criteria for the quality of measurement -- validity -- and develop an understanding of the relationships between reliability and validity in measurement.

- **True Score Theory**



True Score Theory is a theory about measurement. Like all theories, you need to recognize that it is not proven -- it is postulated as a model of how the world operates. Like many very powerful model, the true score theory is a very simple one. Essentially, true score theory maintains that every measurement is an additive composite of two components: **true ability** (or the true level) of the respondent on that measure; and **random error**. We observe the measurement -- the score on the test, the total for a self-esteem instrument, the scale value for a person's weight. We don't observe what's on the right side of the equation (only God knows what those values are!), we assume that there are two components to the right side.

The simple equation of $\mathbf{X = T + e_X}$ has a parallel equation at the level of the variance or variability of a measure. That is, across a set of scores, we assume that:

$$\mathbf{var(X) = var(T) + var(e_X)}$$

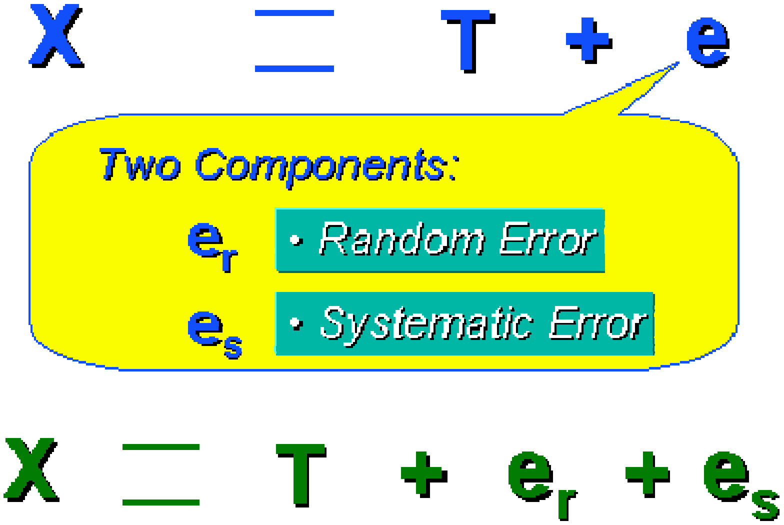
In more human terms this means that the variability of your measure is the sum of the variability due to true score and the variability due to random error. This will have important implications when we consider some of the more advanced models for adjusting for errors in measurement.

Why is true score theory important? For one thing, it is a simple yet powerful model for measurement. It reminds us that most measurement has an error component. Second, true score theory is the foundation of reliability theory. A measure that has no random error (i.e., is all true score) is perfectly reliable; a measure that has no true score (i.e., is all random error) has zero reliability. Third, true score theory can be used in computer simulations as the basis for generating "observed" scores with certain known properties.

You should know that the true score model is not the only measurement model available. measurement theorists continue to come up with more and more complex models that they think represent reality even better. But these models are complicated enough that they lie outside the boundaries of this document. In any event, true score theory should give you an idea of why measurement models are important at all and how they can be used as the basis for defining key research ideas.

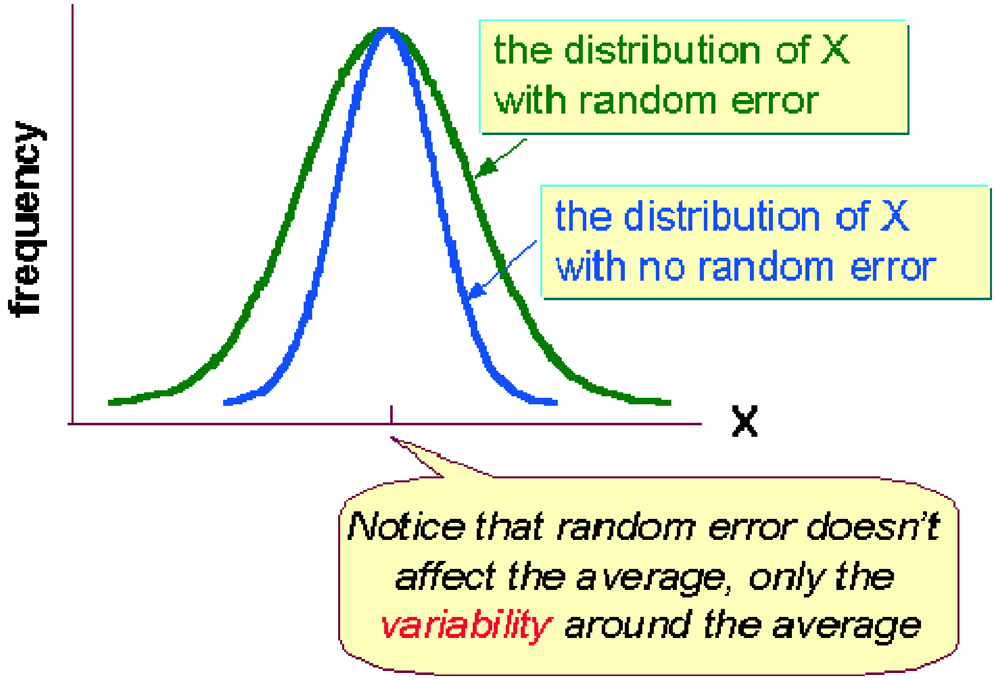
- **Measurement Error**

The true score theory is a good simple model for measurement, but it may not always be an accurate reflection of reality. In particular, it assumes that any observation is composed of the true value plus some random error value. But is that reasonable? What if all error is not random? Isn't it possible that some errors are systematic, that they hold across most or all of the members of a group? One way to deal with this notion is to revise the simple true score model by dividing the error component into two subcomponents, **random error** and **systematic error**. here, we'll look at the differences between these two types of errors and try to diagnose their effects on our research.



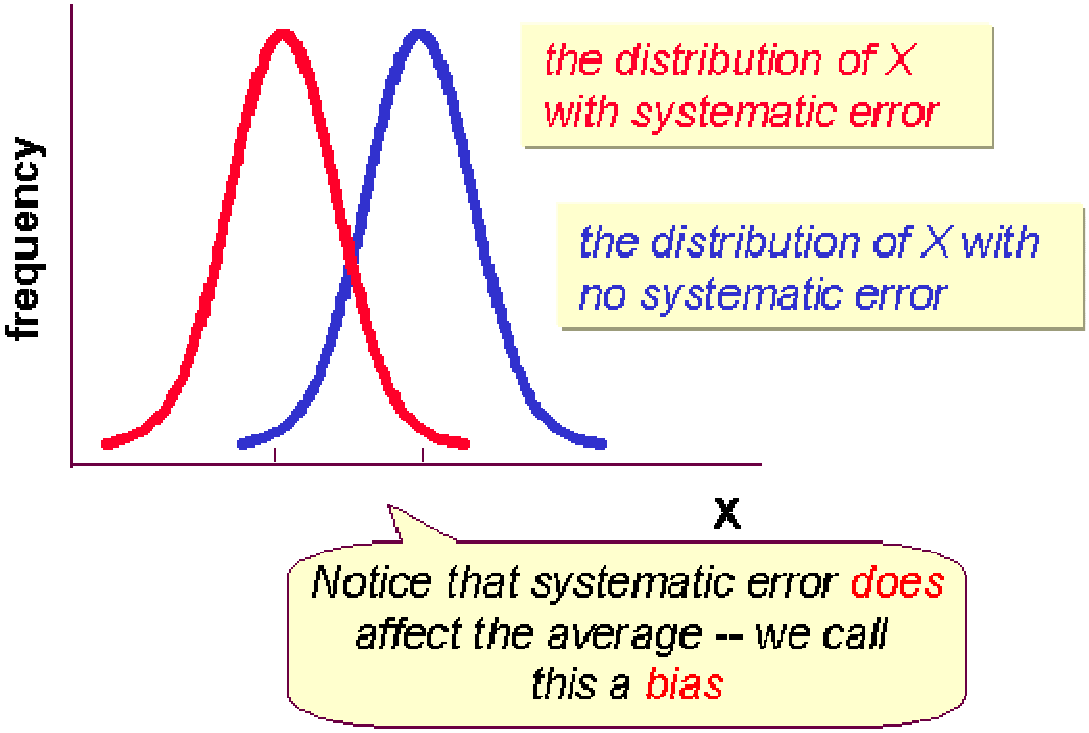
What is Random Error?

Random error is caused by any factors that randomly affect measurement of the variable across the sample. For instance, each person's mood can inflate or deflate their performance on any occasion. In a particular testing, some children may be feeling in a good mood and others may be depressed. If mood affects their performance on the measure, it may artificially inflate the observed scores for some children and artificially deflate them for others. The important thing about random error is that it does not have any consistent effects across the entire sample. Instead, it pushes observed scores up or down randomly. This means that if we could see all of the random errors in a distribution they would have to sum to 0 -- there would be as many negative errors as positive ones. The important property of random error is that it adds variability to the data but does not affect average performance for the group. Because of this, random error is sometimes considered ***noise***.



What is Systematic Error?

Systematic error is caused by any factors that systematically affect measurement of the variable across the sample. For instance, if there is loud traffic going by just outside of a classroom where students are taking a test, this noise is liable to affect all of the children's scores -- in this case, systematically lowering them. Unlike random error, systematic errors tend to be consistently either positive or negative -- because of this, systematic error is sometimes considered to be ***bias*** in measurement.



Reducing Measurement Error

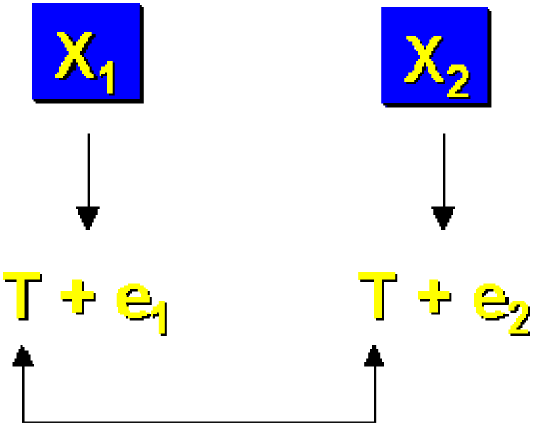
So, how can we reduce measurement errors, random or systematic? One thing you can do is to pilot test your instruments, getting feedback from your respondents regarding how easy or hard the measure was and information about how the testing environment affected their performance. Second, if you are gathering measures using people to collect the data (as interviewers or observers) you should make sure you train them thoroughly so that they aren't inadvertently introducing error. Third, when you collect the data for your study you should double-check the data thoroughly. All data entry for computer analysis should be "double-punched" and verified. This means that you enter the data twice, the second time having your data entry machine check that you are typing the exact same data you did the first time. Fourth, you can use statistical procedures to adjust for measurement error. These range from rather simple formulas you can apply directly to your data to very complex modeling procedures for modeling the error and its effects. Finally, one of the best things you can do to deal with measurement errors, especially systematic errors, is to use multiple measures of the same construct. Especially if the different measures don't share the same systematic errors, you will be able to **triangulate** across the multiple measures and get a more accurate sense of what's going on.

- **Theory of Reliability**

What is **reliability**? We hear the term used a lot in research contexts, but what does it really mean? If you think about how we use the word "reliable" in everyday language, you might get a hint. For instance, we often speak about a machine as reliable: "I have a reliable car." Or, news people talk about a "usually reliable source". In both cases, the word reliable usually means "dependable" or "trustworthy." In research, the term "reliable" also means dependable in a general sense, but that's not a precise enough definition. What does it mean to have a dependable measure or observation in a research context? The reason "dependable" is not a good enough description is that it can be confused too easily with the idea of a valid measure (see Measurement Validity). Certainly, when we speak of a dependable measure, we mean one that is both reliable and valid. So we have to be a little more precise when we try to define reliability.

In research, the term reliability means "repeatability" or "consistency". A measure is considered reliable if it would give us the same result over and over again (assuming that what we are measuring isn't changing!).

Let's explore in more detail what it means to say that a measure is "repeatable" or "consistent". We'll begin by defining a measure that we'll arbitrarily label **X**. It might be a person's score on a math achievement test or a measure of severity of illness. It is the value (numerical or otherwise) that we observe in our study. Now, to see how repeatable or consistent an observation is, we can measure it twice. We'll use subscripts to indicate the first and second observation of the same measure. If we assume that what we're measuring doesn't change between the time of our first and second observation, we can begin to understand how we get at reliability. While we observe a score for what we're measuring, we usually think of that score as consisting of two parts, the 'true' score or actual level for the person on that measure, and the 'error' in measuring it (see True Score Theory).



It's important to keep in mind that we observe the **X** score -- we never actually see the true (**T**) or error (**e**) scores. For instance, a student may get a score of **85** on a math achievement test. That's the score we observe, an **X** of **85**. But the reality might be that the student is actually better at math than that score indicates. Let's say the student's true math ability is **89** (i.e., **T=89**). That means that the error for that student is **-4**. What does this mean? Well, while the student's true math ability may be **89**, he/she may have had a bad day, may not have had breakfast, may have had an argument, or may have been distracted while taking the test. Factors like these can contribute to errors in measurement that make the student's observed ability appear lower than their true or actual ability.

OK, back to reliability. If our measure, **X**, is reliable, we should find that if we measure or observe it twice on the same persons that the scores are pretty much the same. But why would they be the same? If you look at the figure you should see that the only thing that the two observations have in common is their true scores, **T**. How do you know that? Because the error scores ($\mathbf{e_1}$ and $\mathbf{e_2}$) have different subscripts indicating that they are different values. But the true score symbol **T** is the same for both observations. What does this mean? That the two observed scores, $\mathbf{X_1}$ and $\mathbf{X_2}$ are related only to the degree that the observations share true score. You should remember that the error score is assumed to be random. Sometimes errors will lead you to perform better on a test than your true ability (e.g., you had a good day guessing!) while other times it will lead you to score worse. But the true score -- your true ability on that measure -- would be the same on both observations (assuming, of course, that your true ability didn't change between the two measurement occasions).

With this in mind, we can now define reliability more precisely. Reliability is a **ratio** or fraction. In layperson terms we might define this ratio as:

$$\frac{\textbf{True level on the measure}}{\textbf{The entire measure}}$$

You might think of reliability as the proportion of "truth" in your measure. Now, we don't speak of the reliability of a measure for an individual -- reliability is a characteristic of a measure that's taken across individuals. So, to get closer to a more formal definition, let's restate the definition above in terms of a set of observations. The easiest way to do this is to speak of the variance of the scores. Remember that the variance is a measure of the spread or distribution of a *set* of scores. So, we can now state the definition as:

$$\frac{\textbf{The variance of the true score}}{\textbf{The variance of the measure}}$$

We might put this into slightly more technical terms by using the abbreviated name for the variance and our variable names:

$$\frac{\textbf{var(T)}}{\textbf{var(X)}}$$

We're getting to the critical part now. If you look at the equation above, you should recognize that we can easily determine or calculate the bottom part of the reliability ratio -- it's just the variance of the set of scores we observed (You remember how to calculate the variance, don't you? It's just the sum of the squared deviations of the scores from their mean, divided by the number of scores). But how do we calculate the variance of the true scores. We can't see the true scores (we only see X)! Only God knows the true score for a specific observation. And, if we can't calculate the variance of the true scores, we can't compute our ratio, which means ***we can't compute reliability***! Everybody got that? The bottom line is...

we can't compute reliability because we can't calculate the variance of the true scores

Great. So where does that leave us? If we can't compute reliability, perhaps the best we can do is to *estimate* it. Maybe we can get an estimate of the variability of the true scores. How do we do that? Remember our two observations, X_1 and X_2? We assume (using true score theory) that these two observations would be related to each other to the degree that they share true scores. So, let's calculate the correlation between X_1 and X_2. Here's a simple formula for the correlation:

$$\frac{\mathbf{Covariance(X_1, X_2)}}{\mathbf{sd(X_1) * sd(X_2)}}$$

where the 'sd' stands for the standard deviation (which is the square root of the variance). If we look carefully at this equation, we can see that the covariance, which simply measures the "shared" variance between measures must be an indicator of the variability of the true scores because the true scores in X_1 and X_2 are the only thing the two observations share! So, the top part is essentially an estimate of **var(T)** in this context. And, since the bottom part of the equation multiplies the standard deviation of one observation with the standard deviation of the same measure at another time, we would expect that these two values would be the same (it is the same measure we're taking) and that this is essentially the same thing as squaring the standard deviation for either observation. But, the square of the standard deviation is the same thing as the variance of the measure. So, the bottom part of the equation becomes the variance of the measure (or **var(X)**). If you read this paragraph carefully, you should see that the correlation between two observations of the same measure is an estimate of reliability.

It's time to reach some conclusions. We know from this discussion that we cannot calculate reliability because we cannot measure the true score component of an observation. But we also know that we can *estimate* the true score component as the covariance between two observations of the same measure. With that in mind, we can estimate the reliability as the correlation between two observations of the same measure. It turns out that there are several ways we can estimate this reliability correlation. These are discussed in Types of Reliability.

There's only one other issue I want to address here. How big is an estimate of reliability? To figure this out, let's go back to the equation given earlier:

$$\frac{\mathbf{var(T)}}{\mathbf{var(X)}}$$

and remember that because X = T + e, we can substitute in the bottom of the ratio:

$$\frac{\mathbf{var(T)}}{}$$

$$\mathrm{var(T) + var(e)}$$

With this slight change, we can easily determine the range of a reliability estimate. If a measure is perfectly reliable, there is no error in measurement -- everything we observe is true score. Therefore, for a perfectly reliable measure, the equation would reduce to:

$$\frac{\mathrm{var(T)}}{\mathrm{var(T)}}$$

and reliability = 1. Now, if we have a perfectly unreliable measure, there is no true score -- the measure is entirely error. In this case, the equation would reduce to:

$$\frac{0}{\mathrm{var(e)}}$$

and the reliability = 0. From this we know that reliability will always range between 0 and 1. The value of a reliability estimate tells us the proportion of variability in the measure attributable to the true score. A reliability of .5 means that about half of the variance of the observed score is attributable to truth and half is attributable to error. A reliability of .8 means the variability is about 80% true ability and 20% error. And so on.

- **Types of Reliability**

You learned in the Theory of Reliability that it's not possible to calculate reliability exactly. Instead, we have to estimate reliability, and this is always an imperfect endeavor. Here, I want to introduce the major reliability estimators and talk about their strengths and weaknesses.

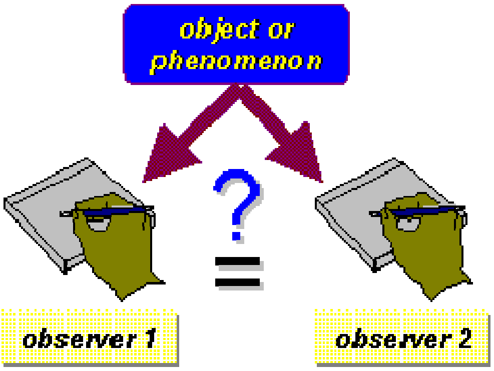
There are four *general classes of reliability estimates*, each of which estimates reliability in a different way. They are:

- **Inter-Rater or Inter-Observer Reliability**
  Used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
- **Test-Retest Reliability**
  Used to assess the consistency of a measure from one time to another.
- **Parallel-Forms Reliability**
  Used to assess the consistency of the results of two tests constructed in the same way from the same content domain.

- **Internal Consistency Reliability**
  Used to assess the consistency of results across items within a test.

Let's discuss each of these in turn.

Inter-Rater or Inter-Observer Reliability



Whenever you use humans as a part of your measurement procedure, you have to worry about whether the results you get are reliable or consistent. People are notorious for their inconsistency. We are easily distractible. We get tired of doing repetitive tasks. We daydream. We misinterpret.

So how do we determine whether two observers are being consistent in their observations? You probably should establish inter-rater reliability outside of the context of the measurement in your study. After all, if you use data from your study to establish reliability, and you find that reliability is low, you're kind of stuck. Probably it's best to do this as a side study or pilot study. And, if your study goes on for a long time, you may want to reestablish inter-rater reliability from time to time to assure that your raters aren't changing.

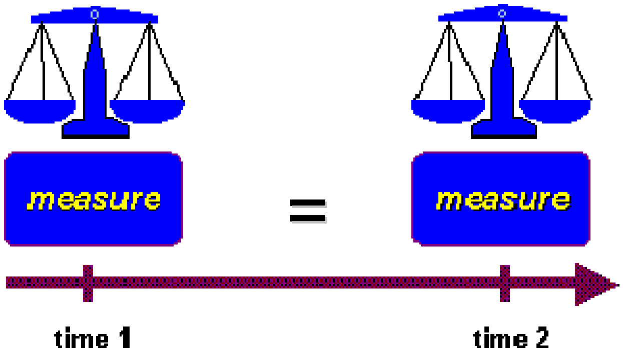
There are two major ways to actually estimate inter-rater reliability. If your measurement consists of categories -- the raters are checking off which category each observation falls in -- you can calculate the percent of agreement between the raters. For instance, let's say you had 100 observations that were being rated by two raters. For each observation, the rater could check one of three categories. Imagine that on 86 of the 100 observations the raters checked the same category. In this case, the percent of agreement would be 86%. OK, it's a crude measure, but it does give an idea of how much agreement exists, and it works no matter how many categories are used for each observation.

The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. There, all you need to do is calculate the correlation between the ratings of the two observers. For instance, they might be rating the overall level of activity in a classroom on a 1-to-7 scale. You could have them give their rating at regular time intervals (e.g., every 30 seconds). The correlation between these ratings would give you an estimate of the reliability or consistency between the raters.

You might think of this type of reliability as "calibrating" the observers. There are other things you could do to encourage reliability between observers, even if you don't estimate it. For instance, I used to work in a psychiatric unit where every morning a nurse had to do a ten-item rating of each patient on the unit. Of course, we couldn't count on the same nurse being present every day, so we had to find a way to assure that any of the nurses would give comparable ratings. The way we did it was to hold weekly "calibration" meetings where we would have all of the nurses ratings for several patients and discuss why they chose the specific values they did. If there were disagreements, the nurses would discuss them and attempt to come up with rules for deciding when they would give a "3" or a "4" for a rating on a specific item. Although this was not an estimate of reliability, it probably went a long way toward improving the reliability between raters.

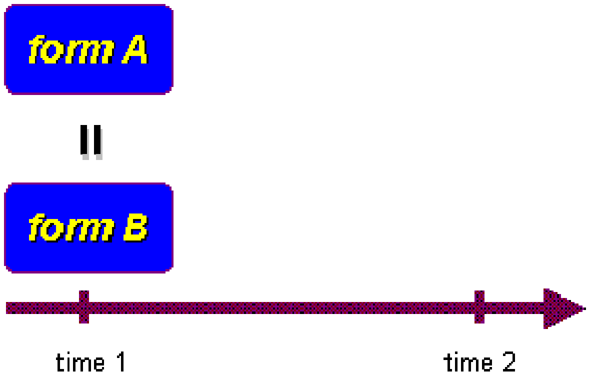
Test-Retest Reliability

We estimate test-retest reliability when we administer the same test to the same sample on two different occasions. This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. We know that if we measure the same thing twice that the correlation between the two observations will depend in part by how much time elapses between the two measurement occasions. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation. This is because the two observations are related over time -- the closer in time we get the more similar the factors that contribute to error. Since this correlation is the test-retest estimate of reliability, you can obtain considerably different estimates depending on the interval.



Parallel-Forms Reliability

In parallel forms reliability you first have to create two parallel forms. One way to accomplish this is to create a large set of questions that address the same construct and then randomly divide the questions into two sets. You administer both instruments to the same sample of people. The correlation between the two parallel forms is the estimate of reliability. One major problem with this approach is that you have to be able to generate lots of items that reflect the same construct. This is often no easy feat. Furthermore, this approach makes the assumption that the randomly divided halves are parallel or equivalent. Even by chance this will sometimes not be the case. The parallel forms approach is very similar to the split-half reliability described below. The major difference is that parallel forms are constructed so that the two forms can be used independent of each other and considered equivalent measures. For instance, we might be concerned about a testing threat to internal validity. If we use Form A for the pretest and Form B for the posttest, we minimize that problem. it would even be better if we randomly assign individuals to receive Form A or B on the pretest and then switch them on the posttest. With split-half reliability we have an instrument that we wish to use as a single measurement instrument and only develop randomly split halves for purposes of estimating reliability.

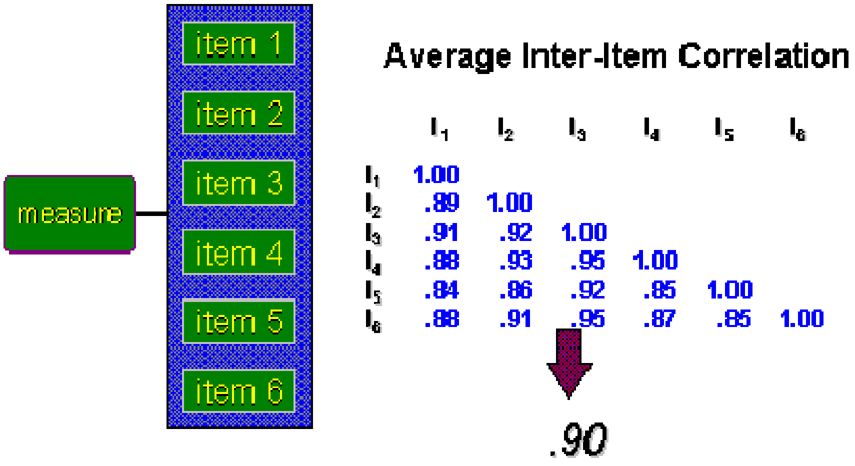


Internal Consistency Reliability

In internal consistency reliability estimation we use our single measurement instrument administered to a group of people on one occasion to estimate reliability. In effect we judge the reliability of the instrument by estimating how well the items that reflect the same construct yield similar results. We are looking at how consistent the results are for different items for the same construct within the measure. There are a wide variety of internal consistency measures that can be used.

Average Inter-item Correlation

The average inter-item correlation uses all of the items on our instrument that are designed to measure the same construct. We first compute the correlation between each pair of items, as illustrated in the figure. For example, if we have six items we will have 15 different item pairings (i.e., 15 correlations). The average interitem correlation is simply the average or mean of all these correlations. In the example, we find an average inter-item correlation of .90 with the individual correlations ranging from .84 to .95.

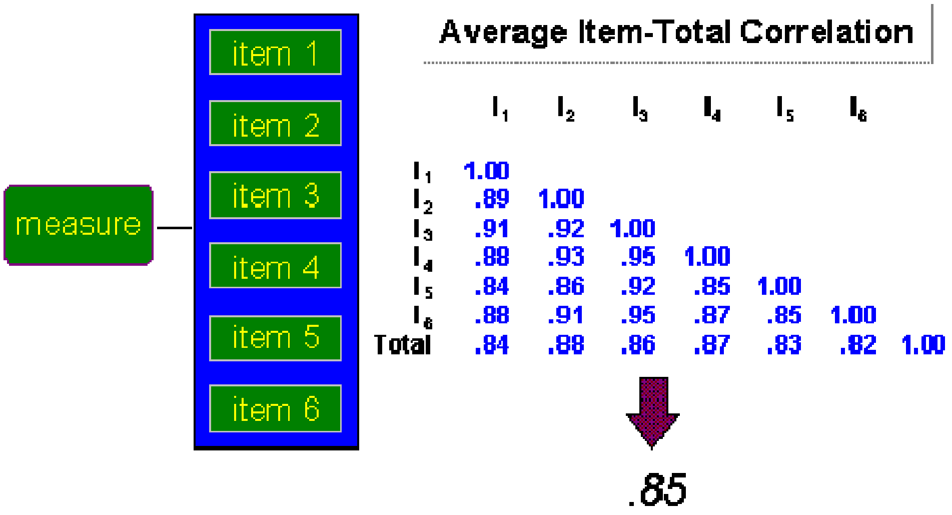


Average Inter-Item Correlation

	I_1	I_2	I_3	I_4	I_5	I_6
I_1	1.00					
I_2	.89	1.00				
I_3	.91	.92	1.00			
I_4	.88	.93	.95	1.00		
I_5	.84	.86	.92	.85	1.00	
I_6	.88	.91	.95	.87	.85	1.00

Average Itemtotal Correlation

This approach also uses the inter-item correlations. In addition, we compute a total score for the six items and use that as a seventh variable in the analysis. The figure shows the six item-to-total correlations at the bottom of the correlation matrix. They range from .82 to .88 in this sample analysis, with the average of these at .85.

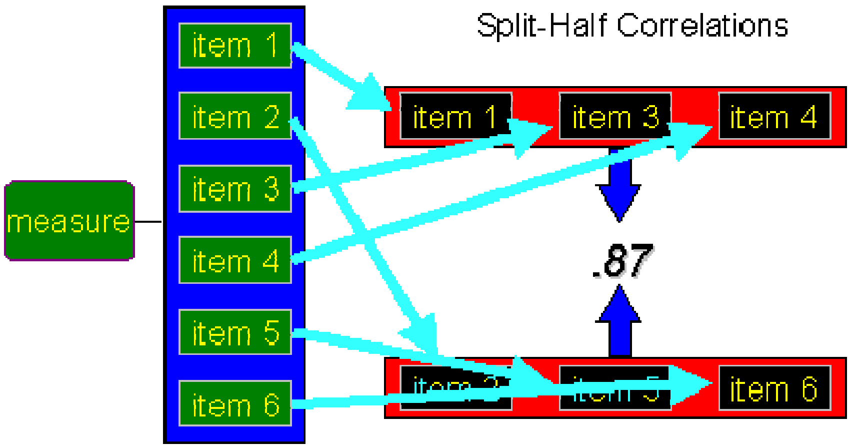


Average Item-Total Correlation

	I_1	I_2	I_3	I_4	I_5	I_6	
I_1	1.00						
I_2	.89	1.00					
I_3	.91	.92	1.00				
I_4	.88	.93	.95	1.00			
I_5	.84	.86	.92	.85	1.00		
I_6	.88	.91	.95	.87	.85	1.00	
Total	.84	.88	.86	.87	.83	.82	1.00

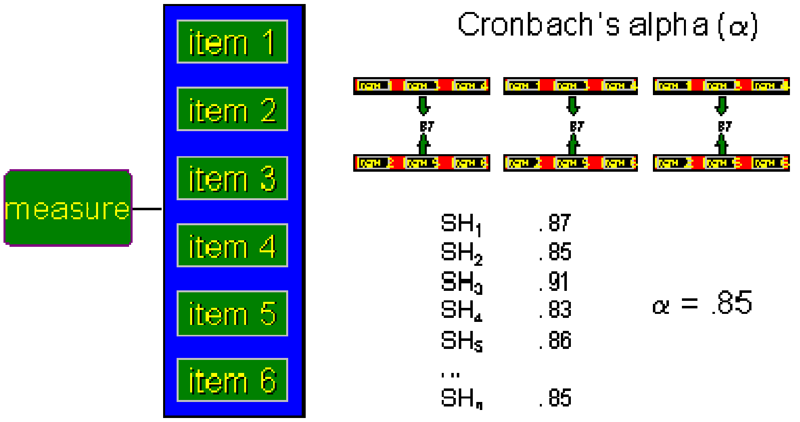
Split-Half Reliability

In split-half reliability we randomly divide all items that purport to measure the same construct into two sets. We administer the entire instrument to a sample of people and calculate the total score for each randomly divided half. the split-half reliability estimate, as shown in the figure, is simply the correlation between these two total scores. In the example it is .87.



Cronbach's Alpha (α)

Imagine that we compute one split-half reliability and then randomly divide the items into another set of split halves and recompute, and keep doing this until we have computed all possible split half estimates of reliability. Cronbach's Alpha is mathematically equivalent to the average of all possible split-half estimates, although that's not how we compute it. Notice that when I say we compute all possible split-half estimates, I don't mean that each time we go an measure a new sample! That would take forever. Instead, we calculate all split-half estimates from the same sample. Because we measured all of our sample on each of the six items, all we have to do is have the computer analysis do the random subsets of items and compute the resulting correlations. The figure shows several of the split-half estimates for our six item example and lists them as SH with a subscript. Just keep in mind that although Cronbach's Alpha is equivalent to the average of all possible split half correlations we would never actually calculate it that way. Some clever mathematician (Cronbach, I presume!) figured out a way to get the mathematical equivalent a lot more quickly.



Comparison of Reliability Estimators

Each of the reliability estimators has certain advantages and disadvantages. Inter-rater reliability is one of the best ways to estimate reliability when your measure is an observation. However, it requires multiple raters or observers. As an alternative, you could look at the correlation of ratings of the same single observer repeated on two different occasions. For example, let's say you collected videotapes of child-mother interactions and had a rater code the videos for how often the mother smiled at the child. To establish inter-rater reliability you could take a sample of videos and have two raters code them independently. To estimate test-retest reliability you could have a single rater code the same videos on two different occasions. You might use the inter-rater approach especially if you were interested in using a team of raters and you wanted to establish that they yielded consistent results. If you get a suitably high inter-rater reliability you could then justify allowing them to work independently on coding different videos. You might use the test-retest approach when you only have a single rater and don't want to train any others. On the other hand, in some studies it is reasonable to do both to help establish the reliability of the raters or observers.

The parallel forms estimator is typically only used in situations where you intend to use the two forms as alternate measures of the same thing. Both the parallel forms and all of the internal consistency estimators have one major constraint -- you have to have multiple items designed to measure the same construct. This is relatively easy to achieve in certain contexts like achievement testing (it's easy, for instance, to construct lots of similar addition problems for a math test), but for more complex or subjective constructs this can be a real challenge. If you do have lots of items, Cronbach's Alpha tends to be the most frequently used estimate of internal consistency.

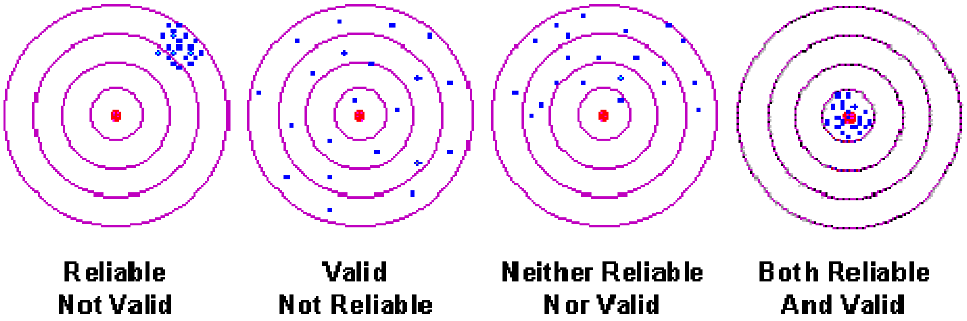
The test-retest estimator is especially feasible in most experimental and quasi-experimental designs that use a no-treatment control group. In these designs you always have a control group that is measured on two occasions (pretest and posttest). the main problem with this approach is that you don't have any information about reliability until you collect the posttest and, if the reliability estimate is low, you're pretty much sunk.

Each of the reliability estimators will give a different value for reliability. In general, the test-retest and inter-rater reliability estimates will be lower in value than the parallel forms and internal consistency ones because they involve measuring at different times or with different raters. Since reliability estimates are often used in statistical analyses of quasi-experimental designs (e.g., the analysis of the nonequivalent group design), the fact that different estimates can differ considerably makes the analysis even more complex.

- **Reliability & Validity**

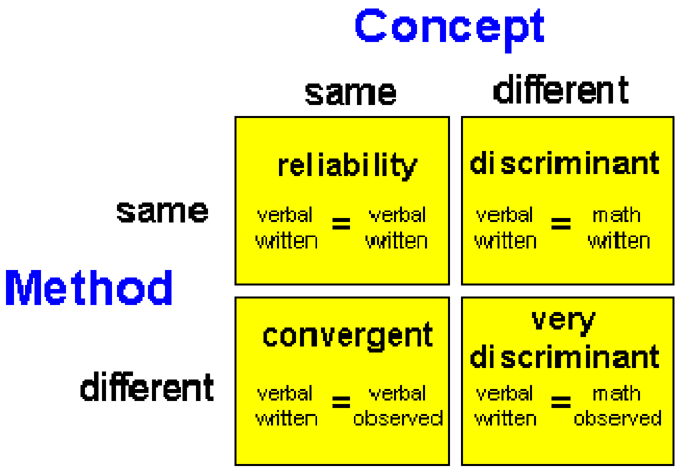
We often think of reliability and validity as separate ideas but, in fact, they're related to each other. Here, I want to show you two ways you can think about their relationship.

One of my favorite metaphors for the relationship between reliability is that of the target. Think of the center of the target as the concept that you are trying to measure. Imagine that for each person you are measuring, you are taking a shot at the target. If you measure the concept perfectly for a person, you are hitting the center of the target. If you don't, you are missing the center. The more you are off for that person, the further you are from the center.



The figure above shows four possible situations. In the first one, you are hitting the target consistently, but you are missing the center of the target. That is, you are consistently and systematically measuring the wrong value for all respondents. This measure is reliable, but no valid (that is, it's consistent but wrong). The second, shows hits that are randomly spread across the target. You seldom hit the center of the target but, on average, you are getting the right answer for the group (but not very well for individuals). In this case, you get a valid group estimate, but you are inconsistent. Here, you can clearly see that reliability is directly related to the variability of your measure. The third scenario shows a case where your hits are spread across the target and you are consistently missing the center. Your measure in this case is neither reliable nor valid. Finally, we see the "Robin Hood" scenario -- you consistently hit the center of the target. Your measure is both reliable and valid (I bet you never thought of Robin Hood in those terms before).

Another way we can think about the relationship between reliability and validity is shown in the figure below. Here, we set up a 2x2 table. The columns of the table indicate whether you are trying to measure the same or different concepts. The rows show whether you are using the same or different methods of measurement. Imagine that we have two concepts we would like to measure, student verbal and math ability. Furthermore, imagine that we can measure each of these in two ways. First, we can use a written, paper-and-pencil exam (very much like the SAT or GRE exams). Second, we can ask the student's classroom teacher to give us a rating of the student's ability based on their own classroom observation.


Concept
same
different
Method
same
different
reliability
verbal written = verbal written
discriminant
verbal written = math written
convergent
verbal written = verbal observed
very discriminant
verbal written = math observed

The first cell on the upper left shows the comparison of the verbal written test score with the verbal written test score. But how can we compare the same measure with itself? We could do this by estimating the reliability of the written test through a test-retest correlation, parallel forms, or an internal consistency measure (See Types of Reliability). What we are estimating in this cell is the reliability of the measure.

The cell on the lower left shows a comparison of the verbal written measure with the verbal teacher observation rating. Because we are trying to measure the same concept, we are looking at convergent validity (See Measurement Validity Types).

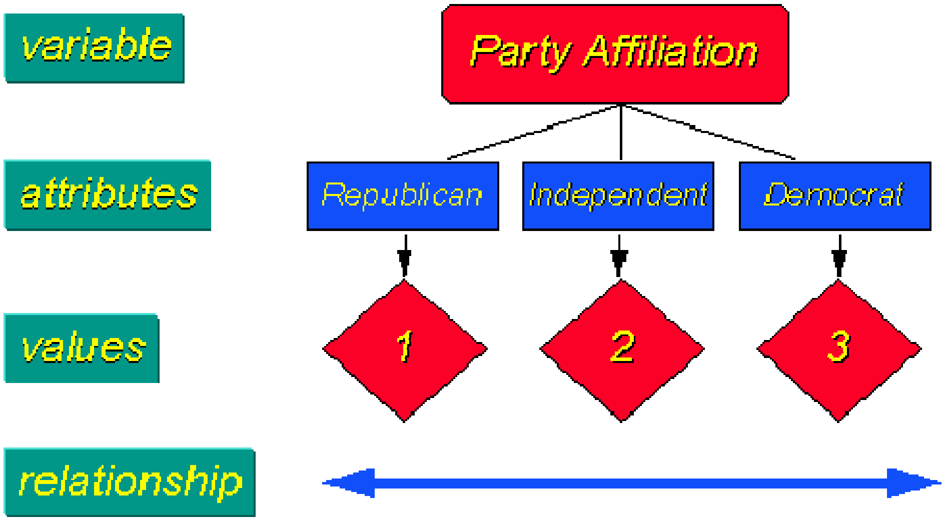
The cell on the upper right shows the comparison of the verbal written exam with the math written exam. Here, we are comparing two different concepts (verbal versus math) and so we would expect the relationship to be lower than a comparison of the same concept with itself (e.g., verbal versus verbal or math versus math). Thus, we are trying to discriminate between two concepts and we would consider this discriminant validity.

Finally, we have the cell on the lower right. Here, we are comparing the verbal written exam with the math teacher observation rating. Like the cell on the upper right, we are also trying to compare two different concepts (verbal versus math) and so this is a discriminant validity estimate. But here, we are also trying to compare two different methods of measurement (written exam versus teacher observation rating). So, we'll call this *very* discriminant to indicate that we would expect the relationship in this cell to be even lower than in the one above it.

The four cells incorporate the different values that we examine in the multitrait-multimethod approach to estimating construct validity.

When we look at reliability and validity in this way, we see that, rather than being distinct, they actually form a continuum. On one end is the situation where the concepts and methods of measurement are the same (reliability) and on the other is the situation where concepts and methods of measurement are different (*very* discriminant validity).

Levels of Measurement



The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable. What does that mean? Begin with the idea of the variable, in this example "party affiliation." That variable has a number of attributes. Let's assume that in this particular election context the only relevant attributes are "republican", "democrat", and "independent". For purposes of analyzing the results of this variable, we arbitrarily assign the values 1, 2 and 3 to the three attributes. The ***level of measurement*** describes the relationship among these three values. In this case, we simply are using the numbers as shorter placeholders for the lengthier text terms. We don't assume that higher values mean "more" of something and lower numbers signify "less". We don't assume the the value of 2 means that democrats are twice something that republicans are. We don't assume that republicans are in first place or have the highest priority just because they have the value of 1. In this case, we only use the values as a shorter name for the attribute. Here, we would describe the level of measurement as "nominal".

Why is Level of Measurement Important?

First, knowing the level of measurement helps you decide how to interpret the data from that variable. When you know that a measure is nominal (like the one just described), then you know that the numerical values are just short codes for the longer names. Second, knowing the level of measurement helps you decide what statistical analysis is appropriate on the values that were assigned. If a measure is nominal, then you know that you would never average the data values or do a t-test on the data.

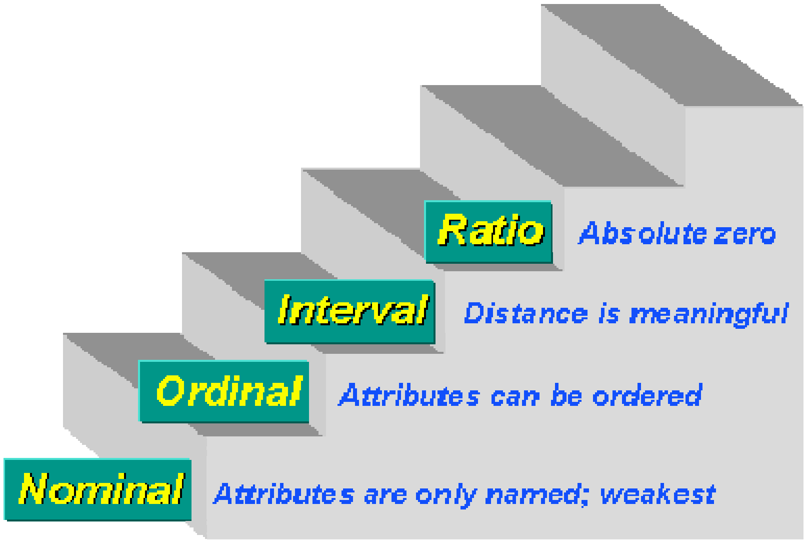
There are typically four levels of measurement that are defined:

- Nominal
- Ordinal

- Interval
- Ratio

In **nominal** measurement the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15, and is certainly not twice whatever number 15 is.

In **ordinal** measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0=less than H.S.; 1=some H.S.; 2=H.S. degree; 3=some college; 4=college degree; 5=post college. In this measure, higher numbers mean *more* education. But is distance from 0 to 1 same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.



In **interval** measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30-40 is same as distance from 70-80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. But note that in interval measurement ratios don't make any sense - 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).

Finally, in **ratio** measurement there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most "count" variables are ratio, for example, the number of clients in past six months. Why? Because you can have zero clients and because it is meaningful to say that "...we had twice as many clients in the past six months as we did in the previous six months."

It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (e.g., interval or ratio) rather than a lower one (nominal or ordinal).

Survey Research

Survey research is one of the most important areas of measurement in applied social research. The broad area of survey research encompasses any measurement procedures that involve asking questions of respondents. A "survey" can be anything form a short paper-and-pencil feedback form to an intensive one-on-one in-depth interview.

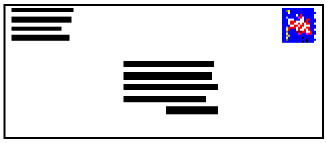
We'll begin by looking at the different types of surveys that are possible. These are roughly divided into two broad areas: Questionnaires and Interviews. Next, we'll look at how you select the survey method that is best for your situation. Once you've selected the survey method, you have to construct the survey itself. Here, we will be address a number of issues including: the different types of questions; decisions about question content; decisions about question wording; decisions about response format; and, question placement and sequence in your instrument. We turn next to some of the special issues involved in administering a personal interview. Finally, we'll consider some of the advantages and disadvantages of survey methods.

- **Types of Surveys**

Surveys can be divided into two broad categories: the **questionnaire** and the **interview**. Questionnaires are usually paper-and-pencil instruments that the respondent completes. Interviews are completed by the interviewer based on the respondent says. Sometimes, it's hard to tell the difference between a questionnaire and an interview. For instance, some people think that questionnaires always ask short closed-ended questions while interviews always ask broad open-ended ones. But you will see questionnaires with open-ended questions (although they do tend to be shorter than in interviews) and there will often be a series of closed-ended questions asked in an interview.

Survey research has changed dramatically in the last ten years. We have automated telephone surveys that use random dialing methods. There are computerized kiosks in public places that allows people to ask for input. A whole new variation of group interview has evolved as focus group methodology. Increasingly, survey research is tightly integrated with the delivery of service. Your hotel room has a survey on the desk. Your waiter presents a short customer satisfaction survey with your check. You get a call for an interview several days after your last call to a computer company for technical assistance. You're asked to complete a short survey when you visit a web site. Here, I'll describe the major types of questionnaires and interviews, keeping in mind that technology is leading to rapid evolution of methods. We'll discuss the relative advantages and disadvantages of these different survey types in Advantages and Disadvantages of Survey Methods.

Questionnaire



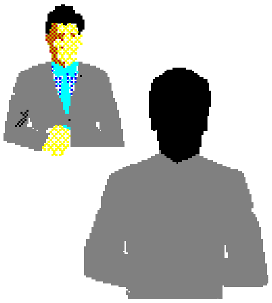
When most people think of questionnaires, they think of the **mail survey**. All of us have, at one time or another, received a questionnaire in the mail. There are many advantages to mail surveys. They are relatively inexpensive to administer. You can send the exact same instrument to a wide number of people. They allow the respondent to fill it out at their own convenience. But there are some disadvantages as well. Response rates from mail surveys are often very low. And, mail questionnaires are not the best vehicles for asking for detailed written responses.

A second type is the **group administered questionnaire**. A sample of respondents is brought together and asked to respond to a structured sequence of questions. Traditionally, questionnaires were administered in group settings for convenience. The researcher could give the questionnaire to those who were present and be fairly sure that there would be a high response rate. If the respondents were unclear about the meaning of a question they could ask for clarification. And, there were often organizational settings where it was relatively easy to assemble the group (in a company or business, for instance).

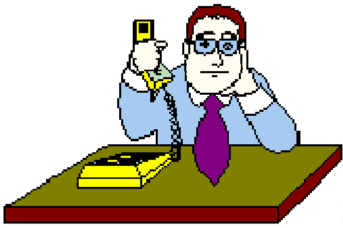
What's the difference between a group administered questionnaire and a group interview or focus group? In the group administered questionnaire, each respondent is *handed an instrument* and asked to complete it while in the room. Each respondent completes an instrument. In the group interview or focus group, the interviewer facilitates the session. People work as a group, listening to each other's comments and answering the questions. Someone takes notes for the entire group -- people don't complete an interview individually.

A less familiar type of questionnaire is the **household drop-off** survey. In this approach, a researcher goes to the respondent's home or business and hands the respondent the instrument. In some cases, the respondent is asked to mail it back or the interview returns to pick it up. This approach attempts to blend the advantages of the mail survey and the group administered questionnaire. Like the mail survey, the respondent can work on the instrument in private, when it's convenient. Like the group administered questionnaire, the interviewer makes personal contact with the respondent -- they don't just send an impersonal survey instrument. And, the respondent can ask questions about the study and get clarification on what is to be done. Generally, this would be expected to increase the percent of people who are willing to respond.

Interviews



Interviews are a far more personal form of research than questionnaires. In the **personal interview**, the interviewer works directly with the respondent. Unlike with mail surveys, the interviewer has the opportunity to probe or ask follow-up questions. And, interviews are generally easier for the respondent, especially if what is sought is opinions or impressions. Interviews can be very time consuming and they are resource intensive. The interviewer is considered a part of the measurement instrument and interviewers have to be well trained in how to respond to any contingency.



Almost everyone is familiar with the **telephone interview**. Telephone interviews enable a researcher to gather information rapidly. Most of the major public opinion polls that are reported were based on telephone interviews. Like personal interviews, they allow for some personal contact between the interviewer and the respondent. And, they allow the interviewer to ask follow-up questions. But they also have some major disadvantages. Many people don't have publicly-listed telephone numbers. Some don't have telephones. People often don't like the intrusion of a call to their homes. And, telephone interviews have to be relatively short or people will feel imposed upon.

- **Selecting the Survey Method**

Selecting the type of survey you are going to use is one of the most critical decisions in many social research contexts. You'll see that there are very few simple rules that will make the decision for you -- you have to use your judgment to balance the advantages and disadvantages of different survey types. Here, all I want to do is give you a number of questions you might ask that can help guide your decision.

Population Issues

The first set of considerations have to do with the population and its accessibility.

- **Can the population be enumerated?**

For some populations, you have a complete listing of the units that will be sampled. For others, such a list is difficult or impossible to compile. For instance, there are complete listings of registered voters or person with active drivers licenses. But no one keeps a complete list of homeless people. If you are doing a study that requires input from homeless persons, you are very likely going to need to go and find the respondents personally. In such contexts, you can pretty much rule out the idea of mail surveys or telephone interviews.

- **Is the population literate?**

Questionnaires require that your respondents can read. While this might seem initially like a reasonable assumption for many adult populations, we know from recent research that the instance of adult illiteracy is alarmingly high. And, even if your respondents can read to some degree, your questionnaire may contain difficult or technical vocabulary. Clearly, there are some populations that you would expect to be illiterate. Young children would not be good targets for questionnaires.

- **Are there language issues?**

We live in a multilingual world. Virtually every society has members who speak other than the predominant language. Some countries (like Canada) are officially multilingual. And, our increasingly global economy requires us to do research that spans countries and language groups. Can you produce multiple versions of your questionnaire? For mail instruments, can you know in advance the language your respondent speaks, or do you send multiple translations of your instrument? Can you be confident that important connotations in your instrument are not culturally specific? Could some of the important nuances get lost in the process of translating your questions?

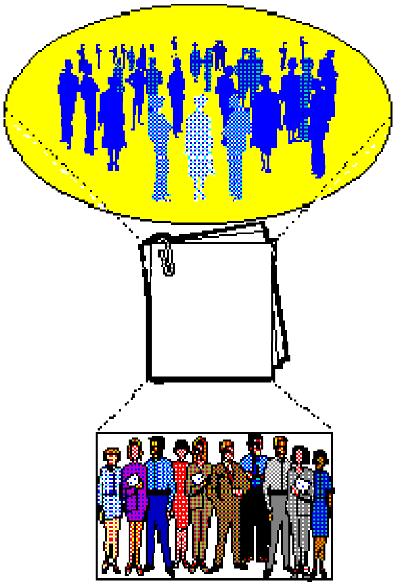
- **Will the population cooperate?**

People who do research on immigration issues have a difficult methodological problem. They often need to speak with undocumented immigrants or people who may be able to identify others who are. Why would we expect those respondents to cooperate? Although the researcher may mean no harm, the respondents are at considerable risk legally if information they divulge should get into the hand of the authorities. The same can be said for any target group that is engaging in illegal or unpopular activities.

- **What are the geographic restrictions?**

Is your population of interest dispersed over too broad a geographic range for you to study feasibly with a personal interview? It may be possible for you to send a mail instrument to a nationwide sample. You may be able to conduct phone interviews with them. But it will almost certainly be less feasible to do research that requires interviewers to visit directly with respondents if they are widely dispersed.

Sampling Issues



The sample is the actual group you will have to contact in some way. There are several important sampling issues you need to consider when doing survey research.

- **What data is available?**

What information do you have about your sample? Do you know their current addresses? Their current phone numbers? Are your contact lists up to date?

- **Can respondents be found?**

Can your respondents be located? Some people are very busy. Some travel a lot. Some work the night shift. Even if you have an accurate phone or address, you may not be able to locate or make contact with your sample.

- **Who is the respondent?**

Who is the respondent in your study? Let's say you draw a sample of households in a small city. A household is not a respondent. Do you want to interview a specific individual? Do you want to talk only to the "head of household" (and how is that person defined)? Are you willing to talk to any member of the household? Do you state that you will speak to the first adult member of the household who opens the door? What if that person is unwilling to be interviewed but someone else in the house is willing? How do you deal with multi-family households? Similar problems arise when you sample groups, agencies, or companies. Can you survey any member of the organization? Or, do you only want to speak to the Director of Human Resources? What if the person you would like to interview is unwilling or unable to participate? Do you use another member of the organization?

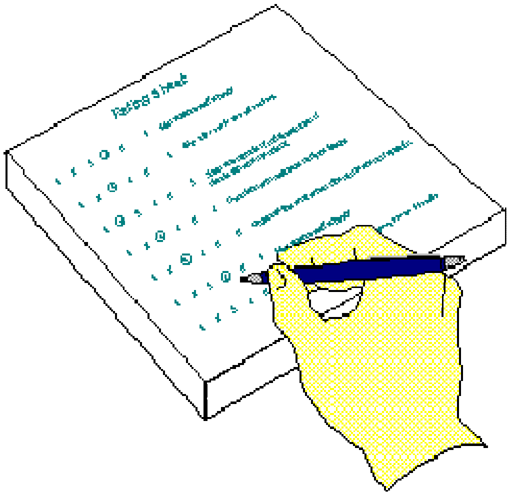
- **Can all members of population be sampled?**

If you have an incomplete list of the population (i.e., sampling frame) you may not be able to sample every member of the population. Lists of various groups are extremely hard to keep up to date. People move or change their names. Even though they are on your sampling frame listing, you may not be able to get to them. And, it's possible they are not even on the list.

- **Are response rates likely to be a problem?**

Even if you are able to solve all of the other population and sampling problems, you still have to deal with the issue of response rates. Some members of your sample will simply refuse to respond. Others have the best of intentions, but can't seem to find the time to send in your questionnaire by the due date. Still others misplace the instrument or forget about the appointment for an interview. Low response rates are among the most difficult of problems in survey research. They can ruin an otherwise well-designed survey effort.

Question Issues



Sometimes the nature of what you want to ask respondents will determine the type of survey you select.

- **What types of questions can be asked?**

Are you going to be asking personal questions? Are you going to need to get lots of detail in the responses? Can you anticipate the most frequent or important types of responses and develop reasonable closed-ended questions?

- **How complex will the questions be?**

Sometimes you are dealing with a complex subject or topic. The questions you want to ask are going to have multiple parts. You may need to branch to sub-questions.

- **Will screening questions be needed?**

A screening question may be needed to determine whether the respondent is qualified to answer your question of interest. For instance, you wouldn't want to ask someone their opinions about a specific computer program without first "screening" them to find out whether they have any experience using the program. Sometimes you have to screen on several variables (e.g., age, gender, experience). The more complicated the screening, the less likely it is that you can rely on paper-and-pencil instruments without confusing the respondent.

- **Can question sequence be controlled?**

Is your survey one where you can construct in advance a reasonable sequence of questions? Or, are you doing an initial exploratory study where you may need to ask lots of follow-up questions that you can't easily anticipate?

- **Will lengthy questions be asked?**

If your subject matter is complicated, you may need to give the respondent some detailed background for a question. Can you reasonably expect your respondent to sit still long enough in a phone interview to ask your question?

- **Will long response scales be used?**

If you are asking people about the different computer equipment they use, you may have to have a lengthy response list (CD-ROM drive, floppy drive, mouse, touch pad, modem, network connection, external speakers, etc.). Clearly, it may be difficult to ask about each of these in a short phone interview.

Content Issues

The content of your study can also pose challenges for the different survey types you might utilize.

- **Can the respondents be expected to know about the issue?**

If the respondent does not keep up with the news (e.g., by reading the newspaper, watching television news, or talking with others), they may not even know about the news issue you want to ask them about. Or, if you want to do a study of family finances and you are talking to the spouse who doesn't pay the bills on a regular basis, they may not have the information to answer your questions.

- **Will respondent need to consult records?**

Even if the respondent understands what you're asking about, you may need to allow them to consult their records in order to get an accurate answer. For instance, if you ask them how much money they spent on food in the past month, they may need to look up their personal check and credit card records. In this case, you don't want to be involved in an interview where they would have to go look things up while they keep you waiting (they wouldn't be comfortable with that).

Bias Issues

People come to the research endeavor with their own sets of biases and prejudices. Sometimes, these biases will be less of a problem with certain types of survey approaches.

- **Can social desirability be avoided?**

Respondents generally want to "look good" in the eyes of others. None of us likes to look like we don't know an answer. We don't want to say anything that would be embarrassing. If you ask people about information that may put them in this kind of position, they may not tell you the truth, or they may "spin" the response so that it makes them look better. This may be more of a problem in an interview situation where they are face-to face or on the phone with a live interviewer.

- **Can interviewer distortion and subversion be controlled?**

Interviewers may distort an interview as well. They may not ask questions that make them uncomfortable. They may not listen carefully to respondents on topics for which they have strong opinions. They may make the judgment that they already know what the respondent would say to a question based on their prior responses, even though that may not be true.

- **Can false respondents be avoided?**

With mail surveys it may be difficult to know who actually responded. Did the head of household complete the survey or someone else? Did the CEO actually give the responses or instead pass the task off to a subordinate? Is the person you're speaking with on the phone actually who they say they are? At least with personal interviews, you have a reasonable chance of knowing who you are speaking with. In mail surveys or phone interviews, this may not be the case.

Administrative Issues

Last, but certainly not least, you have to consider the feasibility of the survey method for your study.

- **costs**

Cost is often the major determining factor in selecting survey type. You might prefer to do personal interviews, but can't justify the high cost of training and paying for the interviewers. You may prefer to send out an extensive mailing but can't afford the postage to do so.

- **facilities**

Do you have the facilities (or access to them) to process and manage your study? In phone interviews, do you have well-equipped phone surveying facilities? For focus groups, do you have a comfortable and accessible room to host the group? Do you have the equipment needed to record and transcribe responses?

- **time**

Some types of surveys take longer than others. Do you need responses immediately (as in an overnight public opinion poll)? Have you budgeted enough time for your study to send out mail surveys and follow-up reminders, and to get the responses back by mail? Have you allowed for enough time to get enough personal interviews to justify that approach?

- **personnel**

Different types of surveys make different demands of personnel. Interviews require interviewers who are motivated and well-trained. Group administered surveys require people who are trained in group facilitation. Some studies may be in a technical area that requires some degree of expertise in the interviewer.

Clearly, there are lots of issues to consider when you are selecting which type of survey you wish to use in your study. And there is no clear and easy way to make this decision in many contexts. There may not be one approach which is clearly the best. You may have to make tradeoffs of advantages and disadvantages. There is judgment involved. Two expert researchers may, for the very same problem or issue, select entirely different survey methods. But, if you select a method that isn't appropriate or doesn't fit the context, you can doom a study before you even begin designing the instruments or questions themselves.

- ## Constructing the Survey

Constructing a survey instrument is an art in itself. There are numerous small decisions that must be made -- about content, wording, format, placement -- that can have important consequences for your entire study. While there's no one perfect way to accomplish this job, we do have lots of advice to offer that might increase your chances of developing a better final product.

First of all you'll learn about the two major types of surveys that exist, the questionnaire and the interview and the different varieties of each. Then you'll see how to write questions for surveys. There are three areas involved in writing a question:

- determining the question content, scope and purpose
- choosing the response format that you use for collecting information from the respondent
- figuring out how to word the question to get at the issue of interest

Finally, once you have your questions written, there is the issue of how best to place them in your survey.

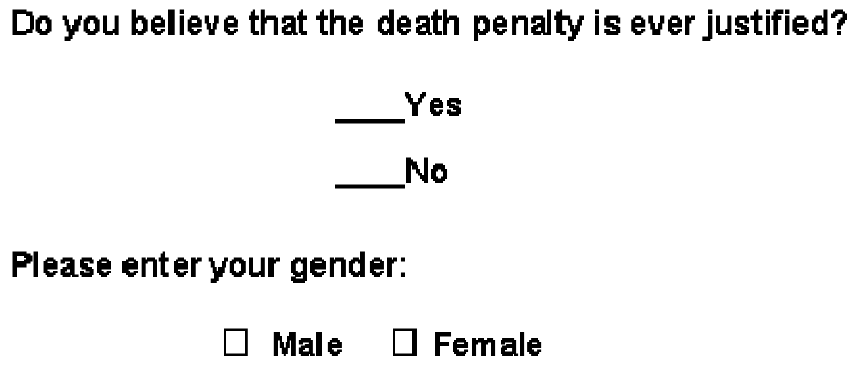
You'll see that although there are many aspects of survey construction that are just common sense, if you are not careful you can make critical errors that have dramatic effects on your results.

❖ Types Of Questions

Survey questions can be divided into two broad types: **structured** and **unstructured**. From an instrument design point of view, the structured questions pose the greater difficulties (see Decisions About the Response Format). From a content perspective, it may actually be more difficult to write good unstructured questions. Here, I'll discuss the variety of structured questions you can consider for your survey (we'll discuss unstructured questioning more under Interviews).

Dichotomous Questions

When a question has two possible responses, we consider it **dichotomous**. Surveys often use dichotomous questions that ask for a Yes/No, True/False or Agree/Disagree response. There are a variety of ways to lay these questions out on a questionnaire:



Do you believe that the death penalty is ever justified?

____Yes

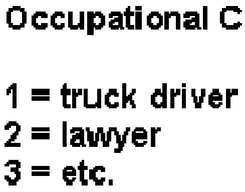
____No

Please enter your gender:

□ Male □ Female

Questions Based on Level Of Measurement

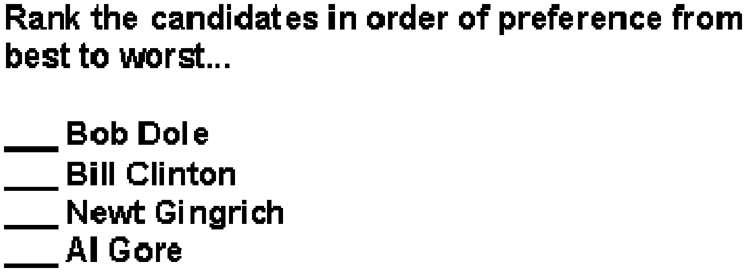
We can also classify questions in terms of their level of measurement. For instance, we might measure occupation using a **nominal** question. Here, the number next to each response has no meaning except as a placeholder for that response. The choice of a "2" for a lawyer and a "1" for a truck driver is arbitrary -- from the numbering system used we can't infer that a lawyer is "twice" something that a truck driver is.



Occupational Class:

1 = truck driver
2 = lawyer
3 = etc.

We might ask respondents to rank order their preferences for presidential candidates using an **ordinal** question:

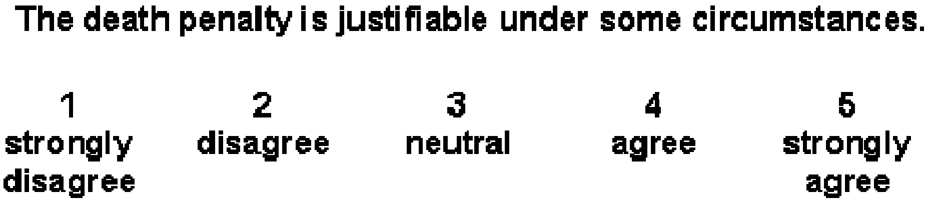


Rank the candidates in order of preference from best to worst...

___ Bob Dole
___ Bill Clinton
___ Newt Gingrich
___ Al Gore

We want the respondent to put a 1, 2, 3 or 4 next to the candidate, where 1 is the respondent's first choice. Note that this could get confusing. We might want to state the prompt more explicitly so the respondent knows we want a number from one to 4 (the respondent might check their favorite candidate, or assign higher numbers to candidates they prefer more instead of understanding that we want rank ordering).

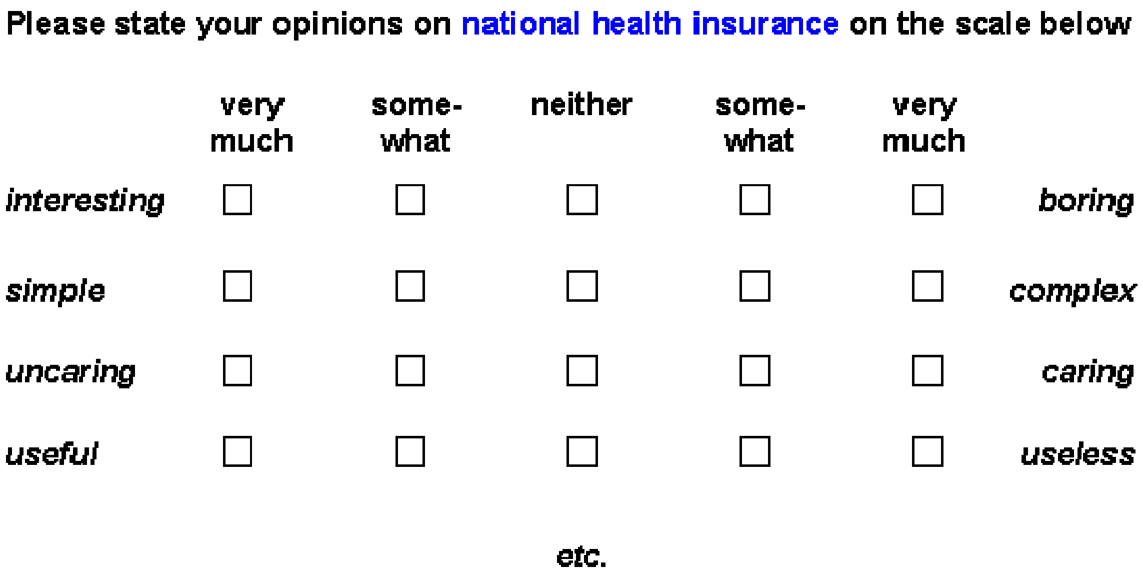
We can also construct survey questions that attempt to measure on an **interval** level. One of the most common of these types is the traditional 1-to-5 rating (or 1-to-7, or 1-to-9, etc.). This is sometimes referred to as a **Likert response scale** (see Likert Scaling). Here, we see how we might ask an opinion question on a 1-to-5 bipolar scale (it's called bipolar because there is a neutral point and the two ends of the scale are at opposite positions of the opinion):



The death penalty is justifiable under some circumstances.

1	2	3	4	5
strongly disagree	disagree	neutral	agree	strongly agree

Another interval question uses an approach called the **semantic differential**. Here, an object is assessed by the respondent on a set of bipolar adjective pairs (using 5-point rating scale):

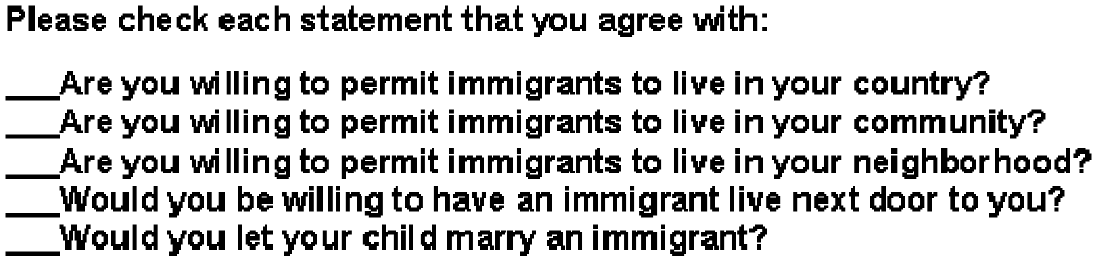


Please state your opinions on national health insurance on the scale below

	very much	some-what	neither	some-what	very much	
interesting	□	□	□	□	□	*boring*
simple	□	□	□	□	□	*complex*
uncaring	□	□	□	□	□	*caring*
useful	□	□	□	□	□	*useless*

etc.

Finally, we can also get at interval measures by using what is called a **cumulative or Guttman scale** (see Guttman Scaling). Here, the respondent checks each item with which they agree. The items themselves are constructed so that they are cumulative -- if you agree to one, you probably agree to all of the ones above it in the list:

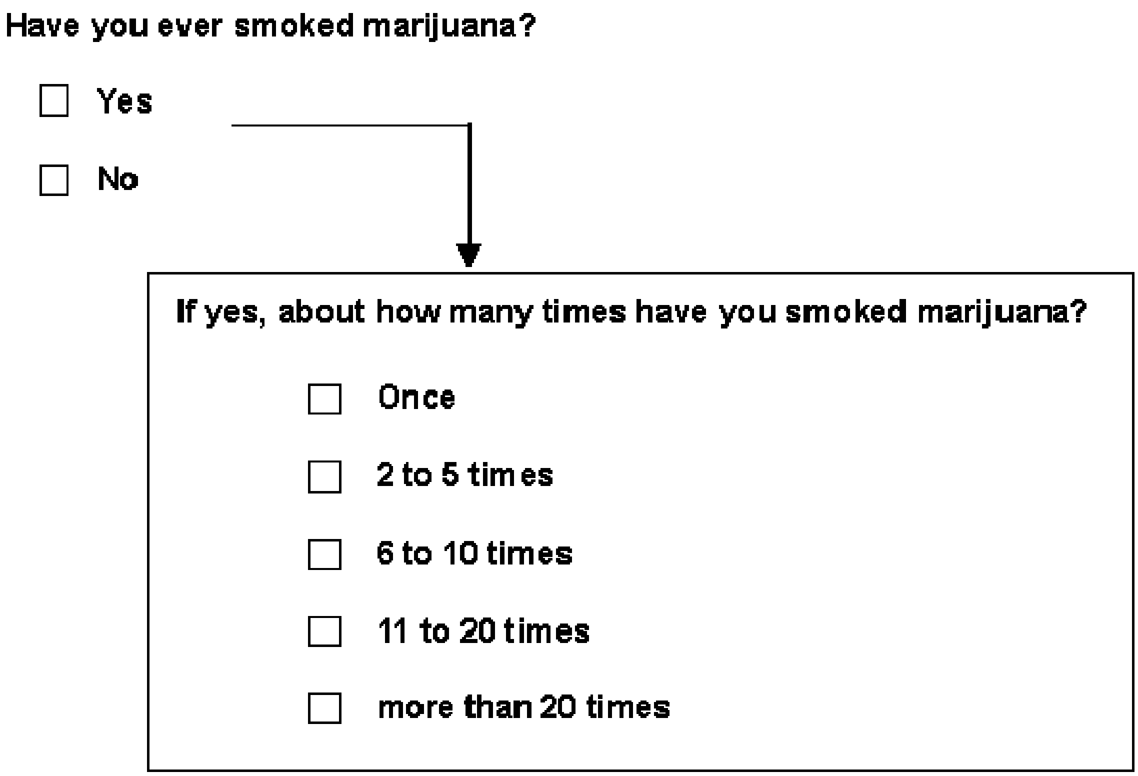


Please check each statement that you agree with:

___Are you willing to permit immigrants to live in your country?
___Are you willing to permit immigrants to live in your community?
___Are you willing to permit immigrants to live in your neighborhood?
___Would you be willing to have an immigrant live next door to you?
___Would you let your child marry an immigrant?

Filter or Contingency Questions

Sometimes you have to ask the respondent one question in order to determine if they are qualified or experienced enough to answer a subsequent one. This requires using a **filter or contingency question**. For instance, you may want to ask one question if the respondent has ever smoked marijuana and a different question if they have not. in this case, you would have to construct a filter question to determine whether they've ever smoked marijuana:



Have you ever smoked marijuana?

□ Yes

□ No

If yes, about how many times have you smoked marijuana?

□ Once

□ 2 to 5 times

□ 6 to 10 times

□ 11 to 20 times

□ more than 20 times

Filter questions can get very complex. Sometimes, you have to have multiple filter questions in order to direct your respondents to the correct subsequent questions. There are a few conventions you should keep in mind when using filters:

- **try to avoid having more than three levels (two jumps) for any question**

Too many jumps will confuse the respondent and may discourage them from continuing with the survey.

- **if only two levels, use graphic to jump (e.g., arrow and box)**

The example above shows how you can make effective use of an arrow and box to help direct the respondent to the correct subsequent question.

- **if possible, jump to a new page**

If you can't fit the response to a filter on a single page, it's probably best to be able to say something like "If YES, please turn to page 4" rather that "If YES, please go to Question 38" because the respondent will generally have an easier time finding a page than a specific question.

❖ Question Content

For each question in your survey, you should ask yourself how well it addresses the content you are trying to get at. Here are some content-related questions you can ask about your survey questions.

Is the Question Necessary/Useful?

Examine each question to see if you need to ask it at all and if you need to ask it at the level of detail you currently have.

- Do you need the age of *each* child or just the *number of children under 16*?
- Do you need to *ask income* or can you *estimate*?

Are Several Questions Needed?

This is the classic problem of the **double-barreled question**. You should think about splitting each of the following questions into two separate ones. You can often spot these kinds of problems by looking for the conjunction "and" in your question.

- What are your feelings towards African-Americans *and* Hispanic-Americans?
- What do you think of proposed changes in benefits *and* hours?

Another reason you might need more than one question is that the question you ask **does not cover all possibilities**. For instance, if you ask about earnings, the respondent might not mention all income (e.g., dividends, gifts). Or, if you ask the respondents if they're in favor of public TV, they might not understand that you're asking generally. They may not be in favor of public TV for themselves (they never watch it), but might favor it very much for their children (who watch *Sesame Street* regularly). You might be better off asking two questions, one for their own viewing and one for other members of their household.

Sometimes you need to ask additional questions because your question **does not give you enough context** to interpret the answer. For instance, if you ask about attitudes towards Catholics, can you interpret this without finding out about their attitudes towards religion in general, or other religious groups?

At times, you need to ask additional questions because your question **does not determine the intensity** of the respondent's attitude or belief. For example, if they say they support public TV, you probably should also ask them whether they ever watch it or if they would be willing to have their tax dollars spent on it. It's one thing for a respondent to tell you they support something. But the intensity of that response is greater if they are willing to back their sentiment of support with their behavior.

Do Respondents Have the Needed Information?

Look at each question in your survey to see whether the respondent is likely to have the necessary information to be able to answer the question. For example, let's say you want to ask the question:

Do you think Dean Rusk acted correctly in the Bay of Pigs crisis?

The respondent won't be able to answer this question if they have no idea who Dean Rusk was or what the Bay of Pigs crisis was. In surveys of television viewing, you cannot expect that the respondent can answer questions about shows they have never watched. You should ask a filter question first (e.g., Have you ever watched the show *ER*?) before asking them their opinions about it.

Does the Question Need to be More Specific?

Sometimes we ask our questions too generally and the information we obtain is more difficult to interpret. For example, let's say you want to find out respondent's opinions about a specific book. You could ask them

How well did you like the book?

on some scale ranging from "Not At All" to "Extremely Well." But what would their response mean? What does it mean to say you liked a book ***very well***? Instead, you might as questions designed to be more specific like:

Did you recommend the book to others?

or

Did you look for other books by that author?

Is Question Sufficiently General?

You can err in the other direction as well by being too specific. For instance, if you ask someone to list the televisions program they liked best in the past week, you could get a very different answer than if you asked them which show they've enjoyed most over the past year. Perhaps a show they don't usually like had a great episode in the past week, or their show was preempted by another program.

Is Question Biased or Loaded?

One danger in question-writing is that your own biases and blind-spots may affect the wording (see Decisions About Question Wording). For instance, you might generally be in favor of tax cuts. If you ask a question like:

What do you see as the benefits of a tax cut?

you're only asking about one side of the issue. You might get a very different picture of the respondents' positions if you also asked about the disadvantages of tax cuts. The same thing could occur if you are in favor of public welfare and you ask:

What do you see as the disadvantages of eliminating welfare?

without also asking about the potential benefits.

Will Respondent Answer Truthfully?

For each question on your survey, ask yourself whether the respondent will have any difficulty answering the question truthfully. If there is some reason why they may not, consider rewording the question. For instance, some people are sensitive about answering questions about their exact age or income. In this case, you might give them **response brackets** to choose from (e.g., between 30 and 40 years old, between $50,000 and $100,000 annual income). Sometimes even bracketed responses won't be enough. Some people do not like to share how much money they give to charitable causes (they may be afraid of being solicited even more). No matter how you word the question, they would not be likely to tell you their contribution rate. But sometimes you can do this by posing the question in terms of a **hypothetical projective respondent** (a little bit like a projective test). In this case, you might get reasonable estimates if you ask the respondent how much money "people you know" typically give in a year to charitable causes. Finally, you can sometimes dispense with asking a question at all if you can obtain the answer unobtrusively (see Unobtrusive Measures). If you are interested in finding out what magazines the respondent reads, you might instead tell them you are collecting magazines for a recycling drive and ask if they have any old ones to donate (of course, you have to consider the ethical implications of such deception!).

❖ Response Format

The response format is how you collect the answer from the respondent. Let's start with a simple distinction between what we'll call **unstructured** response formats and **structured response formats**. [*On this page, I'll use standard web-based form fields to show you how various response formats might look on the web. If you want to see how these are generated, select the View Source option on your web browser.*]

Structured Response Formats

Structured formats help the respondent to respond more easily and help the researcher to accumulate and summarize responses more efficiently. But, they can also constrain the respondent and limit the researcher's ability to understand what the respondent really means. There are many different structured response formats, each with its own strengths and weaknesses. We'll review the major ones here.

Fill-In-The-Blank. One of the simplest response formats is a blank line. A blank line can be used for a number of different response types. For instance:

Please enter your gender:

_____ Male

_____ Female

Here, the respondent would probably put a check mark or an X next to the response. This is also an example of a **dichotomous** response, because it only has two possible values. Other common dichotomous responses are True/False and Yes/No. Here's another common use of a fill-in-the-blank response format:

Please enter your preference for the following candidates where '1' = your first choice, '2' = your second choice, and so on.

_____ Robert Dole

_____ Colin Powell

_____ Bill Clinton

_____ Al Gore

In this example, the respondent writes a number in each blank. Notice that here, we expect the respondent to place a number on every blank, whereas in the previous example, we expect to respondent to choose only one. Then, of course, there's the classic:

NAME: ________________________

And here's the same fill-in-the-blank response item in web format:

NAME:

Of course, there's always the classic fill-in-the-blank test item:

One of President Lincoln's most famous speeches, the ▭ Address, only lasted a few minutes when delivered.

Check The Answer. The respondent places a check next to the response(s). The simplest form would be the example given above where we ask the person to indicate their gender. Sometimes, we supply a box that the person can fill in with an 'X' (which is sort of a variation on the check mark. Here's a web version of the checkbox:

Please check if you have the following item on the computer you use most:

□ **modem**

□ **printer**

□ **CD-ROM drive**

□ **joystick**

□ **scanner**

Notice that in this example, it is possible for you to check more than one response. By convention, we usually use the checkmark format when we want to allow the respondent to select multiple items.

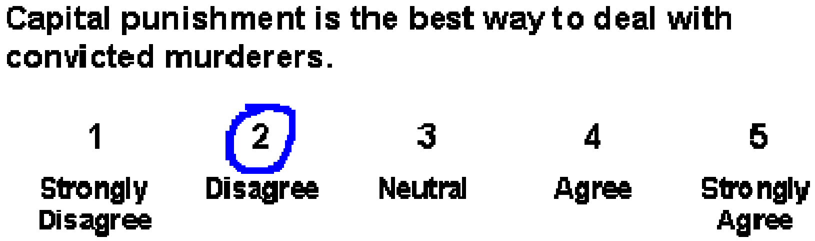
We sometimes refer to this as a **multi-option variable**. You have to be careful when you analyze data from a multi-option variable. Because the respondent can select any of the options, you have to treat this type of variable in your analysis *as though each option is a separate variable*. For instance, for each option we would normally enter either a '0' if the respondent did not check it or a '1' if the respondent did check it. For the example above, if the respondent had only a modem and CD-ROM drive, we would enter the sequence 1, 0, 1, 0, 0. There is a very important reason why you should code this variable as either 0 or 1 when you enter the data. If you do, and you want to determine what percent of your sample has a modem, all you have to do is compute the average of the 0's and 1's for the modem variable. For instance, if you have 10 respondents and only 3 have a modem, the average would be 3/10 = .30 or 30%, which is the percent who checked that item.

The example above is also a good example of a checklist item. Whenever you use a checklist, you want to be sure that you ask the following questions:

- Are all of the alternatives covered?
- Is the list of reasonable length?
- Is the wording impartial?
- Is the form of the response easy, uniform?

Sometimes you may not be sure that you have covered all of the possible responses in a checklist. If that is the case, you should probably allow the respondent to write in any other options that may apply.

Circle The Answer. Sometimes the respondent is asked to circle an item to indicate their response. Usually we are asking them to circle a number. For instance, we might have the following:



Capital punishment is the best way to deal with convicted murderers.

1	[✓ 2]	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

In computer contexts, it's not feasible to have respondents circle a response. In this case, we tend to use an option button:

Capital punishment is the best way to deal with convicted murderers.

○	○	○	○	○
Strongly Disagree	**Disagree**	**Neutral**	**Agree**	Strongly **Agree**

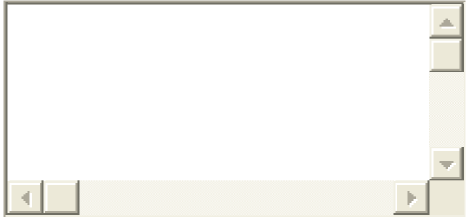
Notice that you can only check one option at a time. The rule of thumb is that you ask someone to circle an item or click on a button when you only want them to be able to select one of the options. In contrast to the multi-option variable described above, we refer to this type of item as a **single-option variable** -- even though the respondent has multiple *choices*, they can only select one of them. We would analyze this as a single variable that can take the integer values from 1 to 5.

Unstructured Response Formats

While there is a wide variety of structured response formats, there are relatively few unstructured ones. What is an unstructured response format? Generally, it's written text. If the respondent (or interviewer) writes down text as the response, you've got an unstructured response format. These can vary from short comment boxes to the transcript of an interview.

In almost every short questionnaire, there's one or more short text field questions. One of the most frequent goes something like this:

Please add any other comments:



Actually, there's really not much more to text-based response formats of this type than writing the prompt and allowing enough space for a reasonable response.

Transcripts are an entirely different matter. There, the transcriber has to decide whether to transcribe every word or only record major ideas, thoughts, quotes, etc. In detailed transcriptions, you may also need to distinguish different speakers (e.g., the interviewer and respondent) and have a standard convention for indicating comments about what's going on in the interview, including non-conversational events that take place and thoughts of the interviewer.

❖ Question Wording

One of the major difficulty in writing good survey questions is getting the wording right. Even slight wording differences can confuse the respondent or lead to incorrect interpretations of the question. Here, I outline some questions you can ask about how you worded each of your survey questions.

Can the Question be Misunderstood?

The survey author has to always be on the lookout for questions that could be misunderstood or confusing. For instance, if you ask a person for their nationality, it might not be clear what you want (Do you want someone from Malaysia to say Malaysian, Asian, or Pacific Islander?). Or, if you ask for marital status, do you want someone to say simply that they are either married or no married? Or, do you want more detail (like divorced, widow/widower, etc.)?

Some terms are just to vague to be useful. For instance, if you ask a question about the "mass media," what do you mean? The newspapers? Radio? Television?

Here's one of my favorites. Let's say you want to know the following:

What kind of headache remedy do you use?

Do you want to know what brand name medicine they take? Do you want to know about "home" remedies? Are you asking whether they prefer a pill, capsule or caplet?

What Assumptions Does the Question Make?

Sometimes we don't stop to consider how a question will appear from the respondent's point-of-view. We don't think about the assumptions behind our questions. For instance, if you ask what social class someone's in, you assume that they know what social class is and that they think of themselves as being in one. In this kind of case, you may need to use a filter question first to determine whether either of these assumptions is true.

Is the time frame specified?

Whenever you use the words "will", "could", "might", or "may" in a question, you might suspect that the question asks a time-related question. Be sure that, if it does, you have specified the time frame precisely. For instance, you might ask:

Do you think Congress will cut taxes?

or something like

Do you think Congress could successfully resist tax cuts?

Neither of these questions specifies a time frame.

How personal is the wording?

With a change of just a few words, a question can go from being relatively impersonal to probing into your private perspectives. Consider the following three questions, each of which asks about the respondent's satisfaction with working conditions:

- **Are working conditions satisfactory or not satisfactory in the plant where you work?**
- **Do you feel that working conditions satisfactory or not satisfactory in the plant where you work?**
- **Are you personally satisfied with working conditions in the plant where you work?**

The first question is stated from a fairly detached, objective viewpoint. The second asks how you "feel." The last asks whether you are "personally satisfied." Be sure the questions in your survey are at an appropriate level for your context. And, be sure there is consistency in this across questions in your survey.

Is the wording too direct?

There are times when asking a question too directly may be too threatening or disturbing for respondents. For instance, consider a study where you want to discuss battlefield experiences with former soldiers who experienced trauma. Examine the following three question options:

- **How did you feel about being in the war?**
- **How well did the equipment hold up in the field?**
- **How well were new recruits trained?**

The first question may be too direct. For this population it may elicit powerful negative emotions based on their recollections. The second question is a less direct one. It asks about equipment in the field, but, for this population, may also lead the discussion toward more difficult issues to discuss directly. The last question is probably the least direct and least threatening. Bashing the new recruits is standard protocol in almost any social context. The question is likely to get the respondent talking, recounting anecdotes, without eliciting much stress. Of course, all of this may simply be begging the question. If you are doing a study where the respondents may experience high levels of stress because of the questions you ask, you should reconsider the ethics of doing the study.

Other Wording Issues

The nuances of language guarantee that the task of the question writer will be endlessly complex. Without trying to generate an exhaustive list, here are a few other questions to keep in mind:

- **Does the question contain difficult or unclear terminology?**
- **Does the question make each alternative explicit?**
- **Is the wording objectionable?**
- **Is the wording loaded or slanted?**

- ❖ **Question Placement**

Decisions About Placement

One of the most difficult tasks facing the survey designer involves the ordering of questions. Which topics should be introduced early in the survey, and which later? If you leave your most important questions until the end, you may find that your respondents are too tired to give them the kind of attention you would like. If you introduce them too early, they may not yet be ready to address the topic, especially if it is a difficult or disturbing one. There are no easy answers to these problems - you have to use your judgment. Whenever you think about question placement, consider the following questions:

- **Is the answer influenced by prior questions?**
- **Does question come too early or too late to arouse interest?**
- **Does the question receive sufficient attention?**

The Opening Questions

Just as in other aspects of life, first impressions are important in survey work. The first few questions you ask will determine the tone for the survey, and can help put your respondent at ease. With that in mind, the opening few questions should, in general, be easy to answer. You might start with some simple descriptive questions that will get the respondent rolling. You should never begin your survey with sensitive or threatening questions.

Sensitive Questions

In much of our social research, we have to ask respondents about difficult or uncomfortable subjects. Before asking such questions, you should attempt to develop some trust or rapport with the respondent. Often, preceding the sensitive questions with some easier warm-up ones will help. But, you have to make sure that the sensitive material does not come up abruptly or appear unconnected with the rest of the survey. It is often helpful to have a transition sentence between sections of your instrument to give the respondent some idea of the kinds of questions that are coming. For instance, you might lead into a section on personal material with the transition:

In this next section of the survey, we'd like to ask you about your personal relationships. Remember, we do not want you to answer any questions if you are uncomfortable doing so.

A Checklist of Considerations

There are lots of conventions or rules-of-thumb in the survey design business. Here's a checklist of some of the most important items. You can use this checklist to review your instrument:

□ start with easy, nonthreatening questions

□ put more difficult, threatening questions near end

□ never start a mail survey with an open-ended question

□ for historical demographics, follow chronological order

□ ask about one topic at a time

□ when switching topics, use a transition

□ reduce response set (the tendency of respondent to just keep checking the same response)

□ for filter or contingency questions, make a flowchart

The Golden Rule

You are imposing in the life of your respondent. You are asking for their time, their attention, their trust, and often, for personal information. Therefore, you should always keep in mind the "golden rule" of survey research (and, I hope, for the rest of your life as well!):

Do unto your respondents as you would have them do unto you!

To put this in more practical terms, you should keep the following in mind:

- **Thank the respondent at the beginning for allowing you to conduct your study**
- **Keep your survey as short as possible -- only include what is absolutely necessary**
- **Be sensitive to the needs of the respondent**
- **Be alert for any sign that the respondent is uncomfortable**
- **Thank the respondent at the end for participating**
- **Assure the respondent that you will send a copy of the final results**

- **Interviews**

Interviews are among the most challenging and rewarding forms of measurement. They require a personal sensitivity and adaptability as well as the ability to stay within the bounds of the designed protocol. Here, I describe the preparation you need to do for an interview study and the process of conducting the interview itself.

Preparation

The Role of the Interviewer

The interviewer is really the "jack-of-all-trades" in survey research. The interviewer's role is complex and multifaceted. It includes the following tasks:

- **Locate and enlist cooperation of respondents**

The interviewer has to find the respondent. In door-to-door surveys, this means being able to locate specific addresses. Often, the interviewer has to work at the least desirable times (like immediately after dinner or on weekends) because that's when respondents are most readily available.

- **Motivate respondents to do good job**

If the interviewer does not take the work seriously, why would the respondent? The interviewer has to be motivated and has to be able to communicate that motivation to the respondent. Often, this means that the interviewer has to be convinced of the importance of the research.

- **Clarify any confusion/concerns**

Interviewers have to be able to think on their feet. Respondents may raise objections or concerns that were not anticipated. The interviewer has to be able to respond candidly and informatively.

- **Observe quality of responses**

Whether the interview is personal or over the phone, the interviewer is in the best position to judge the quality of the information that is being received. Even a verbatim transcript will not adequately convey how seriously the respondent took the task, or any gestures or body language that were evident.

- **Conduct a good interview**

Last, and certainly not least, the interviewer has to conduct a good interview! Every interview has a life of its own. Some respondents are motivated and attentive, others are distracted or disinterested. The interviewer also has good or bad days. Assuring a consistently high-quality interview is a challenge that requires constant effort.

Training the Interviewers

One of the most important aspects of any interview study is the training of the interviewers themselves. In many ways the interviewers are your measures, and the quality of the results is totally in their hands. Even in small studies involving only a single researcher-interviewer, it is important to organize in detail and rehearse the interviewing process before beginning the formal study.

Here are some of the major topics that should be included in interviewer training:

- **Describe the entire study**

Interviewers need to know more than simply how to conduct the interview itself. They should learn about the background for the study, previous work that has been done, and why the study is important.

- **State who is sponsor of research**

Interviewers need to know who they are working for. They -- and their respondents -- have a right to know not just what agency or company is conducting the research, but also, who is paying for the research.

- **Teach enough about survey research**

While you seldom have the time to teach a full course on survey research methods, the interviewers need to know enough that they respect the survey method and are motivated. Sometimes it may not be apparent why a question or set of questions was asked in a particular way. The interviewers will need to understand the rationale for how the instrument was constructed.

- **Explain the sampling logic and process**

Naive interviewers may not understand why sampling is so important. They may wonder why you go through all the difficulties of selecting the sample so carefully. You will have to explain that sampling is the basis for the conclusions that will be reached and for the degree to which your study will be useful.

- **Explain interviewer bias**

Interviewers need to know the many ways that they can inadvertently bias the results. And, they need to understand why it is important that they not bias the study. This is especially a problem when you are investigating political or moral issues on which people have strongly held convictions. While the interviewer may think they are doing good for society by slanting results in favor of what they believe, they need to recognize that doing so could jeopardize the entire study in the eyes of others.

- **"Walk through" the interview**

When you first introduce the interview, it's a good idea to walk through the entire protocol so the interviewers can get an idea of the various parts or phases and how they interrelate.

- **Explain respondent selection procedures, including**

- **reading maps**

It's astonishing how many adults don't know how to follow directions on a map. In personal interviews, the interviewer may need to locate respondents who are spread over a wide geographic area. And, they often have to navigate by night (respondents tend to be most available in evening hours) in neighborhoods they're not familiar with. Teaching basic map reading skills and confirming that the interviewers can follow maps is essential.

- **identifying households**

In many studies it is impossible in advance to say whether every sample household meets the sampling requirements for the study. In your study, you may want to interview only people who live in single family homes. It may be impossible to distinguish townhouses and apartment buildings in your sampling frame. The interviewer must know how to identify the appropriate target household.

- **identify respondents**

Just as with households, many studies require respondents who meet specific criteria. For instance, your study may require that you speak with a male head-of-household between the ages of 30 and 40 who has children under 18 living in the same household. It may be impossible to obtain statistics in advance to target such respondents. The interviewer may have to ask a series of filtering questions before determining whether the respondent meets the sampling needs.

- **Rehearse interview**

You should probably have several rehearsal sessions with the interviewer team. You might even videotape rehearsal interviews to discuss how the trainees responded in difficult situations. The interviewers should be very familiar with the entire interview before ever facing a respondent.

- **Explain supervision**

In most interview studies, the interviewers will work under the direction of a supervisor. In some contexts, the supervisor may be a faculty advisor; in others, they may be the "boss." In order to assure the quality of the responses, the supervisor may have to observe a subsample of interviews, listen in on phone interviews, or conduct follow-up assessments of interviews with the respondents. This can be very threatening to the interviewers. You need to develop an atmosphere where everyone on the research team -- interviewers and supervisors -- feel like they're working together towards a common end.

- **Explain scheduling**

The interviewers have to understand the demands being made on their schedules and why these are important to the study. In some studies it will be imperative to conduct the entire set of interviews within a certain time period. In most studies, it's important to have the interviewers available when it's convenient for the respondents, not necessarily the interviewer.

The Interviewer's Kit

It's important that interviewers have all of the materials they need to do a professional job. Usually, you will want to assemble an interviewer kit that can be easily carried and includes all of the important materials such as:

- **a "professional-looking" 3-ring notebook (this might even have the logo of the company or organization conducting the interviews)**
- **maps**
- **sufficient copies of the survey instrument**
- **official identification (preferable a picture ID)**
- **a cover letter from the Principal Investigator or Sponsor**
- **a phone number the respondent can call to verify the interviewer's authenticity**

The Interview

So all the preparation is complete, the training done, the interviewers ready to proceed, their "kits" in hand. It's finally time to do an actual interview. Each interview is unique, like a small work of art (and sometimes the art may not be very good). Each interview has its own ebb and flow -- its own pace. To the outsider, an interview looks like a fairly standard, simple, prosaic effort. But to the interviewer, it can be filled with special nuances and interpretations that aren't often immediately apparent. Every interview includes some common components. There's the opening, where the interviewer gains entry and establishes the rapport and tone for what follows. There's the middle game, the heart of the process, that consists of the protocol of questions and the improvisations of the probe. And finally, there's the endgame, the wrap-up, where the interviewer and respondent establish a sense of closure. Whether it's a two-minute phone interview or a personal interview that spans hours, the interview is a bit of theater, a mini-drama that involves real lives in real time.

Opening Remarks

In many ways, the interviewer has the same initial problem that a salesperson has. You have to get the respondent's attention initially for a long enough period that you can sell them on the idea of participating in the study. Many of the remarks here assume an interview that is being conducted at a respondent's residence. But the analogies to other interview contexts should be straightforward.

- **Gaining entry**

The first thing the interviewer must do is gain entry. Several factors can enhance the prospects. Probably the most important factor is your initial appearance. The interviewer needs to dress professionally and in a manner that will be comfortable to the respondent. In some contexts a business suit and briefcase may be appropriate. In others, it may intimidate. The way the interviewer appears initially to the respondent has to communicate some simple messages -- that you're trustworthy, honest, and non-threatening. Cultivating a manner of professional confidence, the sense that the respondent has nothing to worry about because you know what you're doing -- is a difficult skill to teach and an indispensable skill for achieving initial entry.

- **Doorstep technique**

You're standing on the doorstep and someone has opened the door, even if only halfway. You need to smile. You need to be brief. State why you are there and suggest what you would like the respondent to do. Don't ask -- suggest what you want. Instead of saying "May I come in to do an interview?", you might try a more imperative approach like " I'd like to take a few minutes of your time to interview you for a very important study."

- **Introduction**

If you've gotten this far without having the door slammed in your face, chances are you will be able to get an interview. Without waiting for the respondent to ask questions, you should move to introducing yourself. You should have this part of the process memorized so you can deliver the essential information in 20-30 seconds at most. State your name and the name of the organization you represent. Show your identification badge and the letter that introduces you. You want to have as legitimate an appearance as possible. If you have a three-ring binder or clipboard with the logo of your organization, you should have it out and visible. You should assume that the respondent will be interested in participating in your important study -- assume that you will be doing an interview here.

- **Explaining the study**

At this point, you've been invited to come in (After all, you're standing there in the cold, holding an assortment of materials, clearly displaying your credentials, and offering the respondent the chance to participate in an interview -- to many respondents, it's a rare and exciting event. They hardly ever get asked their views about anything, and yet they know that important decisions are made all the time based on input from others.). Or, the respondent has continued to listen long enough that you need to move onto explaining the study. There are three rules to this critical explanation: 1) Keep it short; 2) Keep it short; and 3) Keep it short! The respondent doesn't have to or want to know all of the neat nuances of this study, how it came about, how you convinced your thesis committee to buy into it, and so on. You should have a one or two sentence description of the study memorized. No big words. No jargon. No detail. There will be more than enough time for that later (and you should bring some written materials you can leave at the end for that purpose). This is the "25 words or less" description. What you ***should*** spend some time on is assuring the respondent that you are interviewing them confidentially, and that their participation is voluntary.

Asking the Questions

You've gotten in. The respondent has asked you to sit down and make yourself comfortable. It may be that the respondent was in the middle of doing something when you arrived and you may need to allow them a few minutes to finish the phone call or send the kids off to do homework. Now, you're ready to begin the interview itself.

- **Use questionnaire carefully, but informally**

The questionnaire is your friend. It was developed with a lot of care and thoughtfulness. While you have to be ready to adapt to the needs of the setting, your first instinct should always be to trust the instrument that was designed. But you also need to establish a rapport with the respondent. If you have your face in the instrument and you read the questions, you'll appear unprofessional and disinterested. Even though you may be nervous, you need to recognize that your respondent is most likely even more nervous. If you memorize the first few questions, you can refer to the instrument only occasionally, using eye contact and a confident manner to set the tone for the interview and help the respondent get comfortable.

- **Ask questions exactly as written**

Sometimes an interviewer will think that they could improve on the tone of a question by altering a few words to make it simpler or more "friendly." DON'T. You should ask the questions as they are on the instrument. If you had a problem with a question, the time to raise it was during the training and rehearsals, not during the actual interview. It is important that the interview be as standardized as possible across respondents (this is true except in certain types of exploratory or interpretivist research where the explicit goal is to avoid any standardizing). You may think the change you made was inconsequential when, in fact, it may change the entire meaning of the question or response.

- **Follow the order given**

Once you know an interview well, you may see a respondent bring up a topic that you know will come up later in the interview. You may be tempted to jump to that section of the interview while you're on the topic. DON'T. You are more likely to lose your place. You may omit questions that build a foundation for later questions.

- **Ask every question**

Sometimes you'll be tempted to omit a question because you thought you already heard what the respondent will say. Don't assume that. For example, let's say you were conducting an interview with college age women about the topic of date rape. In an earlier question, the respondent mentioned that she knew of a woman on her dormitory floor who had been raped on a date within the past year. A few questions later, you are supposed to ask "Do you know of anyone personally who was raped on a date?" You figure you already know that the answer is yes, so you decide to skip the question. Instead, you might say something like "I know you may have already mentioned this, but do you know of anyone personally who was raped on a date?" At this point, the respondent may say something like "Well, in addition to the woman who lived down the hall in my dorm, I know of a friend from high school who experienced date rape." If you hadn't asked the question, you would never have discovered this detail.

- **Don't finish sentences**

I don't know about you, but I'm one of those people who just hates to be left hanging. I like to keep a conversation moving. Once I know where a sentence seems to be heading, I'm aching to get to the next sentence. I finish people's sentences all the time. If you're like me, you should practice the art of patience (and silence) before doing any interviewing. As you'll see below, silence is one of the most effective devices for encouraging a respondent to talk. If you finish their sentence for them, you imply that what they had to say is transparent or obvious, or that you don't want to give them the time to express themselves in their own language.

Obtaining Adequate Responses - The Probe

OK, you've asked a question. The respondent gives a brief, cursory answer. How do you elicit a more thoughtful, thorough response? You *probe*.

- **Silent probe**

The most effective way to encourage someone to elaborate is to do nothing at all - just pause and wait. This is referred to as the "silent" probe. It works (at least in certain cultures) because the respondent is uncomfortable with pauses or silence. It suggests to the respondent that you are waiting, listening for what they will say next.

- **Overt encouragement**

At times, you can encourage the respondent directly. Try to do so in a way that does not imply approval or disapproval of what they said (that could bias their subsequent results). Overt encouragement could be as simple as saying "Uh-huh" or "OK" after the respondent completes a thought.

- **Elaboration**

You can encourage more information by asking for elaboration. For instance, it is appropriate to ask questions like "Would you like to elaborate on that?" or "Is there anything else you would like to add?"

- **Ask for clarification**

Sometimes, you can elicit greater detail by asking the respondent to clarify something that was said earlier. You might say, "A minute ago you were talking about the experience you had in high school. Could you tell me more about that?"

- **Repetition**

This is the old psychotherapist trick. You say something without really saying anything new. For instance, the respondent just described a traumatic experience they had in childhood. You might say "What I'm hearing you say is that you found that experience very traumatic." Then, you should pause. The respondent is likely to say something like "Well, yes, and it affected the rest of my family as well. In fact, my younger sister..."

Recording the Response

Although we have the capability to record a respondent in audio and/or video, most interview methodologists don't think it's a good idea. Respondents are often uncomfortable when they know their remarks will be recorded word-for-word. They may strain to only say things in a socially acceptable way. Although you would get a more detailed and accurate record, it is likely to be distorted by the very process of obtaining it. This may be more of a problem in some situations than in others. It is increasingly common to be told that your conversation may be recorded during a phone interview. And most focus group methodologies use unobtrusive recording equipment to capture what's being said. But, in general, personal interviews are still best when recorded by the interviewer using pen and paper. Here, I assume the paper-and-pencil approach.

- **Record responses immediately**

The interviewer should record responses as they are being stated. This conveys the idea that you are interested enough in what the respondent is saying to write it down. You don't have to write down every single word -- you're not taking stenography. But you may want to record certain key phrases or quotes verbatim. You need to develop a system for distinguishing what the respondent says verbatim from what you are characterizing (how about quotations, for instance!).

- **Include all probes**

You need to indicate every single probe that you use. Develop a shorthand for different standard probes. Use a clear form for writing them in (e.g., place probes in the left margin).

- **Use abbreviations where possible**

Abbreviations will help you to capture more of the discussion. Develop a standardized system (e.g., R=respondent; DK=don't know). If you create an abbreviation on the fly, have a way of indicating its origin. For instance, if you decide to abbreviate Spouse with an 'S', you might make a notation in the right margin saying "S=Spouse."

Concluding the Interview

When you've gone through the entire interview, you need to bring the interview to closure. Some important things to remember:

- **Thank the respondent**

Don't forget to do this. Even if the respondent was troublesome or uninformative, it is important for you to be polite and thank them for their time.

- **Tell them when you expect to send results**

I hate it when people conduct interviews and then don't send results and summaries to the people who they get the information from. You owe it to your respondent to show them what you learned. Now, they may not want your entire 300-page dissertation. It's common practice to prepare a short, readable, jargon-free summary of interviews that you can send to the respondents.

- **Don't be brusque or hasty**

Allow for a few minutes of winding down conversation. The respondent may want to know a little bit about you or how much you like doing this kind of work. They may be interested in how the results will be used. Use these kinds of interests as a way to wrap up the conversation. As you're putting away your materials and packing up to go, engage the respondent. You don't want the respondent to feel as though you completed the interview and then rushed out on them -- they may wonder what they said that was wrong. On the other hand, you have to be careful here. Some respondents may want to keep on talking long after the interview is over. You have to find a way to politely cut off the conversation and make your exit.

- **Immediately after leaving -- write down any notes about how the interview went**

Sometimes you will have observations about the interview that you didn't want to write down while you were with the respondent. You may have noticed them get upset at a question, or you may have detected hostility in a response. Immediately after the interview you should go over your notes and make any other comments and observations -- but be sure to distinguish these from the notes made during the interview (you might use a different color pen, for instance).

- **Plus & Minus of Survey Methods**

It's hard to compare the advantages and disadvantages of the major different survey types. Even though each type has some general advantages and disadvantages, there are exceptions to almost every rule. Here's my general assessment. Perhaps you would differ in your ratings here or there, but I think you'll generally agree.

Issue	Questionnaire			Interview	
	Group	Mail	Drop-Off	Personal	Phone
Are Visual Presentations Possible?	Yes	Yes	Yes	Yes	No
Are Long Response Categories Possible?	Yes	Yes	Yes	???	No
Is Privacy A Feature?	No	Yes	No	Yes	???
Is the Method Flexible?	No	No	No	Yes	Yes
Are Open-ended Questions Feasible?	No	No	No	Yes	Yes
Is Reading & Writing Needed?	???	Yes	Yes	No	No
Can You Judge Quality of Response?	Yes	No	???	Yes	???
Are High Response Rates Likely?	Yes	No	Yes	Yes	No
Can You Explain Study in Person?	Yes	No	Yes	Yes	???

Is It Low Cost?	Yes	Yes	No	No	No
Are Staff & Facilities Needs Low?	Yes	Yes	No	No	No
Does It Give Access to Dispersed Samples?	No	Yes	No	No	No
Does Respondent Have Time to Formulate Answers?	No	Yes	Yes	No	No
Is There Personal Contact?	Yes	No	Yes	Yes	No
Is A Long Survey Feasible?	No	No	No	Yes	No
Is There Quick Turnaround?	No	Yes	No	No	Yes

Scaling

Scaling is the branch of measurement that involves the construction of an instrument that associates qualitative constructs with quantitative metric units. Scaling evolved out of efforts in psychology and education to measure "unmeasurable" constructs like authoritarianism and self esteem. In many ways, scaling remains one of the most arcane and misunderstood aspects of social research measurement. And, it attempts to do one of the most difficult of research tasks -- measure abstract concepts.

Most people don't even understand what scaling is. The basic idea of scaling is described in General Issues in Scaling, including the important distinction between a scale and a response format. Scales are generally divided into two broad categories: unidimensional and multidimensional. The unidimensional scaling methods were developed in the first half of the twentieth century and are generally named after their inventor. We'll look at three types of unidimensional scaling methods here:

- **Thurstone or Equal-Appearing Interval Scaling**
- **Likert or "Summative" Scaling**
- **Guttman or "Cumulative" Scaling**

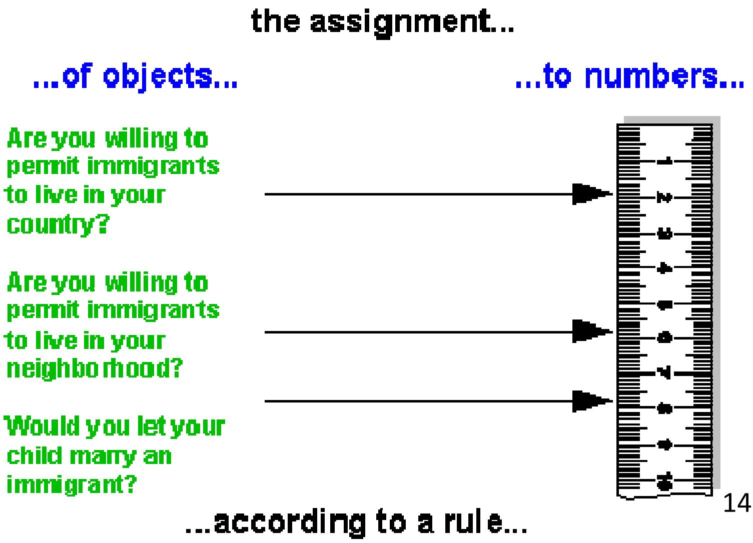
In the late 1950s and early 1960s, measurement theorists developed more advanced techniques for creating multidimensional scales. Although these techniques are not considered here, you may want to look at the method of concept mapping that relies on that approach to see the power of these multivariate methods.

- **General Issues in Scaling**

S.S. Stevens came up with what I think is the simplest and most straightforward definition of scaling. He said:

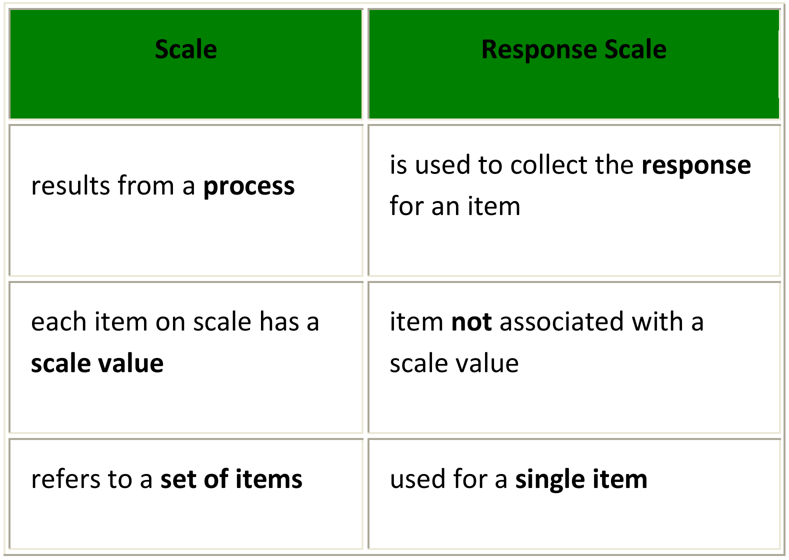
> Scaling is the assignment of objects to numbers according to a rule.

But what does that mean? In most scaling, the objects are text statements, usually statements of attitude or belief. The figure shows an example. There are three statements describing attitudes towards immigration. To scale these statements, we have to assign


the assignment...
...of objects...
...to numbers...
Are you willing to permit immigrants to live in your country?
Are you willing to permit immigrants to live in your neighborhood?
Would you let your child marry an immigrant?
...according to a rule...

numbers to them. Usually, we would like the result to be on at least an interval scale (see Levels of Measurement) as indicated by the ruler in the figure. And what does "according to a rule" mean? If you look at the statements, you can see that as you read down, the attitude towards immigration becomes more restrictive -- if a person agrees with a statement on the list, it's likely that they will also agree with all of the statements higher on the list. In this case, the "rule" is a *cumulative* one. So what is scaling? It's how we get numbers that can be meaningfully assigned to objects -- it's a set of procedures. We'll present several different approaches below.

But first, I have to clear up one of my pet peeves. People often confuse the idea of a scale and a response scale. A response scale is the way you collect responses from people on an instrument. You might use a dichotomous response scale like Agree/Disagree, True/False, or Yes/No. Or, you might use an interval response scale like a 1-to-5 or 1-to-7 rating. But, if all you are doing is attaching a response scale to an object or statement, you can't call that scaling. As you will see, scaling involves procedures that you do independent of the respondent so that you can come up with a numerical value for the object. In true scaling research, you use a scaling procedure to develop your instrument (scale) and you also use a response scale to collect the responses from participants. But just assigning a 1-to-5 response scale for an item is **not** scaling! The differences are illustrated in the table below.



Scale	Response Scale
results from a **process**	is used to collect the **response** for an item
each item on scale has a **scale value**	item **not** associated with a scale value
refers to a **set of items**	used for a **single item**

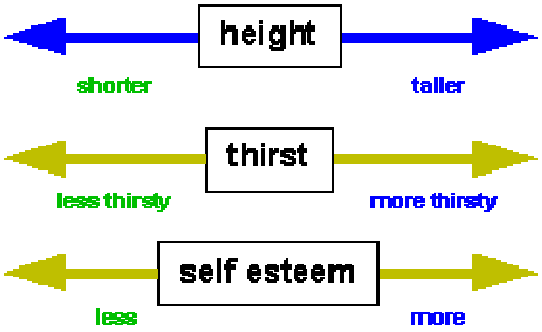
Purposes of Scaling

Why do we do scaling? Why not just create text statements or questions and use response formats to collect the answers? First, sometimes we do scaling to test a hypothesis. We might want to know whether the construct or concept is a single dimensional or multidimensional one (more about dimensionality later). Sometimes, we do scaling as part of exploratory research. We want to know what dimensions underlie a set of ratings. For instance, if you create a set of questions, you can use scaling to determine how well they "hang together" and whether they measure one concept or multiple concepts. But probably the most common reason for doing scaling is for scoring purposes. When a participant gives their responses to a set of items, we often would like to assign a single number that represents that's person's overall attitude or belief.

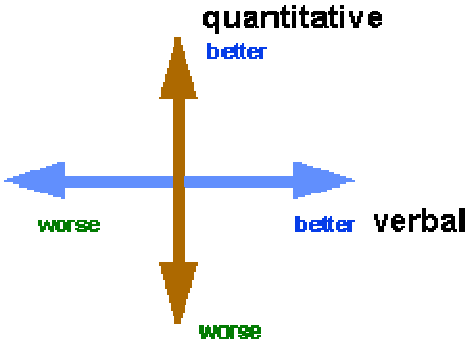
For the figure above, we would like to be able to give a single number that describes a person's attitudes towards immigration, for example.

Dimensionality

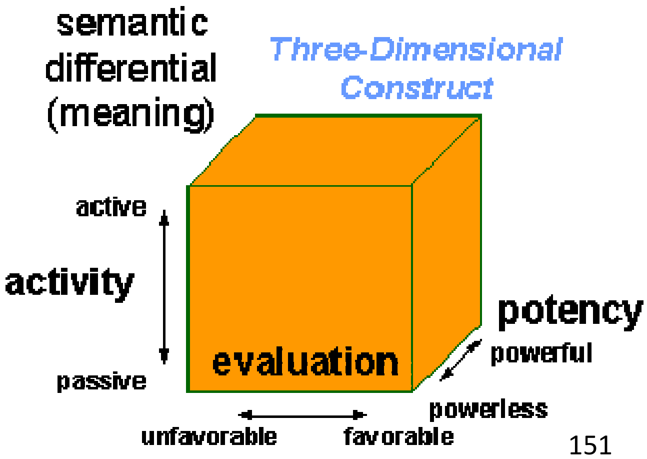
A scale can have any number of dimensions in it. Most scales that we develop have only a few dimensions. What's a dimension? Think of a dimension as a number line. If we want to measure a construct, we have to decide whether the construct can be measured well with one number line or whether it may need more. For instance, height is a concept that is unidimensional or one-dimensional. We can measure the concept of height very well with only a single number line (e.g., a ruler). Weight is also unidimensional -- we can measure it with a scale. Thirst might also bee considered a unidimensional concept -- you are either more or less thirsty at any given time. It's easy to see that height and weight are unidimensional. But what about a concept like self esteem? If you think you can measure a person's self esteem well with a single ruler that goes from low to high, then you probably have a unidimensional construct.



What would a two-dimensional concept be? Many models of intelligence or achievement postulate two major dimensions -- mathematical and verbal ability. In this type of two-dimensional model, a person can be said to possess two types of achievement. Some people will be high in verbal skills and lower in math. For others, it will be the reverse. But, if a concept is truly two-dimensional, it is not possible to depict a person's level on it using only a single number line. In other words, in order to describe achievement you would need to locate a person as a point in two dimensional (x,y) space.



OK, let's push this one step further: how about a three-dimensional concept? Psychologists who study the idea of meaning theorized that the meaning of a term could be well described in three dimensions. Put in other terms, any objects can be distinguished or differentiated from each other along three dimensions. They labeled these three dimensions *activity*, *evaluation*, and *potency*. They called this



general theory of meaning the **semantic differential**. Their theory essentially states that you can rate any object along those three dimensions. For instance, think of the idea of "ballet." If you like the ballet, you would probably rate it high on activity, favorable on evaluation, and powerful on potency. On the other hand, think about the concept of a "book" like a novel. You might rate it low on activity (it's passive), favorable on evaluation (assuming you like it), and about average on potency. Now, think of the idea of "going to the dentist." Most people would rate it low on activity (it's a passive activity), unfavorable on evaluation, and powerless on potency (there are few routine activities that make you feel as powerless!). The theorists who came up with the idea of the semantic differential thought that the meaning of any concepts could be described well by rating the concept on these three dimensions. In other words, in order to describe the meaning of an object you have to locate it as a dot somewhere within the cube (three-dimensional space).

Unidimensional or Multidimensional?

What are the advantages of using a unidimensional model? Unidimensional concepts are generally easier to understand. You have either more or less of it, and that's all. You're either taller or shorter, heavier or lighter. It's also important to understand what a unidimensional scale is as a foundation for comprehending the more complex multidimensional concepts. But the best reason to use unidimensional scaling is because you believe the concept you are measuring really is unidimensional in reality. As you've seen, many familiar concepts (height, weight, temperature) are actually unidimensional. But, if the concept you are studying is in fact multidimensional in nature, a unidimensional scale or number line won't describe it well. If you try to measure academic achievement on a single dimension, you would place every person on a single line ranging from low to high achievers. But how do you score someone who is a high math achiever and terrible verbally, or vice versa? A unidimensional scale can't capture that type of achievement.

The Major Unidimensional Scale Types

There are three major types of unidimensional scaling methods. They are similar in that they each measure the concept of interest on a number line. But they differ considerably in how they arrive at scale values for different items. The three methods are Thurstone or Equal-Appearing Interval Scaling, Likert or "Summative" Scaling, and Guttman or "Cumulative" Scaling.

- **Thurstone Scaling**

Thurstone was one of the first and most productive scaling theorists. He actually invented three different methods for developing a unidimensional scale: the **method of equal-appearing intervals**; the **method of successive intervals**; and, the **method of paired comparisons**. The three methods differed in how the scale values for items were constructed, but in all three cases, the resulting scale was rated the same way by respondents. To illustrate Thurstone's approach, I'll show you the easiest method of the three to implement, the method of equal-appearing intervals.

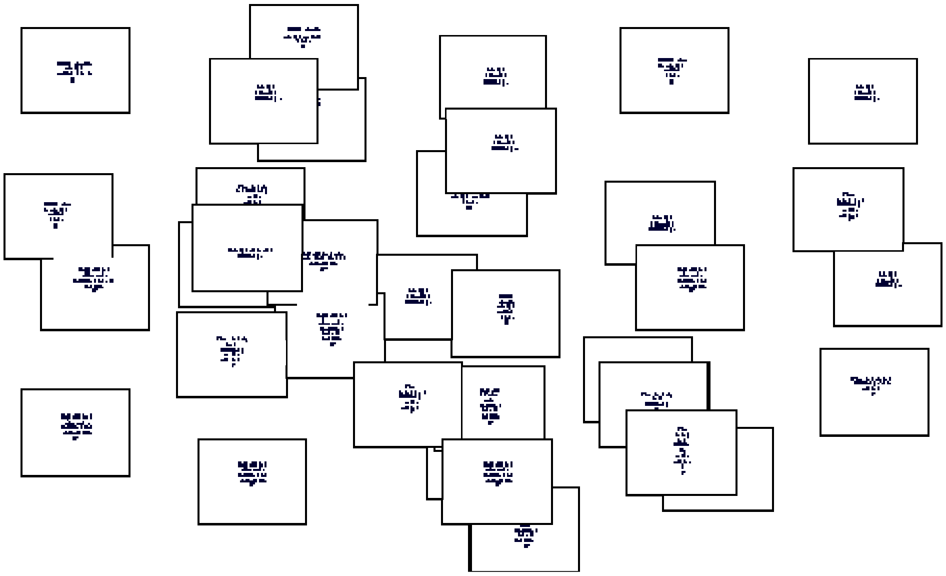
The Method of Equal-Appearing Intervals

Developing the Focus. The Method of Equal-Appearing Intervals starts like almost every other scaling method -- with a large set of statements. Oops! I did it again! You can't start with the set of statements -- you have to first define the focus for the scale you're trying to develop. Let this be a warning to all of you: methodologists like me often start our descriptions with the first objective methodological step (in this case, developing a set of statements) and forget to mention critical foundational issues like the development of the focus for a project. So, let's try this again...

The Method of Equal-Appearing Intervals starts like almost every other scaling method -- with the development of the focus for the scaling project. Because this is a unidimensional scaling method, we assume that the concept you are trying to scale is reasonably thought of as one-dimensional. The description of this concept should be as clear as possible so that the person(s) who are going to create the statements have a clear idea of what you are trying to measure. I like to state the focus for a scaling project in the form of a command -- the command you will give to the people who will create the statements. For instance, you might start with the focus command:

Generate statements that describe specific attitudes that people might have towards persons with AIDS.

You want to be sure that everyone who is generating statements has some idea of what you are after in this focus command. You especially want to be sure that technical language and acronyms are spelled out and understood (e.g., what is AIDS?).

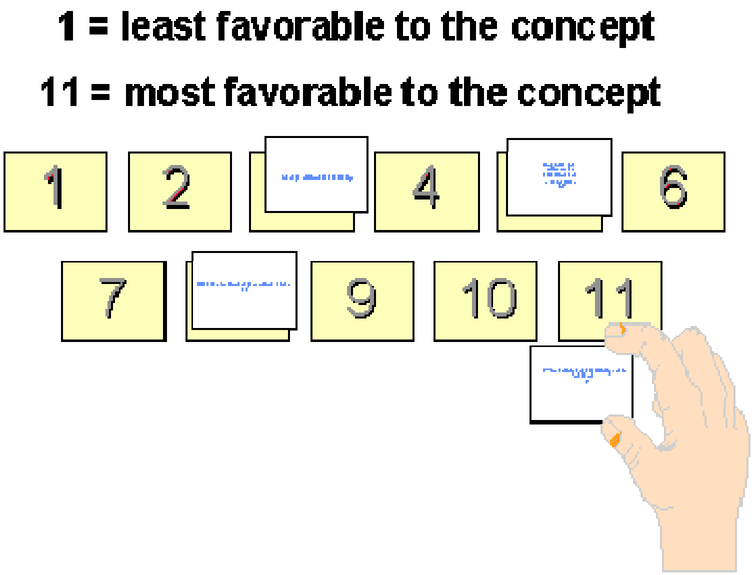


Generating Potential Scale Items. Now, you're ready to create statements. You want a large set of candidate statements (e.g., 80 -- 100) because you are going to select your final scale items from this pool. You also want to be sure that all of the statements are worded similarly -- that they don't differ in grammar or structure. For instance, you might want them each to be worded as a statement which you cold agree or disagree with. You don't want some of them to be statements while others are questions.

For our example focus on developing an AIDS attitude scale, we might generate statements like the following (these statements came from a class exercise I did in my Spring 1997 undergrad class):

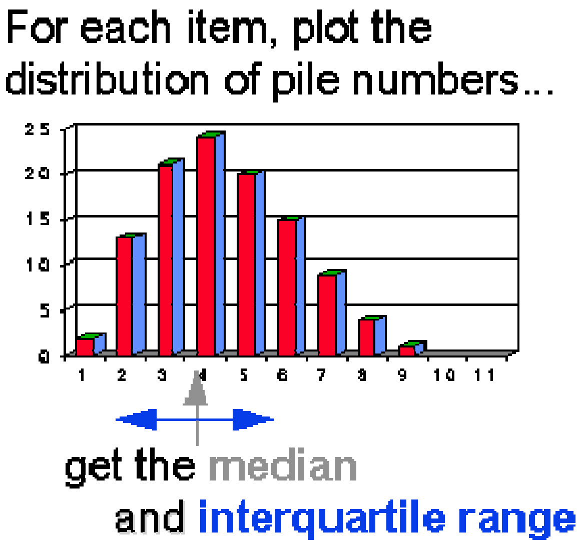
- people get AIDS by engaging in immoral behavior
- you can get AIDS from toilet seats
- AIDS is the wrath of God
- anybody with AIDS is either gay or a junkie
- AIDS is an epidemic that affects us all
- people with AIDS are bad
- people with AIDS are real people
- AIDS is a cure, not a disease
- you can get AIDS from heterosexual sex
- people with AIDS are like my parents
- you can get AIDS from public toilets
- women don't get AIDS
- I treat everyone the same, regardless of whether or not they have AIDS
- AIDS costs the public too much
- AIDS is something the other guy gets
- living with AIDS is impossible
- children cannot catch AIDS
- AIDS is a death sentence
- because AIDS is preventable, we should focus our resources on prevention instead of curing
- People who contract AIDS deserve it
- AIDS doesn't have a preference, anyone can get it.
- AIDS is the worst thing that could happen to you.
- AIDS is good because it will help control the population.
- If you have AIDS, you can still live a normal life.
- People with AIDS do not need or deserve our help
- By the time I would get sick from AIDS, there will be a cure
- AIDS will never happen to me
- you can't get AIDS from oral sex
- AIDS is spread the same way colds are
- AIDS does not discriminate
- You can get AIDS from kissing
- AIDS is spread through the air
- Condoms will always prevent the spread of AIDS
- People with AIDS deserve what they got
- If you get AIDS you will die within a year
- Bad people get AIDS and since I am a good person I will never get AIDS
- I don't care if I get AIDS because researchers will soon find a cure for it.
- AIDS distracts from other diseases that deserve our attention more
- bringing AIDS into my family would be the worst thing I could do
- very few people have AIDS, so it's unlikely that I'll ever come into contact with a sufferer
- if my brother caught AIDS I'd never talk to him again
- People with AIDS deserve our understanding, but not necessarily special treatment
- AIDS is a omnipresent, ruthless killer that lurks around dark alleys, silently waiting for naive victims to wander passed so that it might pounce.

- I can't get AIDS if I'm in a monogamous relationship
- the nation's blood supply is safe
- universal precautions are infallible
- people with AIDS should be quarantined to protect the rest of society
- because I don't live in a big city, the threat of AIDS is very small
- I know enough about the spread of the disease that I would have no problem working in a health care setting with patients with AIDS
- the AIDS virus will not ever affect me
- Everyone affected with AIDS deserves it due to their lifestyle
- Someone with AIDS could be just like me
- People infected with AIDS did not have safe sex
- Aids affects us all.
- People with AIDS should be treated just like everybody else.
- AIDS is a disease that anyone can get if there are not careful.
- It's easy to get AIDS.
- The likelihood of contracting AIDS is very low.
- The AIDS quilt is an emotional reminder to remember those who did not deserve to die painfully or in vain
- The number of individuals with AIDS in Hollywood is higher than the general public thinks
- It is not the AIDS virus that kills people, it is complications from other illnesses (because the immune system isn't functioning) that cause death
- AIDS is becoming more a problem for heterosexual women and their offsprings than IV drug users or homosexuals
- A cure for AIDS is on the horizon
- A cure for AIDS is on the horizon
- Mandatory HIV testing should be established for all pregnant women


1 = least favorable to the concept
11 = most favorable to the concept
1
2
4
6
7
9
10
11

Rating the Scale Items. OK, so now you have a set of statements. The next step is to have your participants (i.e., judges) rate each statement on a 1-to-11 scale in terms of how much each statement indicates a ***favorable*** attitude towards people with AIDS. Pay close attention here! You

DON'T want the participants to tell you what their attitudes towards AIDS are, or whether they would agree with the statements. You want them to rate the "favorableness" of each statement in terms of an attitude towards AIDS, where 1 = "extremely unfavorable attitude towards people with AIDS" and 11 = "extremely favorable attitude towards people with AIDS.". (Note that I could just as easily had the judges rate how much each statement represents a negative attitude towards AIDS. If I did, the scale I developed would have higher scale values for people with more negative attitudes).



Computing Scale Score Values for Each Item. The next step is to analyze the rating data. For each statement, you need to compute the Median and the Interquartile Range. The median is the value above and below which 50% of the ratings fall. The first quartile (Q1) is the value below which 25% of the cases fall and above which 75% of the cases fall -- in other words, the 25th percentile. The median is the 50th percentile. The third quartile, Q3, is the 75th percentile. The Interquartile Range is the difference between third and first quartile, or Q3 - Q1. The figure above shows a histogram for a single item and indicates the median and Interquartile Range. You can compute these values easily with any introductory statistics program or with most spreadsheet programs. To facilitate the final selection of items for your scale, you might want to sort the table of medians and Interquartile Range in ascending order by Median and, within that, in descending order by Interquartile Range. For the items in this example, we got a table like the following:

Statement Number	Median	Q1	Q3	Interquartile Range
23	1	1	2.5	1.5
8	1	1	2	1
12	1	1	2	1

34	1	1	2	1
39	1	1	2	1
54	1	1	2	1
56	1	1	2	1
57	1	1	2	1
18	1	1	1	0
25	1	1	1	0
51	1	1	1	0
27	2	1	5	4
45	2	1	4	3
16	2	1	3.5	2.5
42	2	1	3.5	2.5
24	2	1	3	2
44	2	2	4	2
36	2	1	2.5	1.5
43	2	1	2.5	1.5
33	3	1	5	4
48	3	1	5	4
20	3	1.5	5	3.5
28	3	1.5	5	3.5
31	3	1.5	5	3.5

19	3	1	4	3
22	3	1	4	3
37	3	1	4	3
41	3	2	5	3
6	3	1.5	4	2.5
21	3	1.5	4	2.5
32	3	2	4.5	2.5
9	3	2	3.5	1.5
1	4	3	7	4
26	4	1	5	4
47	4	1	5	4
30	4	1.5	5	3.5
13	4	2	5	3
11	4	2	4.5	2.5
15	4	3	5	2
40	5	4.5	8	3.5
2	5	4	6.5	2.5
14	5	4	6	2
17	5.5	4	8	4
49	6	5	9.75	4.75
50	8	5.5	11	5.5

35	8	6.25	10	3.75
29	9	5.5	11	5.5
38	9	5.5	10.5	5
3	9	6	10	4
55	9	7	11	4
10	10	6	10.5	4.5
7	10	7.5	11	3.5
46	10	8	11	3
5	10	8.5	11	2.5
53	11	9.5	11	1.5
4	11	10	11	1

Selecting the Final Scale Items. Now, you have to select the final statements for your scale. You should select statements that are at equal intervals across the range of medians. In our example, we might select one statement for each of the eleven median values. Within each value, you should try to select the statement that has the smallest Interquartile Range. This is the statement with the least amount of variability across judges. You don't want the statistical analysis to be the only deciding factor here. Look over the candidate statements at each level and select the statement that makes the most sense. If you find that the best statistical choice is a confusing statement, select the next best choice.

When we went through our statements, we came up with the following set of items for our scale:

- People with AIDS are like my parents (6)
- Because AIDS is preventable, we should focus our resources on prevention instead of curing (5)
- People with AIDS deserve what they got. (1)
- Aids affects us all (10)
- People with AIDS should be treated just like everybody else. (11)
- AIDS will never happen to me. (3)
- It's easy to get AIDS (5)
- AIDS doesn't have a preference, anyone can get it (9)
- AIDS is a disease that anyone can get if they are not careful (9)
- If you have AIDS, you can still lead a normal life (8)
- AIDS is good because it helps control the population. (2)

- I can't get AIDS if I'm in a monogamous relationship. (4)

The value in parentheses after each statement is its scale value. Items with higher scale values should, in general, indicate a more favorable attitude towards people with AIDS. Notice that we have randomly scrambled the order of the statements with respect to scale values. Also, notice that we do not have an item with scale value of 7 and that we have two with values of 5 and of 9 (one of these pairs will average out to a 7).

Administering the Scale. You now have a scale -- a yardstick you can use for measuring attitudes towards people with AIDS. You can give it to a participant and ask them to agree or disagree with each statement. To get that person's total scale score, you average the scale scores of all the items that person agreed with. For instance, let's say a respondent completed the scale as follows:

● Agree	○ Disagree	People with AIDS are like my parents.
● Agree	○ Disagree	Because AIDS is preventable, we should focus our resources on prevention instead of curing.
○ Agree	● Disagree	People with AIDS deserve what they got.
● Agree	○ Disagree	Aids affects us all.
● Agree	○ Disagree	People with AIDS should be treated just like everybody else.
○ Agree	● Disagree	AIDS will never happen to me.
○ Agree	● Disagree	It's easy to get AIDS.
● Agree	○ Disagree	AIDS doesn't have a preference, anyone can get it.
●	○	AIDS is a disease that anyone can get if they are

Agree	Disagree	not careful.
◉ Agree	○ Disagree	If you have AIDS, you can still lead a normal life.
○ Agree	◉ Disagree	AIDS is good because it helps control the population.
◉ Agree	○ Disagree	I can't get AIDS if I'm in a monogamous relationship.

If you're following along with the example, you should see that the respondent checked eight items as Agree. When we take the average scale values for these eight items, we get a final value for this respondent of 7.75. This is where this particular respondent would fall on our "yardstick" that measures attitudes towards persons with AIDS. Now, let's look at the responses for another individual:

○ Agree	◉ Disagree	People with AIDS are like my parents.
○ Agree	◉ Disagree	Because AIDS is preventable, we should focus our resources on prevention instead of curing.
◉ Agree	○ Disagree	People with AIDS deserve what they got.
○ Agree	◉ Disagree	Aids affects us all.
○ Agree	◉ Disagree	People with AIDS should be treated just like everybody else.
◉ Agree	○ Disagree	AIDS will never happen to me.

○ Agree	● Disagree	It's easy to get AIDS.
○ Agree	● Disagree	AIDS doesn't have a preference, anyone can get it.
○ Agree	● Disagree	AIDS is a disease that anyone can get if they are not careful.
○ Agree	● Disagree	If you have AIDS, you can still lead a normal life.
● Agree	○ Disagree	AIDS is good because it helps control the population.
● Agree	○ Disagree	I can't get AIDS if I'm in a monogamous relationship.

In this example, the respondent only checked four items, all of which are on the negative end of the scale. When we average the scale items for the statements with which the respondent agreed we get an average score of 2.5, considerably lower or more negative in attitude than the first respondent.

The Other Thurstone Methods

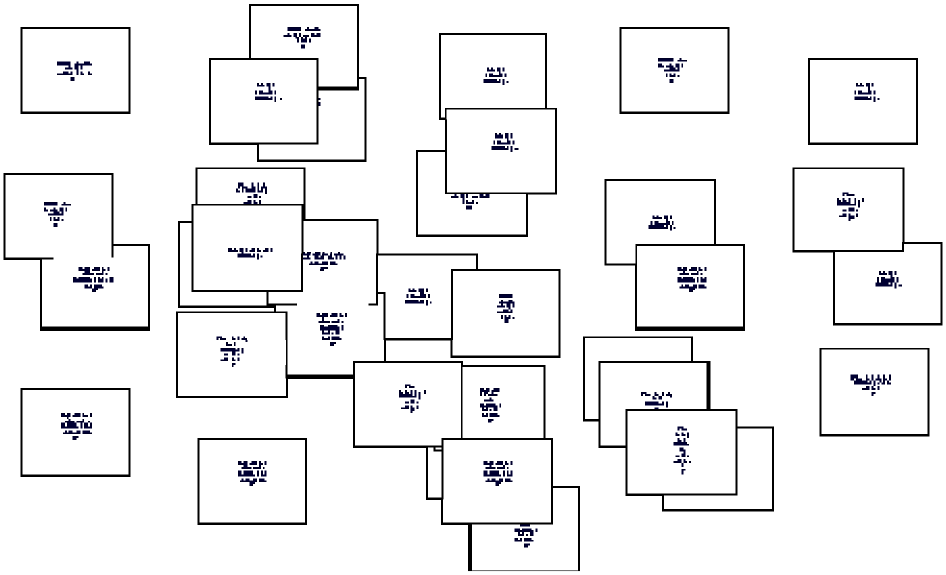
The other Thurstone scaling methods are similar to the Method of Equal-Appearing Intervals. All of them begin by focusing on a concept that is assumed to be unidimensional and involve generating a large set of potential scale items. All of them result in a scale consisting of relatively few items which the respondent rates on Agree/Disagree basis. The major differences are in how the data from the judges is collected. For instance, the method of paired comparisons requires each judge to make a judgement about each pair of statements. With lots of statements, this can become very time consuming indeed. With 57 statements in the original set, there are 1,596 unique pairs of statements that would have to be compared! Clearly, the paired comparison method would be too time consuming when there are lots of statements initially.

Thurstone methods illustrate well how a simple unidimensional scale might be constructed. There are other approaches, most notably Likert or Summative Scales and Guttman or Cumulative Scales.

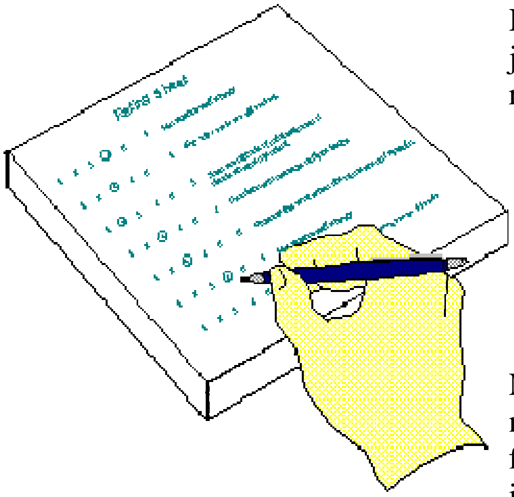
- **Likert Scaling**

Like Thurstone or Guttman Scaling, Likert Scaling is a unidimensional scaling method. Here, I'll explain the basic steps in developing a Likert or "Summative" scale.

Defining the Focus. As in all scaling methods, the first step is to define what it is you are trying to measure. Because this is a unidimensional scaling method, it is assumed that the concept you want to measure is one-dimensional in nature. You might operationalize the definition as an instruction to the people who are going to create or generate the initial set of candidate items for your scale.



Generating the Items. next, you have to create the set of potential scale items. These should be items that can be rated on a 1-to-5 or 1-to-7 Disagree-Agree response scale. Sometimes you can create the items by yourself based on your intimate understanding of the subject matter. But, more often than not, it's helpful to engage a number of people in the item creation step. For instance, you might use some form of brainstorming to create the items. It's desirable to have as large a set of potential items as possible at this stage, about 80-100 would be best.



Rating the Items. The next step is to have a group of judges rate the items. Usually you would use a 1-to-5 rating scale where:

1. = strongly unfavorable to the concept
2. = somewhat unfavorable to the concept
3. = undecided
4. = somewhat favorable to the concept
5. = strongly favorable to the concept

Notice that, as in other scaling methods, the judges are not telling you what they believe -- they are judging how favorable each item is with respect to the construct of interest.

Selecting the Items. The next step is to compute the intercorrelations between all pairs of items, based on the ratings of the judges. In making judgements about which items to retain for the final scale there are several analyses you can do:

- Throw out any items that have a low correlation with the total (summed) score across all items

  In most statistics packages it is relatively easy to compute this type of Item-Total correlation. First, you create a new variable which is the sum of all of the individual items for each respondent. Then, you include this variable in the correlation matrix computation (if you include it as the last variable in the list, the resulting Item-Total correlations will all be the last line of the correlation matrix and will be easy to spot). How low should the correlation be for you to throw out the item? There is no fixed rule here -- you might eliminate all items with a correlation with the total score less that .6, for example.

- For each item, get the average rating for the top quarter of judges and the bottom quarter. Then, do a t-test of the differences between the mean value for the item for the top and bottom quarter judges.

  Higher t-values mean that there is a greater difference between the highest and lowest judges. In more practical terms, items with higher t-values are better discriminators, so you want to keep these items. In the end, you will have to use your judgement about which items are most sensibly retained. You want a relatively small number of items on your final scale (e.g., 10-15) and you want them to have high Item-Total correlations and high discrimination (e.g., high t-values).

Administering the Scale. You're now ready to use your Likert scale. Each respondent is asked to rate each item on some response scale. For instance, they could rate each item on a 1-to-5 response scale where:

1. = strongly disagree
2. = disagree
3. = undecided
4. = agree
5. = strongly agree

There are a variety possible response scales (1-to-7, 1-to-9, 0-to-4). All of these odd-numbered scales have a middle value is often labeled Neutral or Undecided. It is also possible to use a forced-choice response scale with an even number of responses and no middle neutral or undecided choice. In this situation, the respondent is forced to decide whether they lean more towards the agree or disagree end of the scale for each item.

The final score for the respondent on the scale is the sum of their ratings for all of the items (this is why this is sometimes called a "summated" scale). On some scales, you will have items that are reversed in meaning from the overall direction of the scale. These are called **reversal items**. You will need to reverse the response value for each of these items before summing for the total. That is, if the respondent gave a 1, you make it a 5; if they gave a 2 you make it a 4; 3 = 3; 4 = 2; and, 5 = 1.

Example: The Employment Self Esteem Scale

Here's an example of a ten-item Likert Scale that attempts to estimate the level of self esteem a person has on the job. Notice that this instrument has no center or neutral point -- the respondent has to declare whether he/she is in agreement or disagreement with the item.

INSTRUCTIONS: Please rate how strongly you agree or disagree with each of the following statements by placing a check mark in the appropriate box.

Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	1. I feel good about my work on the job.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	2. On the whole, I get along well with others at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	3. I am proud of my ability to cope with difficulties at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	4. When I feel uncomfortable at work, I know how to handle it.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	5. I can tell that other people at work are glad to have me there.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	6. I know I'll be able to cope with work for as long as I want.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	7. I am proud of my relationship with my supervisor at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	8. I am confident that I can handle my job without constant assistance.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	9. I feel like I make a useful contribution at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	10. I can tell that my coworkers respect me.

- **Guttman Scaling**

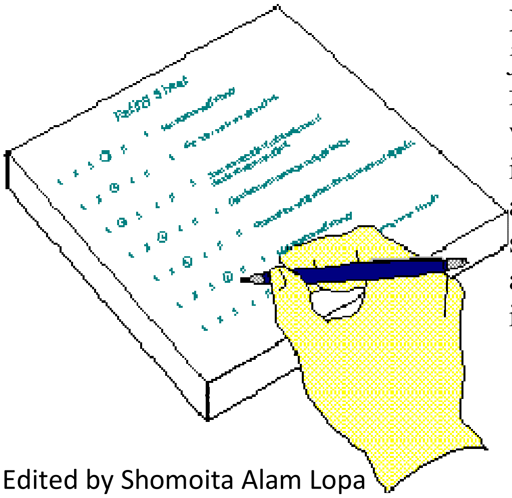
Guttman scaling is also sometimes known as **cumulative scaling** or **scalogram analysis**. The purpose of Guttman scaling is to establish a one-dimensional continuum for a concept you wish to measure. What does that mean? Essentially, we would like a set of items or statements so that a respondent who agrees with any specific question in the list will also agree with all previous questions. Put more formally, we would like to be able to predict item responses perfectly knowing only the total score for the respondent. For example, imagine a ten-item cumulative scale. If the respondent scores a four, it should mean that he/she agreed with the first four statements. If the respondent scores an eight, it should mean they agreed with the first eight. The object is to find a set of items that perfectly matches this pattern. In practice, we would seldom expect to find this cumulative pattern perfectly. So, we use scalogram analysis to examine how closely a set of items corresponds with this idea of cumulativeness. Here, I'll explain how we develop a Guttman scale.

Define the Focus. As in all of the scaling methods. we begin by defining the focus for our scale. Let's imagine that you wish to develop a cumulative scale that measures U.S. citizen attitudes towards immigration. You would want to be sure to specify in your definition whether you are talking about any type of immigration (legal and illegal) from anywhere (Europe, Asia, Latin and South America, Africa).

Develop the Items. Next, as in all scaling methods, you would develop a large set of items that reflect the concept. You might do this yourself or you might engage a knowledgeable group to help. Let's say you came up with the following statements:

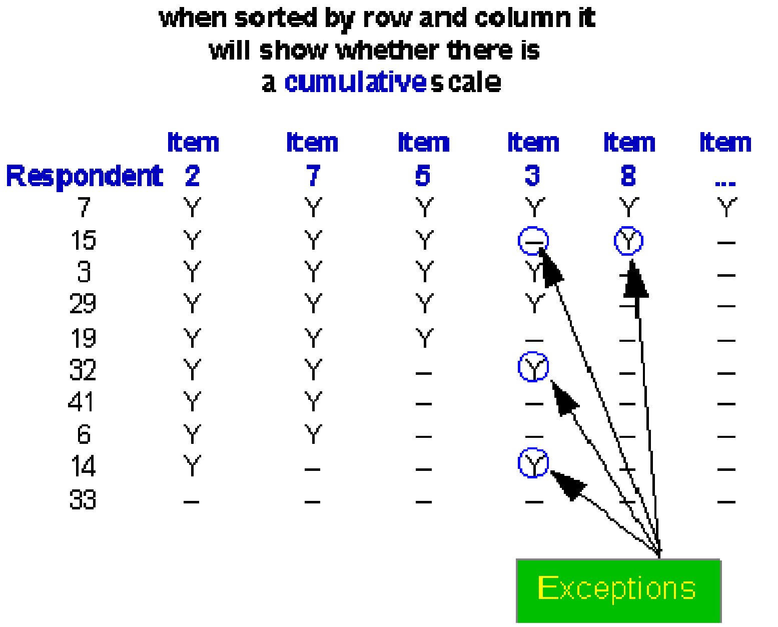
- I would permit a child of mine to marry an immigrant.
- I believe that this country should allow more immigrants in.
- I would be comfortable if a new immigrant moved next door to me.
- I would be comfortable with new immigrants moving into my community.
- It would be fine with me if new immigrants moved onto my block.
- I would be comfortable if my child dated a new immigrant.

Of course, we would want to come up with many more statements (about 80-100 would be desirable).



Rate the Items. Next, we would want to have a group of judges rate the statements or items in terms of how favorable they are to the concept of immigration. They would give a Yes if the item was favorable toward immigration and a No if it is not. Notice that we are not asking the judges whether they personally agree with the statement. Instead, we're asking them to make a judgment about how the statement is related to the construct of interest.

Develop the Cumulative Scale. The key to Guttman scaling is in the analysis. We construct a matrix or table that shows the responses of all the respondents on all of the items. We then sort this matrix so that respondents who agree with more statements are listed at the top and those agreeing with fewer are at the bottom. For respondents with the same number of agreements, we sort the statements from left to right from those that most agreed to to those that fewest agreed to. We might get a table something like the figure. Notice that the scale is very nearly cumulative when you read from left to right across the columns (items). Specifically if someone agreed with Item 7, they always agreed with Item 2. And, if someone agreed with Item 5, they always agreed with Items 7 and 2. The matrix shows that the cumulativeness of the scale is not perfect, however. While in general, a person agreeing with Item 3 tended to also agree with 5, 7 and 2, there are several exceptions to that rule.



when sorted by row and column it will show whether there is a cumulative scale

Respondent	Item 2	Item 7	Item 5	Item 3	Item 8	Item ...
7	Y	Y	Y	Y	Y	Y
15	Y	Y	Y	–	Y	–
3	Y	Y	Y	Y	–	–
29	Y	Y	Y	Y	–	–
19	Y	Y	Y	–	–	–
32	Y	Y	–	Y	–	–
41	Y	Y	–	–	–	–
6	Y	Y	–	–	–	–
14	Y	–	–	Y	–	–
33	–	–	–	–	–	–

Exceptions

While we can examine the matrix if there are only a few items in it, if there are lots of items, we need to use a data analysis called **scalogram analysis** to determine the subsets of items from our pool that best approximate the cumulative property. Then, we review these items and select our final scale elements. There are several statistical techniques for examining the table to find a cumulative scale. Because there is seldom a perfectly cumulative scale we usually have to test how good it is. These statistics also estimate a scale score value for each item. This scale score is used in the final calculation of a respondent's score.

Administering the Scale. Once you've selected the final scale items, it's relatively simple to administer the scale. You simply present the items and ask the respondent to check items with which they agree. For our hypothetical immigration scale, the items might be listed in cumulative order as:

- I believe that this country should allow more immigrants in.
- I would be comfortable with new immigrants moving into my community.
- It would be fine with me if new immigrants moved onto my block.
- I would be comfortable if a new immigrant moved next door to me.
- I would be comfortable if my child dated a new immigrant.
- I would permit a child of mine to marry an immigrant.

Of course, when we give the items to the respondent, we would probably want to mix up the order. Our final scale might look like:

INSTRUCTIONS: Place a check next to each statement you agree with.

_____ I would permit a child of mine to marry an immigrant.

_____ I believe that this country should allow more immigrants in.

_____ I would be comfortable if a new immigrant moved next door to me.

_____ I would be comfortable with new immigrants moving into my community.

_____ It would be fine with me if new immigrants moved onto my block.

_____ I would be comfortable if my child dated a new immigrant.

Each scale item has a scale value associated with it (obtained from the scalogram analysis). To compute a respondent's scale score we simply sum the scale values of every item they agree with. In our example, their final value should be an indication of their attitude towards immigration.

Qualitative Measures

Qualitative research is a vast and complex area of methodology that can easily take up whole textbooks on its own. The purpose of this section is to introduce you to the idea of qualitative research (and how it is related to quantitative research) and give you some orientation to the major types of qualitative research data, approaches and methods.

There are a number of important questions you should consider before undertaking qualitative research:

- **Do you want to generate new theories or hypotheses?**

One of the major reasons for doing qualitative research is to become more experienced with the phenomenon you're interested in. Too often in applied social research (especially in economics and psychology) we have our graduate students jump from doing a literature review on a topic of interest to writing a research proposal complete with theories and hypotheses based on current thinking. What gets missed is the direct experience of the phenomenon. We should probably require of all students that before they mount a study they spend some time living with the phenomenon. Before doing that multivariate analysis of gender-based differences in wages, go observe several work contexts and see how gender tends to be perceived and seems to affect wage allocations. Before looking at the effects of a new psychotropic drug for the mentally ill, go spend some time visiting several mental health treatment contexts to observe what goes on. If you do, you are likely to approach the existing literature on the topic with a fresh perspective born of your direct experience. You're likely to begin to formulate your own ideas about what causes what else to happen. This is where most of the more interesting and valuable new theories and hypotheses will originate. Of course, there's a need for balance here as in anything else. If this advice was followed literally, graduate school would be prolonged even more than is currently the case. We need to use qualitative research as the basis for direct experience, but we also need to know when and how to move on to formulate some tentative theories and hypotheses that can be explicitly tested.

- **Do you need to achieve a deep understanding of the issues?**

I believe that qualitative research has special value for investigating complex and sensitive issues. For example, if you are interested in how people view topics like God and religion, human sexuality, the death penalty, gun control, and so on, my guess is that you would be hard-pressed to develop a quantitative methodology that would do anything more than summarize a few key positions on these issues. While this does have its place (and its done all the time), if you really want to try to achieve a deep understanding of how people think about these topics, some type of in-depth interviewing is probably called for.

- **Are you willing to trade detail for generalizability?**

Qualitative research certainly excels at generating information that is very detailed. Of course, there are quantitative studies that are detailed also in that they involve collecting lots of numeric data. But in detailed quantitative research, the data themselves tend to both shape and limit the analysis. For example, if you collect a simple interval-level quantitative measure, the analyses you are likely to do with it are fairly delimited (e.g., descriptive statistics, use in correlation, regression or multivariate models, etc.). And, generalizing tends to be a fairly straightforward endeavor in most quantitative research. After all, when you collect the same variable from everyone in your sample, all you need to do to generalize to the sample as a whole is to compute some aggregate statistic like a mean or median.

Things are not so simple in most qualitative research. The data are more "raw" and are seldom pre-categorized. Consequently, you need to be prepared to organize all of that raw detail. And there are almost an infinite number of ways this could be accomplished. Even generalizing across a sample of interviews or written documents becomes a complex endeavor.

The detail in most qualitative research is both a blessing and a curse. On the positive side, it enables you to describe the phenomena of interest in great detail, in the original language of the research participants. In fact, some of the best "qualitative" research is often published in book form, often in a style that almost approaches a narrative story. One of my favorite writers (and, I daresay, one of the finest qualitative researchers) is Studs Terkel. He has written intriguing accounts of the Great Depression (Hard Times), World War II (The Good War) and socioeconomic divisions in America (The Great Divide), among others. In each book he follows a similar qualitative methodology, identifying informants who directly experienced the phenomenon in question, interviewing them at length, and then editing the interviews heavily so that they "tell a story" that is different from what any individual interviewee might tell but addresses the question of interest. If you haven't read one of Studs' works yet, I highly recommend them.

On the negative side, when you have that kind of detail, it's hard to determine what the generalizable themes may be. In fact, many qualitative researchers don't even care about generalizing -- they're content to generate rich descriptions of their phenomena.

That's why there is so much value in mixing qualitative research with quantitative. Quantitative research excels at summarizing large amounts of data and reaching generalizations based on statistical projections. Qualitative research excels at "telling the story" from the participant's viewpoint, providing the rich descriptive detail that sets quantitative results into their human context.

- **Is funding available for this research?**

I hate to be crass, but in most social research we do have to worry about how it will get paid for. There is little point in proposing any research that would be unable to be carried out for lack of funds. For qualitative research this is an often especially challenging issue. Because much qualitative research takes an enormous amount of time, is very labor intensive, and yields results that may not be as generalizable for policy-making or decision-making, many funding sources view it as a "frill" or as simply too expensive.

There's a lot that you can (and shouldn't) do in proposing qualitative research that will often enhance its fundability. My pet peeve with qualitative research proposals is when the author says something along these lines (Of course, I'm paraphrasing here. No good qualitative researcher would come out and say something like this directly.):

This study uses an emergent, exploratory, inductive qualitative approach. Because the basis of such an approach is that one does not predetermine or delimit the directions the investigation might take, there is no way to propose specific budgetary or time estimates.

Of course, this is just silly! There is always a way to estimate (especially when we view an estimate as simply an educated guess!). I've reviewed proposals that say almost this kind of thing and let me assure you that I and other reviewers don't judge the researcher's credibility as very high under these circumstances. As an alternative that doesn't hem you in or constrain the methodology, you might reword the same passage something like:

This study uses an emergent, exploratory, inductive qualitative approach. Because the basis of such an approach is that one does not predetermine or delimit the directions the investigation might take, it is especially important to detail the specific stages that this research will follow in addressing the research questions. [Inset detailed description of data collection, coding, analysis, etc. Especially note where there may be iterations of the phases.]. Because of the complexities involved in this type of research, the proposal is divided into several broad stages with funding and time estimates provided for each. [Provide detail].

Notice that the first approach is almost an insult to the reviewer. In the second, the author acknowledges the unpredictability of qualitative research but does as reasonable a job as possible to anticipate the course of the study, its costs, and milestones. Certainly more fundable.

- **The Qualitative Debate**

The Qualitative-Quantitative Debate

There has probably been more energy expended on debating the differences between and relative advantages of qualitative and quantitative methods than almost any other methodological topic in social research. The "qualitative-quantitative debate" as it is sometimes called is one of those hot-button issues that almost invariably will trigger an intense debate in the hotel bar at any social research convention. I've seen friends and colleagues degenerate into academic enemies faster than you can say "last call."

After years of being involved in such verbal brawling, as an observer and direct participant, the only conclusion I've been able to reach is that this debate is "much ado about nothing." To say that one or the other approach is "better" is, in my view, simply a trivializing of what is a far more complex topic than a dichotomous choice can settle. Both quantitative and qualitative research rest on rich and varied traditions that come from multiple disciplines and both have been employed to address almost any research topic you can think of. In fact, in almost every applied social research project I believe there is value in consciously combining both qualitative and quantitative methods in what is referred to as a "mixed methods" approach.

I find it useful when thinking about this debate to distinguish between the general *assumptions* involved in undertaking a research project (qualitative, quantitative or mixed) and the *data* that are collected. At the level of the data, I believe that there is little difference between the qualitative and the quantitative. But at the level of the assumptions that are made, the differences can be profound and irreconcilable (which is why there's so much fighting that goes on).

Qualitative and Quantitative Data

It may seem odd that I would argue that there is little difference between qualitative and quantitative *data*. After all, qualitative data typically consists of words while quantitative data consists of numbers. Aren't these fundamentally different? I don't think so, for the following reasons:

- **All qualitative data can be coded quantitatively.**

What I mean here is very simple. Anything that is qualitative can be assigned meaningful numerical values. These values can then be manipulated to help us achieve greater insight into the meaning of the data and to help us examine specific hypotheses. Let's consider a simple example. Many surveys have one or more short open-ended questions that ask the respondent to supply text responses. The simplest example is probably the "Please add any additional comments" question that is often tacked onto a short survey. The immediate responses are text-based and qualitative. But we can always (and usually will) perform some type of simple classification of the text responses. We might sort the responses into simple categories, for instance. Often, we'll give each category a short label that represents the theme in the response.

What we don't often recognize is that even the simple act of categorizing can be viewed as a quantitative one as well. For instance, let's say that we develop five themes that each respondent could express in their open-ended response. Assume that we have ten respondents. We could easily set up a simple coding table like the one in the figure below to represent the coding of the ten responses into the five themes.

Person	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
1	✓	✓		✓	
2	✓		✓		
3	✓	✓		✓	
4		✓		✓	

5		✓		✓	✓
6	✓	✓			✓
7			✓	✓	✓
8		✓		✓	
9			✓		✓
10				✓	✓

This is a simple qualitative thematic coding analysis. But, we can represent exactly the same information quantitatively as in the following table:

Person	Theme 1	Theme 2	Theme 3	Theme 4	Theme 5	Totals
1	1	1	0	1	0	3
2	1	0	1	0	0	2
3	1	1	0	1	0	3
4	0	1	0	1	0	2
5	0	1	0	1	1	3
6	1	1	0	0	1	3
7	0	0	1	1	1	3
8	0	1	0	1	0	2
9	0	0	1	0	1	2
10	0	0	0	1	1	2
Totals	4	6	3	7	5	

Notice that this is the exact same data. The first would probably be called a qualitative coding while the second is clearly quantitative. The quantitative coding gives us additional useful information and makes it possible to do analyses that we couldn't do with the qualitative coding. For instance, from just the table above we can say that Theme 4 was the most frequently mentioned and that all respondents touched on two or three of the themes. But we can do even more. For instance, we could look at the similarities among the themes based on which respondents addressed them. How? Well, why don't we do a simple correlation matrix for the table above. Here's the result:

	Theme 1	**Theme 2**	**Theme 3**	**Theme 4**
Theme 2	**0.250**			
Theme 3	**-0.089**	**-0.802**		
Theme 4	**-0.356**	**0.356**	**-0.524**	
Theme 5	**-0.408**	**-0.408**	**0.218**	**-0.218**

The analysis shows that Themes 2 and 3 are strongly negatively correlated -- People who said Theme 2 seldom said Theme 3 and vice versa (check it for yourself). We can also look at the similarity among respondents as shown below:

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**
P2	**-0.167**								
P3	**1.000**	**-0.167**							
P4	**0.667**	**-0.667**	**0.667**						
P5	**0.167**	**-1.000**	**0.167**	**0.667**					
P6	**0.167**	**-0.167**	**0.167**	**-0.167**	**0.167**				
P7	**-0.667**	**-0.167**	**-0.667**	**-0.167**	**0.167**	**-0.667**			
P8	**0.667**	**-0.667**	**0.667**	**1.000**	**0.667**	**-0.167**	**-0.167**		
P9	**-1.000**	**0.167**	**-1.000**	**-0.667**	**-0.167**	**-0.167**	**0.667**	**-0.667**	
P10	**-0.167**	**-0.667**	**-0.167**	**0.167**	**0.667**	**-0.167**	**0.667**	**0.167**	**0.167**

We can see immediately that Persons 1 and 3 are perfectly correlated (r = +1.0) as are Persons 4 and 8. There are also a few perfect opposites (r = -1.0) -- P1 and P9, P2 and P5, and P3 and P9.

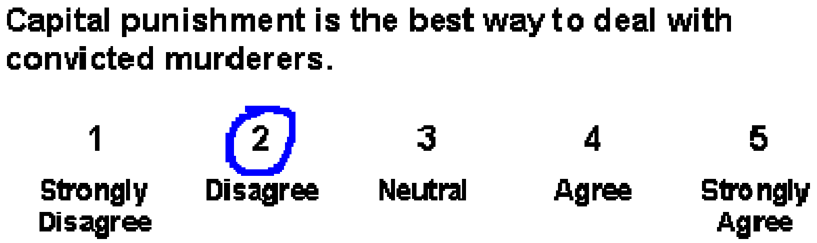
We could do much more. If we had more respondents (and we often would with a survey), we could do some simple multivariate analyses. For instance, we could draw a similarity "map" of the respondents based on their intercorrelations. The map would have one dot per respondent and respondents with more similar responses would cluster closer together.

The point is that the line between qualitative and quantitative is less distinct than we sometimes imagine. All qualitative data can be quantitatively coded in an almost infinite varieties of ways. This doesn't detract from the qualitative information. We can still do any kinds of judgmental syntheses or analyses we want. But recognizing the similarities between qualitative and quantitative information opens up new possibilities for interpretation that might otherwise go unutilized.

Now to the other side of the coin...

- **All quantitative data is based on qualitative judgment.**

Numbers in and of themselves can't be interpreted without understanding the assumptions which underlie them. Take, for example, a simple 1-to-5 rating variable:



Capital punishment is the best way to deal with convicted murderers.

1	[✓ 2]	3	4	5
Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Here, the respondent answered 2=Disagree. What does this mean? How do we interpret the value "2" here? We can't really understand this quantitative value unless we dig into some of the judgments and assumptions that underlie it:

- Did the respondent understand the term "capital punishment"?
- Did the respondent understand that a "2" means that they are disagreeing with the statement?
- Does the respondent have any idea about alternatives to capital punishment (otherwise how can they judge what's "best")?
- Did the respondent read carefully enough to determine that the statement was limited only to convicted murderers (for instance, rapists were not included)?
- Does the respondent care or were they just circling anything arbitrarily?
- How was this question presented in the context of the survey (e.g., did the questions immediately before this one bias the response in any way)?
- Was the respondent mentally alert (especially if this is late in a long survey or the respondent had other things going on earlier in the day)?
- What was the setting for the survey (e.g., lighting, noise and other distractions)?
- Was the survey anonymous? Was it confidential?
- In the respondent's mind, is the difference between a "1" and a "2" the same as between a "2" and a "3" (i.e., is this an interval scale?)?

We could go on and on, but my point should be clear. All numerical information involves numerous judgments about what the number means.

The bottom line here is that quantitative and qualitative data are, at some level, virtually inseparable. Neither exists in a vacuum or can be considered totally devoid of the other. To ask which is "better" or more "valid" or has greater "verisimilitude" or whatever ignores the intimate connection between them. To do good research we need to use both the qualitative and the quantitative.

Qualitative and Quantitative Assumptions

To say that qualitative and quantitative data are similar only tells half the story. After all, the intense academic wrangling of the qualitative-quantitative debate must have some basis in reality. My sense is that there are some fundamental differences, but that they lie primarily at the level of assumptions about research (epistemological and ontological assumptions) rather than at the level of the data.

First, let's do away with the most common myths about the differences between qualitative and quantitative research. Many people believe the following:

- Quantitative research is confirmatory and deductive in nature.
- Qualitative research is exploratory and inductive in nature.

I think that while there's a shred of truth in each of these statements, they are not exactly correct. In general, a lot of quantitative research tends to be confirmatory and deductive. But there's lots of quantitative research that can be classified as exploratory as well. And while much qualitative research does tend to be exploratory, it can also be used to confirm very specific deductive hypotheses. The problem I have with these kinds of statements is that they don't acknowledge the richness of both traditions. They don't recognize that both qualitative and quantitative research can be used to address almost any kind of research question.

So, if the difference between qualitative and quantitative is not along the exploratory-confirmatory or inductive-deductive dimensions, then where is it?

My belief is that the heart of the quantitative-qualitative debate is philosophical, not methodological. Many qualitative researchers operate under different **epistemological assumptions** from quantitative researchers. For instance, many qualitative researchers believe that the best way to understand any phenomenon is to view it in its context. They see all quantification as limited in nature, looking only at one small portion of a reality that cannot be split or unitized without losing the importance of the whole phenomenon. For some qualitative researchers, the best way to understand what's going on is to become immersed in it. Move into the culture or organization you are studying and experience what it is like to be a part of it. Be flexible in your inquiry of people in context. Rather than approaching measurement with the idea of constructing a fixed instrument or set of questions, allow the questions to emerge and change as you become familiar with what you are studying. Many qualitative researchers also operate under different **ontological assumptions** about the world. They don't assume that there is a single unitary reality apart from our perceptions. Since each of us experiences from our own point of view, each of us experiences a different reality. Conducting research without taking this into account violates their fundamental view of the individual. Consequently, they may be opposed to methods that attempt to aggregate across individuals on the grounds that each individual is unique. They also argue that the researcher is a unique individual and that all research is essentially biased by each researcher's individual perceptions. There is no point in trying to establish "validity" in any external or objective sense. All that we can hope to do is interpret our view of the world as researchers.

Let me end this brief excursion into the qualitative-quantitative debate with a few personal observations. Any researcher steeped in the qualitative tradition would certainly take issue with my comments above about the similarities between quantitative and qualitative data. They would argue (with some correctness I fear) that it is not possible to separate your research assumptions from the data. Some would claim that my perspective on data is based on assumptions common to the quantitative tradition. Others would argue that it doesn't matter if you can code data thematically or quantitatively because they wouldn't do *either* -- both forms of analysis impose artificial structure on the phenomena and, consequently, introduce distortions and biases. I have to admit that I would see the point in much of this criticism. In fact, I tend to see the point on both sides of the qualitative-quantitative debate.

In the end, people who consider themselves primarily qualitative or primarily quantitative tend to be almost as diverse as those from the opposing camps. There are qualitative researchers who fit comfortably into the post-positivist tradition common to much contemporary quantitative research. And there are quantitative researchers (albeit, probably fewer) who use quantitative information as the basis for exploration, recognizing the inherent limitations and complex assumptions beneath all numbers. In either camp, you'll find intense and fundamental disagreement about both philosophical assumptions and the nature of data. And, increasingly, we find researchers who are interested in blending the two traditions, attempting to get the advantages of each. I don't think there's any resolution to the debate. And, I believe social research is richer for the wider variety of views and methods that the debate generates.

• Qualitative Data

Qualitative data is extremely varied in nature. It includes virtually any information that can be captured that is not numerical in nature. Here are some of the major categories or types:

• In-Depth Interviews

In-Depth Interviews include both individual interviews (e.g., one-on-one) as well as "group" interviews (including focus groups). The data can be recorded in a wide variety of ways including stenography, audio recording, video recording or written notes. In depth interviews differ from direct observation primarily in the nature of the interaction. In interviews it is assumed that there is a questioner and one or more interviewees. The purpose of the interview is to probe the ideas of the interviewees about the phenomenon of interest.

- **Direct Observation**

Direct observation is meant very broadly here. It differs from interviewing in that the observer does not actively query the respondent. It can include everything from field research where one lives in another context or culture for a period of time to photographs that illustrate some aspect of the phenomenon. The data can be recorded in many of the same ways as interviews (stenography, audio, video) and through pictures, photos or drawings (e.g., those courtroom drawings of witnesses are a form of direct observation).

- **Written Documents**

Usually this refers to existing documents (as opposed transcripts of interviews conducted for the research). It can include newspapers, magazines, books, websites, memos, transcripts of conversations, annual reports, and so on. Usually written documents are analyzed with some form of content analysis.

- **Qualitative Approaches**

A qualitative "approach" is a general way of thinking about conducting qualitative research. It describes, either explicitly or implicitly, the purpose of the qualitative research, the role of the researcher(s), the stages of research, and the method of data analysis. here, four of the major qualitative approaches are introduced.

Ethnography

The ethnographic approach to qualitative research comes largely from the field of anthropology. The emphasis in ethnography is on studying an entire culture. Originally, the idea of a culture was tied to the notion of ethnicity and geographic location (e.g., the culture of the Trobriand Islands), but it has been broadened to include virtually any group or organization. That is, we can study the "culture" of a business or defined group (e.g., a Rotary club).

Ethnography is an extremely broad area with a great variety of practitioners and methods. However, the most common ethnographic approach is participant observation as a part of field research. The ethnographer becomes immersed in the culture as an active participant and records extensive field notes. As in grounded theory, there is no preset limiting of what will be observed and no real ending point in an ethnographic study.

Phenomenology

Phenomenology is sometimes considered a philosophical perspective as well as an approach to qualitative methodology. It has a long history in several social research disciplines including psychology, sociology and social work. Phenomenology is a school of thought that emphasizes a focus on people's subjective experiences and interpretations of the world. That is, the phenomenologist wants to understand how the world appears to others.

Field Research

Field research can also be considered either a broad approach to qualitative research or a method of gathering qualitative data. the essential idea is that the researcher goes "into the field" to observe the phenomenon in its natural state or in situ. As such, it is probably most related to the method of participant observation. The field researcher typically takes extensive field notes which are subsequently coded and analyzed in a variety of ways.

Grounded Theory

Grounded theory is a qualitative research approach that was originally developed by Glaser and Strauss in the 1960s. The self-defined purpose of grounded theory is to develop theory about phenomena of interest. But this is not just abstract theorizing they're talking about. Instead the *theory* needs to be *grounded* or rooted in observation -- hence the term.

Grounded theory is a complex *iterative* process. The research begins with the raising of *generative questions* which help to guide the research but are not intended to be either static or confining. As the researcher begins to gather data, *core theoretical concept(s)* are identified. Tentative *linkages* are developed between the theoretical core concepts and the data. This early phase of the research tends to be very open and can take months. Later on the researcher is more engaged in verification and summary. The effort tends to evolve toward one *core category* that is central.

There are several key analytic strategies:

- *Coding* is a process for both categorizing qualitative data and for describing the implications and details of these categories. Initially one does *open coding*, considering the data in minute detail while developing some initial categories. Later, one moves to more *selective coding* where one systematically codes with respect to a core concept.
- *Memoing* is a process for recording the thoughts and ideas of the researcher as they evolve throughout the study. You might think of memoing as extensive marginal notes and comments. Again, early in the process these memos tend to be very open while later on they tend to increasingly focus in on the core concept.
- *Integrative diagrams and sessions* are used to pull all of the detail together, to help make sense of the data with respect to the emerging theory. The diagrams can be any form of graphic that is useful at that point in theory development. They might be concept maps or directed graphs or even simple cartoons that can act as summarizing devices. This integrative work is best done in group sessions where different members of the research team are able to interact and share ideas to increase insight.

Eventually one approaches *conceptually dense theory* as new observation leads to new linkages which lead to revisions in the theory and more data collection. The core concept or category is identified and fleshed out in detail.

When does this process end? One answer is: never! Clearly, the process described above could continue indefinitely. Grounded theory doesn't have a clearly demarcated point for ending a study. Essentially, the project ends when the researcher decides to quit.

What do you have when you're finished? Presumably you have an extremely well-considered explanation for some phenomenon of interest -- the grounded theory. This theory can be explained in words and is usually presented with much of the contextually relevant detail collected.

- **Qualitative Methods**

There are a wide variety of methods that are common in qualitative measurement. In fact, the methods are largely limited by the imagination of the researcher. Here I discuss a few of the more common methods.

Participant Observation

One of the most common methods for qualitative data collection, participant observation is also one of the most demanding. It requires that the researcher become a participant in the culture or context being observed. The literature on participant observation discusses how to enter the context, the role of the researcher as a participant, the collection and storage of field notes, and the analysis of field data. Participant observation often requires months or years of intensive work because the researcher needs to become accepted as a natural part of the culture in order to assure that the observations are of the natural phenomenon.

Direct Observation

Direct observation is distinguished from participant observation in a number of ways. First, a direct observer doesn't typically try to become a participant in the context. However, the direct observer does strive to be as unobtrusive as possible so as not to bias the observations. Second, direct observation suggests a more detached perspective. The researcher is watching rather than taking part. Consequently, technology can be a useful part of direct observation. For instance, one can videotape the phenomenon or observe from behind one-way mirrors. Third, direct observation tends to be more focused than participant observation. The researcher is observing certain sampled situations or people rather than trying to become immersed in the entire context. Finally, direct observation tends not to take as long as participant observation. For instance, one might observe child-mother interactions under specific circumstances in a laboratory setting from behind a one-way mirror, looking especially for the nonverbal cues being used.

Unstructured Interviewing

Unstructured interviewing involves direct interaction between the researcher and a respondent or group. It differs from traditional structured interviewing in several important ways. First, although the researcher may have some initial guiding questions or core concepts to ask about, there is no formal structured instrument or protocol. Second, the interviewer is free to move the conversation in any direction of interest that may come up. Consequently, unstructured interviewing is particularly useful for exploring a topic broadly. However, there is a price for this lack of structure. Because each interview tends to be unique with no predetermined set of questions asked of all respondents, it is usually more difficult to analyze unstructured interview data, especially when synthesizing across respondents.

Case Studies

A case study is an intensive study of a specific individual or specific context. For instance, Freud developed case studies of several individuals as the basis for the theory of psychoanalysis and Piaget did case studies of children to study developmental phases. There is no single way to conduct a case study, and a combination of methods (e.g., unstructured interviewing, direct observation) can be used.

- **Qualitative Validity**

Depending on their philosophical perspectives, some qualitative researchers reject the framework of validity that is commonly accepted in more quantitative research in the social sciences. They reject the basic realist assumption that their is a reality external to our perception of it. Consequently, it doesn't make sense to be concerned with the "truth" or "falsity" of an observation with respect to an external reality (which is a primary concern of validity). These qualitative researchers argue for different standards for judging the quality of research.

For instance, Guba and Lincoln proposed four criteria for judging the soundness of qualitative research and explicitly offered these as an alternative to more traditional quantitatively-oriented criteria. They felt that their four criteria better reflected the underlying assumptions involved in much qualitative research. Their proposed criteria and the "analogous" quantitative criteria are listed in the table.

Traditional Criteria for Judging Quantitative Research	Alternative Criteria for Judging Qualitative Research
internal validity	credibility
external validity	transferability
reliability	dependability
objectivity	confirmability

Credibility

The credibility criteria involves establishing that the results of qualitative research are credible or believable from the perspective of the participant in the research. Since from this perspective, the purpose of qualitative research is to describe or understand the phenomena of interest from the participant's eyes, the participants are the only ones who can legitimately judge the credibility of the results.

Transferability

Transferability refers to the degree to which the results of qualitative research can be generalized or transferred to other contexts or settings. From a qualitative perspective transferability is primarily the responsibility of the one doing the generalizing. The qualitative researcher can enhance transferability by doing a thorough job of describing the research context and the assumptions that were central to the research. The person who wishes to "transfer" the results to a different context is then responsible for making the judgment of how sensible the transfer is.

Dependability

The traditional quantitative view of reliability is based on the assumption of replicability or repeatability. Essentially it is concerned with whether we would obtain the same results if we could observe the same thing twice. But we can't actually measure the same thing twice -- by definition if we are measuring twice, we are measuring two different things. In order to estimate reliability, quantitative researchers construct various hypothetical notions (e.g., true score theory) to try to get around this fact.

The idea of dependability, on the other hand, emphasizes the need for the researcher to account for the ever-changing context within which research occurs. The research is responsible for describing the changes that occur in the setting and how these changes affected the way the research approached the study.

Confirmability

Qualitative research tends to assume that each researcher brings a unique perspective to the study. Confirmability refers to the degree to which the results could be confirmed or corroborated by others. There are a number of strategies for enhancing confirmability. The researcher can document the procedures for checking and rechecking the data throughout the study. Another researcher can take a "devil's advocate" role with respect to the results, and this process can be documented. The researcher can actively search for and describe and *negative instances* that contradict prior observations. And, after he study, one can conduct a *data audit* that examines the data collection and analysis procedures and makes judgements about the potential for bias or distortion.

There has been considerable debate among methodologists about the value and legitimacy of this alternative set of standards for judging qualitative research. On the one hand, many quantitative researchers see the alternative criteria as just a relabeling of the very successful quantitative criteria in order to accrue greater legitimacy for qualitative research. They suggest that a correct reading of the quantitative criteria would show that they are not limited to quantitative research alone and can be applied equally well to qualitative data. They argue that the alternative criteria represent a different philosophical perspective that is subjectivist rather than realist in nature. They claim that research inherently assumes that there is some reality that is being observed and can be observed with greater or less accuracy or validity. if you don't make this assumption, they would contend, you simply are not engaged in research (although that doesn't mean that what you are doing is not valuable or useful).

Perhaps there is some legitimacy to this counter argument. Certainly a broad reading of the traditional quantitative criteria might make them appropriate to the qualitative realm as well. But historically the traditional quantitative criteria have been described almost exclusively in terms of quantitative research. No one has yet done a thorough job of translating how the same criteria might apply in qualitative research contexts. For instance, the discussions of external validity have been dominated by the idea of statistical sampling as the basis for generalizing. And, considerations of reliability have traditionally been inextricably linked to the notion of true score theory.

But qualitative researchers do have a point about the irrelevance of traditional quantitative criteria. How could we judge the external validity of a qualitative study that does not use formalized sampling methods? And, how can we judge the reliability of qualitative data when there is no mechanism for estimating the true score? No one has adequately explained how the operational procedures used to assess validity and reliability in quantitative research can be translated into legitimate corresponding operations for qualitative research.

While alternative criteria may not in the end be necessary (and I personally hope that more work is done on broadening the "traditional" criteria so that they legitimately apply across the entire spectrum of research approaches), and they certainly can be confusing for students and newcomers to this discussion, these alternatives do serve to remind us that qualitative research cannot easily be considered only an extension of the quantitative paradigm into the realm of nonnumeric data.

Unobtrusive Measures

Unobtrusive measures are measures that don't require the researcher to intrude in the research context. Direct and participant observation require that the researcher be physically present. This can lead the respondents to alter their behavior in order to look good in the eyes of the researcher. A questionnaire is an interruption in the natural stream of behavior. Respondents can get tired of filling out a survey or resentful of the questions asked.

Unobtrusive measurement presumably reduces the biases that result from the intrusion of the researcher or measurement instrument. However, unobtrusive measures reduce the degree the researcher has control over the type of data collected. For some constructs there may simply not be any available unobtrusive measures.

Three types of unobtrusive measurement are discussed here.

Indirect Measures

An indirect measure is an unobtrusive measure that occurs naturally in a research context. The researcher is able to collect the data without introducing any formal measurement procedure.

The types of indirect measures that may be available are limited only by the researcher's imagination and inventiveness. For instance, let's say you would like to measure the popularity of various exhibits in a museum. It may be possible to set up some type of mechanical measurement system that is invisible to the museum patrons. In one study, the system was simple. The museum installed new floor tiles in front of each exhibit they wanted a measurement on and, after a period of time, measured the wear-and-tear of the tiles as an indirect measure of patron traffic and interest. We might be able to improve on this approach considerably using electronic measures. We could, for instance, construct an electrical device that senses movement in front of an exhibit. Or we could place hidden cameras and code patron interest based on videotaped evidence.

One of my favorite indirect measures occurred in a study of radio station listening preferences. Rather than conducting an obtrusive survey or interview about favorite radio stations, the researchers went to local auto dealers and garages and checked all cars that were being serviced to see what station the radio was currently tuned to. In a similar manner, if you want to know magazine preferences, you might rummage through the trash of your sample or even stage a door-to-door magazine recycling effort.

These examples illustrate one of the most important points about indirect measures -- you have to be very careful about the ethics of this type of measurement. In an indirect measure you are, by definition, collecting information without the respondent's knowledge. In doing so, you may be violating their right to privacy and you are certainly not using informed consent. Of course, some types of information may be public and therefore not involve an invasion of privacy.

There may be times when an indirect measure is appropriate, readily available and ethical. Just as with all measurement, however, you should be sure to attempt to estimate the reliability and validity of the measures. For instance, collecting radio station preferences at two different time periods and correlating the results might be useful for assessing test-retest reliability. Or, you can include the indirect measure along with other direct measures of the same construct (perhaps in a pilot study) to help establish construct validity.

Content Analysis

Content analysis is the analysis of text documents. The analysis can be quantitative, qualitative or both. Typically, the major purpose of content analysis is to identify patterns in text. Content analysis is an extremely broad area of research. It includes:

- Thematic analysis of text

The identification of themes or major ideas in a document or set of documents. The documents can be any kind of text including field notes, newspaper articles, technical papers or organizational memos.

- Indexing

There are a wide variety of automated methods for rapidly indexing text documents. For instance, Key Words in Context (KWIC) analysis is a computer analysis of text data. A computer program scans the text and indexes all key words. A key word is any term in the text that is not included in an exception dictionary. Typically you would set up an exception dictionary that includes all non-essential words like "is", "and", and "of". All key words are alphabetized and are listed with the text that precedes and follows it so the researcher can see the word in the context in which it occurred in the text. In an analysis of interview text, for instance, one could easily identify all uses of the term "abuse" and the context in which they were used.

- Quantitative descriptive analysis

Here the purpose is to describe features of the text quantitatively. For instance, you might want to find out which words or phrases were used most frequently in the text. Again, this type of analysis is most often done directly with computer programs.

Content analysis has several problems you should keep in mind. First, you are limited to the types of information available in text form. If you are studying the way a news story is being handled by the news media, you probably would have a ready population of news stories from which you could sample. However, if you are interested in studying people's views on capital punishment, you are less likely to find an archive of text documents that would be appropriate. Second, you have to be especially careful with sampling in order to avoid bias. For instance, a study of current research on methods of treatment for cancer might use the published literature as the population. This would leave out both the writing on cancer that did not get published for one reason or another as well as the most recent work that has not yet been published. Finally, you have to be careful about interpreting results of automated content analyses. A computer program cannot determine what someone meant by a term or phrase. It is relatively easy in a large analysis to misinterpret a result because you did not take into account the subtleties of meaning.

However, content analysis has the advantage of being unobtrusive and, depending on whether automated methods exist, can be a relatively rapid method for analyzing large amounts of text.

Secondary Analysis of Data

Secondary analysis, like content analysis, makes use of already existing sources of data. However, secondary analysis typically refers to the re-analysis of quantitative data rather than text.

In our modern world there is an unbelievable mass of data that is routinely collected by governments, businesses, schools, and other organizations. Much of this information is stored in electronic databases that can be accessed and analyzed. In addition, many research projects store their raw data in electronic form in computer archives so that others can also analyze the data. Among the data available for secondary analysis is:

- census bureau data
- crime records
- standardized testing data
- economic data
- consumer data

Secondary analysis often involves combining information from multiple databases to examine research questions. For example, you might join crime data with census information to assess patterns in criminal behavior by geographic location and group.

Secondary analysis has several advantages. First, it is efficient. It makes use of data that were already collected by someone else. It is the research equivalent of recycling. Second, it often allows you to extend the scope of your study considerably. In many small research projects it is impossible to consider taking a national sample because of the costs involved. Many archived databases are already national in scope and, by using them, you can leverage a relatively small budget into a much broader study than if you collected the data yourself.

However, secondary analysis is not without difficulties. Frequently it is no trivial matter to access and link data from large complex databases. Often the researcher has to make assumptions about what data to combine and which variables are appropriately aggregated into indexes. Perhaps more importantly, when you use data collected by others you often don't know what problems occurred in the original data collection. Large, well-financed national studies are usually documented quite thoroughly, but even detailed documentation of procedures is often no substitute for direct experience collecting data.

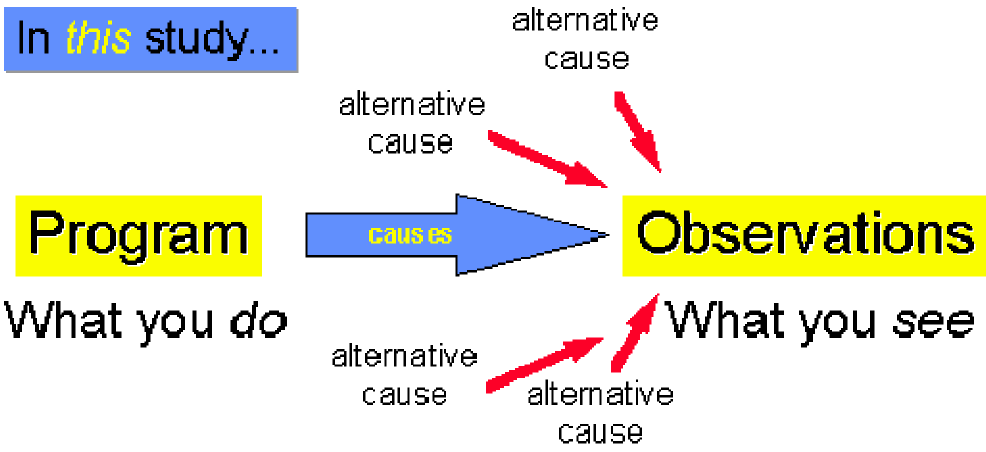
One of the most important and least utilized purposes of secondary analysis is to replicate prior research findings. In any original data analysis there is the potential for errors. In addition, each data analyst tends to approach the analysis from their own perspective using analytic tools they are familiar with. In most research the data are analyzed only once by the original research team. It seems an awful waste. Data that might have taken months or years to collect is only examined once in a relatively brief way and from one analyst's perspective. In social research we generally do a terrible job of documenting and archiving the data from individual studies and making these available in electronic form for others to re-analyze. And, we tend to give little professional credit to studies that are re-analyses. Nevertheless, in the hard sciences the tradition of replicability of results is a critical one and we in the applied social sciences could benefit by directing more of our efforts to secondary analysis of existing data.

Design

Research design provides the glue that holds the research project together. A design is used to structure the research, to show how all of the major parts of the research project -- the samples or groups, measures, treatments or programs, and methods of assignment -- work together to try to address the central research questions. Here, after a brief introduction to research design, I'll show you how we classify the major types of designs. You'll see that a major distinction is between the experimental designs that use random assignment to groups or programs and the quasi-experimental designs that don't use random assignment. [People often confuse what is meant by random selection with the idea of random assignment. You should make sure that you understand the distinction between random selection and random assignment.] Understanding the relationships among designs is important in making design choices and thinking about the strengths and weaknesses of different designs. Then, I'll talk about the heart of the art form of designing designs for research and give you some ideas about how you can think about the design task. Finally, I'll consider some of the more recent advances in quasi-experimental thinking -- an area of special importance in applied social research and program evaluation.

Internal Validity

Internal Validity is the approximate truth about inferences regarding cause-effect or causal relationships. Thus, internal validity is only relevant in studies that try to establish a causal relationship. It's not relevant in most observational or descriptive studies, for instance. But for studies that assess the effects of social programs or interventions, internal validity is perhaps the primary consideration. In those contexts, you would like to be able to conclude that your program or treatment made a difference -- it improved test scores or reduced symptomology. But there may be lots of reasons, other than your program, why test scores may improve or symptoms may reduce. The key question in internal validity is whether observed changes can be attributed to your program or intervention (i.e., the cause) and **not** to other possible causes (sometimes described as "alternative explanations" for the outcome).



One of the things that's most difficult to grasp about internal validity is that it is only relevant to the specific study in question. That is, you can think of internal validity as a "zero generalizability" concern. All that internal validity means is that you have evidence that what you did in the study (i.e., the program) caused what you observed (i.e., the outcome) to happen. It doesn't tell you whether what you did for the program was what you wanted to do or whether what you observed was what you wanted to observe -- those are construct validity concerns. It is possible to have internal validity in a study and not have construct validity. For instance, imagine a study where you are looking at the effects of a new computerized tutoring program on math performance in first grade students. Imagine that the tutoring is unique in that it has a heavy computer game component and you think that's what will really work to improve math performance. Finally, imagine that you were wrong (hard, isn't it?) -- it turns out that math performance did improve, and that it was because of something you did, but that it had nothing to do with the computer program. What caused the improvement was the individual attention that the adult tutor gave to the child -- the computer program didn't make any difference. This study would have internal validity because something that you did affected something that you observed -- you did cause *something* to happen. But the study would not have construct validity, specifically, the label "computer math program" does not accurately describe the actual cause (perhaps better described as "personal adult attention").

Since the key issue in internal validity is the causal one, we'll begin by considering what conditions need to be met in order to establish a causal relationship in your project. Then we'll consider the different threats to internal validity -- the kinds of criticisms your critics will raise when you try to conclude that your program caused the outcome. For convenience, we divide the threats to validity into three categories. The first involve the single group threats -- criticisms that apply when you are only studying a single group that receives your program. The second consists of the multiple group threats -- criticisms that are likely to be raised when you have several groups in your study (e.g., a program and a comparison group). Finally, we'll consider what I call the social threats to internal validity -- threats that arise because social research is conducted in real-world human contexts where people will react to not only what affects them, but also to what is happening to others around them.

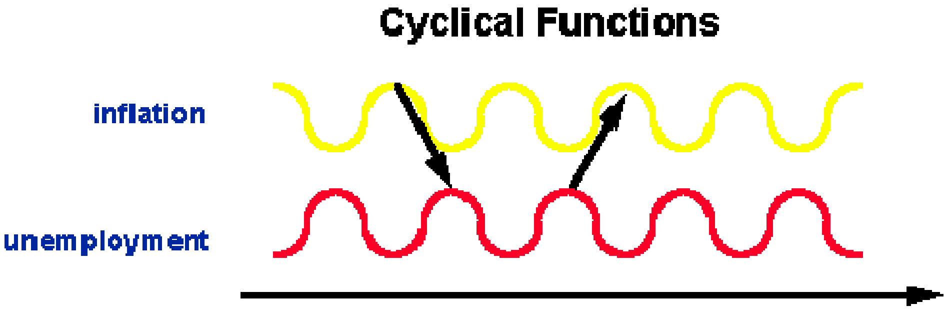
- **Establishing Cause & Effect**

Establishing a Cause-Effect Relationship

How do we establish a cause-effect (causal) relationship? What criteria do we have to meet? Generally, there are three criteria that you must meet before you can say that you have evidence for a causal relationship:

- **Temporal Precedence**

First, you have to be able to show that your cause happened ***before*** your effect. Sounds easy, huh? Of course my cause has to happen before the effect. Did you ever hear of an effect happening before its cause? Before we get lost in the logic here, consider a classic example from economics: does inflation cause unemployment? It certainly seems plausible that as inflation increases, more employers find that in order to meet costs they have to lay off employees. So it seems that inflation could, at least partially, be a cause for unemployment. But both inflation and employment rates are occurring together on an ongoing basis. Is it possible that fluctuations in employment can affect inflation? If we have an increase in the work force (i.e., lower unemployment) we may have more demand for goods, which would tend to drive up the prices (i.e., inflate them) at least until supply can catch up. So which is the cause and which the effect, inflation or unemployment? It turns out that in this kind of cyclical situation involving ongoing processes that interact that both may cause and, in turn, be affected by the other. This makes it very hard to establish a causal relationship in this situation.



- **Covariation of the Cause and Effect**

What does this mean? Before you can show that you have a *causal* relationship you have to show that you have some type of relationship. For instance, consider the syllogism:

if X then Y
if not X then not Y

If you observe that whenever X is present, Y is also present, and whenever X is absent, Y is too, then you have demonstrated that there is a relationship between X and Y. I don't know about you, but sometimes I find it's not easy to think about X's and Y's. Let's put this same syllogism in program evaluation terms:

if program then outcome
if not program then not outcome

Or, in colloquial terms: if you give a program you observe the outcome but if you don't give the program you don't observe the outcome. This provides evidence that the program and outcome are related. Notice, however, that this syllogism doesn't not provide evidence that the program caused the outcome -- perhaps there was some other factor present with the program that caused the outcome, rather than the program. The relationships described so far are rather simple binary relationships. Sometimes we want to know whether different amounts of the program lead to different amounts of the outcome -- a continuous relationship:

if more of the program then more of the outcome
if less of the program then less of the outcome

- **No Plausible Alternative Explanations**

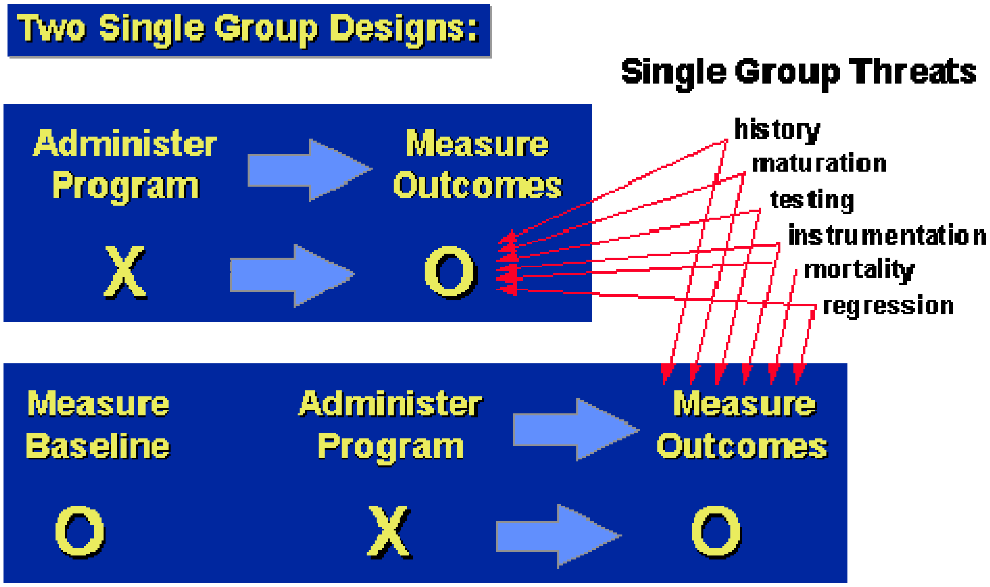
  Just because you show there's a relationship doesn't mean it's a causal one. It's possible that there is some other variable or factor that is causing the outcome. This is sometimes referred to as the "third variable" or "missing variable" problem and it's at the heart of the issue of internal validity. What are some of the possible plausible alternative explanations? Just go look at the threats to internal validity (see single group threats, multiple group threats or social threats) -- each one describes a type of alternative explanation.

  In order for you to argue that you have demonstrated internal validity -- that you have shown there's a causal relationship -- you have to "rule out" the plausible alternative explanations. How do you do that? One of the major ways is with your research design. Let's consider a simple single group threat to internal validity, a *history* threat. Let's assume you measure your program group before they start the program (to establish a baseline), you give them the program, and then you measure their performance afterwards in a posttest. You see a marked improvement in their performance which you would like to infer is caused by your program. One of the plausible alternative explanations is that you have a history threat -- it's not your program that caused the gain but some other specific historical event. For instance, it's not your anti-smoking campaign that caused the reduction in smoking but rather the Surgeon General's latest report that happened to be issued between the time you gave your pretest and posttest. How do you rule this out with your research design? One of the simplest ways would be to incorporate the use of a control group -- a group that is comparable to your program group with the only difference being that they didn't receive the program. But they did experience the Surgeon General's latest report. If you find that they didn't show a reduction in smoking even though they did experience the same Surgeon General report you have effectively "ruled out" the Surgeon General's report as a plausible alternative explanation for why you observed the smoking reduction.

In most applied social research that involves evaluating programs, temporal precedence is not a difficult criterion to meet because you administer the program before you measure effects. And, establishing covariation is relatively simple because you have some control over the program and can set things up so that you have some people who get it and some who don't (if X and if not X). Typically the most difficult criterion to meet is the third -- ruling out alternative explanations for the observed effect. That is why research design is such an important issue and why it is intimately linked to the idea of internal validity.

- **Single Group Threats**

The Single Group Case



What is meant by a "single group" threat? Let's consider two single group designs and then consider the threats that are most relevant with respect to internal validity. The top design in the figure shows a "posttest-only" single group design. Here, a group of people receives your program and afterwards is given a posttest. In the bottom part of the figure we see a "pretest-posttest" single group design. In this case, we give the participants a pretest or baseline measure, give them the program or treatment, and then give them a posttest.

To help make this a bit more concrete, let's imagine that we are studying the effects of a compensatory education program in mathematics for first grade students on a measure of math performance such as a standardized math achievement test. In the post-only design, we would give the first graders the program and then give a math achievement posttest. We might choose not to give them a baseline measure because we have reason to believe they have no prior knowledge of the math skills we are teaching. It wouldn't make sense to pretest them if we expect they would all get a score of zero. In the pre-post design we are not willing to assume that they have no prior knowledge. We measure the baseline in order to determine where the students start out in math achievement. We might hypothesize that the change or gain from pretest to posttest is due to our special math tutoring program. This is a *compensatory* program because it is only given to students who are identified as potentially low in math ability on the basis of some screening mechanism.

The Single Group Threats

With either of these scenarios in mind, consider what would happen if you observe a certain level of posttest math achievement or a change or gain from pretest to posttest. You want to conclude that the outcome is due to your math program. How could you be wrong? Here are some of the ways, some of the threats to interval validity that your critics might raise, some of the plausible alternative explanations for your observed effect:

- **History Threat**

It's not your math program that caused the outcome, it's something else, some historical event that occurred. For instance, we know that lot's of first graders watch the public TV program *Sesame Street*. And, we know that in every *Sesame Street* show they present some very elementary math concepts. Perhaps these shows cause the outcome and not your math program.

- **Maturation Threat**

The children would have had the exact same outcome even if they had never had your special math training program. All you are doing is measuring normal maturation or growth in understanding that occurs as part of growing up -- your math program has no effect. How is this maturation explanation different from a history threat? In general, if we're talking about a specific event or chain of events that could cause the outcome, we call it a history threat. If we're talking about all of the events that typically transpire in your life over a period of time (without being specific as to which ones are the active causal agents) we call it a maturation threat.

- **Testing Threat**

This threat only occurs in the pre-post design. What if taking the pretest made some of the children more aware of that kind of math problem -- it "primed" them for the program so that when you began the math training they were ready for it in a way that they wouldn't have been without the pretest. This is what is meant by a testing threat -- taking the pretest (not getting your program) affects how participants do on the posttest.

- **Instrumentation Threat**

Like the testing threat, this one only operates in the pretest-posttest situation. What if the change from pretest to posttest is due not to your math program but rather to a change in the test that was used? This is what's meant by an instrumentation threat. In many schools when they have to administer repeated testing they don't use the exact same test (in part because they're worried about a testing threat!) but rather give out "alternate forms" of the same tests. These alternate forms were designed to be "equivalent" in the types of questions and level of difficulty, but what if they aren't? Perhaps part or all of any pre-post gain is attributable to the change in instrument, not to your program. Instrumentation threats are especially likely when the "instrument" is a human observer. The observers may get tired over time or bored with the observations. Conversely, they might get better at making the observations as they practice more. In either event, it's the change in instrumentation, not the program, that leads to the outcome.

- **Mortality Threat**

Mortality doesn't mean that people in your study are dying (although if they are, it would be considered a mortality threat!). Mortality is used metaphorically here. It means that people are "dying" with respect to your study. Usually, it means that they are dropping out of the study. What's wrong with that? Let's assume that in our compensatory math tutoring program we have a nontrivial dropout rate between pretest and posttest. And, assume that the kids who are dropping out are the low pretest math achievement test scorers. If you look at the average gain from pretest to posttest using all of the scores available to you at each occasion, you would include these low pretest subsequent dropouts in the pretest and not in the posttest. You'd be dropping out the potential low scorers from the posttest, or, you'd be artificially inflating the posttest average over what it would have been if no students had dropped out. And, you won't necessarily solve this problem by comparing pre-post averages for only those kids who stayed in the study. This subsample would certainly not be representative even of the original entire sample. Furthermore, we know that because of regression threats (see below) these students may appear to actually do worse on the posttest, simply as an artifact of the non-random dropout or mortality in your study. When mortality is a threat, the researcher can often gauge the degree of the threat by comparing the dropout group against the nondropout group on *pretest* measures. If there are no major differences, it may be more reasonable to assume that mortality was happening across the entire sample and is not biasing results greatly. But if the pretest differences are large, one must be concerned about the potential biasing effects of mortality.

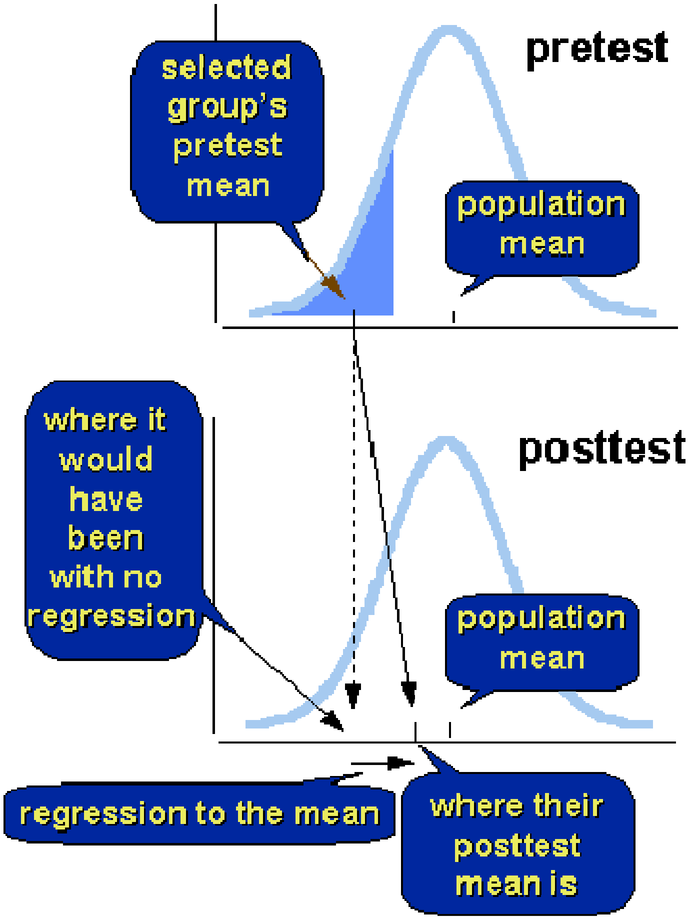
- **Regression Threat**

A regression threat, also known as a "regression artifact" or "regression to the mean" is a statistical phenomenon that occurs whenever you have a nonrandom sample from a population and two measures that are imperfectly correlated. OK, I know that's gibberish. Let me try again. Assume that your two measures are a pretest and posttest (and you can certainly bet these aren't perfectly correlated with each other). Furthermore, assume that your sample consists of low pretest scorers. The regression threat means that the pretest average for the group in your study will appear to increase or improve (relatively to the overall population) even if you don't do anything to them -- even if you never give them a treatment. Regression is a confusing threat to understand at first. I like to think about it as the "*you can only go up from here*" phenomenon. If you include in your program only the kids who constituted the lowest ten percent of the class on the pretest, what are the chances that they would constitute exactly the lowest ten percent on the posttest? Not likely. Most of them would score low on the posttest, but they aren't likely to be the lowest ten percent twice. For instance, maybe there were a few kids on the pretest who got lucky on a few guesses and scored at the eleventh percentile who won't get so lucky next time. No, if you choose the lowest ten percent on the pretest, they can't get any lower than being the lowest -- they can only go up from there, relative to the larger population from which they were selected. This purely statistical phenomenon is what we mean by a regression threat. To see a more detailed discussion of why regression threats occur and how to estimate them, click ***here***.

How do we deal with these single group threats to internal validity? While there are several ways to rule out threats, one of the most common approaches to ruling out the ones listed above is through your research design. For instance, instead of doing a single group study, you could incorporate a control group. In this scenario, you would have two groups: one receives your program and the other one doesn't. In fact, the only difference between these groups should be the program. If that's true, then the control group would experience all the same history and maturation threats, would have the same testing and instrumentation issues, and would have similar rates of mortality and regression to the mean. In other words, a good control group is one of the most effective ways to rule out the single-group threats to internal validity. Of course, when you add a control group, you no-longer have a single group design. And, you will still have to deal with threats two major types of threats to internal validity: the multiple-group threats to internal validity and the social threats to internal validity.

❖ Regression to the Mean

A regression threat, also known as a "regression artifact" or "regression to the mean" is a statistical phenomenon that occurs whenever you have a nonrandom sample from a population and two measures that are imperfectly correlated. The figure shows the regression to the mean phenomenon. The top part of the figure shows the pretest distribution for a population. Pretest scores are "normally" distributed, the frequency distribution looks like a "bell-shaped" curve. Assume that the sample for your study was selected exclusively from the low pretest scorers. You can see on the top part of the figure where their pretest mean is -- clearly, it is considerably below the population average. What would we predict the posttest to look like? First, let's assume that your program or treatment doesn't work at all (the "null" case). Our naive assumption would be that our sample would score just as badly on the posttest as they did on the pretest. But they don't! The bottom of the figure shows where the sample's posttest mean would have been without regression and where it actually is. In actuality, the sample's posttest mean wound up closer to the posttest population mean than their pretest mean was to the pretest population mean. In other words, the sample's mean appears to *regress toward the mean* of the population from pretest to posttest.


pretest
selected group's pretest mean
population mean
posttest
where it would have been with no regression
population mean
regression to the mean
where their posttest mean is

Why Does It Happen?

Let's start with a simple explanation and work from there. To see why regression to the mean happens, consider a concrete case. In your study you select the lowest 10% of the population based on their pretest score. What are the chances that on the posttest that exact group will once again constitute the lowest ten percent? Not likely. Most of them will probably be in the lowest ten percent on the posttest, but if even just a few are not, then their group's mean will have to be closer to the population's posttest than it was to the pretest. The same thing is true on the other end. If you select as your sample the highest ten percent pretest scorers, they aren't likely to be the highest ten percent on the posttest (even though most of them may be in the top ten percent). If even just a few score below the top ten percent on the posttest their group's posttest mean will have to be closer to the population posttest mean than to their pretest mean.

Here are a few things you need to know about the regression to the mean phenomenon:

- **It is a *statistical* phenomenon.**

Regression toward the mean occurs for two reasons. First, it results because you asymmetrically sampled from the population. If you randomly sample from the population, you would observe (subject to random error) that the population and your sample have the same pretest average. Because the sample is already at the population mean on the pretest, it is impossible for them to regress towards the mean of the population any more!

- **It is a *group* phenomenon.**

You cannot tell which way an individual's score will move based on the regression to the mean phenomenon. Even though the group's average will move toward the population's, some individuals in the group are likely to move in the other direction.

- **It happens between *any two variables.***

Here's a common research mistake. You run a program and don't find any overall group effect. So, you decide to look at those who did best on the posttest (your "success" stories!?) and see how much they gained over the pretest. You are selecting a group that is extremely high on the posttest. They won't likely all be the best on the pretest as well (although many of them will be). So, their pretest mean has to be closer to the population mean than their posttest one. You describe this nice "gain" and are almost ready to write up your results when someone suggests you look at your "failure" cases, the people who score worst on your posttest. When you check on how they were doing on the pretest you find that they weren't the worst scorers there. If they had been the worst scorers both times, you would have simply said that your program didn't have any effect on them. But now it looks worse than that -- it looks like your program actually made them worse relative to the population! What will you do? How will you ever get your grant renewed? Or your paper published? Or, heaven help you, how will you ever get tenured?

What you have to realize, is that the pattern of results I just described will happen anytime you measure two measures! It will happen forwards in time (i.e., from pretest to posttest). It will happen backwards in time (i.e., from posttest to pretest)! It will happen across measures collected at the same time (e.g., height and weight)! It will happen even if you don't give your program or treatment.

- **It is a *relative* phenomenon.**

It has nothing to do with overall maturational trends. Notice in the figure above that I didn't bother labeling the x-axis in either the pretest or posttest distribution. It could be that everyone in the population gains 20 points (on average) between the pretest and the posttest. But regression to the mean would still be operating, even in that case. That is, the low scorers would, on average, be gaining more than the population gain of 20 points (and thus their mean would be closer to the population's).

- **You can have regression up or down.**

If your sample consists of below-population-mean scorers, the regression to the mean will make it appear that they move ***up*** on the other measure. But if your sample consists of high scorers, their mean will appear to move ***down*** relative to the population. (Note that even if their mean increases, they could be losing ground to the population. So, if a high-pretest-scoring sample gains five points on the posttest while the overall sample gains 15, we would suspect regression to the mean as an alternative explanation [to our program] for that relatively low change).

- **The more extreme the sample group, the greater the regression to the mean.**

If your sample differs from the population by only a little bit on the first measure, their won't be much regression to the mean because there isn't much room for them to regress -- they're already near the population mean. So, if you have a sample, even a nonrandom one, that is a pretty good subsample of the population, regression to the mean will be inconsequential (although it will be present). But if your sample is very extreme relative to the population (e.g., the lowest or highest x%), their mean is further from the population's and has more room to regress.

- **The less correlated the two variables, the greater the regression to the mean.**

The other major factor that affects the amount of regression to the mean is the correlation between the two variables. If the two variables are *perfectly* correlated -- the highest scorer on one is the highest on the other, next highest on one is next highest on the other, and so on -- there will no be regression to the mean. But this is unlikely to ever occur in practice. We know from measurement theory that there is no such thing as "perfect" measurement -- all measurement is assumed (under the true score model) to have some random error in measurement. It is only when the measure has no random error -- is perfectly reliable -- that we can expect it will be able to correlate perfectly. Since that just doesn't happen in the real world, we have to assume that measures have some degree of unreliability, and that relationships between measures will not be perfect, and that there will appear to be regression to the mean between these two measures, given asymmetrically sampled subgroups.

The Formula for the Percent of Regression to the Mean

You can estimate exactly the percent of regression to the mean in any given situation. The formula is:

$$\mathbf{P_{rm} = 100(1 - r)}$$

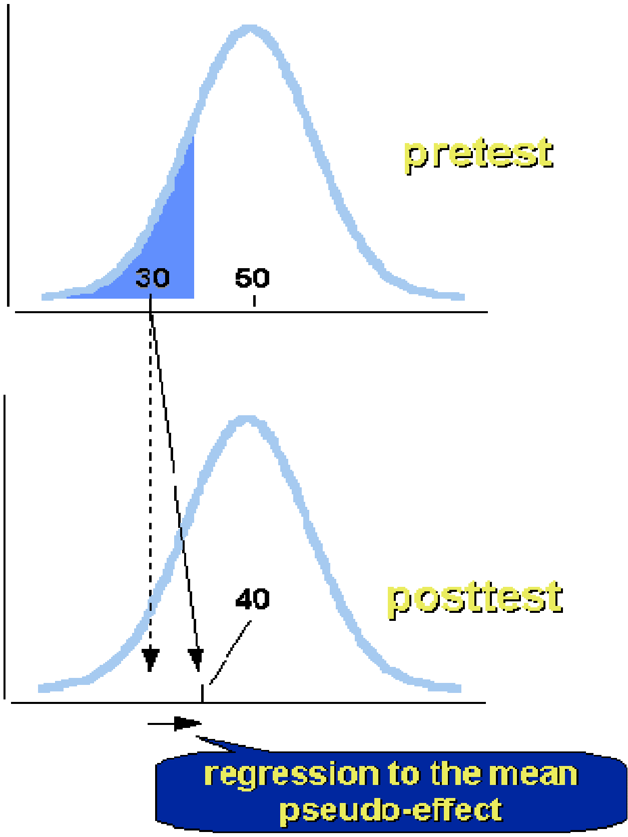
where:

P_{rm} = the percent of regression to the mean
r = the correlation between the two measures

Consider the following four cases:

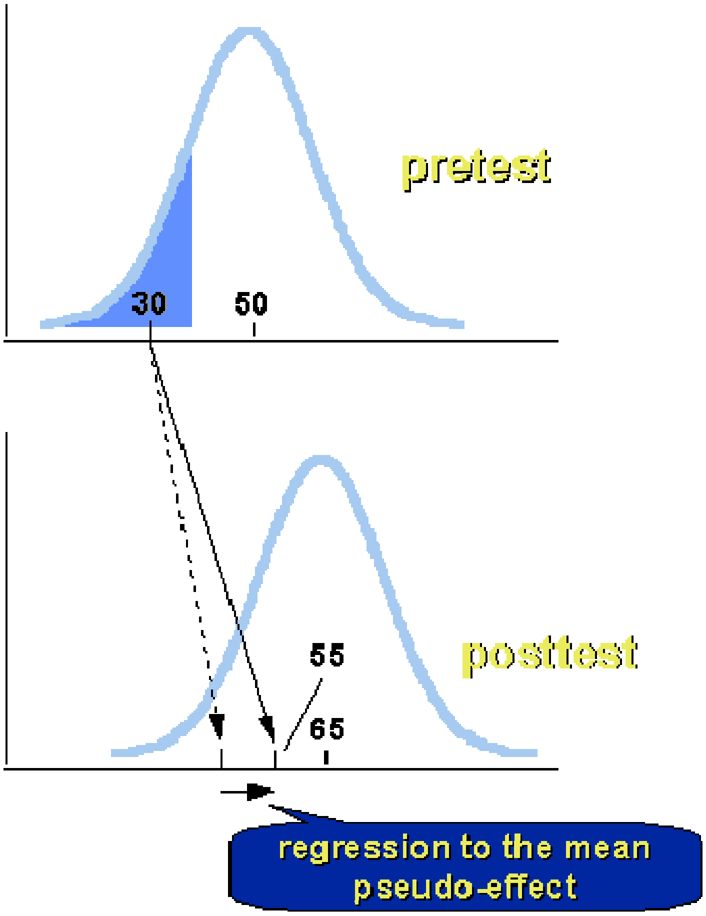
- if r = 1, there is no (i.e., 0%) regression to the mean
- if r = .5, there is 50% regression to the mean
- if r = .2, there is 80% regression to the mean
- if r = 0, there is 100% regression to the mean

In the first case, the two variables are perfectly correlated and there is no regression to the mean. With a correlation of .5, the sampled group moves ***fifty percent*** of the distance from the no-regression point to the mean of the population. If the correlation is a small .20, the sample will regress 80% of the distance. And, if there is no correlation between the measures, the sample will "regress" all the way back to the population mean! It's worth thinking about what this last case means. With zero correlation, knowing a score on one measure gives you absolutely no information about the likely score for that person on the other measure. In that case, your best guess for how any person would perform on the second measure will be the mean of that second measure.



Estimating and Correcting Regression to the Mean

Given our percentage formula, for any given situation we can estimate the regression to the mean. All we need to know is the mean of the sample on the first measure the population mean on both measures, and the correlation between measures. Consider a simple example. Here, we'll assume that the pretest population mean is 50 and that we select a low-pretest scoring sample that has a mean of 30. To begin with, let's assume that we do not give any program or treatment (i.e., the null case) and that the population is not changing over time on the characteristic being measured (i.e., steady-state). Given this, we would predict that the population mean would be 50 and that the sample would get a posttest score of 30 *if there was no regression to the mean*. Now, assume that the correlation is .50 between the pretest and posttest for the population. Given our formula, we would expect that the sampled group would regress 50% of the distance from the no-regression point to the population mean, or 50% of the way from 30 to 50. In this case, we would observe a score of 40 for the sampled group, which would constitute a 10-point pseudo-effect or regression artifact.



Now, let's relax some of the initial assumptions. For instance, let's assume that between the pretest and posttest the population gained 15 points on average (and that this gain was uniform across the entire distribution, that is, the variance of the population stays the same across the two measurement occasions). In this case, a sample that had a pretest mean of 30 would be expected to get a posttest mean of 45 (i.e., 30+15) if there is no regression to the mean (i.e., r=1). But here, the correlation between pretest and posttest is .5 so we expect to see regression to the mean that covers 50% of the distance from the mean of 45 to the population posttest mean of 65. That is, we would observe a posttest average of 55 for our sample, again a pseudo-effect of 10 points.

Regression to the mean is one of the trickiest threats to validity. It is subtle in its effects, and even excellent researchers sometimes fail to catch a potential regression artifact. You might want to learn more about the regression to the mean phenomenon. One good way to do that would be to simulate the phenomenon.

- **Multiple Group Threats**

The Central Issue

A multiple-group design typically involves at least two groups and before-after measurement. Most often, one group receives the program or treatment while the other does not and constitutes the "control" or comparison group. But sometimes one group gets the program and the other gets either the standard program or another program you would like to compare. In this case, you would be comparing two programs for their relative outcomes. Typically you would construct a multiple-group design so that you could compare the groups directly. In such designs, the key internal validity issue is the degree to which the groups are comparable before the study. If they are comparable, and the only difference between them is the program, posttest differences can be attributed to the program. But that's a big *if*. If the groups aren't comparable to begin with, you won't know how much of the outcome to attribute to your program or to the initial differences between groups.

There really is only one multiple group threat to internal validity: that the groups were not comparable before the study. We call this threat a ***selection bias*** or ***selection threat***. A selection threat is *any* factor other than the program that leads to posttest differences between groups. Whenever we suspect that outcomes differ between groups not because of our program but because of prior group differences we are suspecting a selection bias. Although the term 'selection bias' is used as the general category for all prior differences, when we know specifically what the group difference is, we usually hyphenate it with the 'selection' term. The multiple-group selection threats directly parallel the single group threats. For instance, while we have 'history' as a single group threat, we have 'selection-history' as its multiple-group analogue.

As with the single group threats to internal validity, we'll assume a simple example involving a new compensatory mathematics tutoring program for first graders. The design will be a pretest-posttest design, and we will divide the first graders into two groups, one getting the new tutoring program and the other not getting it.

Here are the major multiple-group threats to internal validity for this case:

- **Selection-History Threat**

A selection-history threat is any other event that occurs between pretest and posttest that the groups experience differently. Because this is a selection threat, it means the groups differ in some way. Because it's a 'history' threat, it means that the way the groups differ is with respect to their reactions to history events. For example, what if the children in one group differ from those in the other in their television habits. Perhaps the program group children watch Sesame Street more frequently than those in the control group do. Since Sesame Street is a children's show that presents simple mathematical concepts in interesting ways, it may be that a higher average posttest math score for the program group doesn't indicate the effect of our math tutoring -- it's really an effect of the two groups differentially experiencing a relevant event -- in this case Sesame Street -- between the pretest and posttest.

- **Selection-Maturation Threat**

A selection-maturation threat results from differential rates of normal growth between pretest and posttest for the groups. In this case, the two groups are different in their different rates of maturation with respect to math concepts. It's important to distinguish between history and maturation threats. In general, history refers to a discrete event or series of events whereas maturation implies the normal, ongoing developmental process that would take place. In any case, if the groups are maturing at different rates with respect to the outcome, we cannot assume that posttest differences are due to our program -- they may be selection-maturation effects.

- **Selection-Testing Threat**

A selection-testing threat occurs when there is a *differential* effect between groups on the posttest of taking the pretest. Perhaps the test "primed" the children in each group differently or they may have learned differentially from the pretest. in these cases, an observed posttest difference can't be attributed to the program, they could be the result of selection-testing.

- **Selection-Instrumentation Threat**

Selection-instrumentation refers to any *differential* change in the test used for each group from pretest and posttest. In other words, the test changes differently for the two groups. Perhaps the test consists of observers who rate the class performance of the children. What if the program group observers for example, get better at doing the observations while, over time, the comparison group observers get fatigued and bored. Differences on the posttest could easily be due to this differential instrumentation -- selection-instrumentation -- and not to the program.

- **Selection-Mortality Threat**

Selection-mortality arises when there is *differential* nonrandom dropout between pretest and posttest. In our example, different types of children might drop out of each group, or more may drop out of one than the other. Posttest differences might then be due to the different types of dropouts -- the selection-mortality -- and not to the program.

- **Selection-Regression Threat**

Finally, selection-regression occurs when there are different rates of regression to the mean in the two groups. This might happen if one group is more extreme on the pretest than the other. In the context of our example, it may be that the program group is getting a disproportionate number of low math ability children because teachers think they need the math tutoring more (and the teachers don't understand the need for 'comparable' program and comparison groups!). Since the tutoring group has the more extreme lower scorers, their mean will regress a greater distance toward the overall population mean and they will appear to gain more than their comparison group counterparts. This is not a real program gain -- it's just a selection-regression artifact.

When we move from a single group to a multiple group study, what do we gain from the rather significant investment in a second group? If the second group is a control group and is comparable to the program group, we can rule out the single group threats to internal validity because they will all be reflected in the comparison group and cannot explain why posttest group differences would occur. But the key is that the groups must be comparable. How can we possibly hope to create two groups that are truly "comparable"? The only way we know of doing that is to randomly assign persons in our sample into the two groups -- we conduct a randomized or "true" experiment. But in many applied research settings we can't randomly assign, either because of logistical or ethical factors. In that case, we typically try to assign two groups nonrandomly so that they are as equivalent as we can make them. We might, for instance, have one classroom of first graders assigned to the math tutoring program while the other class is the comparison group. In this case, we would hope the two are equivalent, and we may even have reasons to believe that they are. But because they may not be equivalent and because we did not use a procedure like random assignment to at least assure that they are probabilistically equivalent, we call such designs quasi-experimental designs. If we measure them on a pretest, we can examine whether they appear to be similar on key measures before the study begins and make some judgement about the plausibility that a selection bias exists.

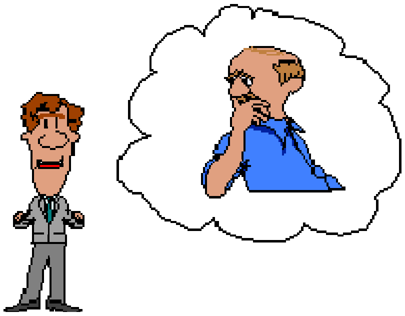
Even if we move to a multiple group design and have confidence that our groups are comparable, we cannot assume that we have strong internal validity. There are a number of social threats to internal validity that arise from the human interaction present in applied social research that we will also need to address.

- **Social Interaction Threats**

What are "Social" Threats?

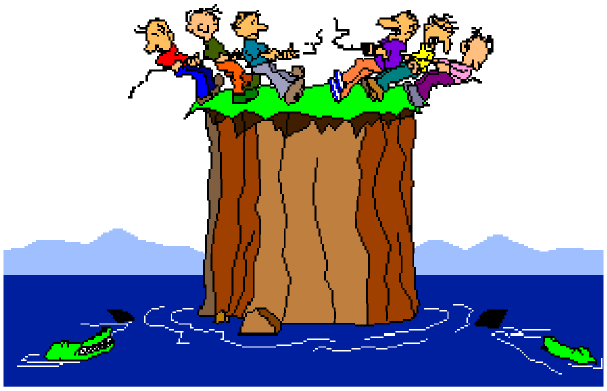
Applied social research is a human activity. And, the results of such research are affected by the human interactions involved. The social threats to internal validity refer to the social pressures in the research context that can lead to posttest differences that are not directly caused by the treatment itself. Most of these threats occur because the various groups (e.g., program and comparison), or key people involved in carrying out the research (e.g., managers and administrators, teachers and principals) are aware of each other's existence and of the role they play in the research project or are in contact with one another. Many of these threats can be minimized by *isolating the two groups from each other*, but this leads to other problems (e.g., it's hard to randomly assign and then isolate; this is likely to reduce generalizability or external validity). Here are the major social interaction threats to internal validity:

- **Diffusion or Imitation of Treatment**

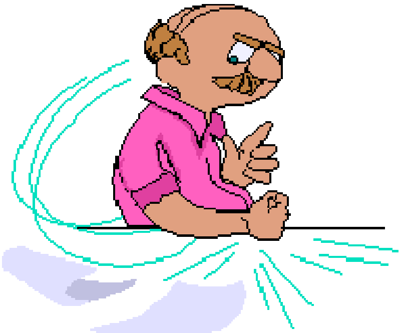


This occurs when a comparison group learns about the program either directly or indirectly from program group participants. In a school context, children from different groups within the same school might share experiences during lunch hour. Or, comparison group students, seeing what the program group is getting, might set up their own experience to try to imitate that of the program group. In either case, if the diffusion of imitation affects the posttest performance of the comparison group, it can have an jeopardize your ability to assess whether your program is causing the outcome. Notice that this threat to validity tend to equalize the outcomes between groups, minimizing the chance of seeing a program effect even if there is one.

- **Compensatory Rivalry**



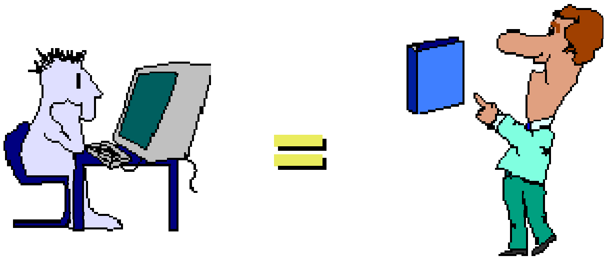
Here, the comparison group knows what the program group is getting and develops a competitive attitude with them. The students in the comparison group might see the special math tutoring program the program group is getting and feel jealous. This could lead them to deciding to compete with the program group "just to show them" how well they can do. Sometimes, in contexts like these, the participants are even encouraged by well-meaning teachers or administrators to compete with each other (while this might make educational sense as a motivation for the students in both groups to work harder, it works against our ability to see the effects of the program). If the rivalry between groups affects posttest performance, it could maker it more difficult to detect the effects of the program. As with diffusion and imitation, this threat generally works to in the direction of equalizing the posttest performance across groups, increasing the chance that you won't see a program effect, even if the program is effective.



- **Resentful Demoralization**

This is almost the opposite of compensatory rivalry. Here, students in the comparison group know what the program group is getting. But here, instead of developing a rivalry, they get discouraged or angry and they give up (sometimes referred to as the "screw you" effect!). Unlike the previous two threats, this one is likely to exaggerate posttest differences between groups, making your program look even more effective than it actually is.

- **Compensatory Equalization of Treatment**



This is the only threat of the four that primarily involves the people who help manage the research context rather than the participants themselves. When program and comparison group participants are aware of each other's conditions they may wish they were in the other group (depending on the perceived desirability of the program it could work either way). Often they or their parents or teachers will put pressure on the administrators to have them reassigned to the other group. The administrators may begin to feel that the allocation of goods to the groups is not "fair" and may be pressured to or independently undertake to compensate one group for the perceived advantage of the other. If the special math tutoring program was being done with state-of-the-art computers, you can bet that the parents of the children assigned to the traditional non-computerized comparison group will pressure the principal to "equalize" the situation. Perhaps the principal will give the comparison group some other good, or let them have access to the computers for other subjects. If these "compensating" programs equalize the groups on posttest performance, it will tend to work against your detecting an effective program even when it does work. For instance, a compensatory program might improve the self-esteem of the comparison group and eliminate your chance to discover whether the math program would cause changes in self-esteem relative to traditional math training.

As long as we engage in applied social research we will have to deal with the realities of human interaction and its effect on the research process. The threats described here can often be minimized by constructing multiple groups that are not aware of each other (e.g., program group from one school, comparison group from another) or by training administrators in the importance of preserving group membership and not instituting equalizing programs. But we will never be able to entirely eliminate the possibility that human interactions are making it more difficult for us to assess cause-effect relationships.

Introduction to Design

What is Research Design?

Research design can be thought of as the *structure* of research -- it is the "glue" that holds all of the elements in a research project together. We often describe a design using a concise notation that enables us to summarize a complex design structure efficiently. What are the "elements" that a design includes? They are:

- **Observations or Measures**

These are symbolized by an '**O**' in design notation. An **O** can refer to a single measure (e.g., a measure of body weight), a single instrument with multiple items (e.g., a 10-item self-esteem scale), a complex multi-part instrument (e.g., a survey), or a whole battery of tests or measures given out on one occasion. If you need to distinguish among specific measures, you can use subscripts with the **O**, as in $\mathbf{O_1, O_2}$, and so on.

- **Treatments or Programs**

These are symbolized with an '**X**' in design notations. The **X** can refer to a simple intervention (e.g., a one-time surgical technique) or to a complex hodgepodge program (e.g., an employment training program). Usually, a no-treatment control or comparison group has no symbol for the treatment (some researchers use **X+** and **X-** to indicate the treatment and control respectively). As with observations, you can use subscripts to distinguish different programs or program variations.

- **Groups**

Each group in a design is given its own line in the design structure. if the design notation has three lines, there are three groups in the design.

- **Assignment to Group**

Assignment to group is designated by a letter at the beginning of each line (i.e., group) that describes how the group was assigned. The major types of assignment are:

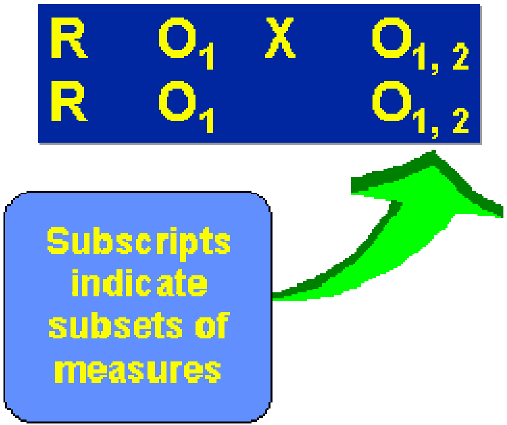
- **R** = random assignment
- **N** = nonequivalent groups
- **C** = assignment by cutoff

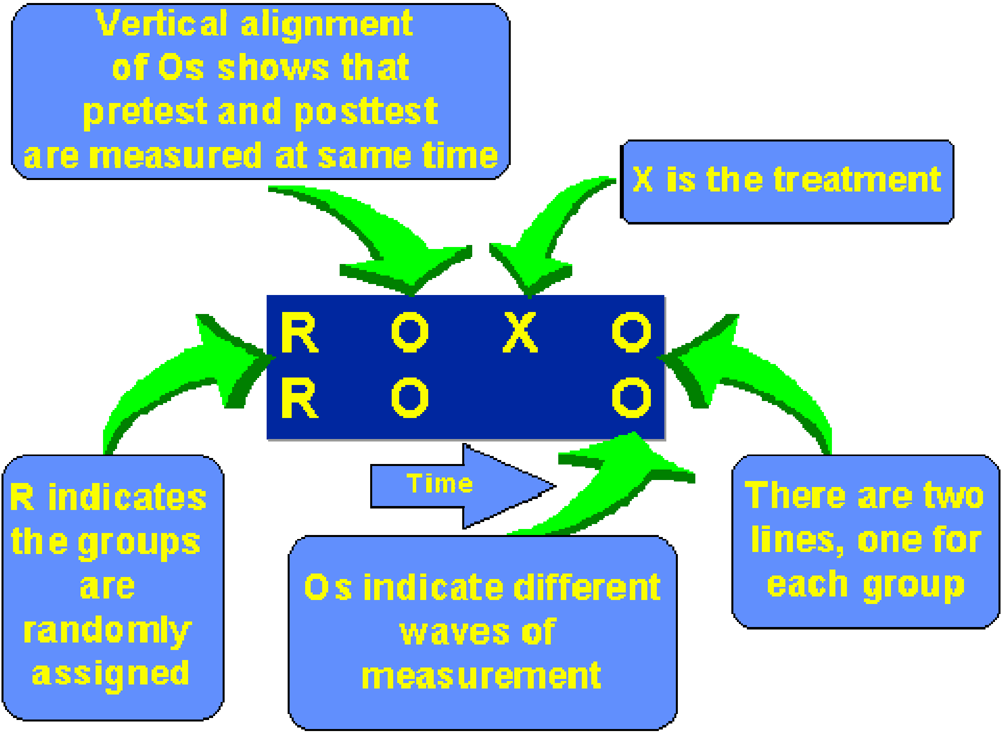
- **Time**

Time moves from left to right. Elements that are listed on the left occur before elements that are listed on the right.

Design Notation Examples

It's always easier to explain design notation through examples than it is to describe it in words. The figure shows the design notation for a ***pretest-posttest (or before-after) treatment versus comparison group randomized experimental design***. Let's go through each of the parts of the design. There are two lines in the notation, so you should realize that the study has two groups. There are four **O**s in the notation, two on each line and two for each group. When the **O**s are stacked vertically on top of each other it means they are collected at the same time. In the notation you can see that we have two **O**s that are taken before (i.e., to the left of) any treatment is given -- the pretest -- and two **O**s taken after the treatment is given -- the posttest. The **R** at the beginning of each line signifies that the two groups are randomly assigned (making it an experimental design). The design is a treatment versus comparison group one

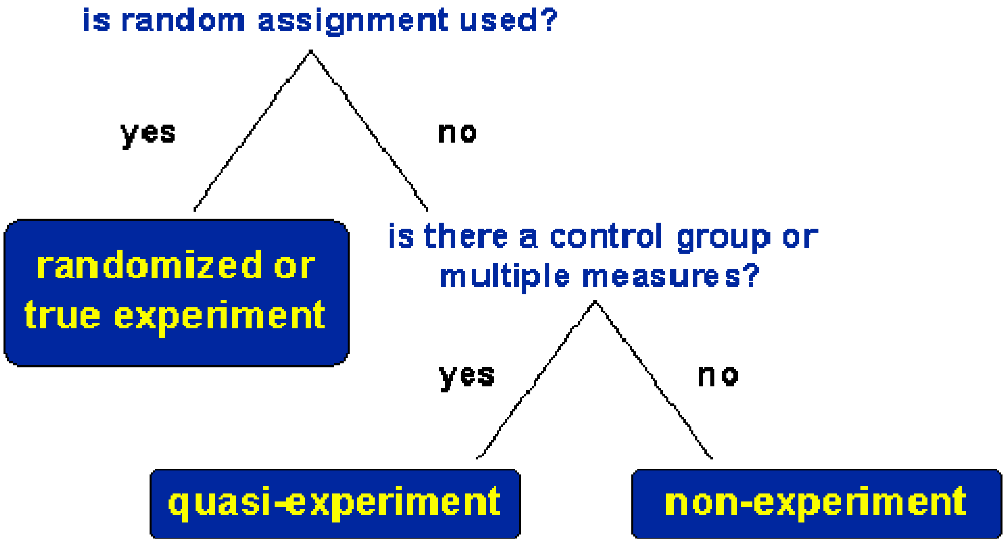



Vertical alignment of Os shows that pretest and posttest are measured at same time
X is the treatment
R O X O
R O O
Time
R indicates the groups are randomly assigned
Os indicate different waves of measurement
There are two lines, one for each group

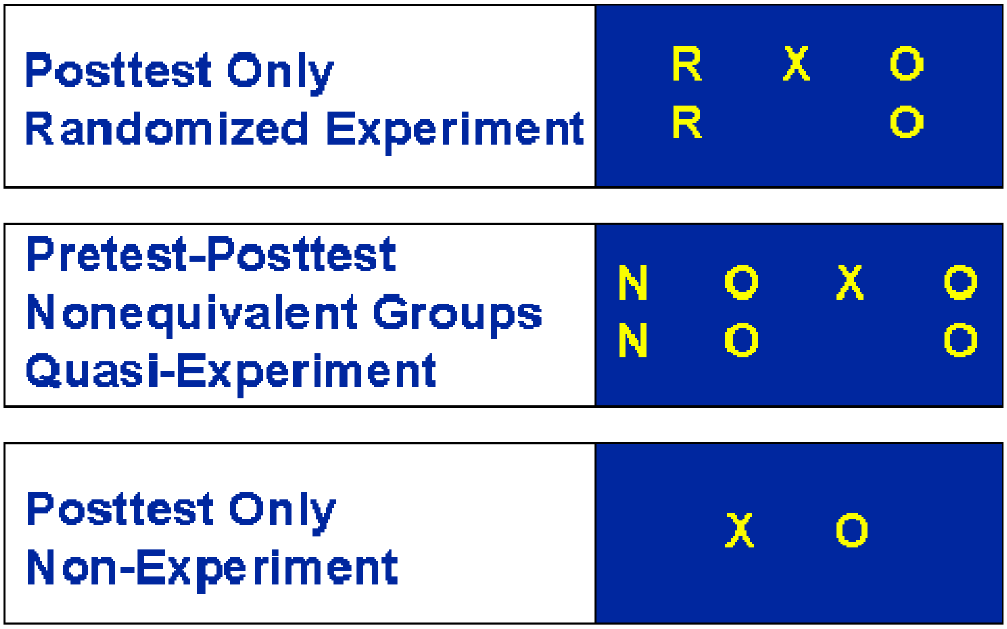
because the top line (treatment group) has an **X** while the bottom line (control group) does not. You should be able to see why many of my students have called this type of notation the "tic-tac-toe" method of design notation -- there are lots of **X**s and **O**s! Sometimes we have to be more specific in describing the **O**s or **X**s than just using a single letter. In the second figure, we have the identical research design with some subscripting of the **O**s. What does this mean? Because all of the **O**s have a subscript of **1**, there is some measure or set of measures that is collected for both groups on both occasions. But the design also has two **O**s with a subscript of **2**, both taken at the posttest. This means that there was some measure or set of measures that were collected ***only*** at the posttest.

With this simple set of rules for describing a research design in notational form, you can concisely explain even complex design structures. And, using a notation helps to show common design sub-structures across different designs that we might not recognize as easily without the notation.

Types of Designs


is random assignment used?
yes
no
randomized or true experiment
is there a control group or multiple measures?
yes
no
quasi-experiment
non-experiment

What are the different major types of research designs? We can classify designs into a simple threefold classification by asking some key questions. First, does the design use random assignment to groups? [Don't forget that random *assignment* is not the same thing as random *selection* of a sample from a population!] If random assignment is used, we call the design a **randomized experiment** or **true experiment**. If random assignment is not used, then we have to ask a second question: Does the design use ***either*** multiple groups or multiple waves of measurement? If the answer is yes, we would label it a **quasi-experimental design**. If no, we would call it a **non-experimental design**. This threefold classification is especially useful for describing the design with respect to internal validity. A randomized experiment generally is the strongest of the three designs when your interest is in establishing a cause-effect relationship. A non-experiment is generally the weakest in this respect. I have to hasten to add here, that I don't mean that a non-experiment is the weakest of the the three designs *overall*, but only with respect to internal validity or causal assessment. In fact, the simplest form of non-experiment is a one-shot survey design that consists of nothing but a single observation **O**. This is probably one of the most common forms of research and, for some research questions -- especially descriptive ones -- is clearly a strong design. When I say that the non-experiment is the weakest with respect to internal validity, all I mean is that it isn't a particularly good method for assessing the cause-effect relationship that you think might exist between a program and its outcomes.



Posttest Only Randomized Experiment	R X O R O
Pretest-Posttest Nonequivalent Groups Quasi-Experiment	N O X O N O O
Posttest Only Non-Experiment	X O

To illustrate the different types of designs, consider one of each in design notation. The first design is a posttest-only randomized experiment. You can tell it's a randomized experiment because it has an R at the beginning of each line, indicating random assignment. The second design is a pre-post nonequivalent groups quasi-experiment. We know it's not a randomized experiment because random assignment wasn't used. And we know it's not a non-experiment because there are both multiple groups and multiple waves of measurement. That means it must be a quasi-experiment. We add the label "nonequivalent" because in this design we do not explicitly control the assignment and the groups may be nonequivalent or not similar to each other (see nonequivalent group designs). Finally, we show a posttest-only nonexperimental design. You might use this design if you want to study the effects of a natural disaster like a flood or tornado and you want to do so by interviewing survivors. Notice that in this design, you don't have a comparison group (e.g., interview in a town down the road the road that didn't have the tornado to see what differences the tornado caused) and you don't have multiple waves of measurement (e.g., a pre-tornado level of how people in the ravaged town were doing before the disaster). Does it make sense to do the non-experimental study? Of course! You could gain lots of valuable information by well-conducted post-disaster interviews. But you may have a hard time establishing which of the things you observed are due to the disaster rather than to other factors like the peculiarities of the town or pre-disaster characteristics.

Experimental Design

Experimental designs are often touted as the most "rigorous" of all research designs or, as the "gold standard" against which all other designs are judged. In one sense, they probably are. If you can implement an experimental design well (and that is a big "if" indeed), then the experiment is probably the strongest design with respect to internal validity. Why? Recall that internal validity is at the center of all causal or cause-effect inferences. When you want to determine whether some program or treatment *causes* some outcome or outcomes to occur, then you are interested in having strong internal validity. Essentially, you want to assess the proposition:

If X, then Y

or, in more colloquial terms:

If the program is given, then the outcome occurs

Unfortunately, it's not enough just to show that when the program or treatment occurs the expected outcome also happens. That's because there may be lots of reasons, other than the program, for why you observed the outcome. To really show that there is a causal relationship, you have to simultaneously address the two propositions:

If X, then Y

and

If *not* X, then *not* Y

Or, once again more colloquially:

If the program is given, then the outcome occurs

and

If the program is *not* given, then the outcome does *not* occur

If you are able to provide evidence for both of these propositions, then you've in effect isolated the program from all of the other potential causes of the outcome. You've shown that when the program is present the outcome occurs and when it's not present, the outcome doesn't occur. That points to the causal effectiveness of the program.

Think of all this like a fork in the road. Down one path, you implement the program and observe the outcome. Down the other path, you don't implement the program and the outcome doesn't occur. But, how do we take *both* paths in the road in the same study? How can we be in two places at once? Ideally, what we want is to have the same conditions -- the same people, context, time, and so on -- and see whether when the program is given we get the outcome and when the program is not given we don't. Obviously, we can never achieve this hypothetical situation. If we give the program to a group of people, we can't simultaneously not give it! So, how do we get out of this apparent dilemma?

Perhaps we just need to think about the problem a little differently. What if we could create two groups or contexts that are as similar as we can possibly make them? If we could be confident that the two situations are comparable, then we could administer our program in one (and see if the outcome occurs) and not give the program in the other (and see if the outcome doesn't occur). And, if the two contexts are comparable, then this is like taking both forks in the road simultaneously! We can have our cake and eat it too, so to speak.

That's exactly what an experimental design tries to achieve. In the simplest type of experiment, we create two groups that are "equivalent" to each other. One group (the program or treatment group) gets the program and the other group (the comparison or control group) does not. In all other respects, the groups are treated the same. They have similar people, live in similar contexts, have similar backgrounds, and so on. Now, if we observe differences in outcomes between these two groups, then the differences must be due to the only thing that differs between them -- that one got the program and the other didn't.

OK, so how do we create two groups that are "equivalent"? The approach used in experimental design is to assign people randomly from a common pool of people into the two groups. The experiment relies on this idea of random assignment to groups as the basis for obtaining two groups that are similar. Then, we give one the program or treatment and we don't give it to the other. We observe the same outcomes in both groups.

The key to the success of the experiment is in the random assignment. In fact, even with random assignment we never expect that the groups we create will be exactly the same. How could they be, when they are made up of different people? We rely on the idea of probability and assume that the two groups are "probabilistically equivalent" or equivalent within known probabilistic ranges.

So, if we randomly assign people to two groups, and we have enough people in our study to achieve the desired probabilistic equivalence, then we may consider the experiment to be strong in internal validity and we probably have a good shot at assessing whether the program causes the outcome(s).

But there are lots of things that can go wrong. We may not have a large enough sample. Or, we may have people who refuse to participate in our study or who drop out part way through. Or, we may be challenged successfully on ethical grounds (after all, in order to use this approach we have to deny the program to some people who might be equally deserving of it as others). Or, we may get resistance from the staff in our study who would like some of their "favorite" people to get the program. Or, they mayor might insist that her daughter be put into the new program in an educational study because it may mean she'll get better grades.

The bottom line here is that experimental design is intrusive and difficult to carry out in most real world contexts. And, because an experiment is often an intrusion, you are to some extent setting up an artificial situation so that you can assess your causal relationship with high internal validity. If so, then you are limiting the degree to which you can generalize your results to real contexts where you haven't set up an experiment. That is, you have reduced your external validity in order to achieve greater internal validity.

In the end, there is just no simple answer (no matter what anyone tells you!). If the situation is right, an experiment can be a very strong design to use. But it isn't automatically so. My own personal guess is that randomized experiments are probably appropriate in no more than 10% of the social research studies that attempt to assess causal relationships.

Experimental design is a fairly complex subject in its own right. I've been discussing the simplest of experimental designs -- a two-group program versus comparison group design. But there are lots of experimental design variations that attempt to accomplish different things or solve different problems. In this section you'll explore the basic design and then learn some of the principles behind the major variations.

- **Two-Group Experimental Designs**

```
R   X   O
R       O
```

The simplest of all experimental designs is the two-group posttest-only randomized experiment. In design notation, it has two lines -- one for each group -- with an R at the beginning of each line to indicate that the groups were randomly assigned. One group gets the treatment or program (the X) and the other group is the comparison group and doesn't get the program (note that this you could alternatively have the comparison group receive the standard or typical treatment, in which case this study would be a relative comparison).

history	✓
maturation	✓
testing	✓
instrumentation	✓
mortality	✓
regression to the mean	✓
selection	✓
selection-history	✓
selection- maturation	✓
selection- testing	✓
selection- instrumentation	✓
selection- mortality	✗
selection- regression	✓
diffusion or imitation	✗
compensatory equalization	✗
compensatory rivalry	✗
resentful demoralization	✗

Notice that a pretest is not required for this design. Usually we include a pretest in order to determine whether groups are comparable prior to the program, but because we are using random assignment we can assume that the two groups are probabilistically equivalent to begin with and the pretest is not required (although you'll see with covariance designs that a pretest may still be desirable in this context).

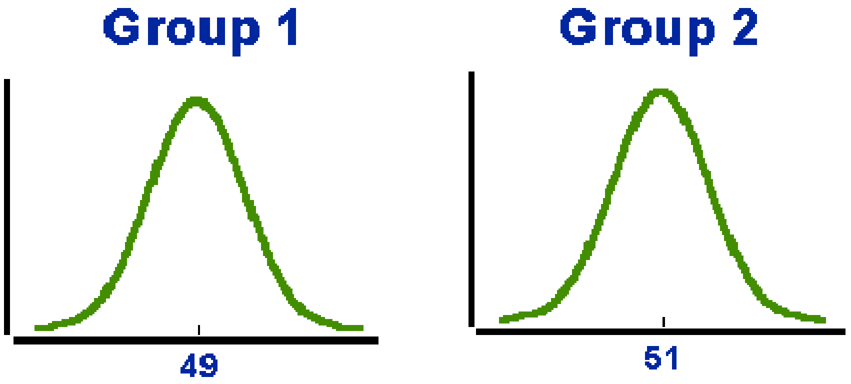
In this design, we are most interested in determining whether the two groups are different after the program. Typically we measure the groups on one or more measures (the Os in notation) and we compare them by testing for the differences between the means using a t-test or one way Analysis of Variance (ANOVA).

The posttest-only randomized experiment is strong against the single-group threats to internal validity because it's not a single group design! (Tricky, huh?) It's strong against the all of the multiple-group threats except for selection-mortality. For instance, it's strong against selection-testing and selection-instrumentation because it doesn't use repeated measurement. The selection-mortality threat is especially salient if there are differential rates of dropouts in the two groups. This could result if the treatment or program is a noxious or negative one (e.g., a painful medical procedure like chemotherapy) or if the control group condition is painful or intolerable. This design is susceptible to all of the social interaction threats to internal validity. Because the design requires random assignment, in some institutional settings (e.g., schools) it is more likely to utilize persons who would be aware of each other and of the conditions they've been assigned to.

The posttest-only randomized experimental design is, despite its simple structure, one of the best research designs for assessing cause-effect relationships. It is easy to execute and, because it uses only a posttest, is relatively inexpensive. But there are many variations on this simple experimental design. You can begin to explore these by looking at how we classify the various experimental designs.

❖ Probabilistic Equivalence

What is Probabilistic Equivalence?



With α = .05, we expect that we will observe a pretest difference 5 times out of 100

What do I mean by the term **probabilistic equivalence**? Well, to begin with, I certainly *don't* mean that two groups are equal to each other. When we deal with human beings it is impossible to ever say that any two individuals or groups are equal or equivalent. Clearly the important term in the phrase is "probabilistic". This means that the type of equivalence we have is based on the notion of probabilities. In more concrete terms, probabilistic equivalence means that we know *perfectly* the odds that we will find a difference between two groups. Notice, it doesn't mean that the means of the two groups will be equal. It just means that we know the odds that they won't be equal. The figure shows two groups, one having a mean of 49 and the other with a mean of 51. Could these two groups be probabilistically equivalent? Certainly!

We achieve probabilistic equivalence through the mechanism of random assignment to groups. When we randomly assign to groups, we can calculate the chance that the two groups will differ just because of the random assignment (i.e., by chance alone). Let's say we are assigning a group of first grade students to two groups. Further, let's assume that the average test scores for these children for a standardized test with a population mean of 50 were 49 and 51 respectively. We might conduct a t-test to see if the means of our two randomly assigned groups are statistically different. We know -- through random assignment and the law of large numbers -- that the chance that they will be different is 5 out of 100 when we set our significance level to .05 (i.e., ***alpha*** = .05). In other words, 5 times out of every 100, when we randomly assign two groups, we can expect to get a significant difference at the .05 level of significance.

When we assign randomly, the only reason the groups can differ is because of chance assignment because their assignment is entirely based on the randomness of assignment. If, by chance, the groups differ on one variable, we have no reason to believe that they will automatically be different on any other. Even if we find that the groups differ on a pretest, we have no reason to suspect that they will differ on a posttest. Why? Because their pretest difference had to be a chance one. So, when we randomly assign, we are able to assume that the groups do have a form of equivalence. We don't expect them to be equal. But we do expect that they are "probabilistically" equal.

❖ Random Selection & Assignment

Random *selection* is how you draw the sample of people for your study from a population. **Random *assignment*** is how you assign the sample that you draw to different groups or treatments in your study.

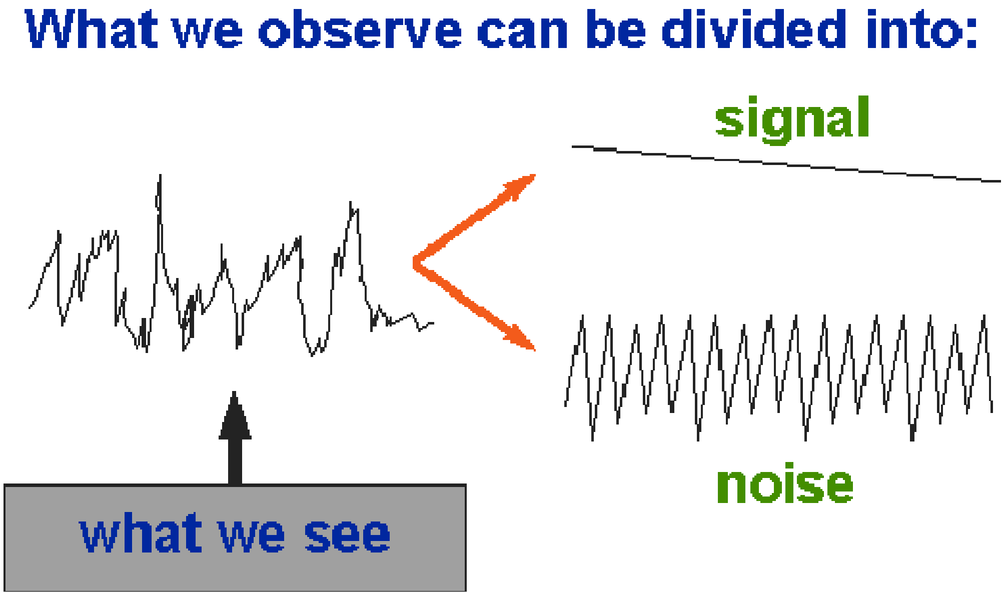
It is possible to have *both* random selection and assignment in a study. Let's say you drew a random sample of 100 clients from a population list of 1000 current clients of your organization. That is random sampling. Now, let's say you randomly assign 50 of these clients to get some new additional treatment and the other 50 to be controls. That's random assignment.

It is also possible to have *only one of these* (random selection or random assignment) but not the other in a study. For instance, if you do not randomly draw the 100 cases from your list of 1000 but instead just take the first 100 on the list, you do not have random selection. But you could still randomly assign this nonrandom sample to treatment versus control. Or, you could randomly select 100 from your list of 1000 and then nonrandomly (haphazardly) assign them to treatment or control.

And, it's possible to have *neither* random selection nor random assignment. In a typical nonequivalent groups design in education you might nonrandomly choose two 5th grade classes to be in your study. This is nonrandom selection. Then, you could arbitrarily assign one to get the new educational program and the other to be the control. This is nonrandom (or nonequivalent) assignment.

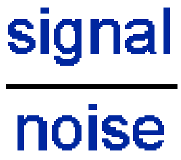
Random selection is related to sampling. Therefore it is most related to the external validity (or generalizability) of your results. After all, we would randomly sample so that our research participants better represent the larger group from which they're drawn. Random assignment is most related to design. In fact, when we randomly assign participants to treatments we have, by definition, an experimental design. Therefore, random assignment is most related to internal validity. After all, we randomly assign in order to help assure that our treatment groups are similar to each other (i.e., equivalent) prior to the treatment.

- **Classifying Experimental Designs**



Although there are a great variety of experimental design variations, we can classify and organize them using a simple signal-to-noise ratio metaphor. In this metaphor, we assume that what we observe or see can be divided into two components, the signal and the noise (by the way, this is directly analogous to the true score theory of measurement). The figure, for instance, shows a time series with a slightly downward slope. But because there is so much variability or noise in the series, it is difficult even to detect the downward slope. When we divide the series into its two components, we can clearly see the slope.

In most research, the signal is related to the key variable of interest -- the construct you're trying to measure, the program or treatment that's being implemented. The noise consists of all of the random factors in the situation that make it harder to see the signal -- the lighting in the room,



local distractions, how people felt that day, etc. We can construct a ratio of these two by dividing the signal by the noise. In research, we want the signal to be high relative to the noise. For instance, if you have a very powerful treatment or program (i.e., strong signal) and very good measurement (i.e., low noise) you will have a better chance of seeing the effect of the program than if you have either a strong program and weak measurement or a weak program and strong measurement.

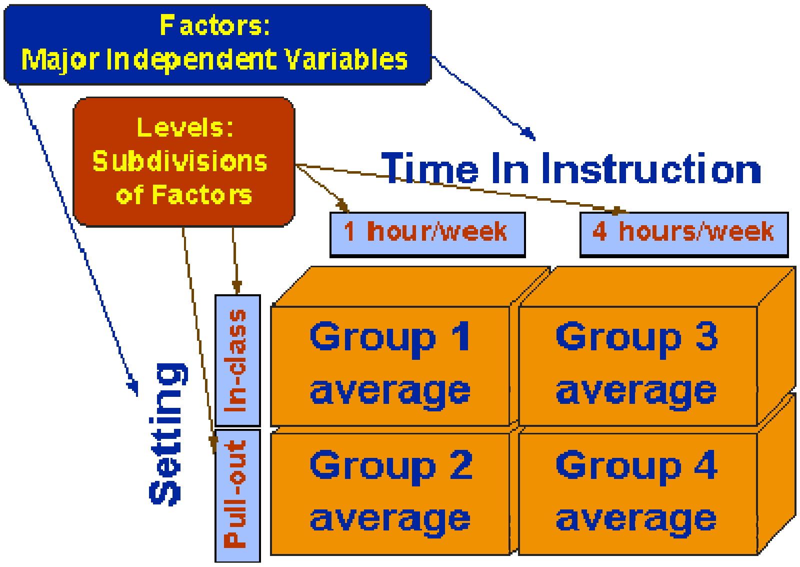
With this in mind, we can now classify the experimental designs into two categories: **signal enhancers** or **noise reducers**. Notice that doing either of these things -- enhancing signal or reducing noise -- improves the quality of the research. The *signal-enhancing experimental designs* are called the factorial designs. In these designs, the focus is almost entirely on the setup of the program or treatment, its components and its major dimensions. In a typical factorial design we would examine a number of different variations of a treatment.

There are two major types of *noise-reducing experimental designs*: covariance designs and blocking designs. In these designs we typically use information about the makeup of the sample or about pre-program variables to remove some of the noise in our study.

- **Factorial Designs**

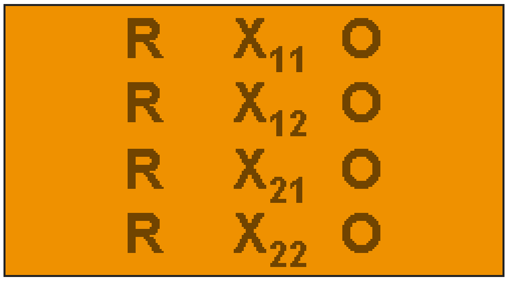
A Simple Example

Probably the easiest way to begin understanding factorial designs is by looking at an example. Let's imagine a design where we have an educational program where we would like to look at a variety of program variations to see which works best. For instance, we would like to vary the amount of time the children receive instruction with one group getting 1 hour of instruction per week and another getting 4 hours per week. And, we'd like to vary the setting with one group getting the instruction in-class (probably pulled off into a corner of the classroom) and the other group being pulled-out of the classroom for instruction in another room. We could think about having four separate groups to do this, but when we are varying the amount of time in instruction, what setting would we use: in-class or pull-out? And,



when we were studying setting, what amount of instruction time would we use: 1 hour, 4 hours, or something else?

With factorial designs, we don't have to compromise when answering these questions. We can have it both ways if we cross each of our two time in instruction conditions with each of our two settings. Let's begin by doing some defining of terms. In factorial designs, a **factor** is a major independent variable. In this example we have two factors: time in instruction and setting. A **level** is a subdivision of a factor. In this example, time in instruction has two levels and setting has two levels. Sometimes we depict a factorial design with a numbering notation. In this example, we can say that we have a 2 x 2 (spoken "two-by-two) factorial design. In this notation, the *number of numbers* tells you how many factors there are and the *number values* tell you how many levels. If I said I had a 3 x 4 factorial design, you would know that I had 2 factors and that one factor had 3 levels while the other had 4. Order of the numbers makes no difference and we could just as easily term this a 4 x 3 factorial design. The number of different treatment groups that we have in any factorial design can easily be determined by multiplying through the number notation. For instance, in our example we have 2 x 2 = 4 groups. In our notational example, we would need 3 x 4 = 12 groups.

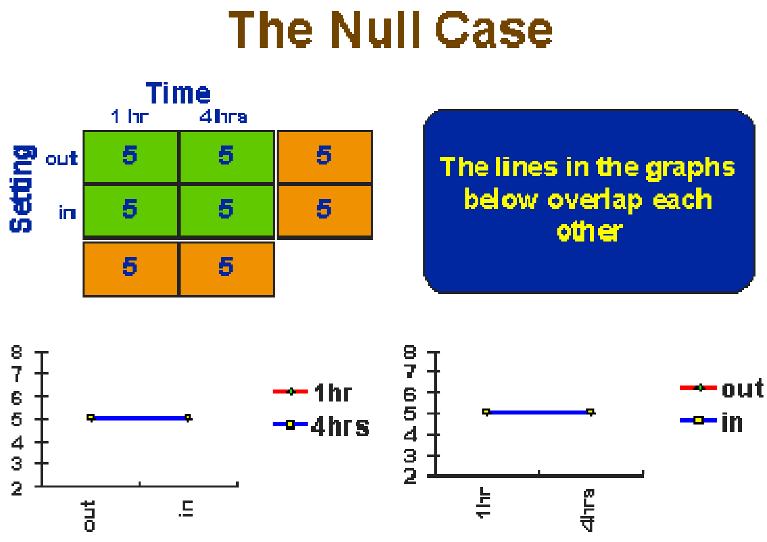


We can also depict a factorial design in design notation. Because of the treatment level combinations, it is useful to use subscripts on the treatment (X) symbol. We can see in the figure that there are four groups, one for each combination of levels of factors. It is also immediately apparent that the groups were randomly assigned and that this is a posttest-only design.

Now, let's look at a variety of different results we might get from this simple 2 x 2 factorial design. Each of the following figures describes a different possible outcome. And each outcome is shown in table form (the 2 x 2 table with the row and column averages) and in graphic form (with each factor taking a turn on the horizontal axis). You should convince yourself that the information in the tables agrees with the information in both of the graphs. You should also convince yourself that the pair of graphs in each figure show the exact same information graphed in two different ways. The lines that are shown in the graphs are technically not necessary -- they are used as a visual aid to enable you to easily track where the averages for a single level go across levels of another factor. Keep in mind that the values shown in the tables and graphs are group averages on the outcome variable of interest. In this example, the outcome might be a test of achievement in the subject being taught. We will assume that scores on this test range from 1 to 10 with higher values indicating greater achievement. You should study carefully the outcomes in each figure in order to understand the differences between these cases.

The Null Outcome

Let's begin by looking at the "null" case. The null case is a situation where the treatments have no effect. This figure assumes that even if we didn't give the training we could expect that students would score a 5 on average on the outcome test. You can see in this hypothetical case that all four groups score an average of 5 and therefore the row and column averages must be 5. You can't see the lines for both levels in the graphs because one line falls right on top of the other.



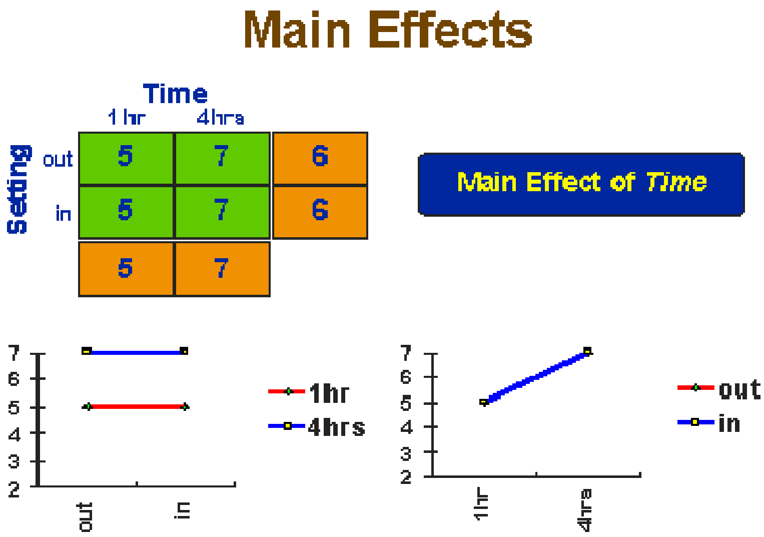
The Null Case

		Time		
		1 hr	4 hrs	
Setting	out	5	5	5
	in	5	5	5
		5	5	

The lines in the graphs below overlap each other

The Main Effects

A **main effect** is an outcome that is a consistent difference between levels of a factor. For instance, we would say there’s a main effect for setting if we find a statistical difference between the averages for the in-class and pull-out groups, *at all levels* of time in instruction. The first figure depicts a main effect of time. For all settings, the 4 hour/week condition worked better than the 1 hour/week one.

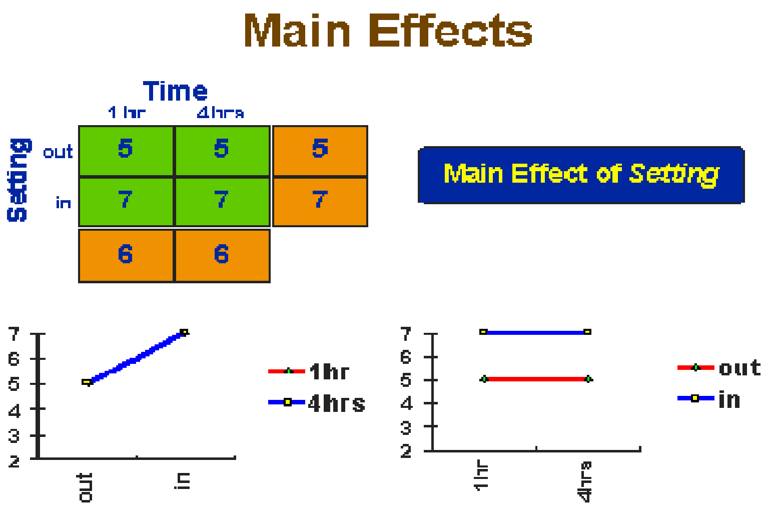


Main Effects

		Time		
		1 hr	4 hrs	
Setting	out	5	7	6
	in	5	7	6
		5	7	

Main Effect of *Time*

It is also possible to have a main effect for setting (and none for time).



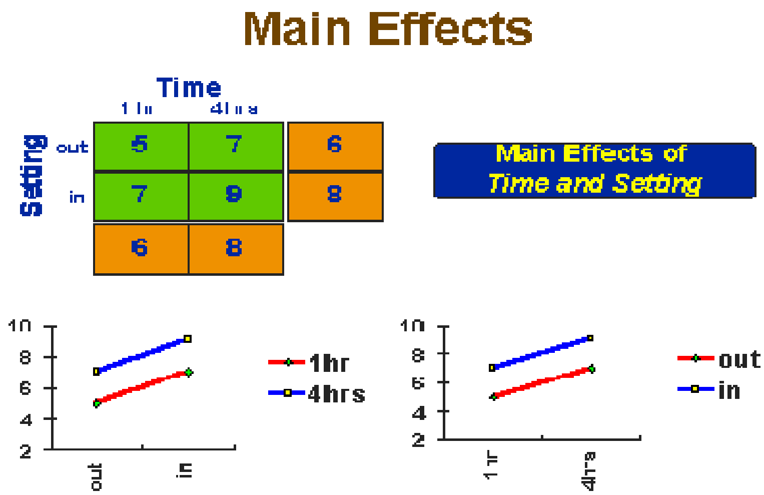
Main Effects

Time

Setting		1 hr	4hrs	
	out	5	5	5
	in	7	7	7
		6	6	

Main Effect of Setting

In the second main effect graph we see that in-class training was better than pull-out training for all amounts of time.



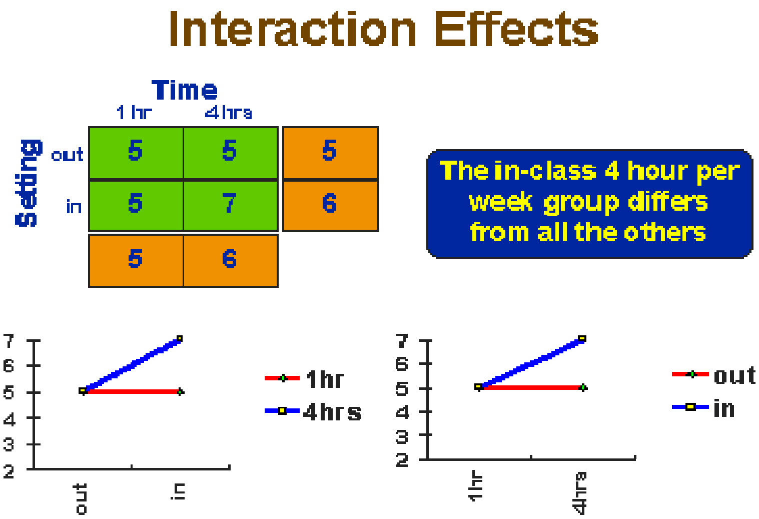
Main Effects

Time

Setting		1 hr	4 hrs	
	out	5	7	6
	in	7	9	8
		6	8	

Main Effects of *Time and Setting*

Finally, it is possible to have a main effect on both variables simultaneously as depicted in the third main effect figure. In this instance 4 hours/week always works better than 1 hour/week and in-class setting always works better than pull-out.



Interaction Effects

Time

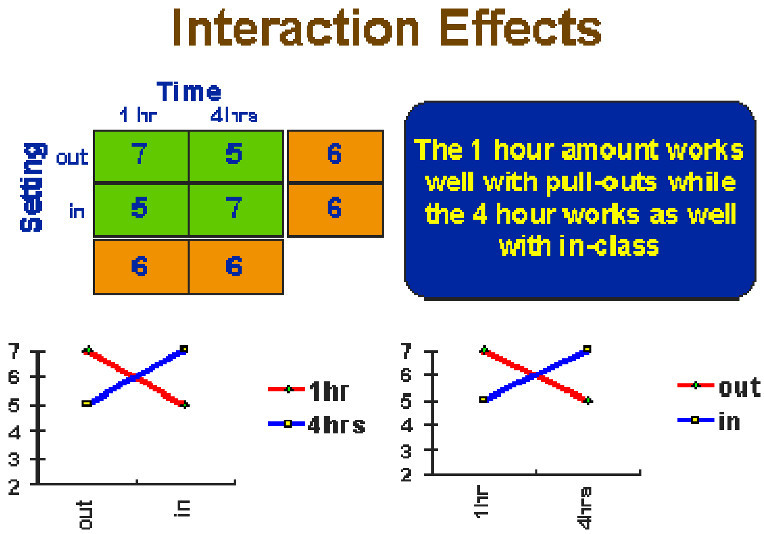
Setting		1 hr	4hrs	
	out	5	5	5
	in	5	7	6
		5	6	

The in-class 4 hour per week group differs from all the others

Interaction Effects

If we could only look at main effects, factorial designs would be useful. But, because of the way we combine levels in factorial designs, they also enable us to examine the **interaction effects** that exist between factors. An *interaction* *effect* exists when differences on one factor depend on the level you are on another factor. It's important to recognize that an interaction is between factors, not levels. We wouldn't say there's an interaction between 4 hours/week and in-class treatment. Instead, we would say that there's an interaction between time and setting, and then we would go on to describe the specific levels involved.

How do you know if there is an interaction in a factorial design? There are three ways you can determine there's an interaction. First, when you run the statistical analysis, the statistical table will report on all main effects and interactions. Second, you know there's an interaction when can't talk about effect on one factor without mentioning the other factor. if you can say at the end of our study that time in instruction makes a difference, then you know that you have a main effect and not an interaction (because you did not have to mention the setting factor when describing the results for time). On the other hand, when you have an interaction it is impossible to describe your results accurately without mentioning both factors. Finally, you can always spot an interaction in the graphs of group means -- whenever there are lines that are not parallel there is an interaction present! If you check out the main effect graphs above, you will notice that all of the lines within a graph are parallel. In contrast, for all of the interaction graphs, you will see that the lines are not parallel.



Setting	Time 1 hr	Time 4hrs	
out	7	5	6
in	5	7	6
	6	6	

The 1 hour amount works well with pull-outs while the 4 hour works as well with in-class

In the first interaction effect graph, we see that one combination of levels -- 4 hours/week and in-class setting -- does better than the other three. In the second interaction we have a more complex "cross-over" interaction. Here, at 1 hour/week the pull-out group does better than the in-class group while at 4 hours/week the reverse is true. Furthermore, the both of these combinations of levels do equally well.

Summary

Factorial design has several important features. First, it has great flexibility for exploring or enhancing the “signal” (treatment) in our studies. Whenever we are interested in examining treatment variations, factorial designs should be strong candidates as the designs of choice. Second, factorial designs are efficient. Instead of conducting a series of independent studies we are effectively able to combine these studies into one. Finally, factorial designs are the only effective way to examine interaction effects.

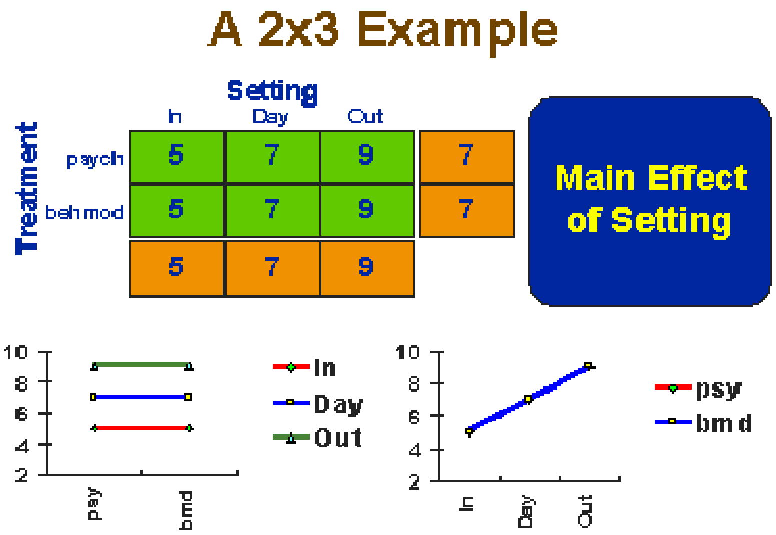
So far, we have only looked at a very simple 2 x 2 factorial design structure. You may want to look at some factorial design variations to get a deeper understanding of how they work. You may also want to examine how we approach the statistical analysis of factorial experimental designs.

❖ Factorial Design Variations

Here, we'll look at a number of different factorial designs. We'll begin with a two-factor design where one of the factors has more than two levels. Then we'll introduce the three-factor design. Finally, we'll present the idea of the incomplete factorial design.

A 2x3 Example

For these examples, let's construct an example where we wish to study of the effect of different treatment combinations for cocaine abuse. Here, the dependent measure is severity of illness rating done by the treatment staff. The outcome ranges from 1 to 10 where higher scores indicate more severe illness: in this case, more severe cocaine addiction. Furthermore, assume that the levels of treatment are:



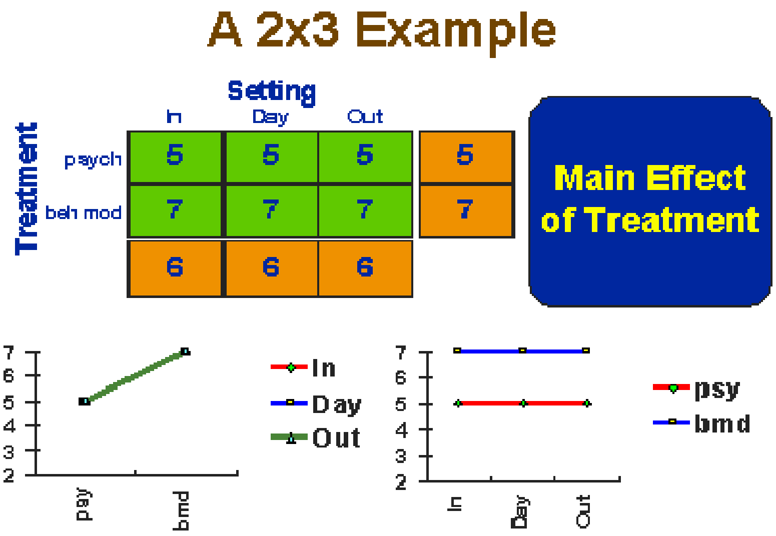
A 2x3 Example

Treatment	Setting: In	Day	Out	
psych	5	7	9	7
beh mod	5	7	9	7
	5	7	9	

Main Effect of Setting

- Factor 1: Treatment
    - psychotherapy
    - behavior modification
- Factor 2: Setting
    - inpatient
    - day treatment
    - outpatient

Note that the setting factor in this example has three levels.



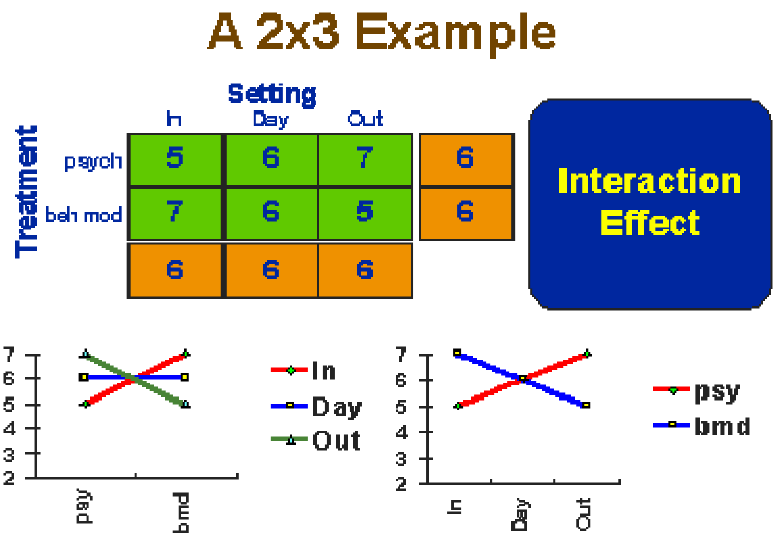
A 2x3 Example

Treatment	Setting: In	Day	Out	
psych	5	5	5	5
beh mod	7	7	7	7
	6	6	6	

Main Effect of Treatment

The first figure shows what an effect for setting outcome might look like. You have to be very careful in interpreting these results because higher scores mean the patient is doing *worse*. It's clear that inpatient treatment works best, day treatment is next best, and outpatient treatment is worst of the three. It's also clear that there is no difference between the two treatment levels (psychotherapy and behavior modification). Even though both graphs in the figure depict the exact same data, I think it's easier to see the main effect for setting in the graph on the lower left where setting is depicted with different lines on the graph rather than at different points along the horizontal axis.

The second figure shows a main effect for treatment with psychotherapy performing better (remember the direction of the outcome variable) in all settings than behavior modification. The effect is clearer in the graph on the lower right where treatment levels are used for the lines. Note that in both this and the previous figure the lines in all graphs are parallel indicating that there are no interaction effects.


A 2x3 Example

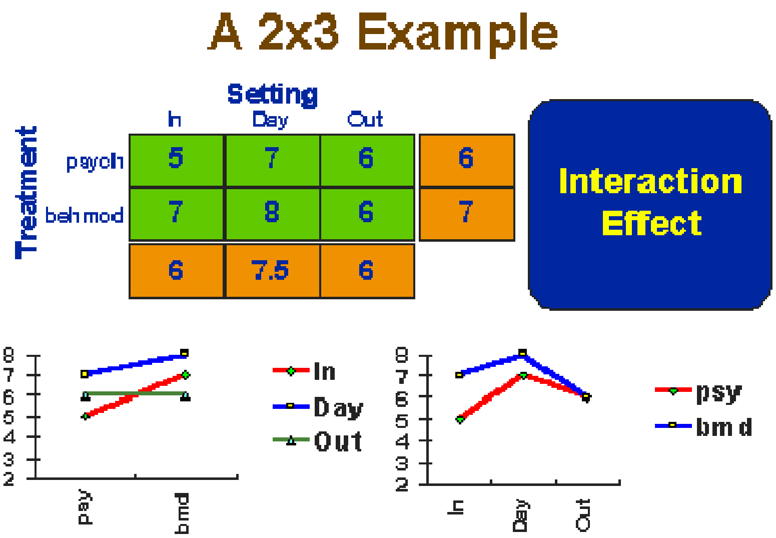
Setting

Treatment	In	Day	Out	
psych	5	6	7	6
beh mod	7	6	5	6
	6	6	6	

Interaction Effect

Now, let's look at a few of the possible interaction effects. In the first case, we see that day treatment is never the best condition. Furthermore, we see that psychotherapy works best with inpatient care and behavior modification works best with outpatient care.

The other interaction effect example is a bit more complicated. Although there may be some main effects mixed in with the interaction, what's important here is that there is a unique combination of levels of factors that stands out as superior: psychotherapy done in the inpatient setting. Once we identify a "best" combination like this, it is almost irrelevant what is going on with main effects.


A 2x3 Example

Setting

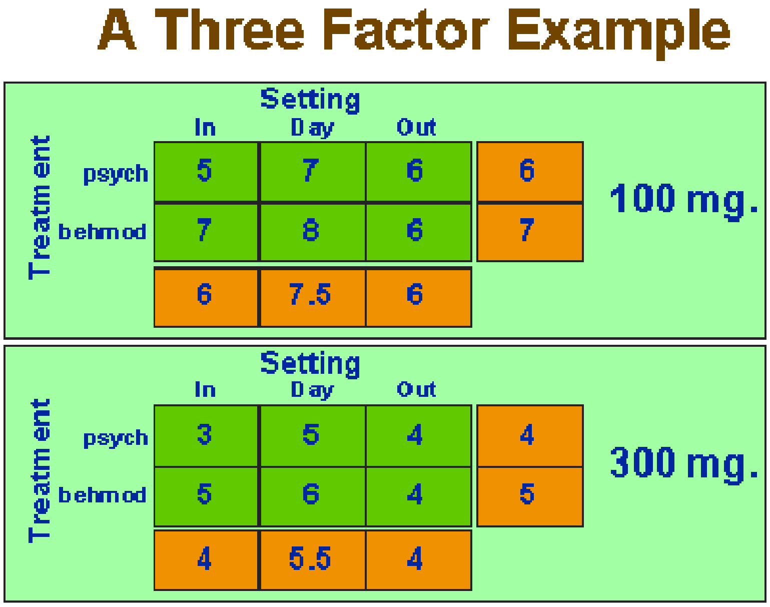
Treatment	In	Day	Out	
psych	5	7	6	6
beh mod	7	8	6	7
	6	7.5	6	

Interaction Effect

A Three-Factor Example

Now let's examine what a three-factor study might look like. We'll use the same factors as above for the first two factors. But here we'll include a new factor for dosage that has two levels. The factor structure in this 2 x 2 x 3 factorial experiment is:

- Factor 1: Dosage
  - 100 mg.
  - 300 mg.
- Factor 2: Treatment
  - psychotherapy
  - behavior modification
- Factor 3: Setting
  - inpatient
  - day treatment
  - outpatient

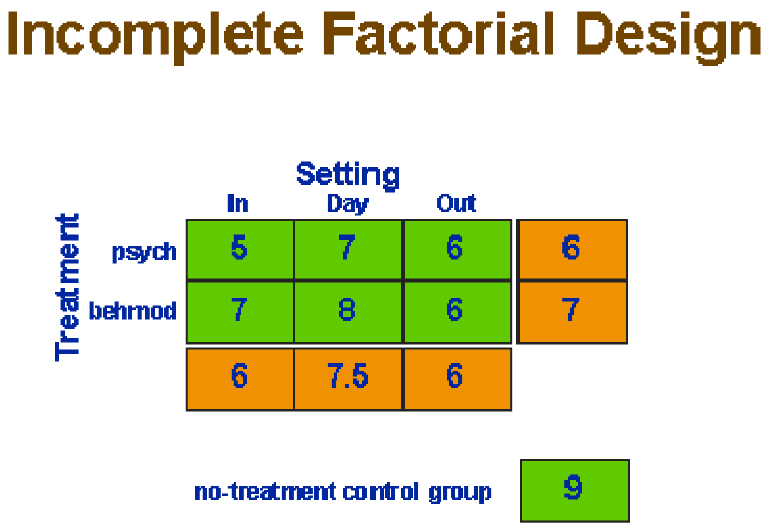


A Three Factor Example

Treatment	Setting In	 Day	 Out		
psych	5	7	6	6	100 mg.
behmod	7	8	6	7	
	6	7.5	6		

Treatment	Setting In	 Day	 Out		
psych	3	5	4	4	300 mg.
behmod	5	6	4	5	
	4	5.5	4		

Notice that in this design we have 2x2x3=12 groups! Although it's tempting in factorial studies to add more factors, the number of groups always increases multiplicatively (is that a real word?). Notice also that in order to even show the tables of means we have to have to tables that each show a two factor relationship. It's also difficult to graph the results in a study like this because there will be a large number of different possible graphs. In the statistical analysis you can look at the main effects for each of your three factors, can look at the three two-way interactions (e.g., treatment vs. dosage, treatment vs. setting, and setting vs. dosage) and you can look at the one three-way interaction. Whatever else may be happening, it is clear that one combination of three levels works best: 300 mg. and psychotherapy in an inpatient setting. Thus, we have a three-way interaction in this study. If you were an administrator having to make a choice among the different treatment combinations you would be best advised to select that one (assuming your patients and setting are comparable to the ones in this study).



Incomplete Factorial Design

Treatment	Setting In	 Day	 Out	
psych	5	7	6	6
behmod	7	8	6	7
	6	7.5	6	

no-treatment control group 9

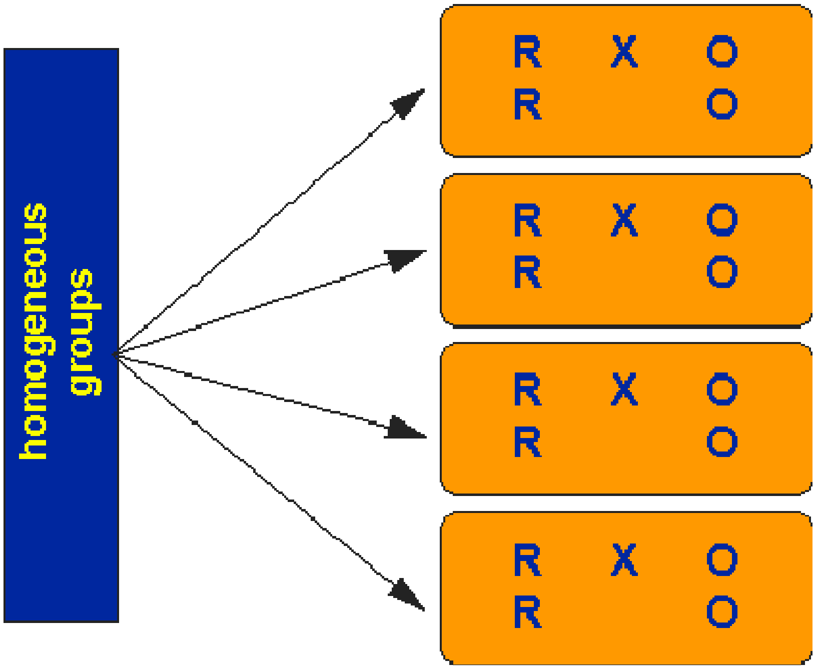
Incomplete Factorial Design

It's clear that factorial designs can become cumbersome and have too many groups even with only a few factors. In much research, you won't be interested in a **fully-crossed factorial design** like the ones we've been showing that pair every combination of levels of factors. Some of the combinations may not make sense from a policy or administrative perspective, or you simply may not have enough funds to implement all combinations. In this case, you may decide to implement an incomplete factorial design. In this variation, some of the cells are intentionally left empty -- you don't assign people to get those combinations of factors.

One of the most common uses of incomplete factorial design is to allow for a control or placebo group that receives no treatment. In this case, it is actually impossible to implement a group that simultaneously has several levels of treatment factors and receives no treatment at all. So, we consider the control group to be its own cell in an incomplete factorial rubric (as shown in the figure). This allows us to conduct both relative and absolute treatment comparisons within a single study and to get a fairly precise look at different treatment combinations.

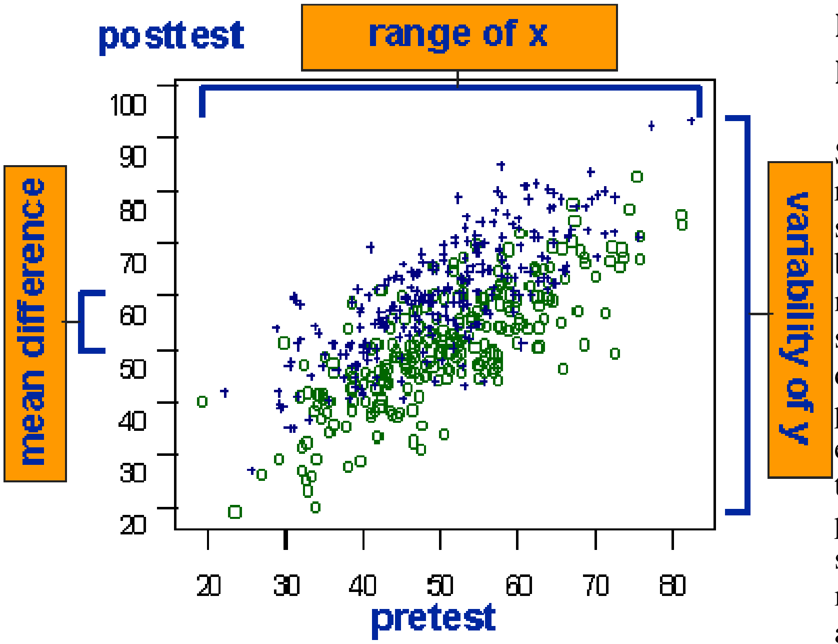
- **Randomized Block Designs**

The Randomized Block Design is research design's equivalent to stratified random sampling. Like stratified sampling, randomized block designs are constructed to reduce noise or variance in the data (see Classifying the Experimental Designs). How do they do it? They require that the researcher divide the sample into relatively homogeneous subgroups or blocks (analogous to "strata" in stratified sampling). Then, the experimental design you want to implement is implemented within each block or homogeneous subgroup. The key idea is that the variability within each block is less than the variability of the entire sample. Thus each estimate of the treatment effect within a block is more efficient than estimates across the entire sample. And, when we pool these more efficient estimates across blocks, we should get an overall more efficient estimate than we would
without blocking.



Here, we can see a simple example. Let's assume that we originally intended to conduct a simple posttest-only randomized experimental design. But, we recognize that our sample has several intact or homogeneous subgroups. For instance, in a study of college students, we might expect that students are relatively homogeneous with respect to class or year. So, we decide to block the sample into four groups: freshman, sophomore, junior, and senior. If our hunch is correct, that the variability within class is less than the variability for the entire sample, we will probably get more powerful estimates of the treatment effect within each block (see the discussion on Statistical Power). Within each of our four blocks, we would implement the simple post-only randomized experiment.

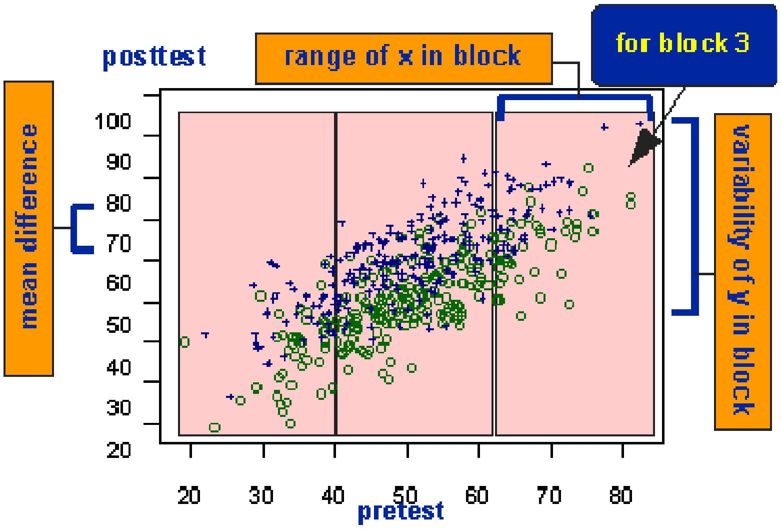
Notice a couple of things about this strategy. First, to an external observer, it may not be apparent that you are blocking. You would be implementing the same design in each block. And, there is no reason that the people in different blocks need to be segregated or separated from each other. In other words, blocking doesn't necessarily affect anything that you do with the research participants. Instead, blocking is a strategy for grouping people in your data analysis in order to reduce noise -- it is an **analysis** strategy. Second, you will only benefit from a blocking design if you are correct in your hunch that the blocks are more homogeneous than the entire sample is. If you are wrong -- if different college-level classes aren't relatively homogeneous with respect to your measures -- you will actually be hurt by blocking (you'll get a less powerful estimate of the treatment effect). How do you know if blocking is a good idea? You need to consider carefully whether the groups are relatively homogeneous. If you are measuring political attitudes, for instance, is it reasonable to believe that freshmen are more like each other than they are like sophomores or juniors? Would they be more homogeneous with respect to measures related to drug abuse? Ultimately the decision to block involves judgment on the part of the researcher.



How Blocking Reduces Noise

So how does blocking work to reduce noise in the data? To see how it works, you have to begin by thinking about the non-blocked study. The figure shows the pretest-posttest distribution for a hypothetical pre-post randomized experimental design. We use the 'X' symbol to indicate a program group case and the 'O' symbol for a comparison group member. You can see that for any specific pretest value, the program group tends to outscore the comparison group by about 10 points on the posttest. That is, there is about a 10-point posttest mean difference.

Now, let's consider an example where we divide the sample into three relatively homogeneous blocks. To see what happens graphically, we'll use the pretest measure to block. This will assure that the groups are very homogeneous. Let's look at what is happening within the third block. Notice that the mean difference is still the same as it was for the entire sample -- about 10 points within each block. But also notice that the variability of the posttest is much less than it was for the entire sample. Remember that the treatment effect estimate is a signal-to-noise ratio. The signal in this case is the mean difference. The noise is the variability. The two figures show that we haven't changed the signal in moving to blocking -- there is still about a 10-point posttest difference. But, we have changed the noise -- the variability on the posttest is much smaller within each block that it is for the entire sample. So, the treatment effect will have less noise for the same signal.



It should be clear from the graphs that the blocking design in this case will yield the stronger treatment effect. But this is true only because we did a good job assuring that the blocks were homogeneous. If the blocks weren't homogeneous -- their variability was as large as the entire sample's -- we would actually get worse estimates than in the simple randomized experimental case. We'll see how to analyze data from a randomized block design in the Statistical Analysis of the Randomized Block Design.

- **Covariance Designs**

Design Notation

R O X O
R O O

The basic Analysis of Covariance Design (ANCOVA or ANACOVA) is a just pretest-posttest randomized experimental design. The notation shown here suggests that the pre-program measure is the same one as the post-program measure (otherwise we would use subscripts to distinguish the two), and so we would call this a pretest. But you should note that the pre-program measure doesn't have to be a pretest -- it can be any variable measured prior to the program intervention. It is also possible for a study to have more than one covariate.

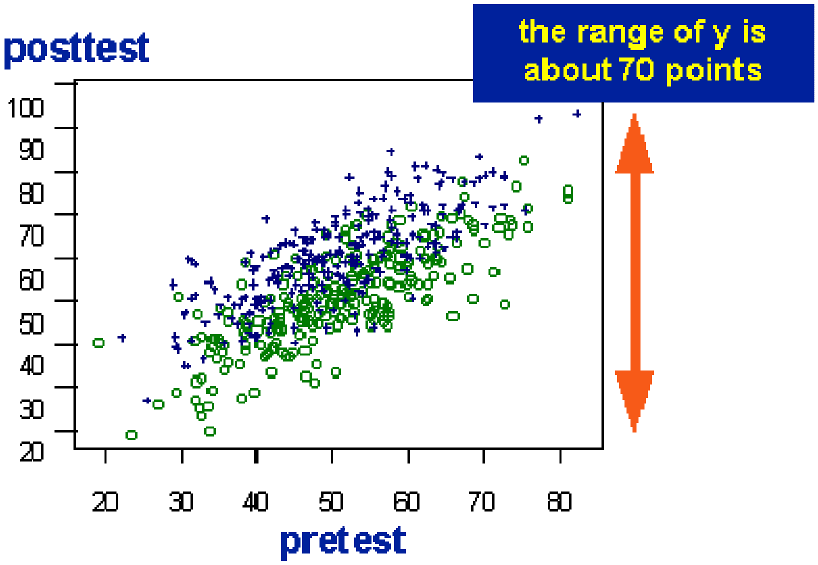
The pre-program measure or pretest is sometimes also called a "covariate" because of the way it's used in the data analysis -- we "covary" it with the outcome variable or posttest in order to remove variability or noise. Thus, the ANCOVA design falls in the class of a "noise reduction" experimental design (see Classifying the Experimental Designs).

In social research we frequently hear about statistical "adjustments" that attempt to control for important factors in our study. For instance, we might read that an analysis "examined posttest performance after **adjusting for** the income and educational level of the participants." In this case, "income" and "education level" are covariates. Covariates are the variables you "adjust for" in your study. Sometimes the language that will be used is that of "removing the effects" of one variable from another. For instance, we might read that an analysis "examined posttest performance after **removing the effect of** income and educational level of the participants."

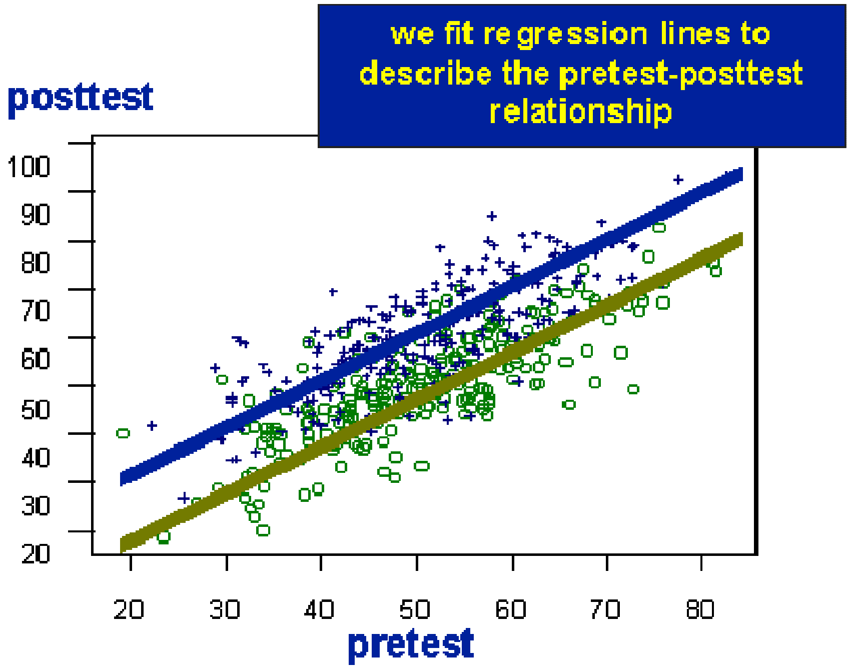
How Does A Covariate Reduce Noise?

One of the most important ideas in social research is how we make a statistical adjustment -- adjust one variable based on its covariance with another variable. If you understand this idea, you'll be well on your way to mastering social research. What I want to do here is to show you a series of graphs that illustrate pictorially what we mean by adjusting for a covariate.

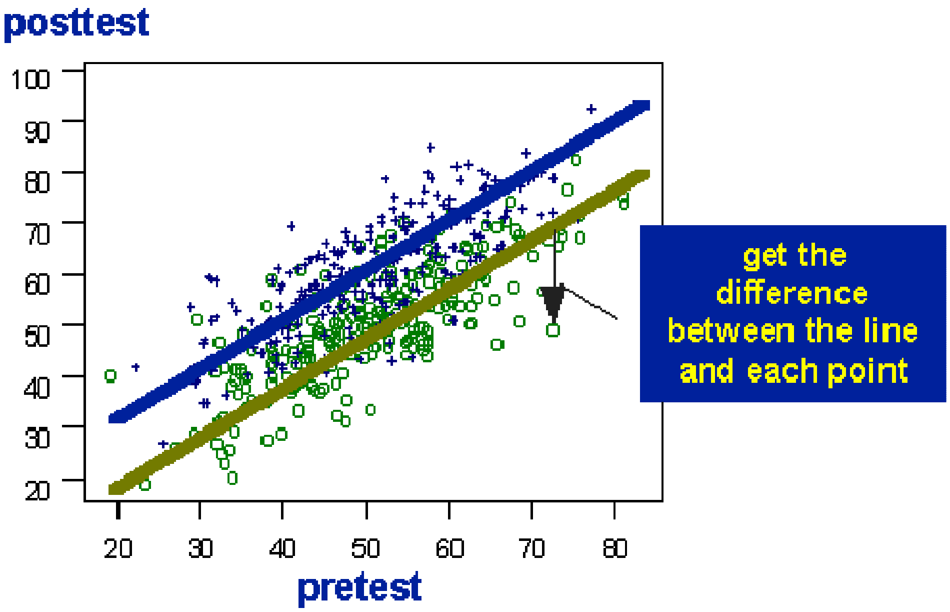
Let's begin with data from a simple ANCOVA design as described above. The first figure shows the pre-post bivariate distribution. Each "dot" on the graph represents the pretest and posttest score for an individual. We use an 'X' to signify a program or treated case and an 'O' to describe a control or comparison case. You should be able to see a few things immediately. First, you should be able to see a whopping treatment effect! It's so obvious that you don't even need statistical analysis to tell you whether there's an effect (although you may want to use statistics to estimate its size and probability). How do I know there's an effect? Look at any pretest value (value on the horizontal axis). Now, look up from that value -- you are looking up the posttest scale from lower to higher posttest scores. Do you see any pattern with respect to the groups? It should be obvious to you that the program cases (the 'X's) tend to score higher on the posttest at any given pretest value. Second, you should see that the posttest variability has a range of about 70 points.



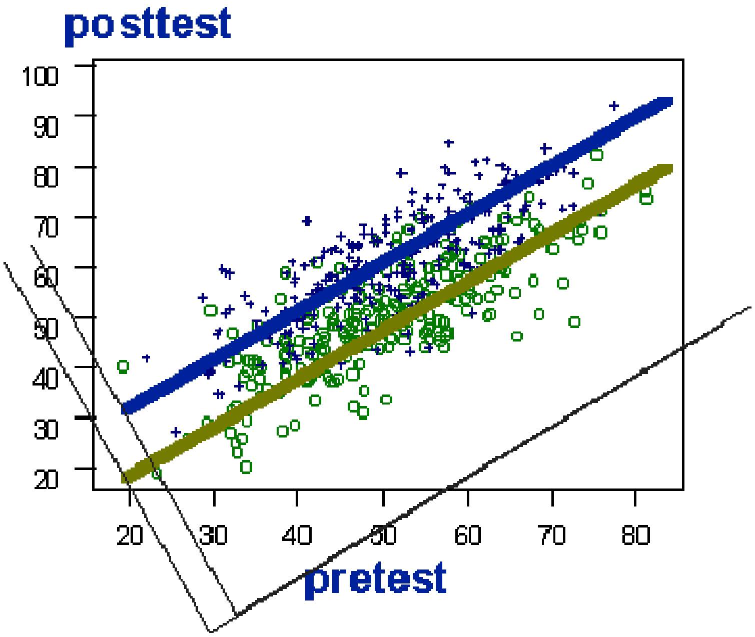
Now, let's fit some straight lines to the data. The lines on the graph are regression lines that describe the pre-post relationship for each of the groups. The regression line shows the expected posttest score for any pretest score. The treatment effect is even clearer with the regression lines. You should see that the line for the treated group is about 10 points higher than the line for the comparison group at any pretest value.



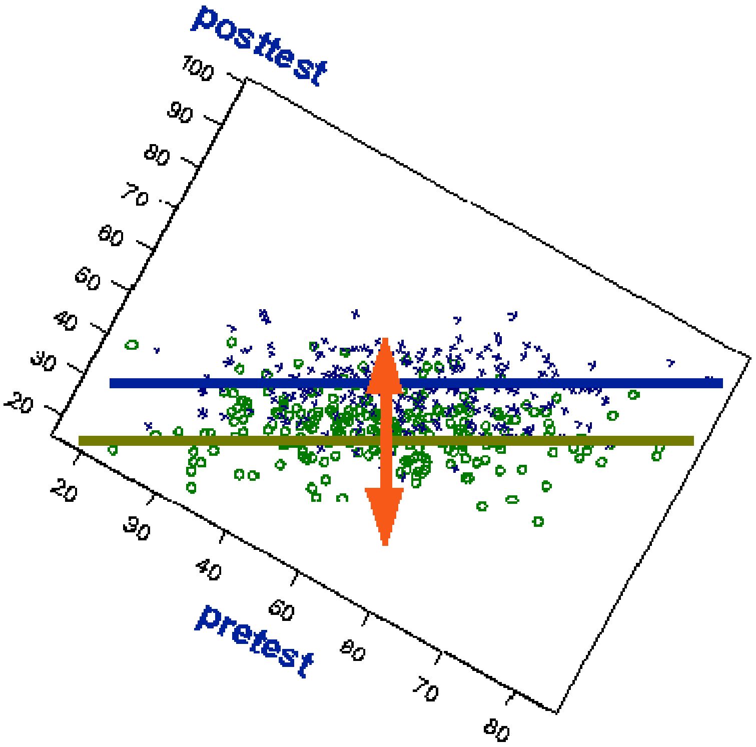
What we want to do is remove some of the variability in the posttest while preserving the difference between the groups. Or, in other terms, we want to "adjust" the posttest scores for pretest variability. In effect, we want to "subtract out" the pretest. You might think of this as subtracting the line from each group from the data for each group. How do we do that? Well, why don't we actually subtract?!? Find the posttest difference between the line for a group and each actual value. We call each of these differences a **residual** -- it's what's left over when you subtract a line from the data.



Now, here comes the tricky part. What does the data look like when we subtract out a line? You might think of it almost like turning the above graph clockwise until the regression lines are horizontal. The figures below show this in two steps. First, I construct and x-y axis system where the x dimension is parallel to the regression lines.

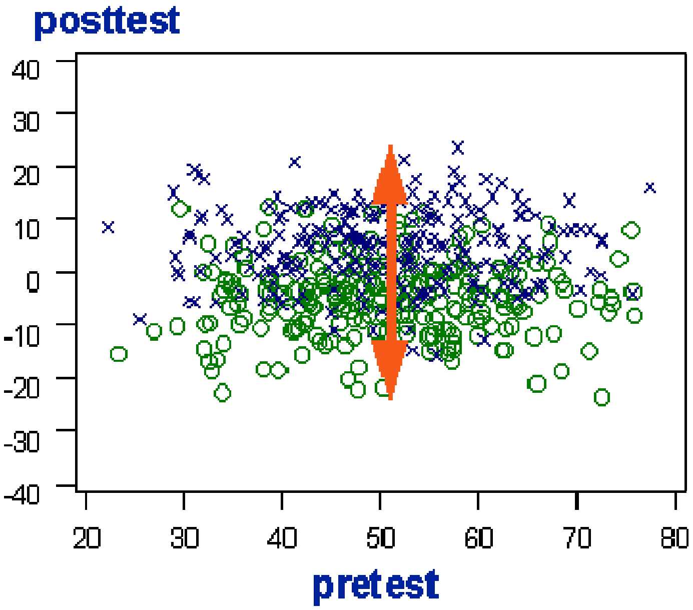


Then, I actually turn the graph clockwise so that the regression lines are now flat horizontally. Now, look at how big the posttest variability or range is in the figure (as indicated by the red double arrow). You should see that the range is considerably smaller that the 70 points we started out with above. You should also see that the difference between the lines is the same as it was before. So, we have in effect reduced posttest variability while maintaining the group difference. We've lowered the noise while keeping the signal at its original strength. The statistical adjustment procedure will result in a more efficient and more powerful estimate of the treatment effect.



You should also note the shape of the pre-post relationship. Essentially, the plot now looks like a zero correlation between the pretest and, in fact, it is. How do I know it's a zero correlation? Because any line that can be fitted through the data well would be horizontal. There's no slope or relationship. And, there shouldn't be. This graph shows the pre-post relationship *after we've removed the pretest*! If we've removed the pretest from the posttest there will be no pre-post correlation left.

Finally, let's redraw the axes to indicate that the pretest has been removed. here, the posttest values are the original posttest values minus the line (the predicted posttest values). That's why we see that the new posttest axis has 0 at it's center. Negative values on the posttest indicate that the original point fell below the regression line on the original axis. Here, we can better estimate that the posttest range is about 50 points instead of the original 70, even though the difference between the regression lines is the same. We've lowered the noise while retaining the signal.



[DISCLAIMER: OK, I know there's some statistical hot-shot out there fuming about the inaccuracy in my description above. My picture rotation is not exactly what we do when we adjust for a covariate. My description suggests that we drop perpendicular lines from the regression line to each point to obtain the subtracted difference. In fact, we drop lines that are perpendicular to the horizontal axis, not the regression line itself (in Least Squares regression we are minimizing the the sum of squares of the residuals on the dependent variable, not jointly on the independent and dependent variable). In any event, while my explanation may not be perfectly accurate from a statistical point of view, it's not very far off, and I think it conveys more clearly the idea of subtracting out a relationship. I thought I'd just put this disclaimer in to let you know I'm not dumb enough to believe that the description above is perfectly accurate.]

The adjustment for a covariate in the ANCOVA design is accomplished with the statistical analysis, not through rotation of graphs. See the Statistical Analysis of the Analysis of Covariance Design for details.

Summary

Some thoughts to conclude this topic. The ANCOVA design is a noise-reducing experimental design. It *"adjusts"* posttest scores for variability on the covariate (pretest). This is what we mean by *"adjusting"* for the effects of one variable on another in social research. You can use *any* continuous variable as a covariate, but the pretest is usually best. Why? Because the pretest is usually the variable that would be most highly correlated with the posttest (a variable should correlate highly with itself, shouldn't it?). Because it's so highly correlated, when you "subtract it out" or "remove' it, you're removing more extraneous variability from the posttest. The rule in selecting covariates is to select the measure(s) that correlate most highly with the outcome and, for multiple covariates, have little intercorrelation (otherwise, you're just adding in redundant covariates and you will actually lose precision by doing that). For example, you probably wouldn't want to use both gross and net income as two covariates in the same analysis because they are highly related and therefore redundant as adjustment variables.

- **Hybrid Experimental Designs**

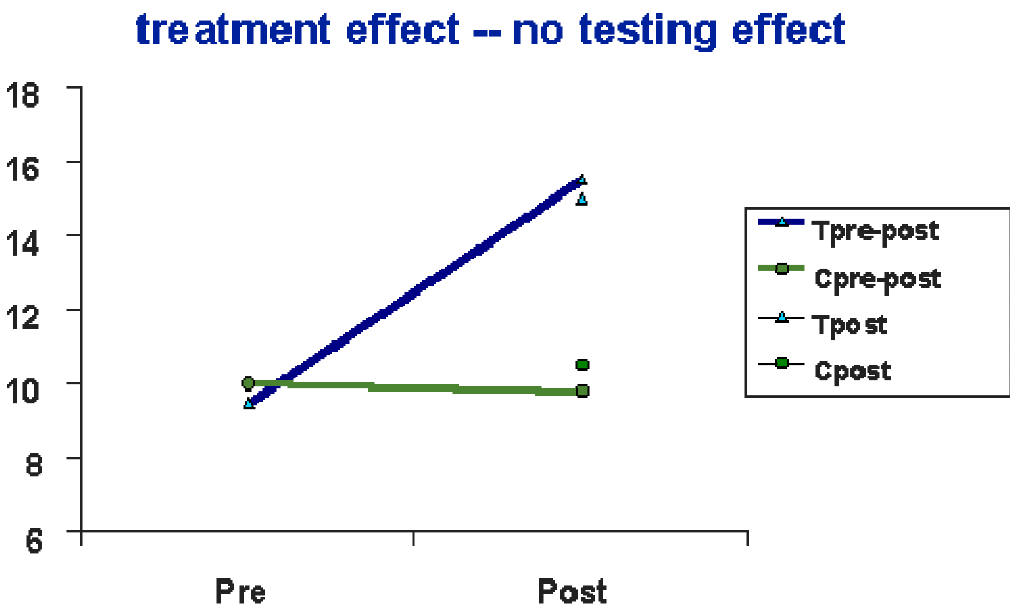
Hybrid experimental designs are just what the name implies -- new strains that are formed by combining features of more established designs. There are lots of variations that could be constructed from standard design features. Here, I'm going to introduce two hybrid designs. I'm featuring these because they illustrate especially well how a design can be constructed to address specific threats to internal validity.

The Solomon Four-Group Design

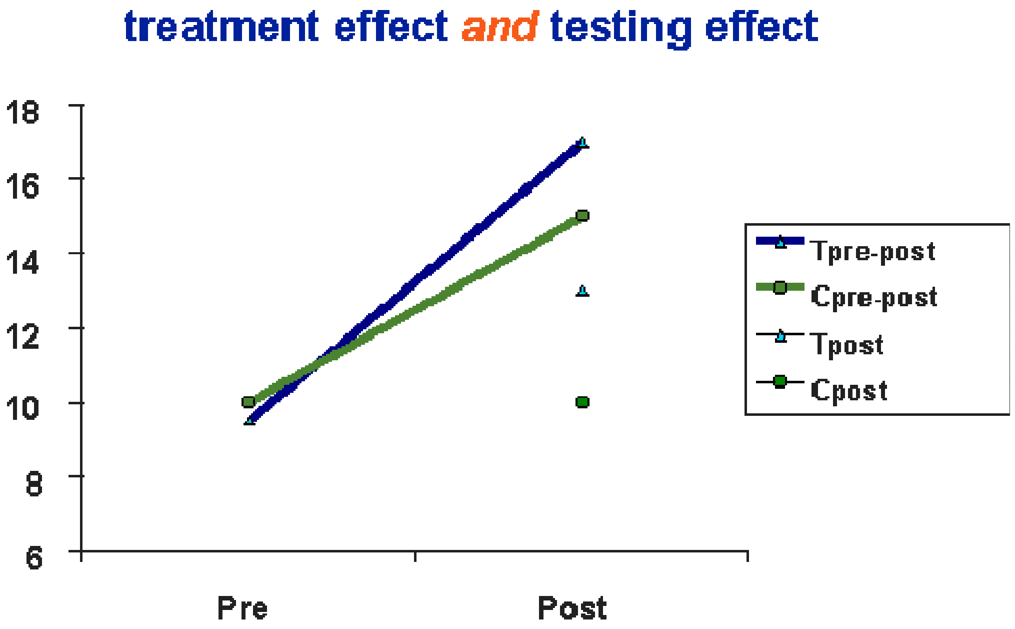
R	O	X	O
R	O		O
R		X	O
R			O

The **Solomon Four-Group Design** is designed to deal with a potential testing threat. Recall that a testing threat occurs when the act of taking a test affects how people score on a retest or posttest. The design notation is shown in the figure. It's probably not a big surprise that this design has four groups. Note that two of the groups receive the treatment and two do not. Further, two of the groups receive a pretest and two do not. One way to view this is as a 2x2 (Treatment Group X Measurement Group) factorial design. Within each treatment condition we have a group that is pretested and one that is not. By explicitly including testing as a factor in the design, we are able to assess experimentally whether a testing threat is operating.

Possible Outcomes. Let's look at a couple of possible outcomes from this design. The first outcome graph shows what the data might look like if there is a treatment or program effect and there is no testing threat. You need to be careful in interpreting this graph to note that there are six dots -- one to represent the average for each O in the design notation. To help you visually see the connection between the pretest and posttest average for the same group, a line is used to connect the dots. The two dots that are not connected by a line represent the two post-only groups. Look first at the two pretest means. They are close to each because the groups were randomly assigned. On the posttest, both treatment groups outscored both controls. Now, look at the posttest values. There appears to be no difference between the treatment groups, even though one got a pretest and the other did not. Similarly, the two control groups scored about the same on the posttest. Thus, the pretest did not appear to affect the outcome. But both treatment groups clearly outscored both controls. There is a main effect for the treatment.



Now, look at a result where there is evidence of a testing threat. In this outcome, the pretests are again equivalent (because the groups were randomly assigned). Each treatment group outscored it's comparable control group. The pre-post treatment outscored the pre-post control. And, the post-only treatment outscored the post-only control. These results indicate that there is a treatment effect. But here, both groups that had the pretest outscored their comparable non-pretest group. That's evidence for a testing threat.



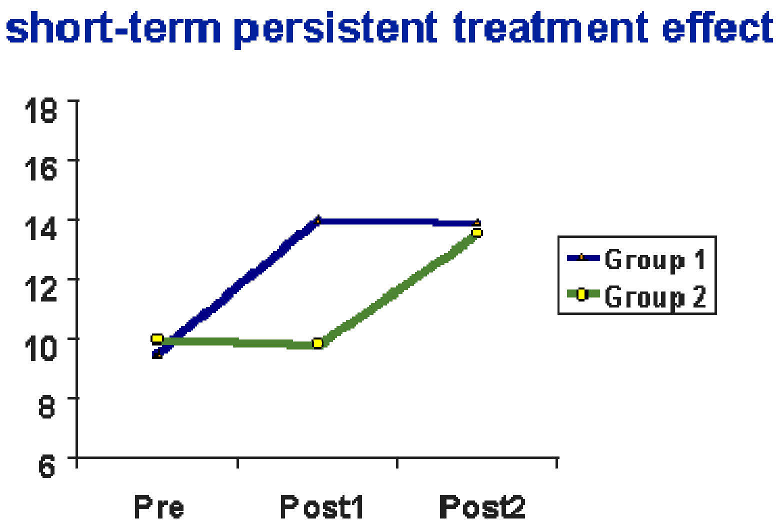
Switching Replications Design

R	O	X	O		O
R	O		O	X	O

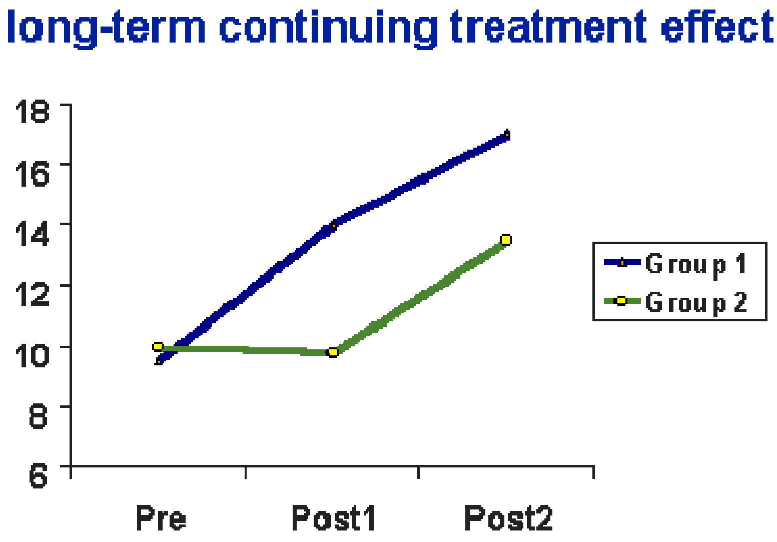
The **Switching Replications** design is one of the strongest of the experimental designs. And, when the circumstances are right for this design, it addresses one of the major problems in experimental designs -- the need to deny the program to some participants through random assignment. The design notation indicates that this is a two group design with three waves of measurement. You might think of this as two pre-post treatment-control designs grafted together. That is, the implementation of the treatment is repeated or ***replicated***. And in the repetition of the treatment, the two groups ***switch*** roles -- the original control group becomes the treatment group in phase 2 while the original treatment acts as the control. By the end of the study all participants have received the treatment.

The switching replications design is most feasible in organizational contexts where programs are repeated at regular intervals. For instance, it works especially well in schools that are on a semester system. All students are pretested at the beginning of the school year. During the first semester, Group 1 receives the treatment and during the second semester Group 2 gets it. The design also enhances organizational efficiency in resource allocation. Schools only need to allocate enough resources to give the program to half of the students at a time.

Possible Outcomes. Let's look at two possible outcomes. In the first example, we see that when the program is given to the first group, the recipients do better than the controls. In the second phase, when the program is given to the original controls, they "catch up" to the original program group. Thus, we have a converge, diverge, reconverge outcome pattern. We might expect a result like this when the program covers specific content that the students master in the short term and where we don't expect that they will continue getting better as a result.



Now, look at the other example result. During the first phase we see the same result as before -- the program group improves while the control does not. And, as before, during the second phase we see the original control group, now the program group, improve as much as did the first program group. But now, during phase two, the original program group continues to increase even though the program is no longer being given them. Why would this happen? It could happen in circumstances where the program has continuing and longer term effects. For instance, if the program focused on learning skills, students might continue to improve even after the formal program period because they continue to apply the skills and improve in them.



I said at the outset that both the Solomon Four-Group and the Switching Replications designs addressed specific threats to internal validity. It's obvious that the Solomon design addressed a testing threat. But what does the switching replications design address? Remember that in randomized experiments, especially when the groups are aware of each other, there is the potential for social threats -- compensatory rivalry, compensatory equalization and resentful demoralization are all likely to be present in educational contexts where programs are given to some students and not to others. The switching replications design helps mitigate these threats because it assures that everyone will eventually get the program. And, it allocates who gets the program first in the fairest possible manner, through the lottery of random assignment.

Quasi-Experimental Design

A quasi-experimental design is one that looks a bit like an experimental design but lacks the key ingredient -- random assignment. My mentor, Don Campbell, often referred to them as "queasy" experiments because they give the experimental purists a queasy feeling. With respect to internal validity, they often appear to be inferior to randomized experiments. But there is something compelling about these designs; taken as a group, they are easily more frequently implemented than their randomized cousins.

I'm not going to try to cover the quasi-experimental designs comprehensively. Instead, I'll present two of the classic quasi-experimental designs in some detail and show how we analyze them. Probably the most commonly used quasi-experimental design (and it may be the most commonly used of all designs) is the nonequivalent groups design. In its simplest form it requires a pretest and posttest for a treated and comparison group. It's identical to the Analysis of Covariance design except that the groups are not created through random assignment. You will see that the lack of random assignment, and the potential nonequivalence between the groups, complicates the statistical analysis of the nonequivalent groups design.

The second design I'll focus on is the regression-discontinuity design. I'm not including it just because I did my dissertation on it and wrote a book about it (although those were certainly factors weighing in its favor!). I include it because I believe it is an important and often misunderstood alternative to randomized experiments because its distinguishing characteristic -- assignment to treatment using a cutoff score on a pretreatment variable -- allows us to assign to the program those who need or deserve it most. At first glance, the regression discontinuity design strikes most people as biased because of regression to the mean. After all, we're assigning low scorers to one group and high scorers to the other. In the discussion of the statistical analysis of the regression discontinuity design, I'll show you why this isn't the case.

Finally, I'll briefly present an assortment of other quasi-experiments that have specific applicability or noteworthy features, including the Proxy Pretest Design, Double Pretest Design, Nonequivalent Dependent Variables Design, Pattern Matching Design, and the Regression Point Displacement design. I had the distinct honor of co-authoring a paper with Donald T. Campbell that first described the Regression Point Displacement Design. At the time of his death in Spring 1996, we had gone through about five drafts each over a five year period. The paper (click here for the entire paper) includes numerous examples of this newest of quasi-experiments, and provides a detailed description of the statistical analysis of the regression point displacement design.

There is one major class of quasi-experimental designs that are not included here -- the interrupted time series designs. I plan to include them in later rewrites of this material.

- **The Nonequivalent Groups Design**

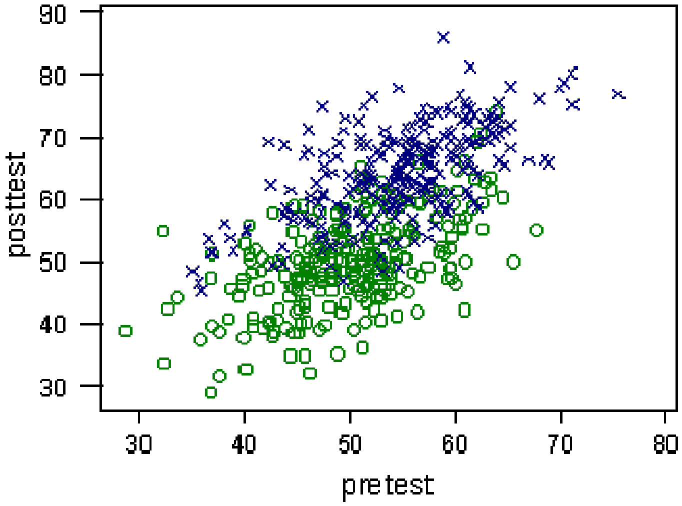
The Basic Design

N	O	X	O
N	O		O

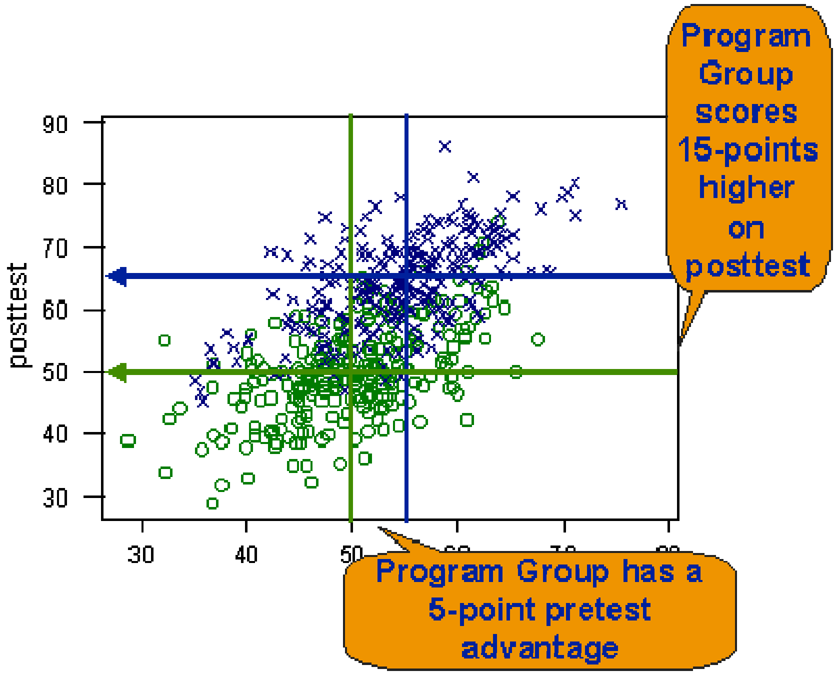
The Non-Equivalent Groups Design (hereafter NEGD) is probably the most frequently used design in social research. It is structured like a pretest-posttest randomized experiment, but it lacks the key feature of the randomized designs -- random assignment. In the NEGD, we most often use intact groups that we think are similar as the treatment and control groups. In education, we might pick two comparable classrooms or schools. In community-based research, we might use two similar communities. We try to select groups that are as similar as possible so we can fairly compare the treated one with the comparison one. But we can never be sure the groups are comparable. Or, put another way, it's unlikely that the two groups would be as similar as they would if we assigned them through a random lottery. Because it's often likely that the groups are not equivalent, this designed was named the **nonequivalent groups** design to remind us.

So, what does the term "nonequivalent" mean? In one sense, it just means that assignment to group was not random. In other words, the researcher did not control the assignment to groups through the mechanism of random assignment. As a result, the groups may be different prior to the study. That is, the NEGD is especially susceptible to the internal validity threat of selection. Any prior differences between the groups may affect the outcome of the study. Under the worst circumstances, this can lead us to conclude that our program didn't make a difference when in fact it did, or that it did make a difference when in fact it didn't.

The Bivariate Distribution



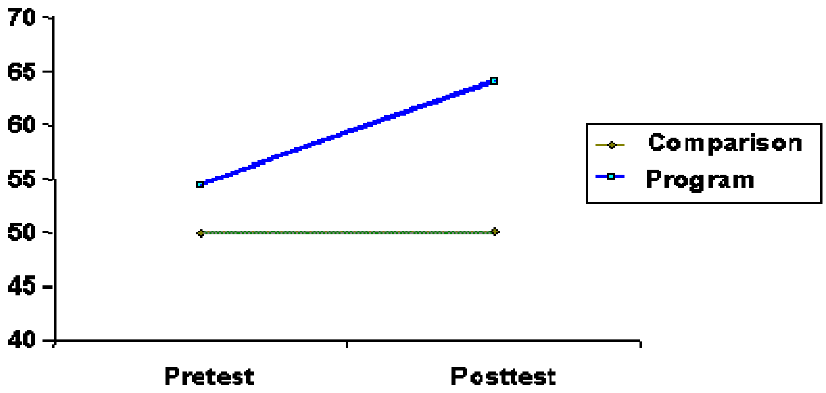
Let's begin our exploration of the NEGD by looking at some hypothetical results. The first figure shows a bivariate distribution in the simple pre-post, two group study. The **treated cases** are indicated with **Xs** while the **comparison cases** are indicated with **Os**. A couple of things should be obvious from the graph. To begin, we don't even need statistics to see that there is a whopping treatment effect (although statistics would help us estimate the size of that effect more precisely). The program cases (**Xs**) consistently score better on the posttest than the comparison cases (**Os**) do. If positive scores on the posttest are "better" then we can conclude that the program improved things. Second, in the NEGD the biggest threat to internal validity is selection -- that the groups differed before the program. Does that appear to be the case here? Although it may be harder to see, the program does appear to be a little further to the right on average. This suggests that they did have an initial advantage and that the positive results may be due in whole or in part to this initial difference.



We can see the initial difference, the selection bias, when we look at the next graph. It shows that the program group scored about five points higher than the comparison group on the pretest. The comparison group had a pretest average of about 50 while the program group averaged about 55. It also shows that the program group scored about fifteen points higher than the comparison group on the posttest. That is, the comparison group posttest score was again about 55, while this time the program group scored around 65. These observations suggest that there is a potential selection threat, although the initial five point difference doesn't explain why we observe a fifteen point difference on the posttest. It may be that there is still a legitimate treatment effect here, even given the initial advantage of the program group.

Possible Outcome #1

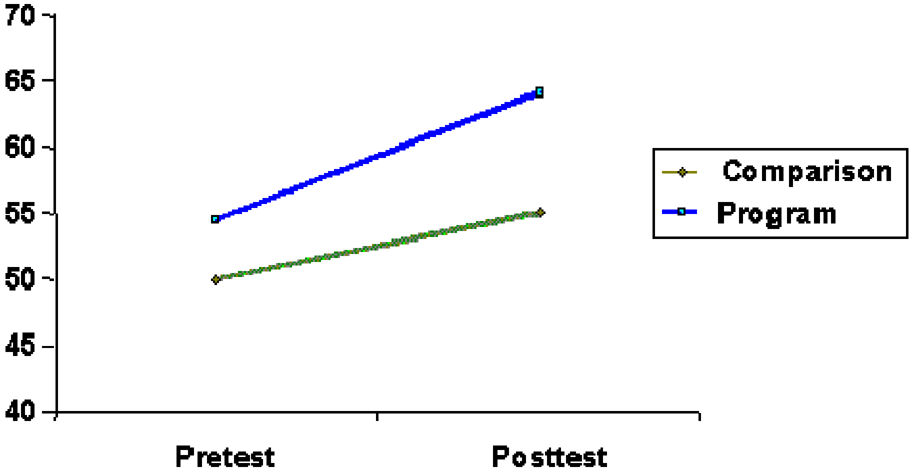
Let's take a look at several different possible outcomes from a NEGD to see how they might be interpreted. The important point here is that each of these outcomes has a different storyline. Some are more susceptible to treats to internal validity than others. Before you read through each of the descriptions, take a good look at the graph and try to figure out how you would explain the results. If you were a critic, what kinds of problems would you be looking for? Then, read the synopsis and see if it agrees with my perception.



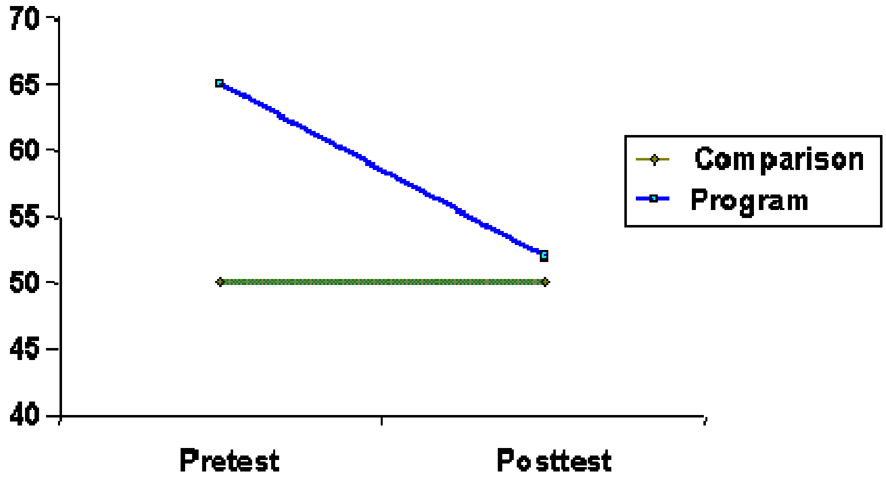
Sometimes it's useful to look at the means for the two groups. The figure shows these means with the pre-post means of the program group joined with a blue line and the pre-post means of the comparison group joined with a green one. This first outcome shows the situation in the two bivariate plots above. Here, we can see much more clearly both the original pretest difference of five points, and the larger fifteen point posttest difference.

How might we interpret these results? To begin, you need to recall that with the NEGD we are usually most concerned about selection threats. Which selection threats might be operating here? The key to understanding this outcome is that the comparison group did not change between the pretest and the posttest. Therefore, it would be hard to argue that that the outcome is due to a selection-maturation threat. Why? Remember that a selection-maturation threat means that the groups are maturing at different rates and that this creates the illusion of a program effect when there is not one. But because the comparison group didn't mature (i.e., change) at all, it's hard to argue that it was differential maturation that produced the outcome. What could have produced the outcome? A selection-history threat certainly seems plausible. Perhaps some event occurred (other than the program) that the program group reacted to and the comparison group didn't. Or, maybe a local event occurred for the program group but not for the comparison group. Notice how much more likely it is that outcome pattern #1 is caused by such a history threat than by a maturation difference. What about the possibility of selection-regression? This one actually works a lot like the selection-maturation threat If the jump in the program group is due to regression to the mean, it would have to be because the program group was below the overall population pretest average and, consequently, regressed upwards on the posttest. But if that's true, it should be even more the case for the comparison group who started with an even lower pretest average. The fact that they don't appear to regress at all helps rule out the possibility the outcome #1 is the result of regression to the mean.

Possible Outcome #2



Our second hypothetical outcome presents a very different picture. Here, both the program and comparison groups gain from pre to post, with the program group gaining at a slightly faster rate. This is almost the definition of a selection-maturation threat. The fact that the two groups differed to begin with suggests that they may already be maturing at different rates. And the posttest scores don't do anything to help rule that possibility out. This outcome might also arise from a selection-history threat. If the two groups, because of their initial differences, react differently to some historical event, we might obtain the outcome pattern shown. Both selection-testing and selection-instrumentation are also possibilities, depending on the nature of the measures used. This pattern could indicate a selection-mortality problem if there are more low-scoring program cases that drop out between testings. What about selection-regression? It doesn't seem likely, for much the same reasoning as for outcome #1. If there was an upwards regression to the mean from pre to post, we would expect that regression to be greater for the comparison group because they have the lower pretest score.

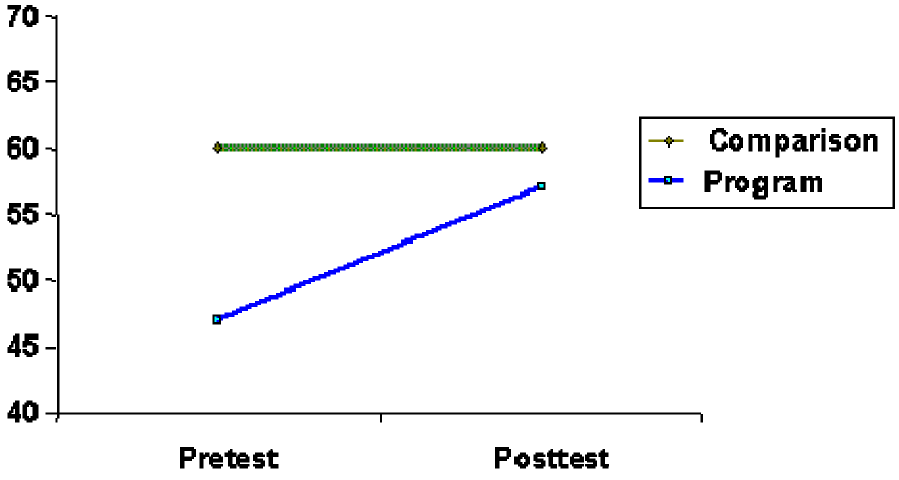


Possible Outcome #3

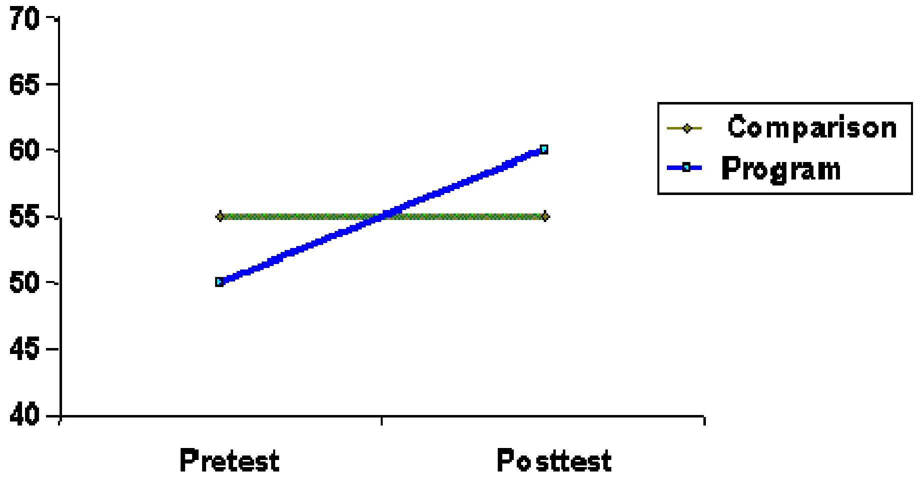
This third possible outcome cries out "selection-regression!" Or, at least it would if it could cry out. The regression scenario is that the program group was selected so that they were extremely high (relative to the population) on the pretest. The fact that they scored lower, approaching the comparison group on the posttest, may simply be due to their regressing toward the population mean. We might observe an outcome like this when we study the effects of giving a scholarship or an award for academic performance. We give the award because students did well (in this case, on the pretest). When we observe their posttest performance, relative to an "average" group of students, they appear to perform a more poorly. Pure regression! Notice how this outcome doesn't suggest a selection-maturation threat. What kind of maturation process would have to occur for the highly advantaged program group to decline while a comparison group evidences no change?

Possible Outcome #4

Our fourth possible outcome also suggests a selection-regression threat. Here, the program group is disadvantaged to begin with. The fact that they appear to pull closer to the program group on the posttest may be due to regression. This outcome pattern may



be suspected in studies of compensatory programs -- programs designed to help address some problem or deficiency. For instance, compensatory education programs are designed to help children who are doing poorly in some subject. They are likely to have lower pretest performance than more average comparison children. Consequently, they are likely to regress to the mean in much the pattern shown in outcome #4.



Possible Outcome #5

This last hypothetical outcome is sometimes referred to as a 'cross-over" pattern. Here, the comparison group doesn't appear to change from pre to post. But the program group does, starting out lower than the comparison group and ending up above them. This is the clearest pattern of evidence for the effectiveness of the program of all five of the hypothetical outcomes. It's hard to come up with a threat to internal validity that would be plausible here. Certainly, there is no evidence for selection maturation here unless you postulate that the two groups are involved in maturational processes that just tend to start and stop and just coincidentally you caught the program group maturing while the comparison group had gone dormant. But, if that was the case, why did the program group actually cross over the comparison group? Why didn't they approach the comparison group and stop maturing? How likely is this outcome as a description of normal maturation? Not very. Similarly, this isn't a selection-regression result. Regression might explain why a low scoring program group approaches the comparison group posttest score (as in outcome #4), but it doesn't explain why they cross over.

Although this fifth outcome is the strongest evidence for a program effect, you can't very well construct your study expecting to find this kind of pattern. It would be a little bit like saying "let's give our program to the toughest cases and see if we can improve them so much that they not only become like 'average' cases, but actually outperform them." That's an awfully big expectation to saddle any program with. Typically, you wouldn't want to subject your program to that kind of expectation. But if you happen to find that kind of result, you really have a program effect that has beat the odds.

- **The Regression-Discontinuity Design**

The regression-discontinuity design. What a terrible name! In everyday language both parts of the term have connotations that are primarily negative. To most people "regression" implies a reversion backwards or a return to some earlier, more primitive state while "discontinuity" suggests an unnatural jump or shift in what might otherwise be a smoother, more continuous process. To a research methodologist, however, the term regression-discontinuity (hereafter labeled "RD") carries no such negative meaning. Instead, the RD design is seen as a useful method for determining whether a program or treatment is effective.

The label "RD design" actually refers to a set of design variations. In its simplest most traditional form, the RD design is a pretest-posttest program-comparison group strategy. The unique characteristic which sets RD designs apart from other pre-post group designs is the method by which research participants are assigned to conditions. In RD designs, participants are assigned to program or comparison groups solely on the basis of a cutoff score on a pre-program measure. Thus the RD design is distinguished from randomized experiments (or randomized clinical trials) and from other quasi-experimental strategies by its unique method of assignment. This cutoff criterion implies the major advantage of RD designs -- they are appropriate when we wish to target a program or treatment to those who most need or deserve it. Thus, unlike its randomized or quasi-experimental alternatives, the RD design does not require us to assign potentially needy individuals to a no-program comparison group in order to evaluate the effectiveness of a program.

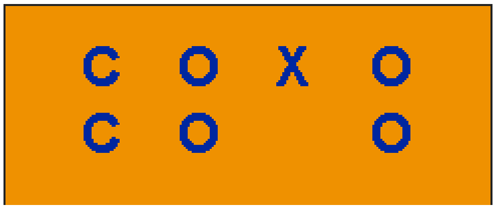
The RD design has not been used frequently in social research. The most common implementation has been in compensatory education evaluation where school children who obtain scores which fall below some predetermined cutoff value on an achievement test are assigned to remedial training designed to improve their performance. The low frequency of use may be attributable to several factors. Certainly, the design is a relative latecomer. Its first major field tests did not occur until the mid-1970s when it was incorporated into the nationwide evaluation system for compensatory education programs funded under Title I of the Elementary and Secondary Education Act (ESEA) of 1965. In many situations, the design has not been used because one or more key criteria were absent. For instance, RD designs force administrators to assign participants to conditions solely on the basis of quantitative indicators thereby often impalatably restricting the degree to which judgment, discretion or favoritism may be used. Perhaps the most telling reason for the lack of wider adoption of the RD design is that at first glance the design doesn't seem to make sense. In most research, we wish to have comparison groups that are equivalent to program groups on pre-program indicators so that post-program differences may be attributed to the program itself. But because of the cutoff criterion in RD designs, program and comparison groups are deliberately and maximally different on pre-program characteristics, an apparently insensible anomaly. An understanding of how the design actually works depends on at least a conceptual familiarity with regression analysis thereby making the strategy a difficult one to convey to nonstatistical audiences.

Despite its lack of use, the RD design has great potential for evaluation and program research. From a methodological point of view, inferences which are drawn from a well-implemented RD design are comparable in internal validity to conclusions from randomized experiments. Thus, the RD design is a strong competitor to randomized designs when causal hypotheses are being investigated. From an ethical perspective, RD designs are compatible with the goal of getting the program to those most in need. It is not necessary to deny the program from potentially deserving recipients simply for the sake of a scientific test. From an administrative viewpoint, the RD design is often directly usable with existing measurement efforts such as the regularly collected statistical information typical of most management information systems. The advantages of the RD design warrant greater educational efforts on the part of the methodological community to encourage its use where appropriate.

The Basic Design

The "basic" RD design is a pretest-posttest two group design. The term "pretest- posttest" implies that the same measure (or perhaps alternate forms of the same measure) is administered before and after some program or treatment. (In fact, the RD design does not require that the pre and post measures are the same.) The term "pretest" implies that the same measure is given twice while the term "pre-program" measure implies more broadly that before and after measures may be the same or different. It is assumed that a cutoff value on the pretest or pre-program measure is being used to assign persons or other units to the program. Two group versions of the RD design might imply either that some treatment or program is being contrasted with a no-program condition or that two alternative programs are being compared. The description of the basic design as a two group design implies that a single pretest cutoff score is used to assign participants to either the program or comparison group. The term "participants" refers to whatever unit is assigned. In many cases, participants are individuals, but they could be any definable units such as hospital wards, hospitals, counties, and so on. The term "program" will be used throughout to refer to any program, treatment or manipulation whose effects we wish to examine. In notational form, the basic RD design might be depicted as shown in the figure where:

- **C indicates that groups are assigned by means of a cutoff score,**
- an O stands for the administration of a measure to a group,
- an X depicts the implementation of a program,
- and each group is described on a single line (i.e., program group on top, control group on the bottom).



```
C  O  X  O
C  O     O
```

To make this initial presentation more concrete, we can imagine a hypothetical study where the interest is in examining the effect of a new treatment protocol for inpatients with a particular diagnosis. For simplicity, we can assume that we wish to try the new protocol on patients who are considered most ill and that for each patient we have a continuous quantitative indicator of health that is a composite rating which can take values from 1 to 100 where high scores indicate greater health. Furthermore, we can assume that a pretest cutoff score of 50 was (more or less arbitrarily) chosen as the assignment criterion or that all those scoring lower than 50 on the pretest are to be given the new treatment protocol while those with scores greater than or equal to 50 are given the standard treatment.

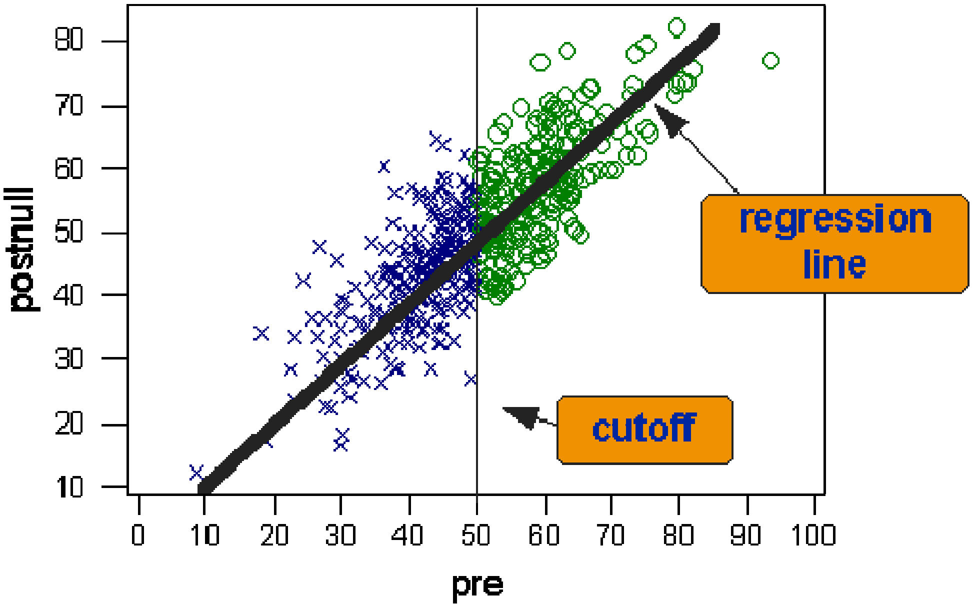


Figure 1. Pre-Post distribution with no treatment effect.

It is useful to begin by considering what the data might look like if we did not administer the treatment protocol but instead only measured all participants at two points in time. Figure 1 shows the hypothetical bivariate distribution for this situation. Each dot on the figure indicates a single person's pretest and posttest scores. The blue Xs to the left of the cutoff show the program cases. They are more severely ill on both the pretest and posttest. The green circles show the comparison group that is comparatively healthy on both measures. The vertical line at the pretest score of 50 indicates the cutoff point (although for Figure 1 we are assuming that no treatment has been given). The solid line through the bivariate distribution is the linear regression line. The distribution depicts a strong positive relationship between the pretest and posttest -- in general, the more healthy a person is at the pretest, the more healthy they'll be on the posttest, and, the more severely ill a person is at the pretest, the more ill they'll be on the posttest.

Now we can consider what the outcome might look like if the new treatment protocol is administered and has a positive effect. For simplicity, we will assume that the treatment had a constant effect which raised each treated person's health score by ten points. This is portrayed in Figure 2.

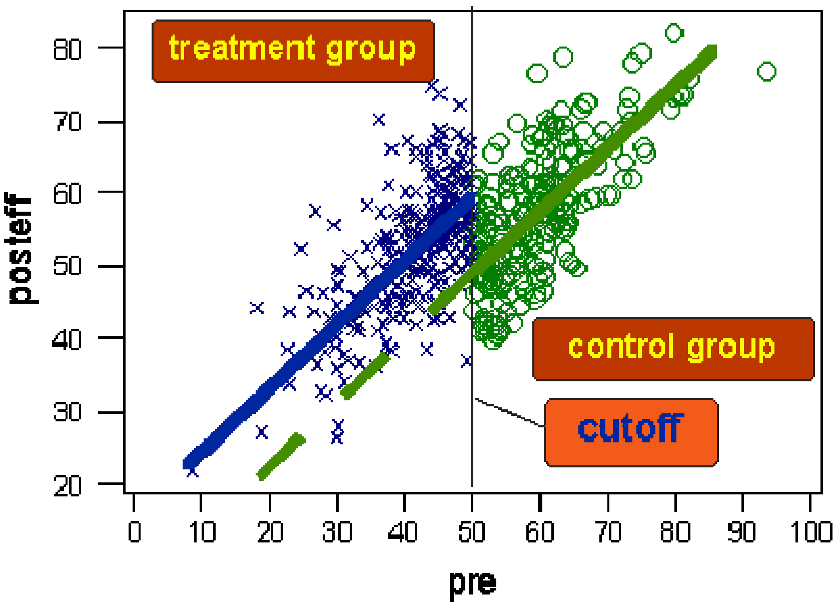


Figure 2. Regression-Discontinuity Design with Ten-point Treatment Effect.

Figure 2 is identical to Figure 1 except that all points to the left of the cutoff (i.e., the treatment group) have been raised by 10 points on the posttest. The dashed line in Figure 2 shows what we would expect the treated group's regression line to look like if the program had no effect (as was the case in Figure 1).

It is sometimes difficult to see the forest for the trees in these types of bivariate plots. So, let's remove the individual data points and look only at the regression lines. The plot of regression lines for the treatment effect case of Figure 2 is shown in Figure 3.

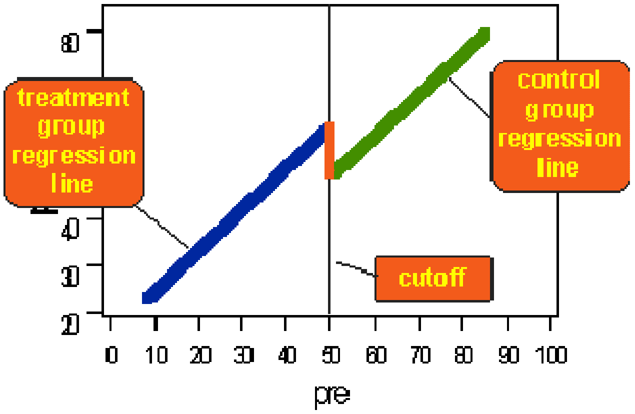


Figure 3. Regression lines for the data shown in Figure 2.

On the basis of Figure 3, we can now see how the RD design got its name - - a program effect is suggested when we observe a "**jump**" or **discontinuity** in the regression lines at the cutoff point. This is illustrated in Figure 4.

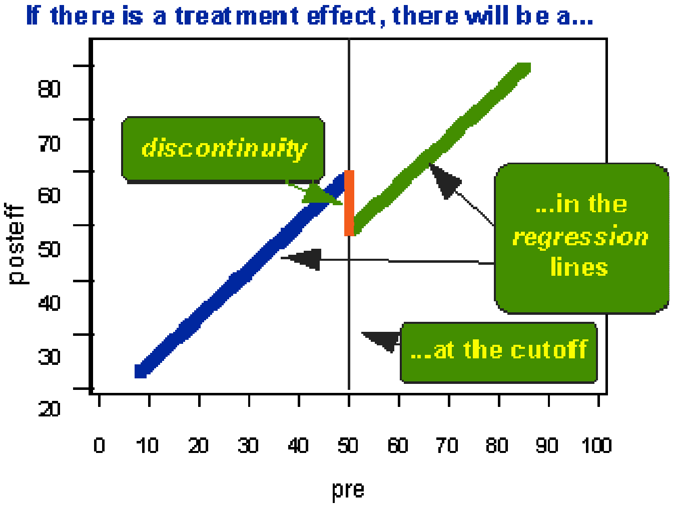


Figure 4. How the Regression-Discontinuity Design got its name.

The Logic of the RD Design

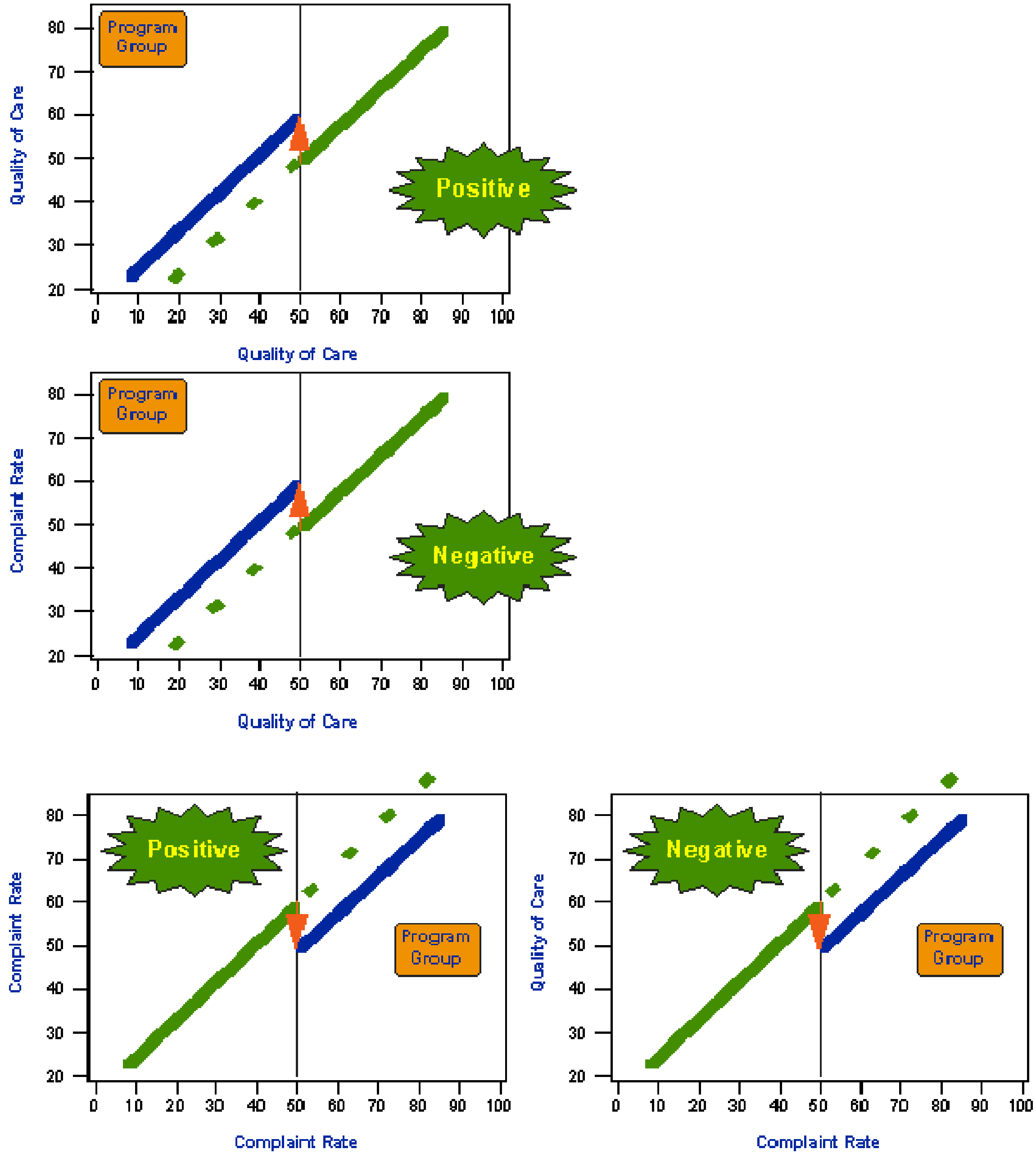
The discussion above indicates what the key feature of the RD design is: **assignment based on a cutoff value on a pre-program measure**. The cutoff rule for the simple two-group case is essentially:

- all persons on one side of the cutoff are assigned to one group...
- all persons on the other side of the cutoff are assigned to the other
- need a continuous quantitative pre-program measure

Selection of the Cutoff. The choice of cutoff value is usually based on one of two factors. It can be made solely on the basis of the program resources that are available. For instance, if a program only has the capability of handling 25 persons and 70 people apply, one can choose a cutoff point that distinguishes the 25 most needy persons from the rest. Alternatively, the cutoff can be chosen on substantive grounds. If the pre-program assignment measure is an indication of severity of illness measured on a 1 to 7 scale and physicians or other experts believe that all patients scoring 5 or more are critical and fit well the criteria defined for program participants then a cutoff value of 5 may be used.

Interpretation of Results.. In order to interpret the results of an RD design, one must know the nature of the assignment variable, who received the program and the nature of the outcome measure. Without this information, there is no distinct outcome pattern which directly indicates whether an effect is positive or negative.

To illustrate this, we can construct a new hypothetical example of an RD design. Let us assume that a hospital administrator would like to improve the quality of patient care through the institution of an intensive quality of care training program for staff. Because of financial constraints, the program is too costly to implement for all employees and so instead it will be administered to the entire staff from specifically targeted units or wards which seem most in need of improving quality of care. Two general measures of quality of care are available. The first is an aggregate rating of quality of care based on observation and rating by an administrative staff member and will be labeled here the QOC rating. The second is the ratio of the number of recorded patient complaints relative to the number of patients in the unit over a fixed period of time and will be termed here the Complaint Ratio. In this scenario, the administrator could use either the QOC rating or Complaint Ratio as the basis for assigning units to receive the training. Similarly, the effects of the training could be measured on either variable. Figure 5 shows four outcomes of alternative RD implementations possible under this scenario.



Only the regression lines are shown in the figure. It is worth noting that even though all four outcomes have the same pattern of regression lines, they do not imply the same result. In Figures

5a and 5b, hospital units were assigned to training because they scored *below* some cutoff score on the QOC rating. In Figures 5c and 5d units were given training because they scored *above* the cutoff score value on the Complaint Ratio measure. In each figure, the dashed line indicates the regression line we would expect to find for the training group if the training had no effect. This dashed line represents the no-discontinuity projection of the comparison group regression line into the region of the program group pretest scores.

We can clearly see that even though the outcome regression lines are the same in all four groups, we would interpret the four graphs differently. Figure 5a depicts a positive effect because training raised the program group regression line on the QOC rating over what would have been expected. Figure 5b however shows a negative effect because the program raised training group scores on the Complaint Ratio indicating increased complaint rates. In Figure 5c we see a positive effect because the regression line has been lowered on the Complaint Ratio relative to what we would have expected. Finally, Figure 5d shows a negative effect where the training resulted in lower QOC ratings than we would expect otherwise. The point here is a simple one. A discontinuity in regression lines indicates a program effect in the RD design. But the discontinuity alone is not sufficient to tell us whether the effect is positive or negative. In order to make this determination, we need to know who received the program and how to interpret the direction of scale values on the outcome measures.

The Role of the Comparison Group in RD Designs. With this introductory discussion of the design in mind, we can now see what constitutes the benchmark for comparison in the RD design. In experimental or other quasi- experimental designs we either assume or try to provide evidence that the program and comparison groups are equivalent prior to the program so that post-program differences can be attributed to the manipulation. The RD design involves no such assumption. Instead, with RD designs we assume that in the absence of the program the pre-post relationship would be equivalent for the two groups. Thus, the strength of the RD design is dependent on two major factors. The first is the assumption that there is no spurious discontinuity in the pre-post relationship which happens to coincide with the cutoff point. The second factor concerns the degree to which we can know and correctly model the pre-post relationship and constitutes the major problem in the statistical analysis of the RD design which will be discussed below.

The Internal Validity of the RD Design. Internal validity refers to whether one can infer that the treatment or program being investigated caused a change in outcome indicators. Internal validity as conceived is not concerned with our ability to generalize but rather focuses on whether a causal relationship can be demonstrated for the immediate research context. Research designs which address causal questions are often compared on their relative ability to yield internally valid results.

In most causal hypothesis tests, the central inferential question is whether any observed outcome differences between groups are attributable to the program or instead to some other factor. In order to argue for the internal validity of an inference, the analyst must attempt to demonstrate that the program -- and not some plausible alternative explanation -- is responsible for the effect. In the literature on internal validity, these plausible alternative explanations or factors are often termed "threats" to internal validity. A number of typical threats to internal validity have been identified. For instance, in a one-group pre-post study a gain from pretest to posttest may be attributable to the program or to other plausible factors such as historical events occurring between pretest and posttest, or natural maturation over time.

Many threats can be ruled out with the inclusion of a control group. Assuming that the control group is equivalent to the program group prior to the study, the control group pre-post gain will provide evidence for the change which should be attributed to all factors other than the program. A different rate of gain in the program group provides evidence for the relative effect of the program itself. Thus, we consider randomized experimental designs to be strong in internal validity because of our confidence in the probabilistic pre-program equivalence between groups which results from random assignment and helps assure that the control group will provide a legitimate reflection of all non-program factors that might affect outcomes.

In designs that do not use random assignment, the central internal validity concern revolves around the possibility that groups may not be equivalent prior to the program. We use the term "selection bias" to refer to the case where pre-program differences between groups are responsible for post-program differences. Any non-program factor which is differentially present across groups can constitute a selection bias or a selection threat to internal validity.

In RD designs, because of the deliberate pre-program differences between groups, there are several selection threats to internal validity which might, at first glance, appear to be a problem. For instance, a selection-maturation threat implies that different rates of maturation between groups might explain outcome differences. For the sake of argument, let's consider a pre-post distribution with a linear relationship having a slope equal to two units. This implies that on the average a person with a given pretest score will have a posttest score two times as high. Clearly there is maturation in this situation, that is, people are getting consistently higher scores over time. If a person has a pretest score of 10 units, we would predict a posttest score of 20 for an absolute gain of 10. But, if a person has a pretest score of 50 we would predict a posttest score of 100 for an absolute gain of 50. Thus the second person naturally gains or matures more in absolute units (although the rate of gain relative to the pretest score is constant). Along these lines, in the RD design we expect that all participants may mature and that in absolute terms this maturation may be different for the two groups on average. Nevertheless, a program effect in the RD design is not indicated by a difference between the posttest averages of the groups, but rather by a change in the pre-post relationship at the cutoff point. In this example, although we expect different absolute levels of maturation, a single continuous regression line with a slope equal to 2 would describe these different maturational rates. More to the point, in order for selection-maturation to be a threat to internal validity in RD designs, it must induce a discontinuity in the pre-post relationship which happens to coincide with the cutoff point -- an unlikely scenario in most studies.

Another selection threat to internal validity which might intuitively seem likely concerns the possibility of differential regression to the mean or a selection-regression threat. The phenomenon of regression to the mean arises when we asymmetrically sample groups from a distribution. On any subsequent measure the obtained sample group mean will be closer to the population mean for that measure (in standardized units) than the sample mean from the original distribution is to its population mean. In RD designs we deliberately create asymmetric samples and consequently expect regression towards the mean in both groups. In general we expect the low-scoring pretest group to evidence a relative gain on the posttest and the high-scoring pretest group to show a relative loss. As with selection-maturation, even though we expect to see differential regression to the mean this poses no problem for the internal validity of the RD design. We don't expect that regression to the mean will result in a discontinuity in the bivariate relationship coincidental with the cutoff point. In fact, the regression to the mean that will occur is expected to be continuous across the range of the pretest scores and is described by the regression line itself. (We should recall that the term "regression" was originally used by Galton to refer to the fact that a regression line describes regression to the mean.)

Although the RD design may initially seem susceptible to selection biases, it is not. The above discussion demonstrates that only factors that would naturally induce a discontinuity in the pre-post relationship could be considered threats to the internal validity of inferences from the RD design. In principle then the RD design is as strong in internal validity as its randomized experimental alternatives. In practice, however, the validity of the RD design depends directly on how well the analyst can model the true pre-post relationship, certainly a nontrivial statistical problem as is discussed in the statistical analysis of the regression-discontinuity design.

The RD Design and Accountability. It makes sense intuitively that the accountability of a program is largely dependent on the explicitness of the assignment or allocation of the program to recipients. Lawmakers and administrators need to recognize that programs are more evaluable and accountable when the allocation of the program is more public and verifiable. The three major pre-post designs -- the Pre-Post Randomized Experiments, the RD Design, and the Nonequivalent Groups Design -- are analogous to the three types of program allocation schemes which legislators or administrators might choose. Randomized experiments are analogous to the use of a lottery for allocating the program. RD designs can be considered explicit, accountable methods for assigning program recipients on the basis of need or merit. Nonequivalent group designs might be considered a type of political allocation because they enable the use of unverifiable, subjective or politically-motivated assignment. Most social programs are politically allocated. Even when programs are allocated primarily on the basis of need or merit, the regulatory agency usually reserves some discretionary capability in deciding who receives the program. Without debating the need for such discretion, it is clear that the methodological community should encourage administrators and legislators who wish their programs to be accountable to make explicit their criteria for program eligibility by either using probabilistically based lotteries or by relying on quantitative eligibility ratings and cutoff values as in the RD design. To the extent that legislators and administrators can be convinced to move toward more explicit assignment criteria, both the potential utility of the RD design and the accountability of the programs will be increased.

Ethics and the RD Design

The discussion above argues that the RD Design is strong in internal validity, certainly stronger than the Nonequivalent Groups Design, and perhaps as strong as the Randomized Experiments. But we know that the RD Designs are not as statistically powerful as the Randomized Experiments. That is, in order to achieve the same level of statistical accuracy, an RD Design needs as much as 2.75 times the participants as a randomized experiment. For instance, if a

Randomized Experiment needs 100 participants to achieve a certain level of power, the RD design might need as many as 275.

So why would we ever use the RD Design instead of a randomized one? The real allure of the RD Design is that it allows us to assign the treatment or program to those who most need or deserve it. Thus, the real attractiveness of the design is ethical -- we don't have to deny the program or treatment to participants who might need it as we do in randomized studies.

- **Other Quasi-Experimental Designs**

- The Proxy Pretest Design
- The Separate Pre-Post Samples Design
- The Double Pretest Design
- The Switching Replications Design
- The Nonequivalent Dependent Variables (NEDV) Design
- The Regression Point Displacement (RPD) Design

There are many different types of quasi-experimental designs that have a variety of applications in specific contexts. Here, I'll briefly present a number of the more interesting or important quasi-experimental designs. By studying the features of these designs, you can come to a deeper understanding of how to tailor design components to address threats to internal validity in your own research contexts.

The Proxy Pretest Design

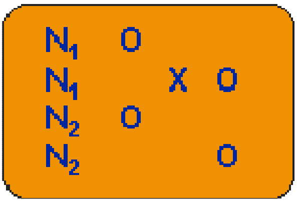
$$\begin{array}{cccc} N & O_1 & X & O_2 \\ N & O_1 & & O_2 \end{array}$$

The proxy pretest design looks like a standard pre-post design. But there's an important difference. The pretest in this design is collected after the program is given! But how can you call it a pretest if it's collected after the program? Because you use a "proxy" variable to estimate where the groups would have been on the pretest. There are essentially two variations of this design. In the first, you ask the participants to estimate where their pretest level would have been. This can be called the "Recollection" Proxy Pretest Design. For instance, you might ask participants to complete your measures "estimating how you would have answered the questions six months ago." This type of proxy pretest is not very good for estimating actual pre-post changes because people may forget where they were at some prior time or they may distort the pretest estimates to make themselves look better. However, there may be times when you are interested not so much in where they were on the pretest but rather in where they think they were. The recollection proxy pretest would be a sensible way to assess participants' perceived gain or change.

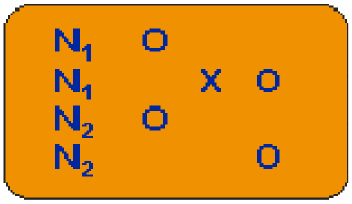
The other proxy pretest design uses archived records to stand in for the pretest. We might call this the "Archived" Proxy Pretest design. For instance, imagine that you are studying the effects of an educational program on the math performance of eighth graders. Unfortunately, you were brought in to do the study after the program had already been started (a too-frequent case, I'm afraid). You are able to construct a posttest that shows math ability after training, but you have no pretest. Under these circumstances, your best bet might be to find a proxy variable that would estimate pretest performance. For instance, you might use the student's grade point average in math from the seventh grade as the proxy pretest.

The proxy pretest design is not one you should ever select by choice. But, if you find yourself in a situation where you have to evaluate a program that has already begun, it may be the best you can do and would almost certainly be better than relying only on a posttest-only design.

The Separate Pre-Post Samples Design

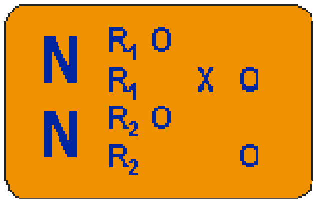

N_1 O
N_1 X O
N_2 O
N_2 O

The basic idea in this design (and its variations) is that the people you use for the pretest are not the same as the people you use for the posttest. Take a close look at the design notation for the first variation of this design. There are four groups (indicated by the four lines) but two of these groups come from a single nonequivalent group and the other two also come from a single nonequivalent group (indicated by the subscripts next to N). Imagine that you have two agencies or organizations that you think are similar. You want to implement your study in one agency and use the other as a control. The program you are looking at is an agency-wide one and you expect that the outcomes will be most noticeable at the agency level. For instance, let's say the program is designed to improve customer satisfaction. Because customers routinely cycle through your agency, you can't measure the same customers pre-post. Instead, you measure customer satisfaction in each agency at one point in time, implement your program, and then measure customer satisfaction in the agency at another point in time after the program. Notice that the customers will be different within each agency for the pre and posttest. This design is not a particularly strong one. Because you cannot match individual participant responses from pre to post, you can only look at the change in average customer satisfaction. Here, you always run the risk that you have nonequivalence not only between the agencies but that within agency the pre and post groups are nonequivalent. For instance, if you have different types of clients at different times of the year, this could bias the results. You could also look at this as having a proxy pretest on a different group of people.


N_1 O
N_1 X O
N_2 O
N_2 O

The second example of the separate pre-post sample design is shown in design notation at the right. Again, there are four groups in the study. This time, however, you are taking random samples from your agency or organization at each point in time. This is essentially the same design as above except for the random sampling. Probably the most sensible use of this design would be in situations where you routinely do sample surveys in an organization or community. For instance, let's assume that every year two similar communities do a community-wide survey of residents to ask about satisfaction with city services. Because of costs, you randomly sample each community each year. In one of the communities you decide to institute a program of community policing and you want to see whether residents feel safer and have changed in their attitudes towards police. You would use the results of last year's survey as the pretest in both communities, and this year's results as the posttest. Again, this is not a particularly strong design. Even though you are taking random samples from each community each year, it may still be the case that the community changes fundamentally from one year to the next and that the random samples within a community cannot be considered "equivalent."

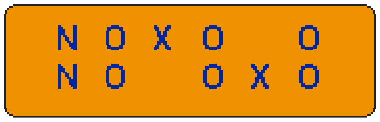
The Double Pretest Design



The Double Pretest is a very strong quasi-experimental design with respect to internal validity. Why? Recall that the Pre-Post Nonequivalent Groups Design (NEGD) is especially susceptible to selection threats to internal validity. In other words, the nonequivalent groups may be different in some way before the program is given and you may incorrectly attribute posttest differences to the program. Although the pretest helps to assess the degree of pre-program similarity, it does not tell us if the groups are changing at similar rates prior to the program. Thus, the NEGD is especially susceptible to selection-maturation threats.

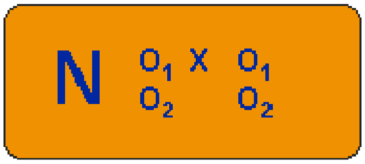
The double pretest design includes two measures prior to the program. Consequently, if the program and comparison group are maturing at different rates you should detect this as a change from pretest 1 to pretest 2. Therefore, this design explicitly controls for selection-maturation threats. The design is also sometimes referred to as a "dry run" quasi-experimental design because the double pretests simulate what would happen in the null case.

The Switching Replications Design



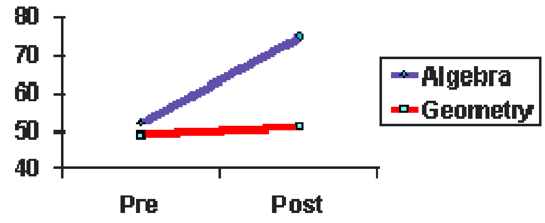
The Switching Replications quasi-experimental design is also very strong with respect to internal validity. And, because it allows for two independent implementations of the program, it may enhance external validity or generalizability. The design has two groups and three waves of measurement. In the first phase of the design, both groups are pretests, one is given the program and both are posttested. In the second phase of the design, the original comparison group is given the program while the original program group serves as the "control". This design is identical in structure to it's randomized experimental version, but lacks the random assignment to group. It is certainly superior to the simple pre-post nonequivalent groups design. In addition, because it assures that all participants eventually get the program, it is probably one of the most ethically feasible quasi-experiments.

The Nonequivalent Dependent Variables (NEDV) Design

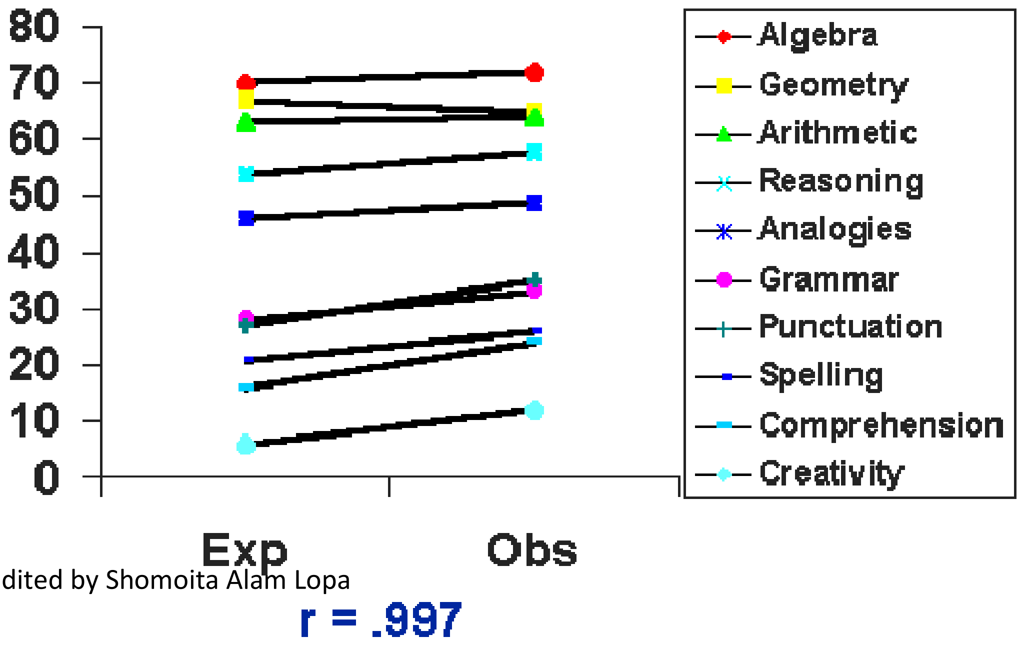


The Nonequivalent Dependent Variables (NEDV) Design is a deceptive one. In its simple form, it is an extremely weak design with respect to internal validity. But in its pattern matching variations, it opens the door to an entirely different approach to causal assessment that is extremely powerful. The design notation shown here is for the simple two-variable case. Notice that this design has only ***a single group of participants***! The two lines in the notation indicate separate variables, not separate groups.

The idea in this design is that you have a program designed to change a specific outcome. For instance, let's assume you are doing training in algebra for first-year high-school students. Your training program is designed to affect algebra scores. But it is not designed to affect geometry scores. And, pre-post geometry performance might be reasonably expected to be affected by other internally validity factors like history or maturation. In this case, the pre-post geometry performance acts like a control group -- it models what would likely have happened to the algebra pre-post scores if the program hadn't been given. The key is that the "control" variable has to be similar enough to the target variable to be affected in the same way by history, maturation, and the other single group internal validity threats, but not so similar that it is affected by the program. The figure shows the results we might get for our two-variable algebra-geometry example. Note that this design only works if the geometry variable is a reasonable proxy for what would have happened on the algebra scores in the absence of the program. The real allure of this design is the possibility that we don't need a control group -- we can give the program to all of our sample! The problem is that in its two-variable simple version, the assumption of the control variable is a difficult one to meet. (Note that a double-pretest version of this design would be considerably stronger).



The Pattern Matching NEDV Design. Although the two-variable NEDV design is quite weak, we can make it considerably stronger by adding multiple outcome variables. In this variation, we need many outcome variables and a theory that tells *how affected* (from most to least) each variable will be by the program. Let's reconsider the example of our algebra program above. Now, instead of having only an algebra and geometry score, we have ten measures that we collect pre and post. We expect that the algebra measure would be most affected by the program (because that's what the program was most designed to affect). But here, we recognize that geometry might also be affected because training in algebra might be relevant, at least tangentially, to geometry skills. On the other hand, we might theorize that creativity would be much less affected, even indirectly, by training in algebra and so our creativity measure is predicted to be least affected of the ten measures.



Now, let's line up our theoretical expectations against our pre-post gains for each variable. The graph we'll use is called a "ladder graph" because if there is a correspondence between expectations and observed results we'll get horizontal lines and a figure that looks a bit like a ladder. You can see in the figure that the expected order of outcomes (on the left) are mirrored well in the actual outcomes (on the right).

Depending on the circumstances, the Pattern Matching NEDV design can be quite strong with respect to internal validity. In general, the design is stronger if you have a larger set of variables and you find that your expectation pattern matches well with the observed results. What are the threats to internal validity in this design? Only a factor (e.g., an historical event or maturational pattern) that would yield the same outcome pattern can act as an alternative explanation. And, the more complex the predicted pattern, the less likely it is that some other factor would yield it. The problem is, the more complex the predicted pattern, the less likely it is that you will find it matches to your observed data as well.

The Pattern Matching NEDV design is especially attractive for several reasons. It requires that the researcher specify expectations prior to institution of the program. Doing so can be a sobering experience. Often we make naive assumptions about how our programs or interventions will work. When we're forced to look at them in detail, we begin to see that our assumptions may be unrealistic. The design also requires a detailed measurement net -- a large set of outcome variables and a detailed sense of how they are related to each other. Developing this level of detail about your measurement constructs is liable to improve the construct validity of your study. Increasingly, we have methodologies that can help researchers empirically develop construct networks that describe the expected interrelationships among outcome variables (see Concept Mapping for more information about how to do this). Finally, the Pattern Matching NEDV is especially intriguing because it suggests that it is possible to assess the effects of programs even if you only have a treated group. Assuming the other conditions for the design are met, control groups are not necessarily needed for causal assessment. Of course, you can also couple the Pattern Matching NEDV design with standard experimental or quasi-experimental control group designs for even more enhanced validity. And, if your experimental or quasi-experimental design already has many outcome measures as part of the measurement protocol, the design might be considerably enriched by generating variable-level expectations about program outcomes and testing the match statistically.

One of my favorite questions to my statistician friends goes to the heart of the potential of the Pattern Matching NEDV design. "Suppose," I ask them, "that you have ten outcome variables in a study and that you find that all ten show no statistically significant treatment effects when tested individually (or even when tested as a multivariate set). And suppose, like the desperate graduate student who finds in their initial analysis that nothing is significant that you decide to look at the direction of the effects across the ten variables. You line up the variables in terms of which should be most to least affected by your program. And, miracle of miracles, you find that there is a strong and statistically significant correlation between the expected and observed *order* of effects even though no individual effect was statistically significant. Is this finding interpretable as a treatment effect?" My answer is "yes." I think the graduate student's desperation-driven intuition to look at order of effects is a sensible one. I would conclude that the reason you did not find statistical effects on the individual variables is that you didn't have sufficient statistical power. Of course, the results will only be interpretable as a treatment effect if you can rule out any other plausible factor that could have caused the ordering of outcomes.

But the more detailed the predicted pattern and the stronger the correlation to observed results, the more likely the treatment effect becomes the most plausible explanation. In such cases, the expected pattern of results is like a unique fingerprint -- and the observed pattern that matches it can only be due to that unique source pattern.

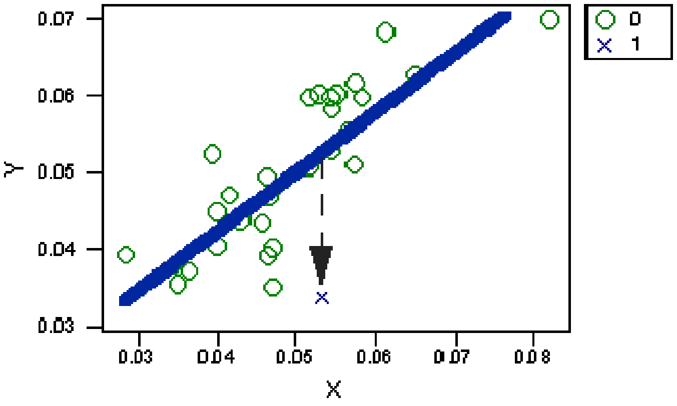
I believe that the pattern matching notion implicit in the NEDV design opens the way to an entirely different approach to causal assessment, one that is closely linked to detailed prior explication of the program and to detailed mapping of constructs. It suggests a much richer model for causal assessment than one that relies only on a simplistic dichotomous treatment-control model. In fact, I'm so convinced of the importance of this idea that I've staked a major part of my career on developing pattern matching models for conducting research!

The Regression Point Displacement (RPD) Design

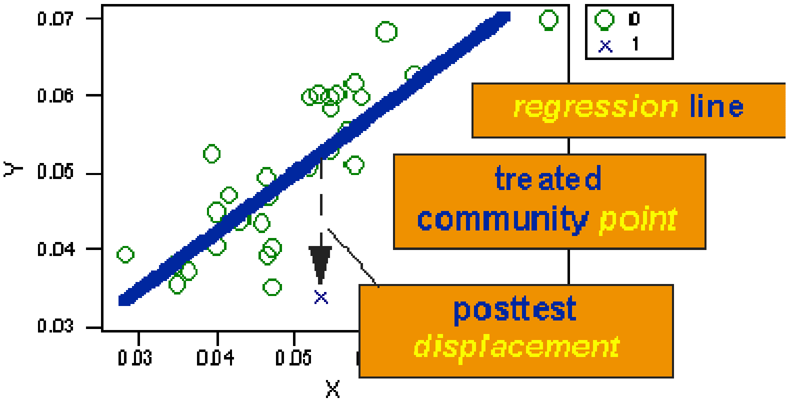
$N_{(n=1)}$	O	X	O
N	O		O

The Regression Point Displacement (RPD) design is a simple quasi-experimental strategy that has important implications, especially for community-based research. The problem with community-level interventions is that it is difficult to do causal assessment, to determine if your program made a difference as opposed to other potential factors. Typically, in community-level interventions, program costs preclude our implementing the program in more than one community. We look at pre-post indicators for the program community and see whether there is a change. If we're relatively enlightened, we seek out another similar community and use it as a comparison. But, because the intervention is at the community level, we only have a single "unit" of measurement for our program and comparison groups.

The RPD design attempts to enhance the single program unit situation by comparing the performance on that single unit with the performance of a large set of comparison units. In community research, we would compare the pre-post results for the intervention community with a large set of other communities. The advantage of doing this is that we don't rely on a single nonequivalent community, we attempt to use results from a heterogeneous set of nonequivalent communities to model the comparison condition, and then compare our single site to this model. For typical community-based research, such an approach may greatly enhance our ability to make causal inferences.



I'll illustrate the RPD design with an example of a community-based AIDS education program. We decide to pilot our new AIDS education program in one particular community in a state, perhaps a county. The state routinely publishes annual HIV positive rates by county for the entire state. So, we use the remaining counties in the state as control counties. But instead of averaging all of the control counties to obtain a single control score, we use them as separate units in the analysis. The first figure shows the bivariate pre-post distribution of HIV positive rates per 1000 people for all the counties in the state. The program county -- the one that gets the AIDS education program -- is shown as an X and the remaining control counties are shown as Os. We compute a regression line for the control cases (shown in blue on the figure). The regression line models our predicted outcome for a count with any specific pretest rate. To estimate the effect of the program we test whether the displacement of the program county from the control county regression line is statistically significant.

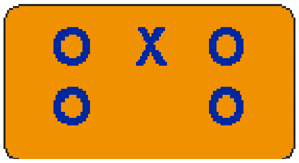


The second figure shows why the RPD design was given its name. In this design, we know we have a treatment effect when there is a significant **displacement** of the program **point** from the control group **regression** line.

The RPD design is especially applicable in situations where a treatment or program is applied in a single geographical unit (e.g., a state, county, city, hospital, hospital unit) instead of an individual, where there are lots of other units available as control cases, and where there is routine measurement (e.g., monthly, annually) of relevant outcome variables.

The analysis of the RPD design turns out to be a variation of the Analysis of Covariance model (see the Statistical Analysis of the Regression Point Displacement Design). I had the opportunity to be the co-developer with Donald T. Campbell of the RPD design. You can view the entire original paper entitled " The Regression Point Displacement Design for Evaluating Community-Based Pilot Programs and Demonstration Projects."

Relationships Among Pre-Post Designs



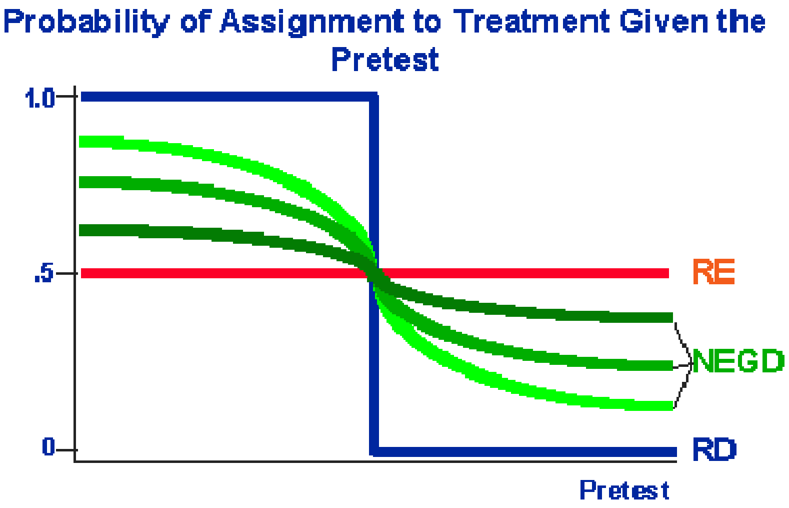
There are three major types of pre-post program-comparison group designs all sharing the basic design structure shown in the notation above:

- The Randomized Experimental (RE) Design
- The Nonequivalent Group (NEGD) Design
- The Regression-Discontinuity (RD) Design

The designs differ in the method by which participants are assigned to the two groups. In the RE, participants are assigned randomly. In the RD design, they are assigned using a cutoff score on the pretest. In the NEGD, assignment of participants is not explicitly controlled -- they may self select into either group, or other unknown or unspecified factors may determine assignment.

Because these three designs differ so critically in their assignment strategy, they are often considered distinct or unrelated. But it is useful to look at them as forming a continuum, both in terms of assignment and in terms of their strength with respect to internal validity.

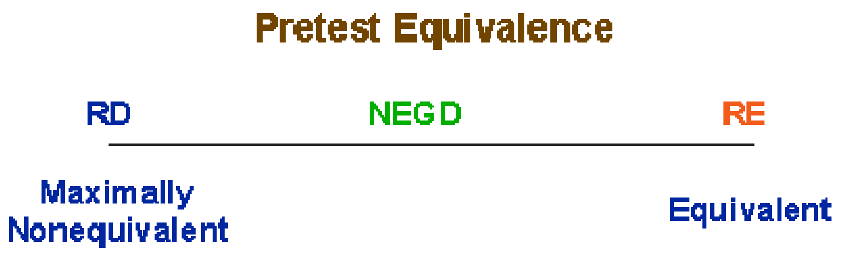
We can look at the similarity three designs in terms of their assignment by graphing their assignment functions with respect to the pretest variable. In the figure, the vertical axis is the probability that a specific unit (e.g., person) will be assigned to the treatment group). These values, because they are probabilities, range from 0 to 1. The horizontal axis is an idealized pretest score.



Let's first examine the assignment function for the simple pre-post randomized experiment. Because units are assigned randomly, we know that the probability that a unit will be assigned to the treatment group is always 1/2 or .5 (assuming equal assignment probabilities are used). This function is indicated by the horizontal red line at .5 in the figure. For the RD design, we arbitrarily set the cutoff value at the midpoint of the pretest variable and assume that we assign units scoring below that value to the treatment and those scoring at or above that value to the control condition (the arguments made here would generalize to the case of high-scoring treatment cases as well). In this case, the assignment function is a simple step function, with the probability of assignment to the treatment = 1 for the pretest scores below the cutoff and = 0 for those above. It is important to note that for both the RE and RD designs it is an easy matter to plot their assignment functions because assignment is explicitly controlled. This is not the case for the NEGD. Here, the idealized assignment function differs depending on the degree to which the groups are nonequivalent on the pretest. If they are extremely nonequivalent (with the treatment group scoring lower on the pretest), the assignment function would approach the step function of the RD design. If the groups are hardly nonequivalent at all, the function would approach the flat-line function of the randomized experiment.

The graph of assignment functions points an important issue about the relationships among these designs -- the designs are not distinct with respect to their assignment functions, they form a continuum. On one end of the continuum is the RE design and at the other is the RD. The NEGD can be viewed as a degraded RD or RE depending on whether the assignment function more closely approximates one or the other.

We can also view the designs on a continuum with respect to the degree to which they generate a pretest difference between the groups.



The figure shows that the RD design induces the maximum possible pretest difference. The RE design induces the smallest pretest difference (the most equivalent). The NEGD fills in the gap between these two extreme cases. If the groups are very nonequivalent, the design is closer to the RD design. If they're very similar, it's closer to the RE design.

Finally, we can also distinguish the three designs in terms of the *a priori* knowledge they give about assignment. It should be clear that in the RE design we know perfectly the probability of assignment to treatment -- it is .5 for each participant. Similarly, with the RD design we also know perfectly the probability of assignment. In this case it is precisely dependent on the cutoff assignment rule. It is dependent on the pretest where the RE design is not. In both these designs, we know the assignment function perfectly, and it is this knowledge that enables us to obtain unbiased estimates of the treatment effect with these designs. This is why we conclude that, with respect to internal validity, the RD design is as strong as the RE design. With the NEGD however, we do not know the assignment function perfectly. Because of this, we need to model this function either directly or indirectly (e.g., through reliability corrections).

The major point is that we should not look at these three designs as entirely distinct. They are related by the nature of their assignment functions and the degree of pretest nonequivalence between groups. This continuum has important implications for understanding the statistical analyses of these designs.

Designing Designs for Research

Reprinted from Trochim, W. and Land, D. (1982). Designing Designs for Research. The Researcher, 1, 1, 1-6.

Much contemporary social research is devoted to examining whether a program, treatment, or manipulation causes some outcome or result. For example, we might wish to know whether a new educational program causes subsequent achievement score gains, whether a special work release program for prisoners causes lower recidivism rates, whether a novel drug causes a reduction in symptoms, and so on. Cook and Campbell (1979) argue that three conditions must be met before we can infer that such a cause-effect relation exists:

1. **Covariation.** Changes in the presumed cause must be related to changes in the presumed effect. Thus, if we introduce, remove, or change the level of a treatment or program, we should observe some change in the outcome measures.
2. **Temporal Precedence.** The presumed cause must occur prior to the presumed effect.
3. **No Plausible Alternative Explanations.** The presumed cause must be the only reasonable explanation for changes in the outcome measures. If there are other factors which could be responsible for changes in the outcome measures we cannot be confident that the presumed cause-effect relationship is correct.

In most social research the third condition is the most difficult to meet. Any number of factors other than the treatment or program could cause changes in outcome measures. Campbell and Stanley (1966) and later, Cook and Campbell (1979) list a number of common plausible alternative explanations (or, threats to internal validity). For example, it may be that some historical event which occurs at the same time that the program or treatment is instituted was responsible for the change in the outcome measures; or, changes in record keeping or measurement systems which occur at the same time as the program might be falsely attributed to the program. The reader is referred to standard research methods texts for more detailed discussions of threats to validity.

This paper is primarily heuristic in purpose. Standard social science methodology textbooks (Cook and Campbell 1979; Judd and Kenny, 1981) typically present an array of research designs and the alternative explanations which these designs rule out or minimize. This tends to foster a "cookbook" approach to research design - an emphasis on the selection of an available design rather than on the construction of an appropriate research strategy. While standard designs may sometimes fit real-life situations, it will often be necessary to "tailor" a research design to minimize specific threats to validity. Furthermore, even if standard textbook designs are used, an understanding of the logic of design construction in general will improve the comprehension of these standard approaches. This paper takes a structural approach to research design. While this is by no means the only strategy for constructing research designs, it helps to clarify some of the basic principles of design logic.

Minimizing Threats to Validity

Good research designs minimize the plausible alternative explanations for the hypothesized cause-effect relationship. But such explanations may be ruled out or minimized in a number of ways other than by design. The discussion which follows outlines five ways to minimize threats to validity, one of which is by research design:

1. **By Argument.** The most straightforward way to rule out a potential threat to validity is to simply argue that the threat in question is not a reasonable one. Such an argument may be made either *a priori* or *a posteriori*, although the former will usually be more convincing than the latter. For example, depending on the situation, one might argue that an instrumentation threat is not likely because the same test is used for pre and post test measurements and did not involve observers who might improve, or other such factors. In most cases, ruling out a potential threat to validity by argument alone will be weaker than the other approaches listed below. As a result, the most plausible threats in a study should not, except in unusual cases, be ruled out by argument only.
2. **By Measurement or Observation.** In some cases it will be possible to rule out a threat by measuring it and demonstrating that either it does not occur at all or occurs so minimally as to not be a strong alternative explanation for the cause-effect relationship. Consider, for example, a study of the effects of an advertising campaign on subsequent sales of a particular product. In such a study, history (i.e., the occurrence of other events which might lead to an increased desire to purchase the product) would be a plausible alternative explanation. For example, a change in the local economy, the removal of a competing product from the market, or similar events could cause an increase in product sales. One might attempt to minimize such threats by measuring local economic indicators and the availability and sales of competing products. If there is no change in these measures coincident with the onset of the advertising campaign, these threats would be considerably minimized. Similarly, if one is studying the effects of special mathematics training on math achievement scores of children, it might be useful to observe everyday classroom behavior in order to verify that students were not receiving any additional math training to that provided in the study.
3. **By Design.** Here, the major emphasis is on ruling out alternative explanations by adding treatment or control groups, waves of measurement, and the like. This topic will be discussed in more detail below.
4. **By Analysis.** There are a number of ways to rule out alternative explanations using statistical analysis. One interesting example is provided by Jurs and Glass (1971). They suggest that one could study the plausibility of an attrition or mortality threat by conducting a two-way analysis of variance. One factor in this study would be the original treatment group designations (i.e., program vs. comparison group), while the other factor would be attrition (i.e., dropout vs. non-dropout group). The dependent measure could be the pretest or other available pre-program measures. A main effect on the attrition factor would be indicative of a threat to external validity or generalizability, while an interaction between group and attrition factors would point to a possible threat to internal validity. Where both effects occur, it is reasonable to infer that there is a threat to both internal and external validity.

   The plausibility of alternative explanations might also be minimized using covariance analysis. For example, in a study of the effects of "workfare" programs on social welfare case loads, one plausible alternative explanation might be the status of local economic conditions. Here, it might be possible to construct a measure of economic conditions and include that measure as a covariate in the statistical analysis. One must be careful when using covariance adjustments of this type -- "perfect" covariates do not exist in most social research and the use of imperfect covariates will not completely adjust for potential alternative explanations. Nevertheless causal assertions are likely to be strengthened by demonstrating that treatment effects occur even after adjusting on a number of good covariates.

5. **By Preventive Action.** When potential threats are anticipated they can often be ruled out by some type of preventive action. For example, if the program is a desirable one, it is likely that the comparison group would feel jealous or demoralized. Several actions can be taken to minimize the effects of these attitudes including offering the program to the comparison group upon completion of the study or using program and comparison groups which have little opportunity for contact and communication. In addition, auditing methods and quality control can be used to track potential experimental dropouts or to insure the standardization of measurement.

The five categories listed above should not be considered mutually exclusive. The inclusion of measurements designed to minimize threats to validity will obviously be related to the design structure and is likely to be a factor in the analysis. A good research plan should, where possible. make use of multiple methods for reducing threats. In general, reducing a particular threat by design or preventive action will probably be stronger than by using one of the other three approaches. The choice of which strategy to use for any particular threat is complex and depends at least on the cost of the strategy and on the potential seriousness of the threat.

Design Construction

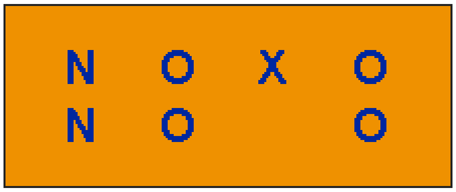
Basic Design Elements. Most research designs can be constructed from four basic elements:

1. **Time.** A causal relationship, by its very nature, implies that some time has elapsed between the occurrence of the cause and the consequent effect. While for some phenomena the elapsed time might be measured in microseconds and therefore might be unnoticeable to a casual observer, we normally assume that the cause and effect in social science arenas do not occur simultaneously, In design notation we indicate this temporal element horizontally - whatever symbol is used to indicate the presumed cause would be placed to the left of the symbol indicating measurement of the effect. Thus, as we read from left to right in design notation we are reading across time. Complex designs might involve a lengthy sequence of observations and programs or treatments across time.
2. **Program(s) or Treatment(s).** The presumed cause may be a program or treatment under the explicit control of the researcher or the occurrence of some natural event or program not explicitly controlled. In design notation we usually depict a presumed cause with the symbol "X". When multiple programs or treatments are being studied using the same design, we can keep the programs distinct by using subscripts such as "X_1" or "X_2". For a comparison group (i.e., one which does not receive the program under study) no "X" is used.
3. **Observation(s) or Measure(s).** Measurements are typically depicted in design notation with the symbol "O". If the same measurement or observation is taken at every point in time in a design, then this "O" will be sufficient. Similarly, if the same set of measures is given at every point in

time in this study, the "O" can be used to depict the entire set of measures. However, if different measures are given at different times it is useful to subscript the "O" to indicate which measurement is being given at which point in time.

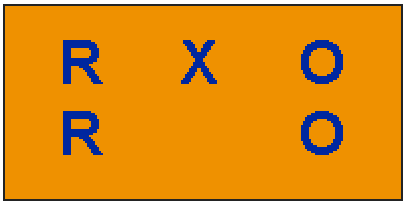
4. **Groups or Individuals.** The final design element consists of the intact groups or the individuals who participate in various conditions. Typically, there will be one or more program and comparison groups. In design notation, each group is indicated on a separate line. Furthermore, the manner in which groups are assigned to the conditions can be indicated by an appropriate symbol at the beginning of each line. Here, "R" will represent a group which was randomly assigned, "N" will depict a group which was nonrandomly assigned (i.e., a nonequivalent group or cohort) and a "C" will indicate that the group was assigned using a cutoff score on a measurement.

Perhaps the easiest way to understand how these four basic elements become integrated into a design structure is to give several examples. One of the most commonly used designs in social research is the two-group pre-post design which can be depicted as:



N	O	X	O
N	O		O

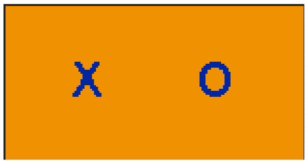
There are two lines in the design indicating that the study was comprised of two groups. The two groups were nonrandomly assigned as indicated by the "N". Both groups were measured before the program or treatment occurred as indicated by the first "O" in each line. Following this preobservation, the group in the first line received a program or treatment while the group in the second line did not. Finally, both groups were measured subsequent to the program. Another common design is the posttest-only randomized experiment. The design can be depicted as:



R	X	O
R		O

Here, two groups are randomly selected with one group receiving the program and one acting as a comparison. Both groups are measured after the program is administered.

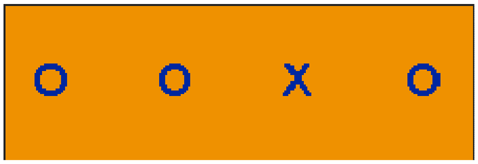
Expanding a Design. We can combine the four basic design elements in a number of ways in order to arrive at a specific design which is appropriate for the setting at hand. One strategy for doing so begins with the basic causal relationship:



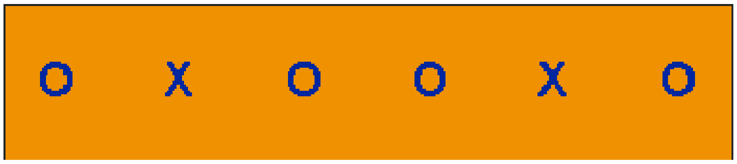
X	O

This is the most simple design in causal research and serves as a starting point for the development of better strategies. When we add to this basic design we are essentially expanding one of the four basic elements described above. Each possible expansion has implications both for the cost of the study and for the threats which might be ruled out.

1. **Expanding Across Time.** We can add to the basic design by including additional observations either before or after the program or, by adding or removing the program or different programs. For example, we might add one or more pre-program measurements and achieve the following design:

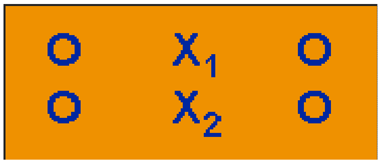


The addition of such pretests provides a "baseline" which, for instance, helps to assess the potential of a maturation or testing threat. If a change occurs between the first and second pre-program measures, it is reasonable to expect that similar change might be seen between the second pretest and the posttest even in the absence of the program. However, if no change occurs between the two pretests, one might be more confident in assuming that maturation or testing is not a likely alternative explanation for the cause-effect relationship which was hypothesized. Similarly, additional postprogram measures could be added. This would be useful for determining whether an immediate program effect decays over time, or whether there is a lag in time between the initiation of the program and the occurrence of an effect. We might also add and remove the program over time:



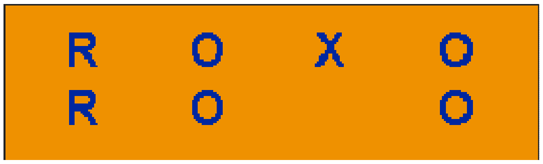
This is one form of the ABAB design which is frequently used in clinical psychology and psychiatry. The design is particularly strong against a history threat. When the program is repeated it is less likely that unique historical events can be responsible for replicated outcome patterns.

2. **Expanding Across Programs.** We have just seen that we can expand the program by adding it or removing it across time. Another way to expand the program would be to partition it into different levels of treatment. For example, in a study of the effect of a novel drug on subsequent behavior. we might use more than one dosage of the drug:



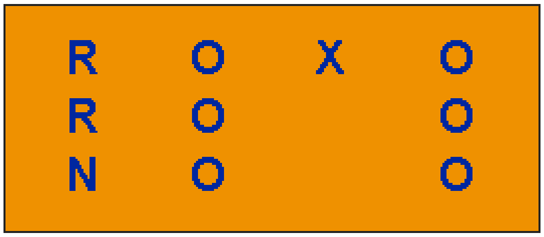
This design is an example of a simple factorial design with one factor having two levels. Notice that group assignment is not specified indicating that any type of assignment might have been used. This is a common strategy in a "sensitivity" or "parametric" study where the primary focus is one the effects obtained at various program levels. In a similar manner, one might expand the program by varying specific components of it across groups. This might be useful if one wishes to study different modes of the delivery of the program, different sets of program materials and the like. Finally, we can expand the program by using theoretically polarized or "opposite" treatments. A comparison group can be considered one example of such a polarization. Another might involve use of a second program which is expected to have an opposite effect on the outcome measures. A strategy of this sort provides evidence that the outcome measure is sensitive enough to differentiate between different programs.

3. **Expanding Across Observations.** At any point in time in a research design it is usually desirable to collect multiple measurements. For example, we might add a number of similar measures in order to determine whether the results of these converge. Or, we might wish to add measurements which theoretically should not be affected by the program in question in order to demonstrate that the program discriminates between effects. Strategies of this type are useful for achieving convergent and discriminant validity of measures as discussed in Campbell and Fiske (1959). Another way to expand the observations is by proxy measurements. Assume that we wish to study a new educational program but neglected to take pre-program measurements. We might use a standardized achievement test for the posttest and grade point average records as a proxy measure of student achievement prior to the initiation of the program. Finally, we might also expand the observations through the use of "recollected" measures. Again, if we were conducting a study and had neglected to administer a pretest or desired information in addition to the pretest information, we might ask participants to recall how they felt or behaved prior to the study and use this information as an additional measure. Different measurement approaches obviously yield data of different quality. What is advocated here is the use of multiple measurements rather than reliance on only a single strategy.
4. **Expanding Across Groups.** Often, it will be to our advantage to add additional groups to a design in order to rule out specific threats to validity. For example, consider the following pre-post two-group randomized experimental design:



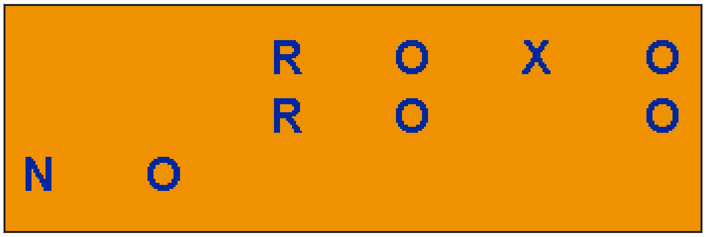
R	O	X	O
R	O		O

If this design were implemented within a single institution where members of the two groups were in contact with each other one might expect that intergroup communication, group rivalry, or demoralization of a group which gets denied a desirable treatment or gains an undesirable one might pose threats to the validity of the causal inference. In such a case. one might add an additional nonequivalent group from a similar institution which consists of persons unaware of the original two groups:



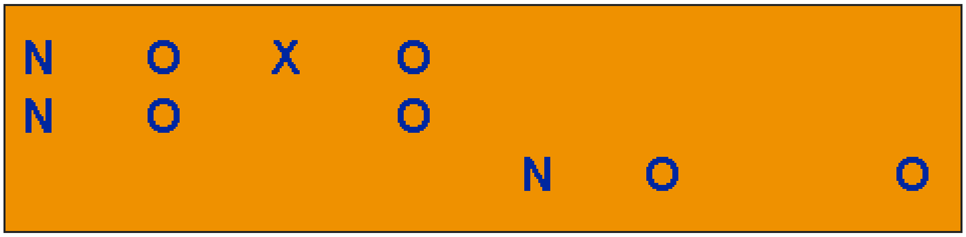
```
R   O   X   O
R   O       O
N   O       O
```

In a similar manner, whenever nonequivalent groups are used in a study it will usually be advantageous to have multiple replications of each group. The use of many nonequivalent groups helps to minimize the potential of a particular selection bias affecting the results. In some cases it may be desirable to include the norm group as an additional group in the design. Norming group averages are available for most standardized achievement tests for example, and might comprise an additional nonequivalent control group. Cohort groups might also be used in a number of ways. For example, one might use a single measure of a cohort group to help rule out a testing threat:



```
        R   O   X   O
        R   O       O
N   O
```

In this design, the randomized groups might be sixth graders from the same school year while the cohort might be the entire sixth grade from the previous academic year. This cohort group did not take the pretest and, if they are similar to the randomly selected control group, would provide evidence for or against the notion that taking the pretest had an effect on posttest scores. We might also use pre-post cohort groups:



```
N   O   X   O
N   O       O
                N   O       O
```

Here, the treatment group consists of sixth graders, the first comparison group of seventh graders in the same year, and the second comparison group consists of the following year's sixth graders (i.e., the fifth graders during the study year). Strategies of this sort are particularly useful in nonequivalent designs where selection bias is a potential problem and where routinely-collected institutional data is available. Finally, one other approach for expanding the groups involves partitioning groups with different assignment strategies. For example, one might randomly divide nonequivalent groups, or select nonequivalent subgroups from randomly assigned groups. An example of this sort involving the combination of random assignment and assignment by a cutoff is discussed in detail below.

A Simple Strategy for Design Construction.

Considering the basic elements of a research design or the possibilities for expansion are not alone sufficient. We need to be able to integrate these elements with an overall strategy. Furthermore we need to decide which potential threats are best handled by design rather than by argument, measurement, analysis, or preventive action.

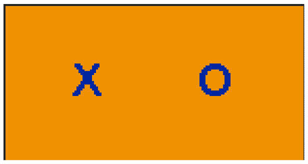
While no definitive approach for designing designs exists, we might suggest a tentative strategy based on the notion of expansion discussed above. First, we begin the designing task by setting forth a design which depicts the simple hypothesized causal relationship. Second, we deliberately over-expand this basic design by expanding across time, program. observations, and groups. At this step, the emphasis is on accounting for as many likely alternative explanations as possible using the design. Finally, we then scale back this over-expanded version considering the effect of eliminating each design component. It is at this point that we face the difficult decisions concerning the costs of each design component and the advantages of ruling out specific threats using other approaches.

There are several advantages which result from using this type of approach to design construction. First, we are forced to be explicit about the decisions which are made. Second. the approach is "conservative" in nature. The strategy minimizes the chance of our overlooking a major threat to validity in constructing our design. Third, we arrive at a design which is "tailored" to the situation at hand. Finally, the strategy is cost-efficient. Threats which can be accounted for by some other, less costly, approach need not be accounted for in the design itself.

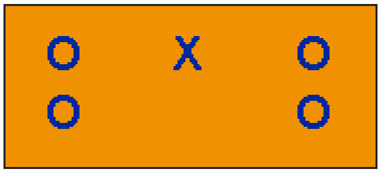
An Example of a Hybrid Design

Some of the ideas discussed above can be illustrated in an example. The design in question is drawn from an earlier discussion by Boruch (1975). To our knowledge, this design has never been used, although it has strong features to commend it.

Let us assume that we wish to study the effects of a new compensatory education program on subsequent student achievement. The program is designed to help students who are poor in reading to improve in those skills. We can begin then with the simple hypothesized cause-effect relationship:

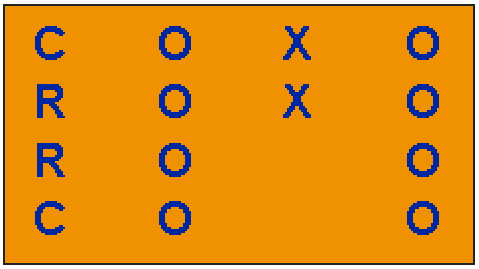


Here, the "X" represents the reading program and the "O" stands for a reading achievement test. We decide that it is desirable to add a pre-program measure so that we might investigate whether the program "improves" reading test scores. We also decide to expand across groups by adding a comparison group. At this point we have the typical:



```
O   X   O
O       O
```

The next problem concerns how the two groups will be assigned. Since the program is specifically designed to help students who need special assistance in reading, we rule out random assignment because it would require denying the program to students in need. We had considered the possibility of offering the program to one randomly assigned group in the first year and to the control group in the second, but ruled that out on the grounds that it would require two years of program expenses and the denial of a potentially helpful program for half of the students for a period of a year. Instead we decide to assign students by means of a cutoff score on the pretest. All students scoring below a preselected percentile on the reading pretest would be given the program while those above that percentile would act as controls (i.e., the regression-discontinuity design). However, previous experience with this strategy (Trochim, 1994) has shown us that it is difficult to adhere to a single cutoff score for assignment to group. We are especially concerned that teachers or administrators will allow students who score slightly above the cutoff point into the program because they have little confidence in the ability of the achievement test to make fine distinctions in reading skills for children who score very close to the cutoff. To deal with this potential problem, we decide to partition the groups using a particular combination of assignment by a cutoff and random assignment:



```
C   O   X   O
R   O   X   O
R   O       O
C   O       O
```

In this design we have set up two cutoff points. All those scoring below a certain percentile are assigned to the treatment group automatically by this cutoff, All those scoring above another higher percentile are automatically assigned to the comparison group by this cutoff. Finally, all those who fall in the interval between the cutoffs on the pretest are randomly assigned to either the treatment or comparison groups.

There are several advantages to this strategy. It directly addresses the concern to teachers and administrators that the test may not be able to discriminate well between students who score immediately above or below a cutoff point. For example, a student whose true ability in reading would place him near the cutoff might have a bad day and therefore might be placed into the treatment or comparison group by chance factors. The design outlined above is defensible. We can agree with the teachers and administrators that the test is fallible. Nevertheless, since we need some criterion to assign students to the program, we can argue that the fairest approach would be to assign borderline cases by lottery. In addition, by combining two excellent strategies (i.e., the randomized experiment and the regression-discontinuity) we can analyze results separately for each and address the possibility that design factors might bias results.

There are many other worthwhile considerations not mentioned in the above scenario. For example, instead of using simple randomized assignment within the cutoff interval, we might use a weighted random assignment so that students scoring lower in the interval have a greater probability of being assigned to the program. In addition, we might consider expanding the design in a number of other ways, by including double.pretests or multiple posttests; multiple measures of reading skills; additional replications of the program or variations of the programs and additional groups such as norming groups, controls from other schools, and the like. Nevertheless, this brief example serves to illustrate the advantages of explicitly constructing a research design to meet the specific needs of a particular situation.

The Nature of Good Design

Throughout the design construction task, it is important to have in mind some endpoint, some criteria which we should try to achieve before finally accepting a design strategy. The criteria discussed below are only meant to be suggestive of the characteristics found in good research design. It is worth noting that all of these criteria point to the need to individually tailor research designs rather than accepting standard textbook strategies as is.

1. **Theory-Grounded.** Good research strategies reflect the theories which are being investigated. Where specific theoretical expectations can be hypothesized these are incorporated into the design. For example, where theory predicts a specific treatment effect on one measure but not on another, the inclusion of both in the design improves discriminant validity and demonstrates the predictive power of the theory.
2. **Situational.** Good research designs reflect the settings of the investigation. This was illustrated above where a particular need of teachers and administrators was explicitly addressed in the design strategy. Similarly, intergroup rivalry, demoralization, and competition might be assessed through the use of additional comparison groups who are not in direct contact with the original group.
3. **Feasible.** Good designs can be implemented. The sequence and timing of events are carefully thought out. Potential problems in measurement, adherence to assignment, database construction and the like, are anticipated. Where needed, additional groups or measurements are included in the design to explicitly correct for such problems.
4. **Redundant.** Good research designs have some flexibility built into them. Often, this flexibility results from duplication of essential design features. For example, multiple replications of a treatment help to insure that failure to implement the treatment in one setting will not invalidate the entire study.
5. **Efficient.** Good designs strike a balance between redundancy and the tendency to overdesign. Where it is reasonable, other, less costly, strategies for ruling out potential threats to validity are utilized.

This is by no means an exhaustive list of the criteria by which we can judge good research design. nevertheless, goals of this sort help to guide the researcher toward a final design choice and emphasize important components which should be included.

The development of a theory of research methodology for the social sciences has largely occurred over the past half century and most intensively within the past two decades. It is not surprising, in such a relatively recent effort, that an emphasis on a few standard research designs has occurred. Nevertheless, by moving away from the notion of "design selection" and towards an emphasis on design construction, there is much to be gained in our understanding of design principles and in the quality of our research.

Advances in Quasi-Experimentation

Reprinted from Trochim, W. (Ed.), (1986). *Editor's Notes*. Advances in quasi-experimental design and analysis. New Directions for Program Evaluation Series, Number 31, San Francisco, CA: Jossey-Bass.

The intent of this volume is to update, perhaps even to alter, our thinking about quasi-experimentation in applied social research and program evaluation. Since Campbell and Stanley (1963) introduced the term quasi-experiment, we have tended to see this area as involving primarily two interrelated topics: the theory of the validity of casual inferences and a taxonomy of the research designs that enable us to examine causal hypotheses. We can see this in the leading expositions of quasi-experimentation (Campbell and Stanley, 1963, 1966; Cook and Campbell, 1979) as well as in the standard textbook presentations of the topic (Kidder and Judd, 1986; Rossi and Freeman, 1985), where it is typical to have separate sections or chapters that discuss validity issues first and then proceed to distinguishable quasi-experimental designs (for example, the pretest-posttest nonequivalent group design, the regression-discontinuity design, the interrupted time series design). My first inclination in editing this volume was to emulate this tradition, beginning the volume with a chapter on validity and following it with a chapter for each of the major quasi-experimental designs that raised the relevant conceptual and analytical issues and discussed recent advances. But, I think, such an approach would have simply contributed to a persistent confusion about the nature of quasi-experimentation and its role in research.

Instead, this volume makes the case that we have moved beyond the traditional thinking on quasi-experiments as a collection of specific designs and threats to validity toward a more integrated, synthetic view of quasi-experimentation as part of a general logical and epistemological framework for research. To support this view that the notion of quasi-experimentation is evolving toward increasing integration, I will discuss a number of themes that seem to characterize our current thinking and that cut across validity typologies and design taxonomies. This list of themes may also be viewed as a tentative description of the advances in our thinking about quasi-experimentation in social research.

The Role of Judgment

One theme that underlies most of the others and that illustrates our increasing awareness of the tentativeness and frailty of quasi-experimentation concerns the importance of human judgment in research. Evidence bearing on a causal relationship emerges from many sources, and it is not a trivial matter to integrate or resolve conflicts or discrepancies. In recognition of this problem of evidence, we are beginning to address causal inference as a psychological issue that can be illuminated by cognitive models of the judgmental process (see Chapter One of this volume and Einhom and Hogarth, 1986). We are also recognizing more clearly the sociological bases of scientific thought (Campbell, 1984) and the fact that science is at root a human enterprise. Thus, a positivist, mechanistic view is all but gone from quasi-experimental thinking, and what remains is a more judgmental and more scientifically sensible perspective.

The Case for Tailored Designs

Early expositions of quasi-experimentation took a largely taxonomic approach, laying out a collection of relatively discrete research designs and discussing how weak or strong they were for valid causal inference. Almost certainly, early proponents recognized that there was a virtual infinity of design variations and that validity was more complexly related to theory and context than their presentations implied. Nonetheless, what seemed to evolve was a "cookbook" approach to quasi-experimentation that involved "choosing" a design that fit the situation and checking off lists of validity threats.

In an important paper on the coupling of randomized and nonrandomized design features, Boruch (1975) explicitly encouraged us to construct research designs as combinations of more elemental units (for example, assignment strategies, measurement occasions) based on the specific contextual needs and plausible alternative explanations for a treatment effect. This move toward hybrid, tailored, or patched-up designs, which involved suggesting how such designs could be accomplished, is one in which I have been a minor participant (Trochim and Land, 1982; Trochim, 1984). It is emphasized by Cordray in Chapter One of this volume. The implication for current practice is that we should focus on the advantages of different combinations of design features rather than on a relatively restricted set of prefabricated designs. In teaching quasi-experimental methods, we need to break away from a taxonomic design mentality and emphasize design principles and issues that cut across the traditional distinctions between true experiments, nonexperiments, and quasi-experiments.

The Crucial Role of Theory

Quasi-experimentation and its randomized experimental parent have been criticized for encouraging an atheoretical "black box" mentality of research (see, for instance, Chen and Rossi, 1984; Cronbach, 1982). Persons are assigned to either complex molar program packages or (often) to equally complex comparison conditions. The machinery of random assignment (or our quasi-experimental attempts to approximate random assignment) are the primary means of defining whether the program has an effect. This *ceteris paribus* mentality is inherently atheoretical and noncontextual: It assumes that the same mechanism works in basically the same way whether we apply it in mental health or criminal justice, income maintenance or education.

There is nothing inherently wrong with this program-group-versus-comparison-group logic. The problem is that it may be a rather crude, uninformative approach. In the two-group case, we are simply creating a dichotomous input into reality. If we observe a posttest difference between groups, it could be explained by this dichotomous program-versus-comparison-group input or by any number of alternative explanations, including differential attrition rates, intergroup rivalry and communication, initial selection differences among groups, or different group histories. We usually try to deal with these alternative explanations by ruling them out through argument, additional measurement, patched-up design features, and auxiliary analysis. Cook and Campbell (1979), Cronbach (1982), and others strongly favor replication of treatment effects as a standard for judging the validity of a causal assertion, but this advice does little to enhance the validity and informativeness within individual studies or program evaluations.

Chen and Rossi (1984, p. 339) approached this issue by advocating increased attention to social science theory: "not the global conceptual schemes of the grand theorists but much more prosaic theories that are concerned with how human organizations work and how social problems are generated." Evaluators have similarly begun to stress the importance of program theory as the basis for causal assessment (for example, Bickman, in press). These developments allow increased emphasis to be placed on the role of pattern matching (Trochim, 1985) through the generation of more complex theory-driven predictions that, if corroborated, allow fewer plausible alternative explanations for the effect of a program. Because appropriate theories may not be readily available, especially for the evaluation of contemporary social programs, we are developing methods and processes that facilitate the articulation of the implicit theories which program administrators and stakeholder groups have in mind and which presumably guide the formation and implementation of the program (Trochim, 1985). This theory-driven perspective is consonant with Mark's emphasis in Chapter Three on the study of causal process and with Cordray's discussion in Chapter One on ruling in the program as opposed to ruling out alternative explanations.

Attention to Program Implementation

A theory-driven approach to quasi-experimentation will be futile unless we can demonstrate that the program was in fact carried out or implemented as the theory intended. Consequently, we have seen the development of program implementation theory (for example, McLaughlin, 1984) that directly addresses the process of program execution. One approach emphasizes the development of organizational procedures and training systems that accurately transmit the program and that anticipate likely institutional sources of resistance. Another strategy involves the assessment of program delivery through program audits, management information systems, and the like. This emphasis on program implementation has further obscured the traditional distinction between process and outcome evaluation. At the least, it is certainly clear that good quasi-experimental outcome evaluation cannot be accomplished without attending to program processes, and we are continuing to develop better notions of how to combine these two efforts.

The Importance of Quality Control

Over and over, our experience with quasi-experimentation has shown that even the best-laid research plans often go awry in practice, sometimes with disastrous results. Thus, over the past decade we have begun to pay increasing attention to the integrity and quality of our research methods in real-world settings. One way of achieving this goal is to incorporate techniques used by other professions -- accounting, auditing, industrial quality control -- that have traditions in data integrity and quality assurance (Trochim and Visco, 1985). For instance, double bookkeeping can be used to keep verifiable records of research participation. Acceptance sampling can be an efficient method for checking accuracy in large data collection efforts, where an exhaustive examination of records is impractical or excessive in cost. These issues are particularly important in quasi-experimentation, where it is incumbent upon the researcher to demonstrate that sampling, measurement, group assignment, and analysis decisions do not interact with program participation in ways that can confound the final interpretation of results.

The Advantages of Multiple Perspectives

We have long recognized the importance of replication and systematic variation in research. In the past few years, Cook (1985) and colleagues Shadish and Houts (Chapter Two in this volume) have articulated a rationale for achieving systematic variation that they term critical multiplism. This perspective rests on the notion that no single realization will ever be sufficient for understanding a phenomenon with validity. Multiple realizations -- of research questions, measures, samples, designs, analyses, replications, and so on -- are essential for convergence on the truth of a matter. However, such a varied approach can become a methodological and epistemological Pandora's box unless we apply critical judgment in deciding which multiples we will emphasize in a study or set of studies (Chapter Two in this volume and Mark and Shotland, 1985).

Evolution of the Concept of Validity

The history of quasi-experimentation is inseparable from the development of the theory of the validity of causal inference. Much of this history has been played out through the ongoing dialogue between Campbell and Cronbach concerning the definition of validity and the relative importance that should be attributed on the one hand to the establishment of a causal relationship and on the other hand to its generalizability. In the most recent major statement in this area, Cronbach (1982) articulated the UTOS model, which conceptually links the units, treatments, observing operations and settings in a study into a framework that can be used for establishing valid causal inference. The dialogue continues in Chapter Four of this volume, where Campbell attempts to dispel persistent confusion about the types of validity by tentatively relabeling internal validity as local molar causal validity and external validity as the principle of proximal similarity. It is reasonable to hope that we might achieve a clearer consensus on this issue, as Mark argues in Chapter Three, where he attempts to resolve several different conceptions of validity, including those of Campbell and Cronbach.

Development of Increasingly Complex Realistic Analytic Models

In the past decade, we have made considerable progress toward complicating our statistical analyses to account for increasingly complex contexts and designs. One such advance involves the articulation of causal models of the sort described by Reichardt and Gollob in Chapter Six, especially models that allow for latent variables and that directly model measurement error Joreskog and Sorbom, 1979).

Another important recent development involves analyses that address the problem of selection bias or group nonequivalence -- a central issue in quasi-experiments because random assignment is not used and there is no assurance that comparison groups are initially equivalent (Rindskopf's discussion in Chapter Five). At the same time, there is increasing recognition of the implications of not attending to the correct unit of analysis when analyzing the data and of the advantages and implications of conducting analyses at multiple levels. Thus, when we assign classrooms to conditions but analyze individual student data rather than classroom aggregates, we are liable to get a different view of program effects than we are when we analyze at the classroom level, as Shadish, Cook, and Houts argue in Chapter Two. Other notable advances that are not explicitly addressed in this volume include the development of log linear, probit, and logit models for the analysis of qualitative or nominal level outcome variables (Feinberg, 1980; Forthofer and Lehnen, 1981) and the increasing proliferation of Bayesian statistical approaches to quasi-experimental contexts (Pollard, 1986).

Parallel to the development of these increasingly complex, realistic analytic models, cynicism has deepened about the ability of any single model or analysis to be sufficient. Thus, in Chapter Six Reichardt and Gollob call for multiple analyses to bracket bias, and in Chapter Five Rindskopf recognizes the assumptive notions of any analytic approach to selection bias. We have virtually abandoned the hope of a single correct analysis, and we have accordingly moved to multiple analyses that are based on systematically distinct assumptional frameworks and that rely in an increasingly direct way on the role of judgment.

Conclusion

All the developments just outlined point to an increasingly realistic and complicated life for quasi-experimentalists. The overall picture that emerges is that all quasi-experimentation is judgmental. It is based on multiple and varied sources of evidence, it should be multiplistic in realization, it must attend to process as well as to outcome, it is better off when theory driven, and it leads ultimately to multiple analyses that attempt to bracket the program effect within some reasonable range.

In one sense, this is hardly a pretty picture. Our views about quasi-experimentation and its role in causal inference are certainly more tentative and critical than they were in 1965 or perhaps even in 1979. But, this more integrated and complex view of quasi-experimentation has emerged directly from our experiences in the conduct of such studies. As such, it realistically represents our current thinking about one of the major strands in the evolution of social research methodology in this century.

Analysis

By the time you get to the analysis of your data, most of the really difficult work has been done. It's much more difficult to: define the research problem; develop and implement a sampling plan; conceptualize, operationalize and test your measures; and develop a design structure. If you have done this work well, the analysis of the data is usually a fairly straightforward affair.

In most social research the data analysis involves three major steps, done in roughly this order:

- **Cleaning and organizing the data for analysis (Data Preparation)**
- **Describing the data (Descriptive Statistics)**
- **Testing Hypotheses and Models (Inferential Statistics)**

Data Preparation involves checking or logging the data in; checking the data for accuracy; entering the data into the computer; transforming the data; and developing and documenting a database structure that integrates the various measures.

Descriptive Statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data. With descriptive statistics you are simply describing what is, what the data shows.

Inferential Statistics investigate questions, models and hypotheses. In many cases, the conclusions from inferential statistics extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population thinks. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

In most research studies, the analysis section follows these three phases of analysis. Descriptions of how the data were prepared tend to be brief and to focus on only the more unique aspects to your study, such as specific data transformations that are performed. The descriptive statistics that you actually look at can be voluminous. In most write-ups, these are carefully selected and organized into summary tables and graphs that only show the most relevant or important information. Usually, the researcher links each of the inferential analyses to specific research questions or hypotheses that were raised in the introduction, or notes any models that were tested that emerged as part of the analysis. In most analysis write-ups it's especially critical to not "miss the forest for the trees." If you present too much detail, the reader may not be able to follow the central line of the results. Often extensive analysis details are appropriately relegated to appendices, reserving only the most critical analysis summaries for the body of the report itself.

Conclusion Validity

Of the four types of validity (see also internal validity, construct validity and external validity) conclusion validity is undoubtedly the least considered and most misunderstood. That's probably due to the fact that it was originally labeled 'statistical' conclusion validity and you know how even the mere mention of the word *statistics* will scare off most of the human race!

In many ways, conclusion validity is the most important of the four validity types because it is relevant whenever we are trying to decide if there is a relationship in our observations (and that's one of the most basic aspects of any analysis). Perhaps we should start with an attempt at a definition:

Conclusion validity is the degree to which conclusions we reach about relationships in our data are reasonable.

For instance, if we're doing a study that looks at the relationship between socioeconomic status (SES) and attitudes about capital punishment, we eventually want to reach some conclusion. Based on our data, we may conclude that there is a positive relationship, that persons with higher SES tend to have a more positive view of capital punishment while those with lower SES tend to be more opposed. Conclusion validity is the degree to which the conclusion we reach is credible or believable.

Although conclusion validity was originally thought to be a statistical inference issue, it has become more apparent that it is also relevant in qualitative research. For example, in an observational field study of homeless adolescents the researcher might, on the basis of field notes, see a pattern that suggests that teenagers on the street who use drugs are more likely to be involved in more complex social networks and to interact with a more varied group of people. Although this conclusion or inference may be based entirely on impressionistic data, we can ask whether it has conclusion validity, that is, whether it is a reasonable conclusion about a relationship in our observations.

Whenever you investigate a relationship, you essentially have two possible conclusions -- either there is a relationship in your data or there isn't. In either case, however, you could be wrong in your conclusion. You might conclude that there is a relationship when in fact there is not, or you might infer that there isn't a relationship when in fact there is (but you didn't detect it!). So, we have to consider all of these possibilities when we talk about conclusion validity.

It's important to realize that conclusion validity is an issue whenever you conclude there is a relationship, even when the relationship is between some program (or treatment) and some outcome. In other words, conclusion validity also pertains to causal relationships. How do we distinguish it from internal validity which is also involved with causal relationships? Conclusion validity is only concerned with whether there is a relationship. For instance, in a program evaluation, we might conclude that there is a positive relationship between our educational program and achievement test scores -- students in the program get higher scores and students not in the program get lower ones. Conclusion validity is essentially whether that relationship is a reasonable one or not, given the data. But it is possible that we will conclude that, while there is a relationship between the program and outcome , the program didn't cause the outcome. Perhaps some other factor, and not our program, was responsible for the outcome in this study. For instance, the observed differences in the outcome could be due to the fact that the program group was smarter than the comparison group to begin with. Our observed posttest differences between these groups could be due to this initial difference and not be the result of our program. This issue -- the possibility that some other factor than our program caused the outcome -- is what internal validity is all about. So, it is possible that in a study we can conclude that our program and outcome are related (conclusion validity) and also conclude that the outcome was caused by some factor other than the program (i.e., we don't have internal validity).

We'll begin this discussion by considering the major threats to conclusion validity, the different reasons you might be wrong in concluding that there is or isn't a relationship. You'll see that there are several key reasons why reaching conclusions about relationships is so difficult. One major problem is that it is often hard to see a relationship because our measures or observations have low reliability -- they are too weak relative to all of the 'noise' in the environment. Another issue is that the relationship we are looking for may be a weak one and seeing it is a bit like looking for a needle in the haystack. Sometimes the problem is that we just didn't collect enough information to see the relationship even if it is there. All of these problems are related to the idea of statistical power and so we'll spend some time trying to understand what 'power' is in this context. One of the most interesting introductions to the idea of statistical power is given in the 'OJ' Page which was created by Rob Becker to illustrate how the decision a jury has to reach (guilty vs. not guilty) is similar to the decision a researcher makes when assessing a relationship. The OJ Page uses the infamous OJ Simpson murder trial to introduce the idea of statistical power and illustrate how manipulating various factors (e.g., the amount of evidence, the "effect size", and the level of risk) affects the validity of the verdict. Finally, we need to recognize that we have some control over our ability to detect relationships, and we'll conclude with some suggestions for improving conclusion validity.

- **Threats to Conclusion Validity**

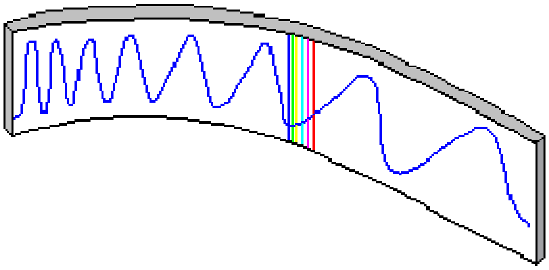
Threats to Conclusion Validity

A threat to conclusion validity is a factor that can lead you to reach an incorrect conclusion about a relationship in your observations. You can essentially make two kinds of errors about relationships:

1. conclude that there is no relationship when in fact there is (you missed the relationship or didn't see it)
2. conclude that there is a relationship when in fact there is not (you're seeing things that aren't there!)

Most threats to conclusion validity have to do with the first problem. Why? Maybe it's because it's so hard in most research to find relationships in our data at all that it's not as big or frequent a problem -- we tend to have more problems finding the needle in the haystack than seeing things that aren't there! So, I'll divide the threats by the type of error they are associated with.

Finding no relationship when there is one (or, "missing the needle in the haystack")



When you're looking for the needle in the haystack you essentially have two basic problems: the tiny needle and too much hay. You can view this as a signal-to-noise ratio problem.The "signal" is the needle -- the relationship you are trying to see. The "noise" consists of all of the factors that make it hard to see the relationship. There are several important sources of noise, each of which is a threat to conclusion validity. One important threat is **low reliability of measures** (see reliability). This can be due to many factors including poor question wording, bad instrument design or layout, illegibility of field notes, and so on. In studies where you are evaluating a program you can introduce noise through **poor reliability of treatment implementation**. If the program doesn't follow the prescribed procedures or is inconsistently carried out, it will be harder to see relationships between the program and other factors like the outcomes. Noise that is caused by **random irrelevancies in the setting** can also obscure your ability to see a relationship. In a classroom context, the traffic outside the room, disturbances in the hallway, and countless other irrelevant events can distract the researcher or the participants. The types of people you have in your study can also make it harder to see relationships. The threat here is due to **random heterogeneity of respondents**. If you have a very diverse group of respondents, they are likely to vary more widely on your measures or observations. Some of their variety may be related to the phenomenon you are looking at, but at least part of it is likely to just constitute individual differences that are irrelevant to the relationship being observed.

All of these threats add variability into the research context and contribute to the "noise" relative to the signal of the relationship you are looking for. But noise is only one part of the problem. We also have to consider the issue of the signal -- the true strength of the relationship. There is one broad threat to conclusion validity that tends to subsume or encompass all of the noise-producing factors above and also takes into account the strength of the signal, the amount of information you collect, and the amount of risk you're willing to take in making a decision about a whether a relationship exists. This threat is called **low statistical power**. Because this idea is so important in understanding how we make decisions about relationships, we have a separate discussion of statistical power.

Finding a relationship when there is not one (or "seeing things that aren't there")

In anything but the most trivial research study, the researcher will spend a considerable amount of time analyzing the data for relationships. Of course, it's important to conduct a thorough analysis, but most people are well aware of the fact that if you play with the data long enough, you can often "turn up" results that support or corroborate your hypotheses. In more everyday terms, you are "fishing" for a specific result by analyzing the data repeatedly under slightly differing conditions or assumptions.

In statistical analysis, we attempt to determine the probability that the finding we get is a "real" one or could have been a "chance" finding. In fact, we often use this probability to decide whether to accept the statistical result as evidence that there is a relationship. In the social sciences, researchers often use the rather arbitrary value known as the 0.05 level of significance to decide whether their result is credible or could be considered a "fluke." Essentially, the value 0.05 means that the result you got could be expected to occur by chance at least 5 times out of every 100 times you run the statistical analysis. The probability assumption that underlies most statistical analyses assumes that each analysis is "independent" of the other. But that may not be true when you conduct multiple analyses of the same data. For instance, let's say you conduct 20 statistical tests and for each one you use the 0.05 level criterion for deciding whether you are observing a relationship. For each test, the odds are 5 out of 100 that you will see a relationship even if there is not one there (that's what it means to say that the result could be "due to chance"). Odds of 5 out of 100 are equal to the fraction 5/100 which is also equal to 1 out of 20. Now, in this example, you conduct 20 separate analyses. Let's say that you find that of the twenty results, only one is statistically significant at the 0.05 level. Does that mean you have found a statistically significant relationship? If you had only done the one analysis, you might conclude that you've found a relationship in that result. But if you did 20 analyses, you would expect to find one of them significant by chance alone, even if there is no real relationship in the data. We call this threat to conclusion validity **fishing and the error rate problem**. The basic problem is that you were "fishing" by conducting multiple analyses and treating each one as though it was independent. Instead, when you conduct multiple analyses, you should adjust the error rate (i.e., significance level) to reflect the number of analyses you are doing. The bottom line here is that you are more likely to see a relationship when there isn't one when you keep reanalyzing your data and don't take that fishing into account when drawing your conclusions.

Problems that can lead to either conclusion error

Every analysis is based on a variety of assumptions about the nature of the data, the procedures you use to conduct the analysis, and the match between these two. If you are not sensitive to the assumptions behind your analysis you are likely to draw erroneous conclusions about relationships. In quantitative research we refer to this threat as the **violated assumptions of statistical tests**. For instance, many statistical analyses assume that the data are distributed normally -- that the population from which they are drawn would be distributed according to a "normal" or "bell-shaped" curve. If that assumption is not true for your data and you use that statistical test, you are likely to get an incorrect estimate of the true relationship. And, it's not always possible to predict what type of error you might make -- seeing a relationship that isn't there or missing one that is.

I believe that the same problem can occur in qualitative research as well. There are assumptions, some of which we may not even realize, behind our qualitative methods. For instance, in interview situations we may assume that the respondent is free to say anything s/he wishes. If that is not true -- if the respondent is under covert pressure from supervisors to respond in a certain way -- you may erroneously see relationships in the responses that aren't real and/or miss ones that are.

The threats listed above illustrate some of the major difficulties and traps that are involved in one of the most basic of research tasks -- deciding whether there is a relationship in your data or observations. So, how do we attempt to deal with these threats? The researcher has a number of strategies for improving conclusion validity through minimizing or eliminating the threats described above.

- **Improving Conclusion Validity**

Improving Conclusion Validity

So you may have a problem assuring that you are reaching credible conclusions about relationships in your data. What can you do about it? Here are some general guidelines you can follow in designing your study that will help improve conclusion validity.

Guidelines for Improving Conclusion Validity

- **Good Statistical Power**. The rule of thumb in social research is that you want statistical power to be greater than 0.8 in value. That is, you want to have at least 80 chances out of 100 of finding a relationship when there is one. As pointed out in the discussion of statistical power, there are several factors that interact to affect power. One thing you can usually do is to collect more information -- use a larger sample size. Of course, you have to weigh the gain in power against the time and expense of having more participants or gathering more data. The second thing you can do is to increase your risk of making a Type I error -- increase the chance that you will find a relationship when it's not there. In practical terms you can do that statistically by raising the alpha level. For instance, instead of using a 0.05 significance level, you might use 0.10 as your cutoff point. Finally, you can increase the effect size. Since the effect size is a ratio of the signal of the relationship to the noise in the context, there are two broad strategies here. To up the signal, you can increase the salience of the relationship itself. This is especially true in experimental contexts where you are looking at the effects of a program or treatment. If you increase the dosage of the program (e.g., increase the hours spent in training or the number of training sessions), it will be easier to see the effect when the treatment is stronger. The other option is to decrease the noise (or, put another way, increase reliability).

- **Good Reliability**. Reliability is related to the idea of noise or "error" that obscures your ability to see a relationship. In general, you can improve reliability by doing a better job of constructing measurement instruments, by increasing the number of questions on an scale or by reducing situational distractions in the measurement context.
- **Good Implementation**. When you are studying the effects of interventions, treatments or programs, you can improve conclusion validity by assuring good implementation. This can be accomplished by training program operators and standardizing the protocols for administering the program.

- **Statistical Power**

One of the most interesting introductions to the idea of statistical power is given in the 'OJ' Page which was created by Rob Becker to illustrate how the decision a jury has to reach (guilty vs. not guilt) is similar to the decision a researcher makes when assessing a relationship. The OJ Page uses the infamous OJ Simpson murder trial to introduce the idea of statistical power and illustrate how manipulating various factors (e.g., the amount of evidence, the "effect size", and the level of risk) affects the validity of the verdict.

There are four interrelated components that influence the conclusions you might reach from a statistical test in a research project. The logic of statistical inference with respect to these components is often difficult to understand and explain. This paper attempts to clarify the four components and describe their interrelationships.

The four components are:

- **sample size**, or the number of units (e.g., people) accessible to the study
- **effect size**, or the salience of the treatment relative to the noise in measurement
- **alpha level** (α, or significance level), or the odds that the observed result is due to chance
- **power**, or the odds that you will observe a treatment effect when it occurs

Given values for any three of these components, it is possible to compute the value of the fourth. For instance, you might want to determine what a reasonable sample size would be for a study. If you could make reasonable estimates of the effect size, alpha level and power, it would be simple to compute (or, more likely, look up in a table) the sample size.

Some of these components will be more manipulable than others depending on the circumstances of the project. For example, if the project is an evaluation of an educational program or counseling program with a specific number of available consumers, the sample size is set or predetermined. Or, if the drug dosage in a program has to be small due to its potential negative side effects, the effect size may consequently be small. The goal is to achieve a balance of the four components that allows the maximum level of power to detect an effect if one exists, given programmatic, logistical or financial constraints on the other components.

Figure 1 shows the basic decision matrix involved in a statistical conclusion. All statistical conclusions involve constructing two mutually exclusive hypotheses, termed the null (labeled H_0) and alternative (labeled H_1) hypothesis. Together, the hypotheses describe all possible outcomes with respect to the inference. The central decision involves determining which hypothesis to accept and which to reject. For instance, in the typical case, the null hypothesis might be:

H_0: Program Effect = 0

while the alternative might be

H_1: Program Effect <> 0

The null hypothesis is so termed because it usually refers to the "no difference" or "no effect" case. Usually in social research we expect that our treatments and programs will make a difference. So, typically, our theory is described in the alternative hypothesis.

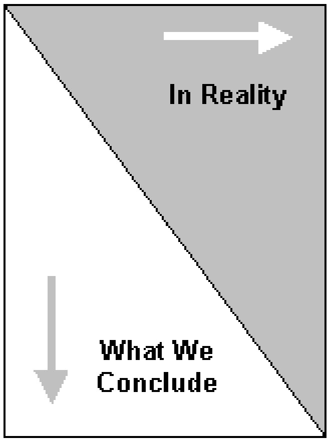
Figure 1 below is a complex figure that you should take some time studying. First, look at the header row (the shaded area). This row depicts reality -- whether there really is a program effect, difference, or gain. Of course, the problem is that you never know for sure what is really happening (unless you're God). Nevertheless, because we have set up mutually exclusive hypotheses, one must be right and one must be wrong. Therefore, consider this the view from God's position, knowing which hypothesis is correct. The first column of the 2x2 table shows the case where our program does not have an effect; the second column shows where it does have an effect or make a difference.

The left header column describes the world we mortals live in. Regardless of what's true, we have to make decisions about which of our hypotheses is correct. This header column describes the two decisions we can reach -- that our program had no effect (the first row of the 2x2 table) or that it did have an effect (the second row).

Now, let's examine the cells of the 2x2 table. Each cell shows the Greek symbol for that cell. Notice that the columns sum to 1 (i.e., $\alpha + (1-\alpha) = 1$ and $\beta + (1-\beta) = 1$). Why can we sum down the columns, but not across the rows? Because if one column is true, the other is irrelevant -- if the program has a real effect (the right column) it can't at the same time not have one. Therefore, the odds or probabilities have to sum to 1 for each column because the two rows in each column describe the only possible decisions (accept or reject the null/alternative) for each possible reality.

Below the Greek symbol is a typical value for that cell. You should especially note the values in the bottom two cells. The value of α is typically set at .05 in the social sciences. A newer, but growing, tradition is to try to achieve a statistical power of at least .80. Below the typical values is the name typically given for that cell (in caps). If you haven't already, you should note that two of the cells describe errors -- you reach the wrong conclusion -- and in the other two you reach the correct conclusion. Sometimes it's hard to remember which error is Type I and which is Type II. If you keep in mind that Type I is the same as the α or significance level, it might help you to remember that it is the odds of finding a difference or effect by chance alone. People are more likely to be susceptible to a Type I error, because they almost always want to conclude that their program works. If they find a statistical effect, they tend to advertise it loudly. On the other hand, people probably check more thoroughly for Type II errors because when you find that the program was not demonstrably effective, you immediately start looking for why (in this case, you might hope to show that you had low power and high β -- that the odds of saying there was no treatment effect even when there was were too high). Following the capitalized common name are several different ways of describing the value of each cell, one in terms of outcomes and one in terms of theory-testing. In italics, we give an example of how to express the numerical value in words.

To better understand the strange relationships between the two columns, think about what happens if you want to increase your power in a study. As you increase power, you increase the chances that you are going to find an effect if it's there (wind up in the bottom row). But, if you increase the chances that you wind up in the bottom row, you must at the same time be increasing the chances of making a Type I error! Although we can't sum to 1 across rows, there is clearly a relationship. Since we usually want high power *and* low Type I Error, you should be able to appreciate that we have a built-in tension here.

	**H_0 (null hypothesis) true** **H_1 (alternative hypothesis) false** **In reality...** • **There is *no* relationship** • **There is *no* difference, no gain** • **Our theory is *wrong***	**H_0 (null hypothesis) false** **H_1 (alternative hypothesis) true** **In reality...** • **There *is* a relationship** • **There *is* a difference or gain** • **Our theory is *correct***
We accept the null hypothesis	**1-α**	β

(H_0) **We reject the alternative hypothesis (H_1)** **We say...** • **"There is no relationship"** • **"There is no difference, no gain"** • **"Our theory is wrong"**	(e.g., .95) **THE CONFIDENCE LEVEL** The odds of saying there is no relationship, difference, gain, when in fact there is none The odds of correctly not confirming our theory *95 times out of 100 when there is no effect, we'll say there is none*	(e.g., .20) **TYPE II ERROR** The odds of saying there is no relationship, difference, gain, when in fact there is one The odds of not confirming our theory when it's true *20 times out of 100, when there is an effect, we'll say there isn't*
We reject the null hypothesis (H_0) **We accept the alternative hypothesis (H_1)** **We say...** • **"There is a relationship"** • **"There is a difference or gain"** • **"Our theory is correct"**	α (e.g., .05) **TYPE I ERROR** **(SIGNIFICANCE LEVEL)** The odds of saying there is an relationship, difference, gain, when in fact there is not The odds of confirming our theory incorrectly *5 times out of 100, when there is no effect, we'll say there is on* We should keep this small when we can't afford/risk wrongly concluding that our program works	**$1-\beta$** (e.g., .80) **POWER** The odds of saying that there is an relationship, difference, gain, when in fact there is one The odds of confirming our theory correctly *80 times out of 100, when there is an effect, we'll say there is* We generally want this to be as large as possible

Figure 1. The Statistical Inference Decision Matrix

We often talk about alpha (α) and beta (β) using the language of "higher" and "lower." For instance, we might talk about the advantages of a higher or lower α-level in a study. You have to be careful about interpreting the meaning of these terms. When we talk about *higher* α-levels, we mean that we are *increasing* the chance of a Type I Error. Therefore, a *lower* α-level actually means that you are conducting a *more rigorous* test. With all of this in mind, let's consider a few common associations evident in the table. You should convince yourself of the following:

- the lower the α, the lower the power; the higher the α, the higher the power
- the lower the α, the less likely it is that you will make a Type I Error (i.e., reject the null when it's true)
- the lower the α, the more "rigorous" the test
- an α of .01 (compared with .05 or .10) means the researcher is being relatively careful, s/he is only willing to risk being wrong 1 in a 100 times in rejecting the null when it's true (i.e., saying there's an effect when there really isn't)
- an α of .01 (compared with .05 or .10) limits one's chances of ending up in the bottom row, of concluding that the program has an effect. This means that both your statistical power and the chances of making a Type I Error are lower.
- an α of .01 means you have a 99% chance of saying there is no difference when there in fact is no difference (being in the upper left box)
- increasing α (e.g., from .01 to .05 or .10) increases the chances of making a Type I Error (i.e., saying there is a difference when there is not), decreases the chances of making a Type II Error (i.e., saying there is no difference when there is) and decreases the rigor of the test
- increasing α (e.g., from .01 to .05 or .10) increases power because one will be rejecting the null more often (i.e., accepting the alternative) and, consequently, when the alternative is true, there is a greater chance of accepting it (i.e., power)

Data Preparation

Data Preparation involves checking or logging the data in; checking the data for accuracy; entering the data into the computer; transforming the data; and developing and documenting a database structure that integrates the various measures.

Logging the Data

In any research project you may have data coming from a number of different sources at different times:

- mail surveys returns
- coded interview data
- pretest or posttest data
- observational data

In all but the simplest of studies, you need to set up a procedure for logging the information and keeping track of it until you are ready to do a comprehensive data analysis. Different researchers differ in how they prefer to keep track of incoming data. In most cases, you will want to set up a database that enables you to assess at any time what data is already in and what is still outstanding. You could do this with any standard computerized database program (e.g., Microsoft Access, Claris Filemaker), although this requires familiarity with such programs. or, you can accomplish this using standard statistical programs (e.g., SPSS, SAS, Minitab, Datadesk) and running simple descriptive analyses to get reports on data status. It is also critical that the data analyst retain the original data records for a reasonable period of time -- returned surveys, field notes, test protocols, and so on. Most professional researchers will retain such records for at least 5-7 years. For important or expensive studies, the original data might be stored in a data archive. The data analyst should always be able to trace a result from a data analysis back to the original forms on which the data was collected. A database for logging incoming data is a critical component in good research record-keeping.

Checking the Data For Accuracy

As soon as data is received you should screen it for accuracy. In some circumstances doing this right away will allow you to go back to the sample to clarify any problems or errors. There are several questions you should ask as part of this initial data screening:

- Are the responses legible/readable?
- Are all important questions answered?
- Are the responses complete?
- Is all relevant contextual information included (e.g., data, time, place, researcher)?

In most social research, quality of measurement is a major issue. Assuring that the data collection process does not contribute inaccuracies will help assure the overall quality of subsequent analyses.

Developing a Database Structure

The database structure is the manner in which you intend to store the data for the study so that it can be accessed in subsequent data analyses. You might use the same structure you used for logging in the data or, in large complex studies, you might have one structure for logging data and another for storing it. As mentioned above, there are generally two options for storing data on computer -- database programs and statistical programs. Usually database programs are the more complex of the two to learn and operate, but they allow the analyst greater flexibility in manipulating the data.

In every research project, you should generate a printed **codebook** that describes the data and indicates where and how it can be accessed. Minimally the codebook should include the following items for each variable:

- variable name
- variable description
- variable format (number, data, text)
- instrument/method of collection
- date collected
- respondent or group
- variable location (in database)
- notes

The codebook is an indispensable tool for the analysis team. Together with the database, it should provide comprehensive documentation that enables other researchers who might subsequently want to analyze the data to do so without any additional information.

Entering the Data into the Computer

There are a wide variety of ways to enter the data into the computer for analysis. Probably the easiest is to just type the data in directly. In order to assure a high level of data accuracy, the analyst should use a procedure called **double entry**. In this procedure you enter the data once. Then, you use a special program that allows you to enter the data a second time and checks each second entry against the first. If there is a discrepancy, the program notifies the user and allows the user to determine the correct entry. This double entry procedure significantly reduces entry errors. However, these double entry programs are not widely available and require some training. An alternative is to enter the data once and set up a procedure for checking the data for accuracy. For instance, you might spot check records on a random basis. Once the data have been entered, you will use various programs to summarize the data that allow you to check that all the data are within acceptable limits and boundaries. For instance, such summaries will enable you to easily spot whether there are persons whose age is 601 or who have a 7 entered where you expect a 1-to-5 response.

Data Transformations

Once the data have been entered it is almost always necessary to transform the raw data into variables that are usable in the analyses. There are a wide variety of transformations that you might perform. Some of the more common are:

- **missing values**

Many analysis programs automatically treat blank values as missing. In others, you need to designate specific values to represent missing values. For instance, you might use a value of -99 to indicate that the item is missing. You need to check the specific program you are using to determine how to handle missing values.

- **item reversals**

On scales and surveys, we sometimes use reversal items to help reduce the possibility of a response set. When you analyze the data, you want all scores for scale items to be in the same direction where high scores mean the same thing and low scores mean the same thing. In these cases, you have to reverse the ratings for some of the scale items. For instance, let's say you had a five point response scale for a self esteem measure where 1 meant strongly disagree and 5 meant strongly agree. One item is "I generally feel good about myself." If the respondent strongly agrees with this item they will put a 5 and this value would be indicative of higher self esteem. Alternatively, consider an item like "Sometimes I feel like I'm not worth much as a person." Here, if a respondent strongly agrees by rating this a 5 it would indicate low self esteem. To compare these two items, we would reverse the scores of one of them (probably we'd reverse the latter item so that high values will always indicate higher self esteem). We want a transformation where if the original value was 1 it's changed to 5, 2 is changed to 4, 3 remains the same, 4 is changed to 2 and 5 is changed to 1. While you could program these changes as separate statements in most program, it's easier to do this with a simple formula like:

New Value = (High Value + 1) - Original Value

In our example, the High Value for the scale is 5, so to get the new (transformed) scale value, we simply subtract each Original Value from 6 (i.e., 5 + 1).

- **scale totals**

Once you've transformed any individual scale items you will often want to add or average across individual items to get a total score for the scale.

- **categories**

For many variables you will want to collapse them into categories. For instance, you may want to collapse income estimates (in dollar amounts) into income ranges.

Descriptive Statistics

Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Descriptive statistics are typically distinguished from inferential statistics. With descriptive statistics you are simply describing what is or what the data shows. With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Descriptive Statistics are used to present quantitative descriptions in a manageable form. In a research study we may have lots of measures. Or we may measure a large number of people on any measure. Descriptive statistics help us to simply large amounts of data in a sensible way. Each descriptive statistic reduces lots of data into a simpler summary. For instance, consider a simple number used to summarize how well a batter is performing in baseball, the batting average. This single number is simply the number of hits divided by the number of times at bat (reported to three significant digits). A batter who is hitting .333 is getting a hit one time in every three at bats. One batting .250 is hitting one time in four. The single number describes a large number of discrete events. Or, consider the scourge of many students, the Grade Point Average (GPA). This single number describes the general performance of a student across a potentially wide range of course experiences.

Every time you try to describe a large set of observations with a single indicator you run the risk of distorting the original data or losing important detail. The batting average doesn't tell you whether the batter is hitting home runs or singles. It doesn't tell whether she's been in a slump or on a streak. The GPA doesn't tell you whether the student was in difficult courses or easy ones, or whether they were courses in their major field or in other disciplines. Even given these limitations, descriptive statistics provide a powerful summary that may enable comparisons across people or other units.

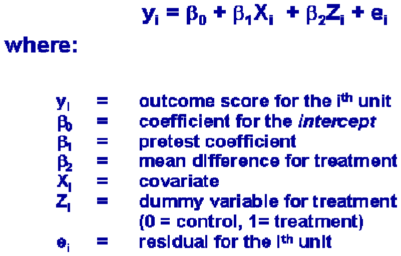
Univariate Analysis

Univariate analysis involves the examination across cases of one variable at a time. There are three major characteristics of a single variable that we tend to look at:

- the distribution
- the central tendency
- the dispersion

In most situations, we would describe all three of these characteristics for each of the variables in our study.

The Distribution. The distribution is a summary of the frequency of individual values or ranges of values for a variable. The simplest distribution would list every value of a variable and the number of persons who had each value. For instance, a typical way to describe the distribution of college students is by year in college, listing the number or percent of students at each of the four years. Or, we describe gender by listing the number or percent of males and females. In these cases, the variable has few enough values that we can list each one and summarize how many sample cases had the value. But what do we do for a variable like income or GPA? With these variables there can be a large number of possible values, with relatively few people having each one. In this case, we group the raw scores into categories according to ranges of values. For instance, we might look at GPA according to the letter grade ranges. Or, we might group income into four or five ranges of income values.



$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$$

where:

y_i	=	outcome score for the i^{th} unit
β_0	=	coefficient for the *intercept*
β_1	=	pretest coefficient
β_2	=	mean difference for treatment
X_i	=	covariate
Z_i	=	dummy variable for treatment (0 = control, 1= treatment)
e_i	=	residual for the i^{th} unit

Table 1. Frequency distribution table.

One of the most common ways to describe a single variable is with a ***frequency distribution***. Depending on the particular variable, all of the data values may be represented, or you may group the values into categories first (e.g., with age, price, or temperature variables, it would usually not be sensible to determine the frequencies for each value. Rather, the value are grouped into ranges and the frequencies determined.). Frequency distributions can be depicted in two ways, as a table or as a graph. Table 1 shows an age frequency distribution with five categories of age ranges defined. The same frequency distribution can be depicted in a graph as shown in Figure 2. This type of graph is often referred to as a histogram or bar chart.

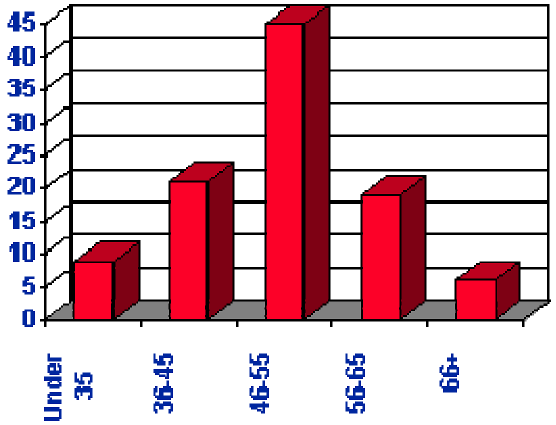


Table 2. Frequency distribution bar chart.

Distributions may also be displayed using percentages. For example, you could use percentages to describe the:

- percentage of people in different income levels
- percentage of people in different age ranges
- percentage of people in different ranges of standardized test scores

Central Tendency. The central tendency of a distribution is an estimate of the "center" of a distribution of values. There are three major types of estimates of central tendency:

- Mean
- Median
- Mode

The **Mean** or average is probably the most commonly used method of describing central tendency. To compute the mean all you do is add up all the values and divide by the number of values. For example, the mean or average quiz score is determined by summing all the scores and dividing by the number of students taking the exam. For example, consider the test score values:

15, 20, 21, 20, 36, 15, 25, 15

The sum of these 8 values is 167, so the mean is 167/8 = 20.875.

The **Median** is the score found at the exact middle of the set of values. One way to compute the median is to list all scores in numerical order, and then locate the score in the center of the sample. For example, if there are 500 scores in the list, score #250 would be the median. If we order the 8 scores shown above, we would get:

15,15,15,20,20,21,25,36

There are 8 scores and score #4 and #5 represent the halfway point. Since both of these scores are 20, the median is 20. If the two middle scores had different values, you would have to interpolate to determine the median.

The **mode** is the most frequently occurring value in the set of scores. To determine the mode, you might again order the scores as shown above, and then count each one. The most frequently occurring value is the mode. In our example, the value 15 occurs three times and is the model. In some distributions there is more than one modal value. For instance, in a bimodal distribution there are two values that occur most frequently.

Notice that for the same set of 8 scores we got three different values -- 20.875, 20, and 15 -- for the mean, median and mode respectively. If the distribution is truly normal (i.e., bell-shaped), the mean, median and mode are all equal to each other.

Dispersion. Dispersion refers to the spread of the values around the central tendency. There are two common measures of dispersion, the range and the standard deviation. The **range** is simply the highest value minus the lowest value. In our example distribution, the high value is 36 and the low is 15, so the range is 36 - 15 = 21.

The **Standard Deviation** is a more accurate and detailed estimate of dispersion because an outlier can greatly exaggerate the range (as was true in this example where the single outlier value of 36 stands apart from the rest of the values. The Standard Deviation shows the relation that set of scores has to the mean of the sample. Again lets take the set of scores:

15,20,21,20,36,15,25,15

to compute the standard deviation, we first find the distance between each value and the mean. We know from above that the mean is 20.875. So, the differences from the mean are:

15 - 20.875 = -5.875
20 - 20.875 = -0.875
21 - 20.875 = +0.125
20 - 20.875 = -0.875
36 - 20.875 = 15.125
15 - 20.875 = -5.875
25 - 20.875 = +4.125
15 - 20.875 = -5.875

Notice that values that are below the mean have negative discrepancies and values above it have positive ones. Next, we square each discrepancy:

-5.875 * -5.875 = 34.515625
-0.875 * -0.875 = 0.765625
+0.125 * +0.125 = 0.015625
-0.875 * -0.875 = 0.765625
15.125 * 15.125 = 228.765625

-5.875 * -5.875 = 34.515625
+4.125 * +4.125 = 17.015625
-5.875 * -5.875 = 34.515625

Now, we take these "squares" and sum them to get the Sum of Squares (SS) value. Here, the sum is 350.875. Next, we divide this sum by the number of scores minus 1. Here, the result is 350.875 / 7 = 50.125. This value is known as the **variance**. To get the standard deviation, we take the square root of the variance (remember that we squared the deviations earlier). This would be SQRT(50.125) = 7.079901129253.

Although this computation may seem convoluted, it's actually quite simple. To see this, consider the formula for the standard deviation:

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{(n - 1)}}$$

where:

$\underline{X}$ = each score
$\bar{X}$ = the mean or average
n = the number of values
Σ means we sum across the values

In the top part of the ratio, the numerator, we see that each score has the the mean subtracted from it, the difference is squared, and the squares are summed. In the bottom part, we take the number of scores minus 1. The ratio is the variance and the square root is the standard deviation. In English, we can describe the standard deviation as:

the square root of the sum of the squared deviations from the mean divided by the number of scores minus one

Although we can calculate these univariate statistics by hand, it gets quite tedious when you have more than a few values and variables. Every statistics program is capable of calculating them easily for you. For instance, I put the eight scores into SPSS and got the following table as a result:

N	8

Mean	20.8750
Median	20.0000
Mode	15.00
Std. Deviation	7.0799
Variance	50.1250
Range	21.00

which confirms the calculations I did by hand above.

The standard deviation allows us to reach some conclusions about specific scores in our distribution. Assuming that the distribution of scores is normal or bell-shaped (or close to it!), the following conclusions can be reached:

- approximately 68% of the scores in the sample fall within one standard deviation of the mean
- approximately 95% of the scores in the sample fall within two standard deviations of the mean
- approximately 99% of the scores in the sample fall within three standard deviations of the mean

For instance, since the mean in our example is 20.875 and the standard deviation is 7.0799, we can from the above statement estimate that approximately 95% of the scores will fall in the range of 20.875-(2*7.0799) to 20.875+(2*7.0799) or between 6.7152 and 35.0348. This kind of information is a critical stepping stone to enabling us to compare the performance of an individual on one variable with their performance on another, even when the variables are measured on entirely different scales.

- **Correlation**

The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. Let's work through an example to show you how this statistic is computed.

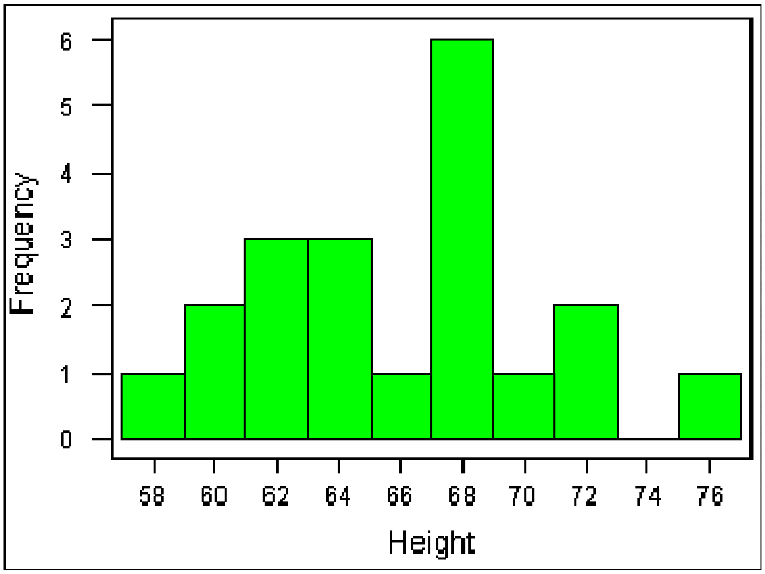
Correlation Example

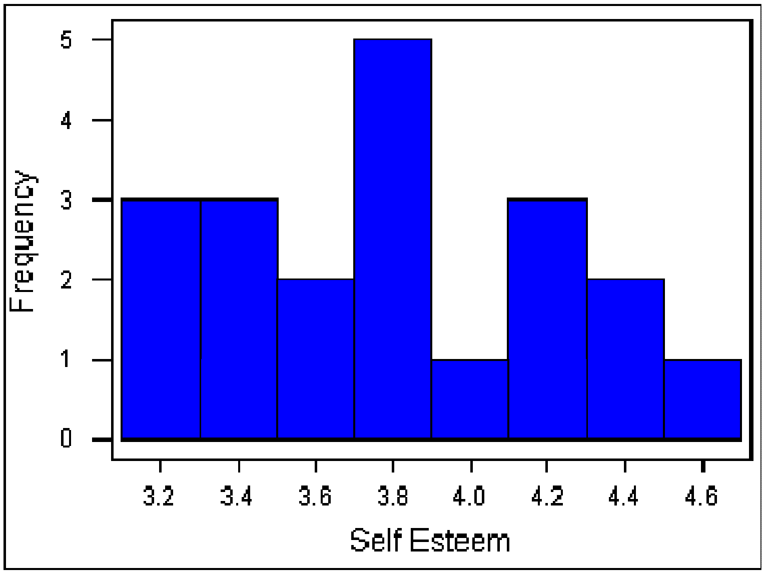
Let's assume that we want to look at the relationship between two variables, height (in inches) and self esteem. Perhaps we have a hypothesis that how tall you are effects your self esteem (incidentally, I don't think we have to worry about the direction of causality here -- it's not likely that self esteem causes your height!). Let's say we collect some information on twenty individuals (all male -- we know that the average height differs for males and females so, to keep this example simple we'll just use males). Height is measured in inches. Self esteem is measured based on the average of 10 1-to-5 rating items (where higher scores mean higher self esteem). Here's the data for the 20 cases (don't take this too seriously -- I made this data up to illustrate what a correlation is):

Person	**Height**	**Self Esteem**
1	68	4.1
2	71	4.6
3	62	3.8
4	75	4.4
5	58	3.2
6	60	3.1
7	67	3.8
8	68	4.1
9	71	4.3
10	69	3.7
11	68	3.5
12	67	3.2
13	63	3.7
14	62	3.3
15	60	3.4
16	63	4.0
17	65	4.1
18	67	3.8
19	63	3.4

20	61	3.6

Now, let's take a quick look at the histogram for each variable:



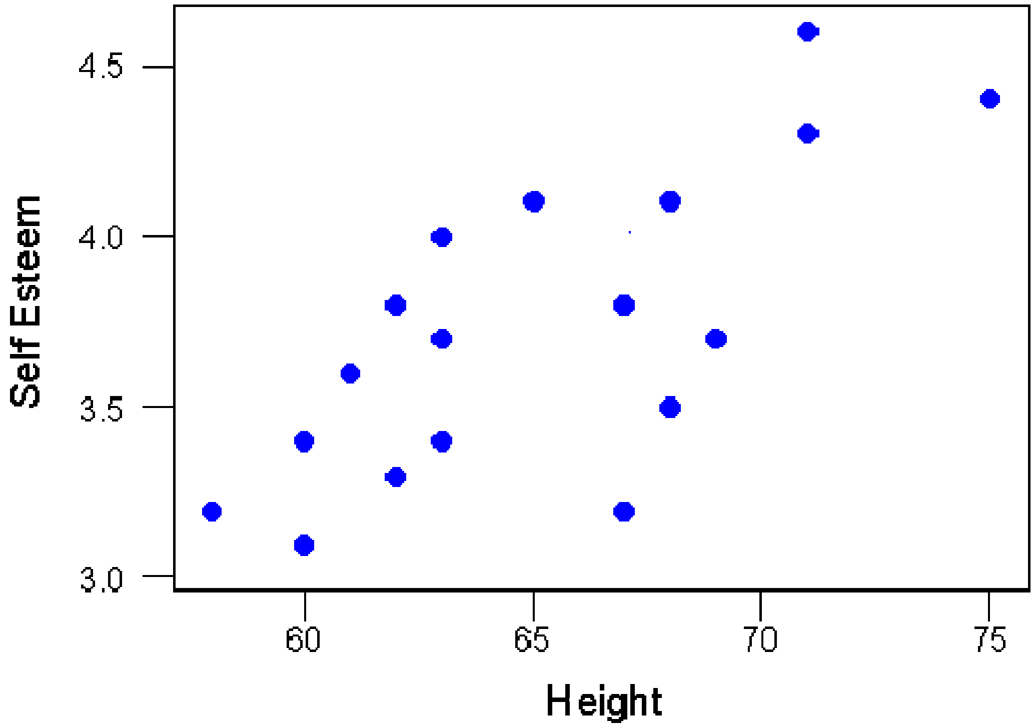


And, here are the descriptive statistics:

Variable	**Mean**	**StDev**	**Variance**	**Sum**	**Minimum**	**Maximum**	**Range**
Height	65.4	4.40574	19.4105	1308	58	75	17
Self	3.755	0.426090	0.181553	75.1	3.1	4.6	1.5

Esteem							

Finally, we'll look at the simple bivariate (i.e., two-variable) plot:



You should immediately see in the bivariate plot that the relationship between the variables is a positive one (if you can't see that, review the section on types of relationships) because if you were to fit a single straight line through the dots it would have a positive slope or move up from left to right. Since the correlation is nothing more than a quantitative estimate of the relationship, we would expect a positive correlation.

What does a "positive relationship" mean in this context? It means that, in general, higher scores on one variable tend to be paired with higher scores on the other and that lower scores on one variable tend to be paired with lower scores on the other. You should confirm visually that this is generally true in the plot above.

Calculating the Correlation

Now we're ready to compute the correlation value. The formula for the correlation is:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N	=	number of pairs of scores
Σxy	=	sum of the products of paired scores
Σx	=	sum of x scores
Σy	=	sum of y scores
Σx^2	=	sum of squared x scores
Σy^2	=	sum of squared y scores

We use the symbol **r** to stand for the correlation. Through the magic of mathematics it turns out that r will always be between -1.0 and +1.0. if the correlation is negative, we have a negative relationship; if it's positive, the relationship is positive. You don't need to know how we came up with this formula unless you want to be a statistician. But you probably will need to know how the formula relates to real data -- how you can use the formula to compute the correlation. Let's look at the data we need for the formula. Here's the original data with the other necessary columns:

Person	**Height (x)**	**Self Esteem (y)**	**x*y**	**x*x**	**y*y**
1	68	4.1	278.8	4624	16.81
2	71	4.6	326.6	5041	21.16
3	62	3.8	235.6	3844	14.44
4	75	4.4	330	5625	19.36
5	58	3.2	185.6	3364	10.24
6	60	3.1	186	3600	9.61
7	67	3.8	254.6	4489	14.44
8	68	4.1	278.8	4624	16.81
9	71	4.3	305.3	5041	18.49
10	69	3.7	255.3	4761	13.69

11	68	3.5	238	4624	12.25
12	67	3.2	214.4	4489	10.24
13	63	3.7	233.1	3969	13.69
14	62	3.3	204.6	3844	10.89
15	60	3.4	204	3600	11.56
16	63	4	252	3969	16
17	65	4.1	266.5	4225	16.81
18	67	3.8	254.6	4489	14.44
19	63	3.4	214.2	3969	11.56
20	61	3.6	219.6	3721	12.96
Sum =	**1308**	**75.1**	**4937.6**	**85912**	**285.45**

The first three columns are the same as in the table above. The next three columns are simple computations based on the height and self esteem data. The bottom row consists of the sum of each column. This is all the information we need to compute the correlation. Here are the values from the bottom row of the table (where N is 20 people) as they are related to the symbols in the formula:

$$\begin{aligned} N &= 20 \\ \Sigma xy &= 4937.6 \\ \Sigma x &= 1308 \\ \Sigma y &= 75.1 \\ \Sigma x^2 &= 85912 \\ \Sigma y^2 &= 285.45 \end{aligned}$$

Now, when we plug these values into the formula given above, we get the following (I show it here tediously, one step at a time):

$$r = \frac{20(4937.6) - (1308)(75.1)}{\sqrt{[20(85912) - (1308*1308)][20(285.45) - (75.1*75.1)]}}$$

$$r = \frac{98752 - 98230.8}{\sqrt{[1718240 - 1710864][5709 - 5640.01]}}$$

$$r = \frac{521.2}{\sqrt{[7376][68.99]}}$$

$$r = \frac{521.2}{\sqrt{508870.2}}$$

$$r = \frac{521.2}{713.3514}$$

$$r = .73$$

So, the correlation for our twenty cases is .73, which is a fairly strong positive relationship. I guess there is a relationship between height and self esteem, at least in this made up data!

Testing the Significance of a Correlation

Once you've computed a correlation, you can determine the probability that the observed correlation occurred by chance. That is, you can conduct a significance test. Most often you are interested in determining the probability that the correlation is a real one and not a chance occurrence. In this case, you are testing the mutually exclusive hypotheses:

Null Hypothesis:	r = 0
Alternative Hypothesis:	r <> 0

The easiest way to test this hypothesis is to find a statistics book that has a table of critical values of r. Most introductory statistics texts would have a table like this. As in all hypothesis testing, you need to first determine the significance level. Here, I'll use the common significance level of alpha = .05. This means that I am conducting a test where the odds that the correlation is a chance occurrence is no more than 5 out of 100. Before I look up the critical value in a table I also have to compute the degrees of freedom or df. The df is simply equal to N-2 or, in this example, is 20-2 = 18. Finally, I have to decide whether I am doing a one-tailed or two-tailed test. In this example, since I have no strong prior theory to suggest whether the relationship between height and self esteem would be positive or negative, I'll opt for the two-tailed test. With these three pieces of information -- the significance level (alpha = .05)), degrees of freedom (df = 18), and type of test (two-tailed) -- I can now test the significance of the correlation I found. When I look up this value in the handy little table at the back of my statistics book I find that the critical value is .4438. This means that if my correlation is greater than .4438 or less than -.4438 (remember, this is a two-tailed test) I can conclude that the odds are less than 5 out of 100 that this is a chance occurrence. Since my correlation 0f .73 is actually quite a bit higher, I conclude that it is not a chance finding and that the correlation is "statistically significant" (given the parameters of the test). I can reject the null hypothesis and accept the alternative.

The Correlation Matrix

All I've shown you so far is how to compute a correlation between two variables. In most studies we have considerably more than two variables. Let's say we have a study with 10 interval-level variables and we want to estimate the relationships among all of them (i.e., between all possible pairs of variables). In this instance, we have 45 unique correlations to estimate (more later on how I knew that!). We could do the above computations 45 times to obtain the correlations. Or we could use just about any statistics program to automatically compute all 45 with a simple click of the mouse.

I used a simple statistics program to generate random data for 10 variables with 20 cases (i.e., persons) for each variable. Then, I told the program to compute the correlations among these variables. Here's the result:

```
          C1       C2       C3       C4       C5       C6       C7       C8
C9      C10
C1     1.000
C2     0.274    1.000
C3    -0.134   -0.269    1.000
C4     0.201   -0.153    0.075    1.000
C5    -0.129   -0.166    0.278   -0.011    1.000
C6    -0.095    0.280   -0.348   -0.378   -0.009    1.000
C7     0.171   -0.122    0.288    0.086    0.193    0.002    1.000
C8     0.219    0.242   -0.380   -0.227   -0.551    0.324   -0.082    1.000
C9     0.518    0.238    0.002    0.082   -0.015    0.304    0.347   -0.013
1.000
C10    0.299    0.568    0.165   -0.122   -0.106   -0.169    0.243    0.014
0.352    1.000
```

This type of table is called a *correlation matrix*. It lists the variable names (C1-C10) down the first column and across the first row. The diagonal of a correlation matrix (i.e., the numbers that go from the upper left corner to the lower right) always consists of ones. That's because these are the correlations between each variable and itself (and a variable is always perfectly correlated with itself). This statistical program only shows the lower triangle of the correlation matrix. In every correlation matrix there are two triangles that are the values below and to the left of the diagonal (lower triangle) and above and to the right of the diagonal (upper triangle). There is no reason to print both triangles because the two triangles of a correlation matrix are always mirror images of each other (the correlation of variable x with variable y is always equal to the correlation of variable y with variable x). When a matrix has this mirror-image quality above and below the diagonal we refer to it as a *symmetric matrix*. A correlation matrix is always a symmetric matrix.

To locate the correlation for any pair of variables, find the value in the table for the row and column intersection for those two variables. For instance, to find the correlation between variables C5 and C2, I look for where row C2 and column C5 is (in this case it's blank because it falls in the upper triangle area) and where row C5 and column C2 is and, in the second case, I find that the correlation is -.166.

OK, so how did I know that there are 45 unique correlations when we have 10 variables? There's a handy simple little formula that tells how many pairs (e.g., correlations) there are for any number of variables:

$$\frac{N * (N - 1)}{2}$$

where N is the number of variables. In the example, I had 10 variables, so I know I have (10 * 9)/2 = 90/2 = 45 pairs.

Other Correlations

The specific type of correlation I've illustrated here is known as the Pearson Product Moment Correlation. It is appropriate when both variables are measured at an interval level. However there are a wide variety of other types of correlations for other circumstances. for instance, if you have two ordinal variables, you could use the Spearman rank Order Correlation (rho) or the Kendall rank order Correlation (tau). When one measure is a continuous interval level one and the other is dichotomous (i.e., two-category) you can use the Point-Biserial Correlation. For other situations, consulting the web-based statistics selection program, ***Selecting Statistics*** at http://trochim.human.cornell.edu/selstat/ssstart.htm

Inferential Statistics

With inferential statistics, you are trying to reach conclusions that extend beyond the immediate data alone. For instance, we use inferential statistics to try to infer from the sample data what the population might think. Or, we use inferential statistics to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance in this study. Thus, we use inferential statistics to make inferences from our data to more general conditions; we use descriptive statistics simply to describe what's going on in our data.

Here, I concentrate on inferential statistics that are useful in experimental and quasi-experimental research design or in program outcome evaluation. Perhaps one of the simplest inferential test is used when you want to compare the average performance of two groups on a single measure to see if there is a difference. You might want to know whether eighth-grade boys and girls differ in math test scores or whether a program group differs on the outcome measure from a control group. Whenever you wish to compare the average performance between two groups you should consider the t-test for differences between groups.

Most of the major inferential statistics come from a general family of statistical models known as the General Linear Model. This includes the t-test, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), regression analysis, and many of the multivariate methods like factor analysis, multidimensional scaling, cluster analysis, discriminant function analysis, and so on. Given the importance of the General Linear Model, it's a good idea for any serious social researcher to become familiar with its workings. The discussion of the General Linear Model here is very elementary and only considers the simplest straight-line model. However, it will get you familiar with the idea of the linear model and help prepare you for the more complex analyses described below.

One of the keys to understanding how groups are compared is embodied in the notion of the "dummy" variable. The name doesn't suggest that we are using variables that aren't very smart or, even worse, that the analyst who uses them is a "dummy"! Perhaps these variables would be better described as "proxy" variables. Essentially a dummy variable is one that uses discrete numbers, usually 0 and 1, to represent different groups in your study. Dummy variables are a simple idea that enable some pretty complicated things to happen. For instance, by including a simple dummy variable in an model, I can model two separate lines (one for each treatment group) with a single equation. To see how this works, check out the discussion on dummy variables.

One of the most important analyses in program outcome evaluations involves comparing the program and non-program group on the outcome variable or variables. How we do this depends on the research design we use. research designs are divided into two major types of designs: experimental and quasi-experimental. Because the analyses differ for each, they are presented separately.

Experimental Analysis. The simple two-group posttest-only randomized experiment is usually analyzed with the simple t-test or one-way ANOVA. The factorial experimental designs are usually analyzed with the Analysis of Variance (ANOVA) Model. Randomized Block Designs use a special form of ANOVA blocking model that uses dummy-coded variables to represent the blocks. The Analysis of Covariance Experimental Design uses, not surprisingly, the Analysis of Covariance statistical model.

Quasi-Experimental Analysis. The quasi-experimental designs differ from the experimental ones in that they don't use random assignment to assign units (e.g., people) to program groups. The lack of random assignment in these designs tends to complicate their analysis considerably. For example, to analyze the Nonequivalent Groups Design (NEGD) we have to adjust the pretest scores for measurement error in what is often called a Reliability-Corrected Analysis of Covariance model. In the Regression-Discontinuity Design, we need to be especially concerned about curvilinearity and model misspecification. Consequently, we tend to use a conservative analysis approach that is based on polynomial regression that starts by overfitting the likely true function and then reducing the model based on the results. The Regression Point Displacement Design has only a single treated unit. Nevertheless, the analysis of the RPD design is based directly on the traditional ANCOVA model.

When you've investigated these various analytic models, you'll see that they all come from the same family -- the General Linear Model. An understanding of that model will go a long way to introducing you to the intricacies of data analysis in applied and social research contexts.

- ## The T-Test

The t-test assesses whether the means of two groups are *statistically* different from each other. This analysis is appropriate whenever you want to compare the means of two groups, and especially appropriate as the analysis for the posttest-only two-group randomized experimental design.

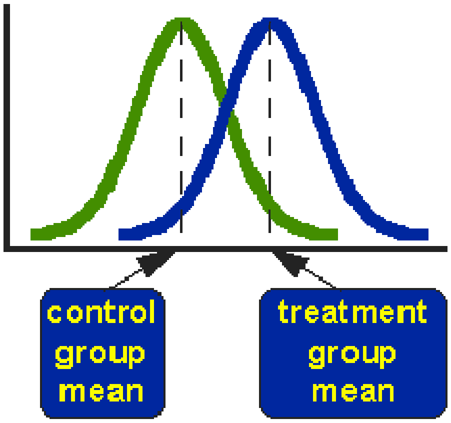


Figure 1. Idealized distributions for treated and comparison group posttest values.

Figure 1 shows the distributions for the treated (blue) and control (green) groups in a study. Actually, the figure shows the idealized distribution -- the actual distribution would usually be depicted with a histogram or bar graph. The figure indicates where the control and treatment group means are located. The question the t-test addresses is whether the means are statistically different.

What does it mean to say that the averages for two groups are statistically different? Consider the three situations shown in Figure 2. The first thing to notice about the three situations is that ***the difference between the means is the same in all three***. But, you should also notice that the three situations don't look the same -- they tell very different stories. The top example shows a case with moderate variability of scores within each group. The second situation shows the high variability case. the third shows the case with low variability. Clearly, we would conclude that the two groups appear most different or distinct in the bottom or low-variability case. Why? Because there is relatively little overlap between the two bell-shaped curves. In the high variability case, the group difference appears least striking because the two bell-shaped distributions overlap so much.

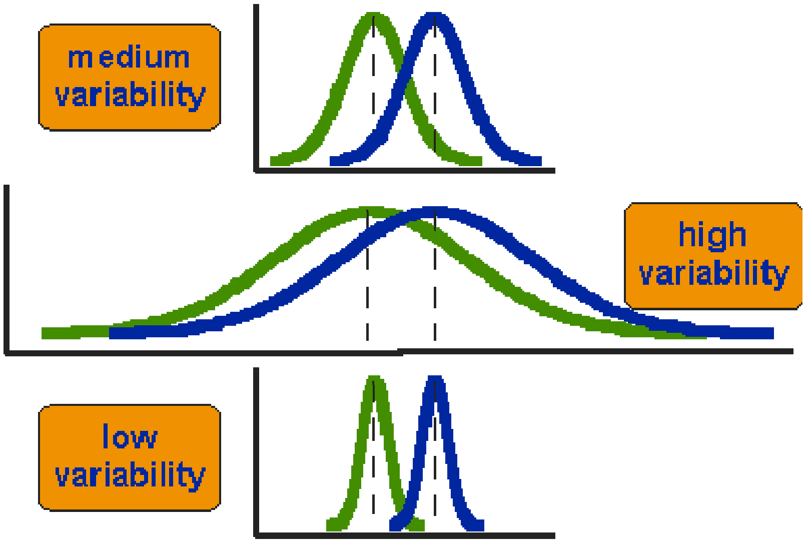


Figure 2. Three scenarios for differences between means.

This leads us to a very important conclusion: when we are looking at the differences between scores for two groups, we have to judge the difference between their means relative to the spread or variability of their scores. The t-test does just this.

Statistical Analysis of the t-test

The formula for the t-test is a ratio. The top part of the ratio is just the difference between the two means or averages. The bottom part is a measure of the variability or dispersion of the scores. This formula is essentially another example of the signal-to-noise metaphor in research: the difference between the means is the signal that, in this case, we think our program or treatment introduced into the data; the bottom part of the formula is a measure of variability that is essentially noise that may make it harder to see the group difference. Figure 3 shows the formula for the t-test and how the numerator and denominator are related to the distributions.

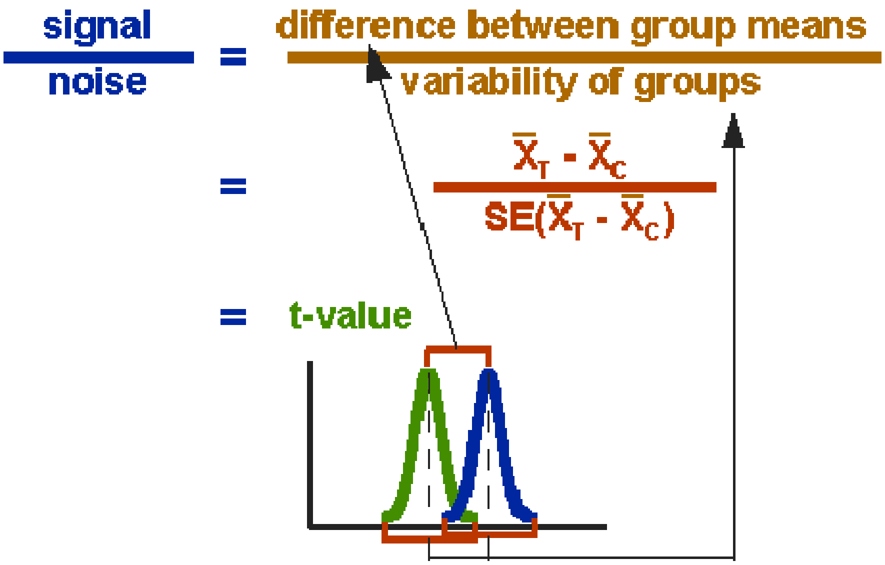


Figure 3. Formula for the t-test.

The top part of the formula is easy to compute -- just find the difference between the means. The bottom part is called the **standard error of the difference**. To compute it, we take the variance for each group and divide it by the number of people in that group. We add these two values and then take their square root. The specific formula is given in Figure 4:

$$SE(\bar{X}_T - \bar{X}_C) = \sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}$$

Figure 4. Formula for the Standard error of the difference between the means.

Remember, that the variance is simply the square of the standard deviation.

The final formula for the t-test is shown in Figure 5:

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

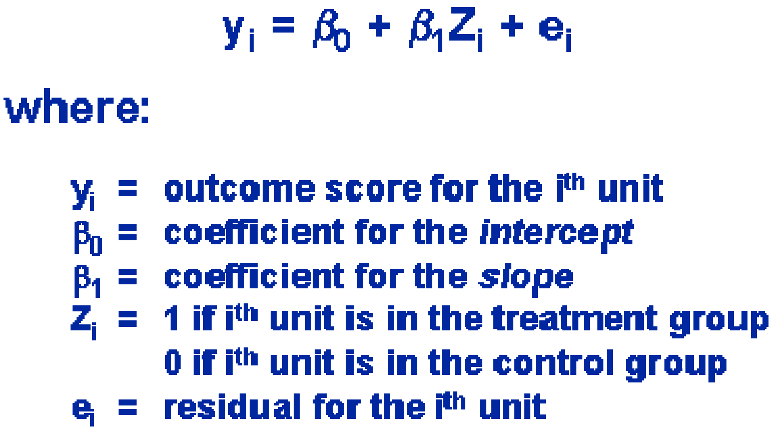
Figure 5. Formula for the t-test.

The t-value will be positive if the first mean is larger than the second and negative if it is smaller. Once you compute the t-value you have to look it up in a table of significance to test whether the ratio is large enough to say that the difference between the groups is not likely to have been a chance finding. To test the significance, you need to set a risk level (called the alpha level). In most social research, the "rule of thumb" is to set the alpha level at .05. This means that five times out of a hundred you would find a statistically significant difference between the means even if there was none (i.e., by "chance"). You also need to determine the degrees of freedom (df) for the test. In the t-test, the degrees of freedom is the sum of the persons in both groups minus 2. Given the alpha level, the df, and the t-value, you can look the t-value up in a standard table of significance (available as an appendix in the back of most statistics texts) to determine whether the t-value is large enough to be significant. If it is, you can conclude that the difference between the means for the two groups is different (even given the variability). Fortunately, statistical computer programs routinely print the significance test results and save you the trouble of looking them up in a table.

The t-test, one-way Analysis of Variance (ANOVA) and a form of regression analysis are mathematically equivalent (see the statistical analysis of the posttest-only randomized experimental design) and would yield identical results.

- **Dummy Variables**

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample in your study. In research design, a dummy variable is often used to distinguish different treatment groups. In the simplest case, we would use a 0,1 dummy variable where a person is given a value of 0 if they are in the control group or a 1 if they are in the treated group. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. This means that we don't need to write out separate equation models for each subgroup. The dummy variables act like **'switches'** that turn various parameters on and off in an equation. Another advantage of a 0,1 dummy-coded variable is that even though it is a nominal-level variable you can treat it statistically like an interval-level variable (if this made no sense to you, you probably should refresh your memory on levels of measurement). For instance, if you take an average of a **0,1** variable, the result is the proportion of **1**s in the distribution.



$$y_i = \beta_0 + \beta_1 Z_i + e_i$$

where:

y_i = outcome score for the i^{th} unit

β_0 = coefficient for the *intercept*

β_1 = coefficient for the *slope*

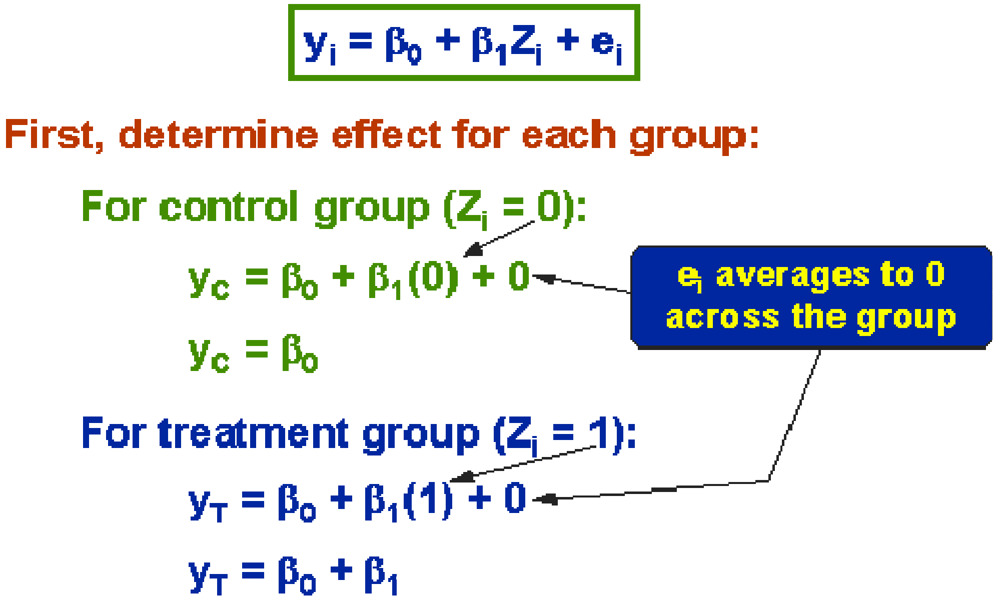
Z_i = 1 if i^{th} unit is in the treatment group

0 if i^{th} unit is in the control group

e_i = residual for the i^{th} unit

To illustrate dummy variables, consider the simple regression model for a posttest-only two-group randomized experiment. This model is essentially the same as conducting a t-test on the posttest means for two groups or conducting a one-way Analysis of Variance (ANOVA). The key term in the model is β_1, the estimate of the difference between the groups. To see how dummy variables work, we'll use this simple model to show you how to use them to pull out the separate sub-equations for each subgroup. Then we'll show how you estimate the difference between the subgroups by subtracting their respective equations. You'll see that we can pack an enormous amount of information into a single equation using dummy variables. All I want to show you here is that β_1 is the difference between the treatment and control groups.

To see this, the first step is to compute what the equation would be for each of our two groups separately. For the control group, Z = 0. When we substitute that into the equation, and recognize that by assumption the error term averages to 0, we find that the predicted value for the control group is β_0, the intercept. Now, to figure out the treatment group line, we substitute the value of 1 for Z, again recognizing that by assumption the error term averages to 0. The equation for the treatment group indicates that the treatment group value is the sum of the two beta values.



$$\boxed{y_i = \beta_0 + \beta_1 Z_i + e_i}$$

First, determine effect for each group:

For control group ($Z_i = 0$):

$$y_C = \beta_0 + \beta_1(0) + 0$$

$$y_C = \beta_0$$

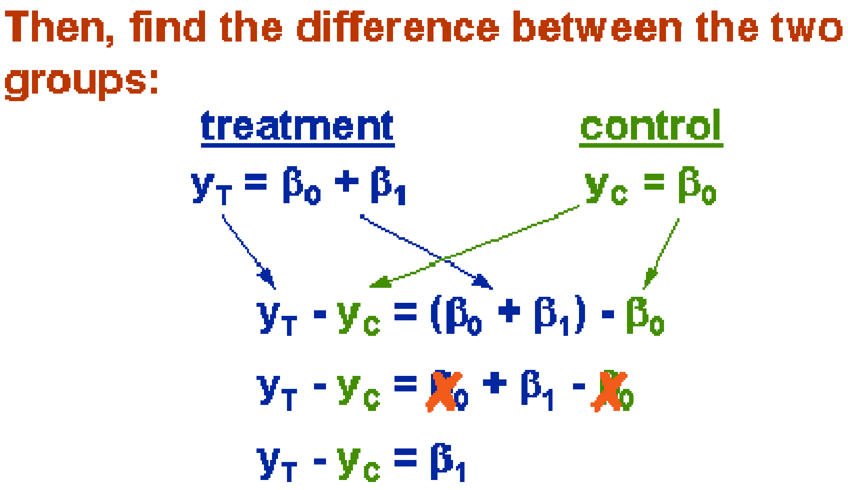
e_i averages to 0 across the group

For treatment group ($Z_i = 1$):

$$y_T = \beta_0 + \beta_1(1) + 0$$

$$y_T = \beta_0 + \beta_1$$

Now, we're ready to move on to the second step -- computing the difference between the groups. How do we determine that? Well, the difference must be the difference between the equations for the two groups that we worked out above. In other word, to find the difference between the groups we just find the difference between the equations for the two groups! It should be obvious from the figure that the difference is β_1. Think about what this means. The difference between the groups is β_1. OK, one more time just for the sheer heck of it. The difference between the groups in this model is β_1!



Then, find the difference between the two groups:

treatment: $y_T = \beta_0 + \beta_1$

control: $y_C = \beta_0$

$$y_T - y_C = (\beta_0 + \beta_1) - \beta_0$$

$$y_T - y_C = \cancel{\beta_0} + \beta_1 - \cancel{\beta_0}$$

$$y_T - y_C = \beta_1$$

Whenever you have a regression model with dummy variables, you can always see how the variables are being used to represent multiple subgroup equations by following the two steps described above:

- create separate equations for each subgroup by substituting the dummy values
- find the difference between groups by finding the difference between their equations

- **General Linear Model**

The General Linear Model (GLM) underlies most of the statistical analyses that are used in applied and social research. It is the foundation for the t-test, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), regression analysis, and many of the multivariate methods including factor analysis, cluster analysis, multidimensional scaling, discriminant function analysis, canonical correlation, and others. Because of its generality, the model is important for students of social research. Although a deep understanding of the GLM requires some advanced statistics training, I will attempt here to introduce the concept and provide a non-statistical description.

The Two-Variable Linear Model

The easiest point of entry into understanding the GLM is with the two-variable case. Figure 1 shows a bivariate plot of two variables. These may be any two continuous variables but, in the discussion that follows we will think of them as a pretest (on the x-axis) and a posttest (on the y-axis). Each dot on the plot represents the pretest and posttest score for an individual. The pattern clearly shows a positive relationship because, in general, people with higher pretest scores also have higher posttests, and vice versa.

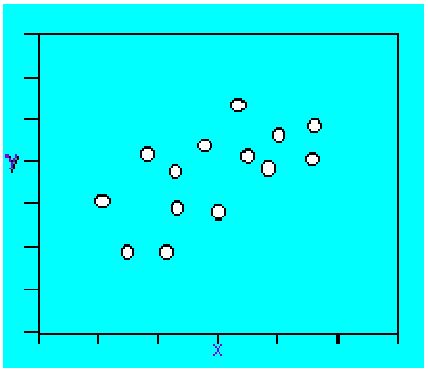


Figure 1. Bivariate plot.

The goal in our data analysis is to summarize or describe accurately what is happening in the data. The bivariate plot shows the data. How might we best summarize these data? Figure 2 shows that a straight line through the "cloud" of data points would effectively describe the pattern in the bivariate plot. Although the line does not perfectly describe any specific point (because no point falls precisely on the line), it does accurately describe the pattern in the data. When we fit a line to data, we are using what we call a **linear model**. The term "linear" refers to the fact that we are fitting a line. The term model refers to the equation that summarizes the line that we fit. A line like the one shown in Figure 2 is often referred to as a **regression line** and the analysis that produces it is often called **regression analysis**.

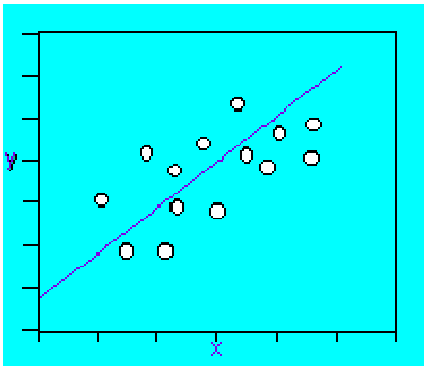


Figure 2. A straight-line summary of the data.

Figure 3 shows the equation for a straight line. You may remember this equation from your high school algebra classes where it is often stated in the form y = mx + b. In this equation, the components are:

y = the y-axis variable, the outcome or posttest
x = the x-axis variable, the pretest
b_0 = the intercept (value of y when x=0)
b_1 = the slope of the line

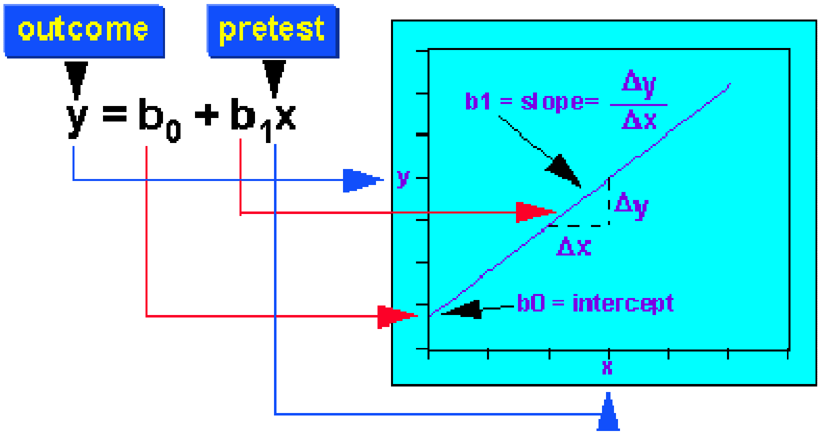


Figure 3. The straight-line model.

The slope of the line is the change in the posttest given in pretest units. As mentioned above, this equation does not perfectly fit the cloud of points in Figure 1. If it did, every point would fall on the line. We need one more component to describe the way this line is fit to the bivariate plot.

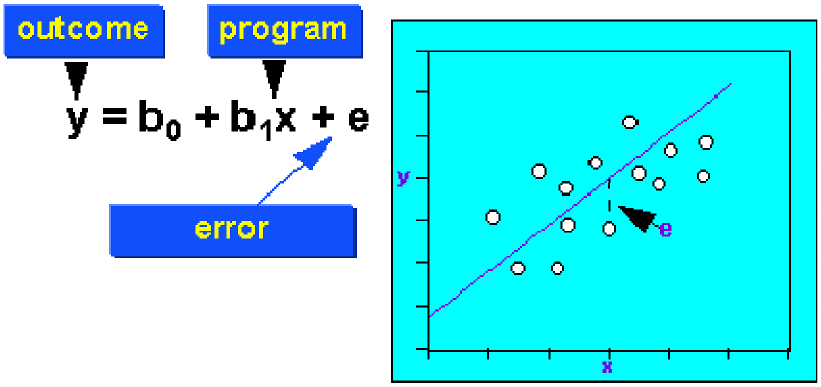


Figure 4. The two-variable linear model.

Figure 4 shows the equation for the two variable or bivariate linear model. The component that we have added to the equation in Figure 3 is an error term, e, that describes the vertical distance from the straight line to each point. This term is called "error" because it is the degree to which the line is in error in describing each point. When we fit the two-variable linear model to our data, we have an x and y score for each person in our study. We input these value pairs into a computer program. The program estimates the b_0 and b_1 values for us as indicated in Figure 5. We will actually get two numbers back that are estimates of those two values.

You can think of the two-variable regression line like any other descriptive statistic -- it is simply describing the relationship between two variables much as a mean describes the central tendency of a single variable. And, just as the mean does not accurately represent every value in a distribution, the regression line does not accurately represent every value in the bivariate distribution. We use these summaries because they show the general patterns in our data and allow us to describe these patterns in more concise ways than showing the entire distribution allows.

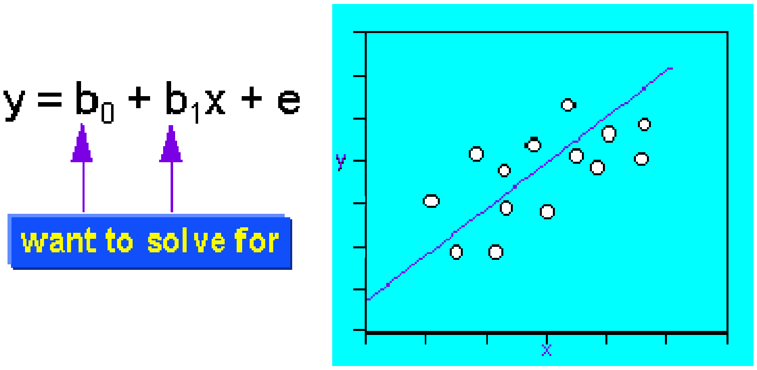


Figure 5. What the model estimates.

The General Linear Model

Given this brief introduction to the two-variable case, we are able to extend the model to its most general case. Essentially the GLM looks the same as the two variable model shown in Figure 4 -- it is just an equation. But the big difference is that each of the four terms in the GLM can represent a set of variables, not just a single one. So, the general linear model can be written:

$$\mathbf{y = b_0 + bx + e}$$

where:

y = a **set** of outcome variables
x = a **set** of pre-program variables or covariates
b0 = the **set** of intercepts (value of each y when each x=0)
b = a **set** of coefficients, one each for each x

You should be able to see that this model allows us to include an enormous amount of information. In an experimental or quasi-experimental study, we would represent the program or treatment with one or more dummy coded variables, each represented in the equation as an additional **x**-value (although we usually use the symbol **z** to indicate that the variable is a dummy-coded **x**). If our study has multiple outcome variables, we can include them as a set of y-values. If we have multiple pretests, we can include them as a set of x-values. For each **x**-value (and each **z**-value) we estimate a **b**-value that represents an **x,y** relationship. The estimates of these **b**-values, and the statistical testing of these estimates, is what enables us to test specific research hypotheses about relationships between variables or differences between groups.

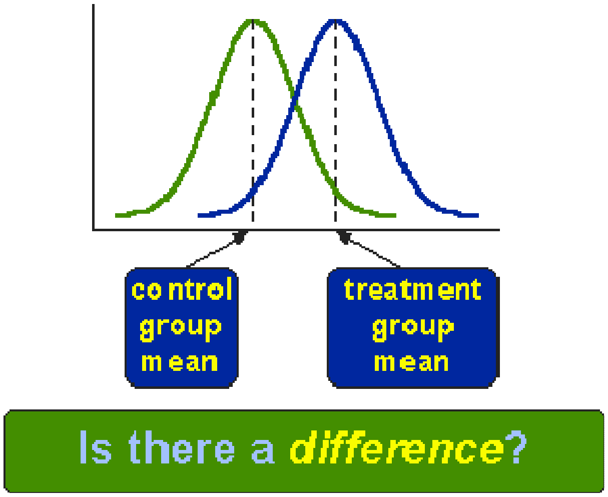
The GLM allows us to summarize a wide variety of research outcomes. The major problem for the researcher who uses the GLM is **model specification**. The researcher is responsible for specifying the exact equation that best summarizes the data for a study. If the model is misspecified, the estimates of the coefficients (the b-values) are likely to be biased (i.e., wrong) and the resulting equation will not describe the data accurately. In complex situations, this model specification problem can be a serious and difficult one (see, for example, the discussion of model specification in the statistical analysis of the regression-discontinuity design).

The GLM is one of the most important tools in the statistical analysis of data. It represents a major achievement in the advancement of social research in the twentieth century.

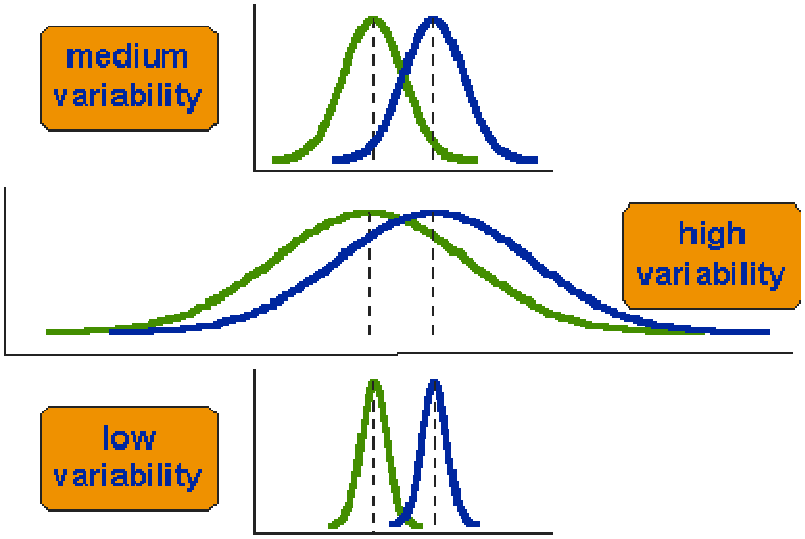
- **Posttest-Only Analysis**

To analyze the two-group posttest-only randomized experimental design we need an analysis that meets the following requirements:

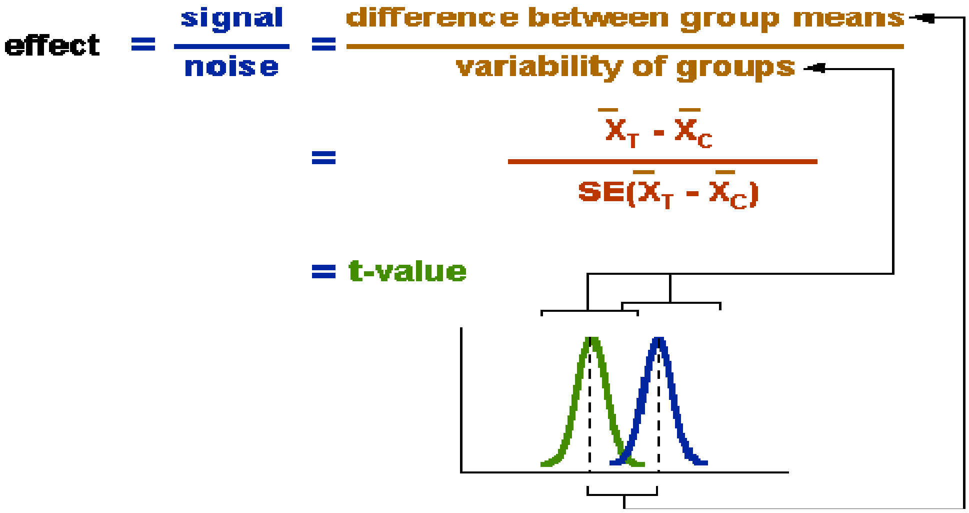
- has two groups
- uses a post-only measure
- has two distributions (measures), each with an average and variation
- assess treatment effect = statistical (i.e., non-chance) difference between the groups



Before we can proceed to the analysis itself, it is useful to understand what is meant by the term "difference" as in "Is there a difference between the groups?" Each group can be represented by a "bell-shaped" curve that describes the group's distribution on a single variable. You can think of the bell curve as a smoothed histogram or bar graph describing the frequency of each possible measurement response. In the figure, we show distributions for both the treatment and control group. The mean values for each group are indicated with dashed lines. The difference between the means is simply the horizontal difference between where the control and treatment group means hit the horizontal axis.



Now, let's look at three different possible outcomes, labeled medium, high and low variability. Notice that the differences between the means in all three situations is exactly the same. The only thing that differs between these is the variability or "spread" of the scores around the means. In which of the three cases would it be easiest to conclude that the means of the two groups are different? If you answered the low variability case, you are correct! Why is it easiest to conclude that the groups differ in that case? Because that is the situation with the least amount of overlap between the bell-shaped curves for the two groups. If you look at the high variability case, you should see that there quite a few control group cases that score in the range of the treatment group and vice versa. Why is this so important? Because, if you want to see if two groups are "different" it's not good enough just to subtract one mean from the other -- you have to take into account the variability around the means! A small difference between means will be hard to detect if there is lots of variability or noise. A large difference will between means will be easily detectable if variability is low. This way of looking at differences between groups is directly related to the signal-to-noise metaphor -- differences are more apparent when the signal is high and the noise is low.



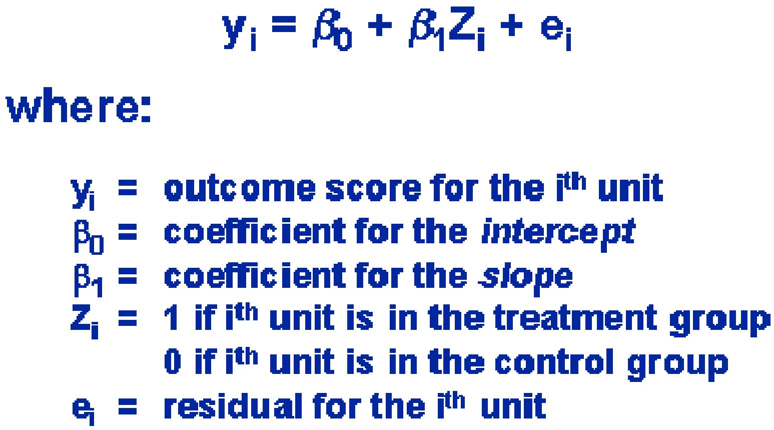
$$\text{effect} = \frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$

With that in mind, we can now examine how we estimate the differences between groups, often called the "effect" size. The top part of the ratio is the actual difference between means, The bottom part is an estimate of the variability around the means. In this context, we would calculate what is known as the standard error of the difference between the means. This standard error incorporates information about the standard deviation (variability) that is in each of the two groups. The ratio that we compute is called a t-value and describes the difference between the groups relative to the variability of the scores in the groups.

There are actually three different ways to estimate the treatment effect for the posttest-only randomized experiment. All three yield mathematically equivalent results, a fancy way of saying that they give you the exact same answer. So why are there three different ones? In large part, these three approaches evolved independently and, only after that, was it clear that they are essentially three ways to do the same thing. So, what are the three ways? First, we can compute an **independent t-test** as described above. Second, we could compute a **one-way Analysis of Variance (ANOVA)** between two independent groups. Finally, we can use **regression analysis** to regress the posttest values onto a dummy-coded treatment variable. Of these three, the regression analysis approach is the most general. In fact, you'll find that I describe the statistical models for all the experimental and quasi-experimental designs in regression model terms. You just need to be aware that the results from all three methods are identical.



$$y_i = \beta_0 + \beta_1 Z_i + e_i$$

where:

y_i = outcome score for the i^{th} unit
β_0 = coefficient for the *intercept*
β_1 = coefficient for the *slope*
Z_i = 1 if i^{th} unit is in the treatment group
0 if i^{th} unit is in the control group
e_i = residual for the i^{th} unit

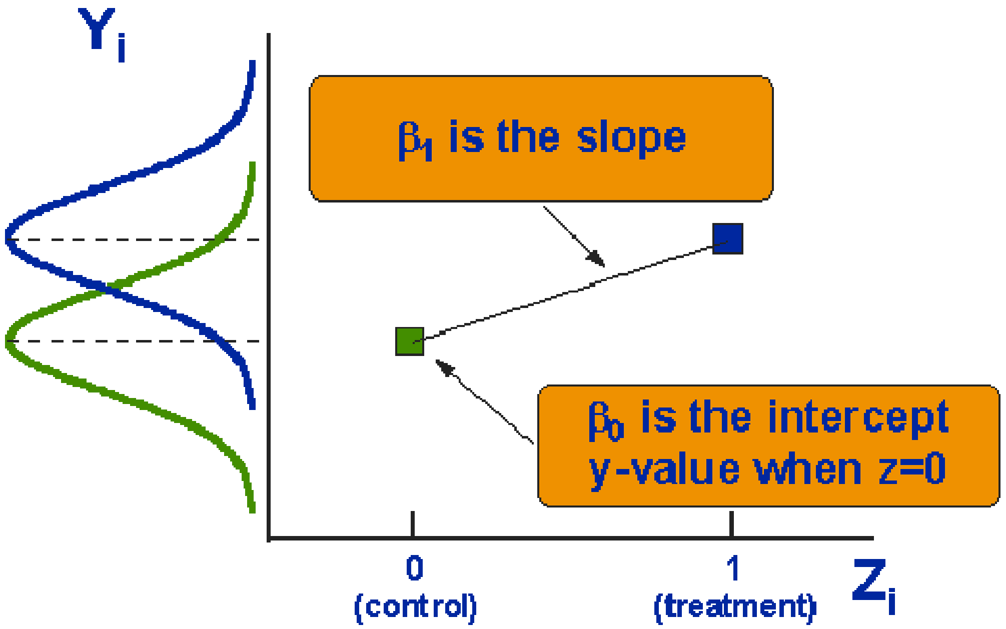
OK, so here's the statistical model in notational form. You may not realize it, but essentially this formula is just the equation for a straight line with a random error term thrown in (e_i). Remember high school algebra? Remember high school? OK, for those of you with faulty memories, you may recall that the equation for a straight line is often given as:

$$\mathbf{y = mx + b}$$

which, when rearranged can be written as:

$$\mathbf{y = b + mx}$$

(The complexities of the commutative property make you nervous? If this gets too tricky you may need to stop for a break. Have something to eat, make some coffee, or take the poor dog out for a walk.). Now, you should see that in the statistical model y_i is the same as y in the straight line formula, β_0 is the same as b, β_1 is the same as m, and Z_i is the same as x. In other words, in the statistical formula, β_0 is the intercept and β_1 is the slope.

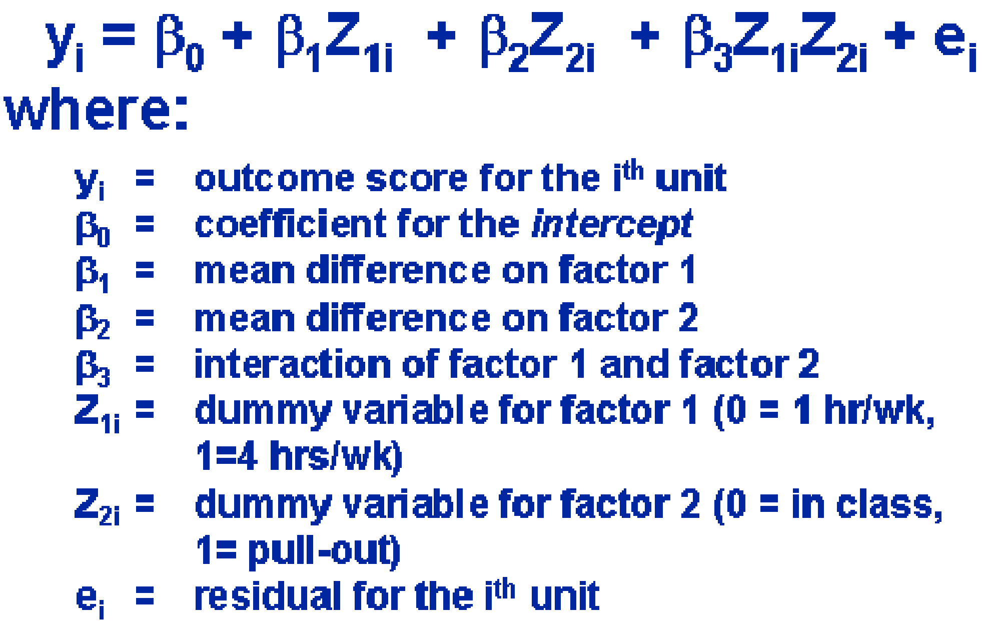


It is critical that you understand that the slope, β_1 is the same thing as the posttest difference between the means for the two groups. How can a slope be a difference between means? To see this, you have to take a look at a graph of what's going on. In the graph, we show the posttest on the vertical axis. This is exactly the same as the two bell-shaped curves shown in the graphs above except that here they're turned on their side. On the horizontal axis we plot the Z variable. This variable only has two values, a 0 if the person is in the control group or a 1 if the person is in the program group. We call this kind of variable a "dummy" variable because it is a "stand in" variable that represents the program or treatment conditions with its two values (note that the term "dummy" is not meant to be a slur against anyone, especially the people participating in your study). The two points in the graph indicate the average posttest value for the control (Z=0) and treated (Z=1) cases. The line that connects the two dots is only included for visual enhancement purposes -- since there are no Z values between 0 and 1 there can be no values plotted where the line is. Nevertheless, we can meaningfully speak about the slope of this line, the line that would connect the posttest means for the two values of Z. Do you remember the definition of slope? (Here we go again, back to high school!). The slope is the change in y over the change in x (or, in this case, Z). But we know that the "change in Z" between the groups is always equal to 1 (i.e., 1 - 0 = 1). So, the slope of the line must be equal to the difference between the average y-values for the two groups. That's what I set out to show (reread the first sentence of this paragraph). β_1 is the same value that you would get if you just subtract the two means from each other (in this case, because we set the treatment group equal to 1, this means we are subtracting the control group out of the treatment group value. A positive value implies that the treatment group mean is higher than the control, a negative means it's lower). But remember at the very beginning of this discussion I pointed out that just knowing the difference between the means was not good enough for estimating the treatment effect because it doesn't take into account the variability or spread of the scores. So how do we do that here? Every regression analysis program will give, in addition to the beta values, a report on whether each beta value is statistically significant. They report a t-value that tests whether the beta value differs from zero. It turns out that the t-value for the β_1 coefficient is the exact same number that you would get if you did a t-test for independent groups. And, it's the same as the square root of the F value in the two group one-way ANOVA (because $t^2 = F$).

Here's a few conclusions from all this:

- the t-test, one-way ANOVA and regression analysis all yield *same* results in this case
- the regression analysis method utilizes a *dummy variable* (Z) for treatment
- regression analysis is the most *general* model of the three.

- **Factorial Design Analysis**



$$y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{1i} Z_{2i} + e_i$$

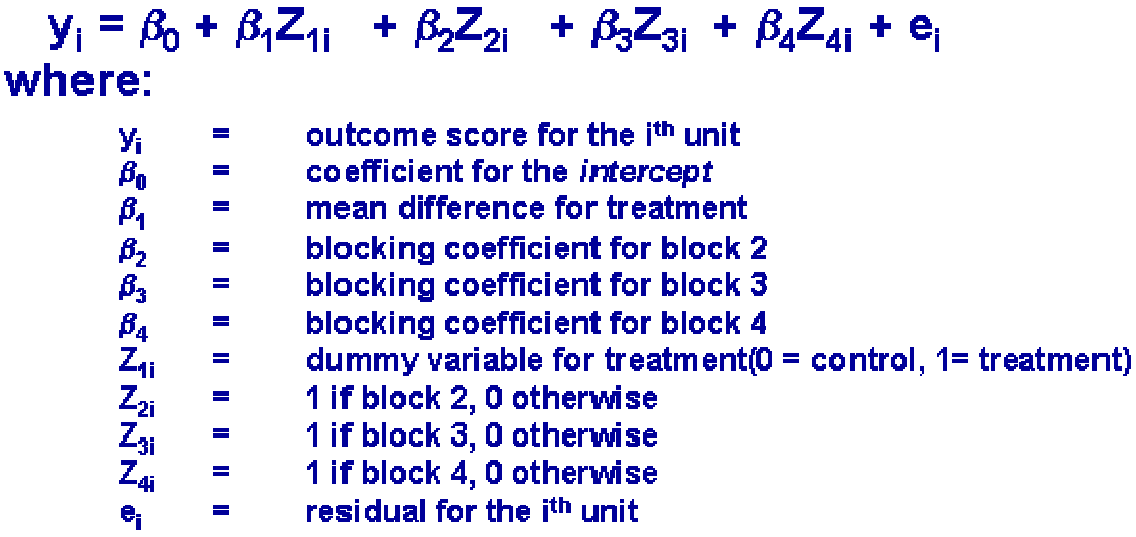
where:

y_i = outcome score for the ith unit
β_0 = coefficient for the *intercept*
β_1 = mean difference on factor 1
β_2 = mean difference on factor 2
β_3 = interaction of factor 1 and factor 2
Z_{1i} = dummy variable for factor 1 (0 = 1 hr/wk, 1=4 hrs/wk)
Z_{2i} = dummy variable for factor 2 (0 = in class, 1= pull-out)
e_i = residual for the ith unit

Here is the regression model statement for a simple 2 x 2 Factorial Design. In this design, we have one factor for time in instruction (1 hour/week versus 4 hours/week) and one factor for setting (in-class or pull-out). The model uses a dummy variable (represented by a Z) for each factor. In two-way factorial designs like this, we have two main effects and one interaction. In this model, the main effects are the statistics associated with the beta values that are adjacent to the Z-variables. The interaction effect is the statistic associated with β_3 (i.e., the t-value for this coefficient) because it is adjacent in the formula to the multiplication of (i.e., interaction of) the dummy-coded Z variables for the two factors. Because there are two dummy-coded variables, each having two values, you can write out 2 x 2 = 4 separate equations from this one general model. You might want to see if you can write out the equations for the four cells. Then, look at some of the differences between the groups. You can also write out two equations for each Z variable. These equations represent the main effect equations. To see the difference between levels of a factor, subtract the equations from each other. If you're confused about how to manipulate these equations, check the section on how dummy variables work.

- **Randomized Block Analysis**

I've decided to present the statistical model for the Randomized Block Design in regression analysis notation. Here is the model for a case where there are four blocks or homogeneous subgroups.



$$y_i = \beta_0 + \beta_1 Z_{1i} + \beta_2 Z_{2i} + \beta_3 Z_{3i} + \beta_4 Z_{4i} + e_i$$

where:

y_i = outcome score for the i^{th} unit

β_0 = coefficient for the *intercept*

β_1 = mean difference for treatment

β_2 = blocking coefficient for block 2

β_3 = blocking coefficient for block 3

β_4 = blocking coefficient for block 4

Z_{1i} = dummy variable for treatment(0 = control, 1= treatment)

Z_{2i} = 1 if block 2, 0 otherwise

Z_{3i} = 1 if block 3, 0 otherwise

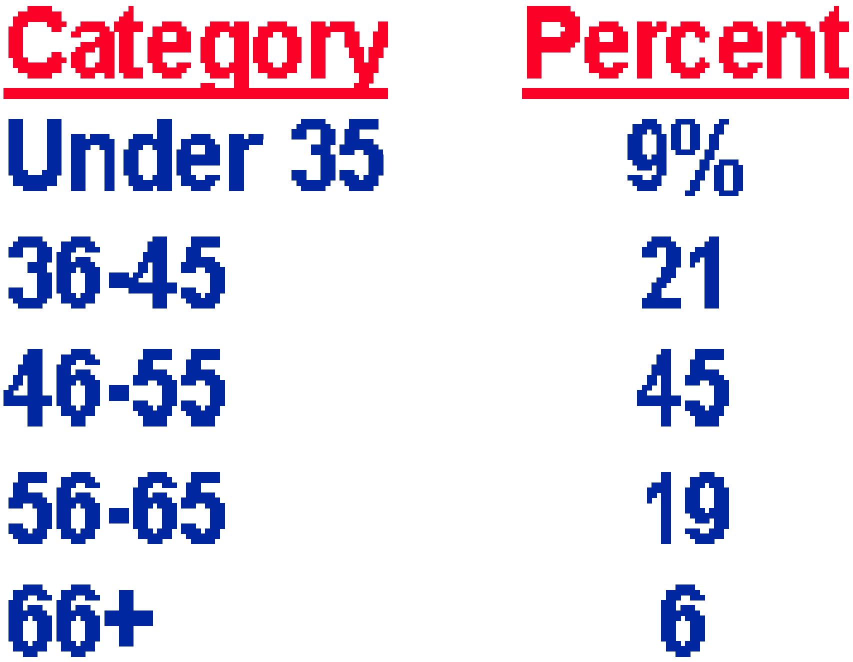
Z_{4i} = 1 if block 4, 0 otherwise

e_i = residual for the i^{th} unit

Notice that we use a number of dummy variables in specifying this model. We use the dummy variable Z_1 to represent the treatment group. We use the dummy variables Z_2, Z_3 and Z_4 to indicate blocks 2, 3 and 4 respectively. Analogously, the beta values (β's) reflect the treatment and blocks 2, 3 and 4. What happened to Block 1 in this model? To see what the equation for the Block 1 comparison group is, fill in your dummy variables and multiply through. In this case, all four Zs are equal to 0 and you should see that the intercept (β_0) is the estimate for the Block 1 control group. For the Block 1 treatment group, $Z_1 = 1$ and the estimate is equal to $\beta_0 + \beta_1$. By substituting the appropriate dummy variable "switches" you should be able to figure out the equation for any block or treatment group.

The data matrix that is entered into this analysis would consist of five columns and as many rows as you have participants: the posttest data, and one column of 0's or 1's for each of the four dummy variables.

- **Analysis of Covariance**



Category	Percent
Under 35	9%
36-45	21
46-55	45
56-65	19
66+	6

I've decided to present the statistical model for the Analysis of Covariance design in regression analysis notation. The model shown here is for a case where there is a single covariate and a treated and control group. We use a dummy variables in specifying this model. We use the dummy variable Z_i to represent the treatment group. The beta values (β's) are the parameters we are estimating. The value β_0 represents the intercept. In this model, it is the predicted posttest value for the control group for a given X value (and, when X=0, it is the intercept for the control group regression line). Why? Because a control group case has a Z=0 and since the Z variable is multiplied with β_2, that whole term would drop out.

The data matrix that is entered into this analysis would consist of three columns and as many rows as you have participants: the posttest data, one column of 0's or 1's to indicate which treatment group the participant is in, and the covariate score.

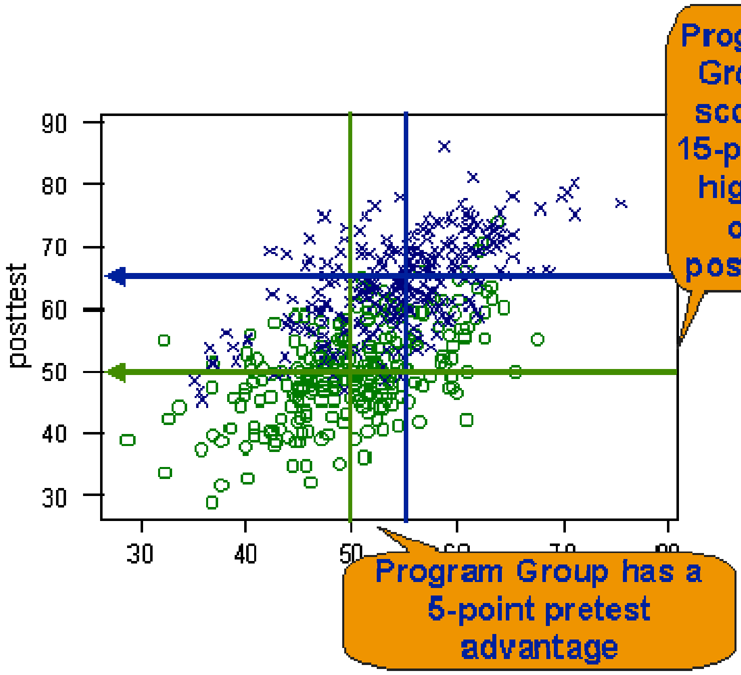
This model assumes that the data in the two groups are well described by straight lines that have the same slope. If this does not appear to be the case, you have to modify the model appropriately.

- **Nonequivalent Groups Analysis**

Analysis Requirements

N	O	X	O
N	O		O

The design notation for the Non-Equivalent Groups Design (NEGD) shows that we have two groups, a program and comparison group, and that each is measured pre and post. The statistical model that we would intuitively expect could be used in this situation would have a pretest variable, posttest variable, and a dummy variable variable that describes which group the person is in. These three variables would be the input for the statistical analysis. We would be interested in estimating the difference between the groups on the posttest after adjusting for differences on the pretest. This is essentially the Analysis of Covariance (ANCOVA) model as described in connection with randomized experiments (see the discussion of Analysis of Covariance and how we adjust for pretest differences). There's only one major problem with this model when used with the NEGD -- it doesn't work! Here, I'll tell you the story of why the ANCOVA model fails and what we can do to adjust it so it works correctly.



Program Group scores 15-points higher on posttest

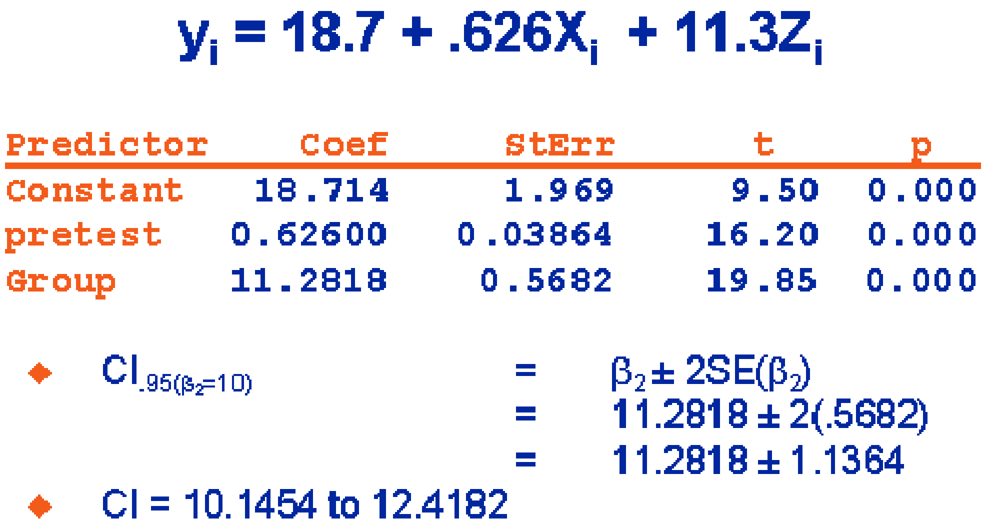
Program Group has a 5-point pretest advantage

A Simulated Example

To see what happens when we use the ANCOVA analysis on data from a NEGD, I created a computer simulation to generate hypothetical data. I created 500 hypothetical persons, with 250 in the program and 250 in the comparison condition. Because this is a nonequivalent design, I made the groups nonequivalent on the pretest by adding five points to each program group person's pretest score. Then, I added 15 points to each program person's posttest score. When we take the initial 5-point advantage into account, we should find a 10 point program effect. The bivariate plot shows the data from this simulation.

I then analyzed the data with the ANCOVA model. Remember that the way I set this up I should observe approximately a 10-point program effect if the ANCOVA analysis works correctly. The results are presented in the table.

In this analysis, I put in three scores for each person: a pretest score (X), a posttest score (Y) and either a 0 or 1 to indicate whether the person was in the program (Z=1) or comparison (Z=0) group. The table shows the equation that the ANCOVA model estimates. The equation has the three values I put in, (X, Y and Z) and the three coefficients that the program estimates. The key coefficient is the one next to the program variable Z. This coefficient estimates the average difference between the program and comparison groups (because it's the coefficient paired with the dummy variable indicating what group the person is in). The value should be 10 because I put in a 10 point difference. In this analysis, the actual value I got was 11.3 (or 11.2818, to be more precise). Well, that's not too bad, you might say. It's fairly close to the 10-point effect I put in. But we need to determine if the obtained value of 11.2818 is statistically different from the true value of 10. To see whether it is, we have to construct a confidence interval around our estimate and examine the difference between 11.2818 and 10 relative to the variability in the data. Fortunately the program does this automatically for us. If you look in the table, you'll see that the third line shows the coefficient associated with the difference between the groups, the standard error for that coefficient (an indicator of variability), the t-value, and the probability value. All the t-value shows is that the coefficient of 11.2818 is statistically different from zero. But we want to know whether it is different from the true treatment effect value of 10. To determine this, we can construct a confidence interval around the t-value, using the standard error. We know that the 95% confidence interval is the coefficient plus or minus two times the standard error value. The calculation shows that the 95% confidence interval for our 11.2818 coefficient is 10.1454 to 12.4182. Any value falling within this range can't be considered different beyond a 95% level from our obtained value of 11.2818. But the true value of 10 points falls outside the range. In other words, our estimate of 11.2818 is significantly different from the true value. In still other words, the results of this analysis are biased -- we got the wrong answer. In this example, our estimate of the program effect is significantly larger than the true program effect (even though the difference between 10 and 11.2818 doesn't seem that much larger, it exceeds chance levels). So, we have a problem when we apply the analysis model that our intuition tells us makes the most sense for the NEGD. To understand why this bias occurs, we have to look a little more deeply at how the statistical analysis works in relation to the NEGD.

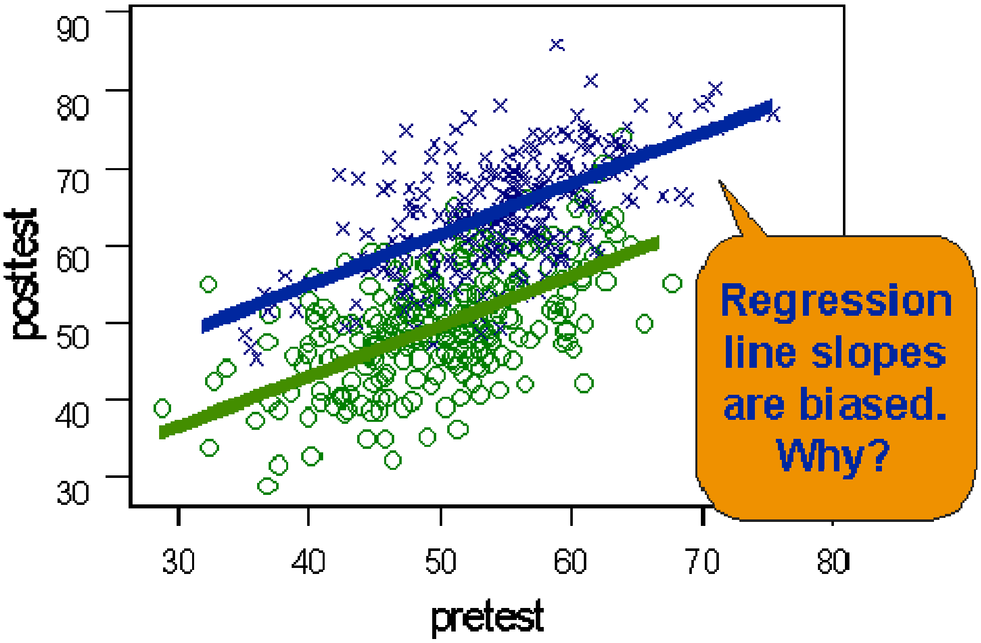


$$y_i = 18.7 + .626X_i + 11.3Z_i$$

Predictor	Coef	StErr	t	p
Constant	18.714	1.969	9.50	0.000
pretest	0.62600	0.03864	16.20	0.000
Group	11.2818	0.5682	19.85	0.000

- $CI_{.95(\beta_2=10)} = \beta_2 \pm 2SE(\beta_2) = 11.2818 \pm 2(.5682) = 11.2818 \pm 1.1364$
- CI = 10.1454 to 12.4182

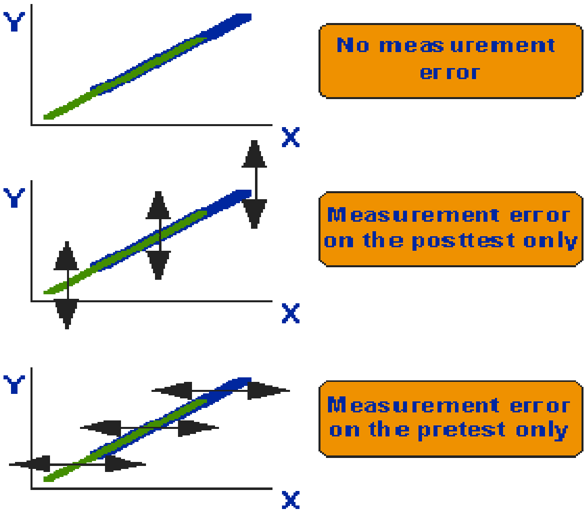
The Problem



Why is the ANCOVA analysis biased when used with the NEGD? And, why isn't it biased when used with a pretest-posttest randomized experiment? Actually, there are several things happening to produce the bias, which is why it's somewhat difficult to understand (and counterintuitive). Here are the two reasons we get a bias:

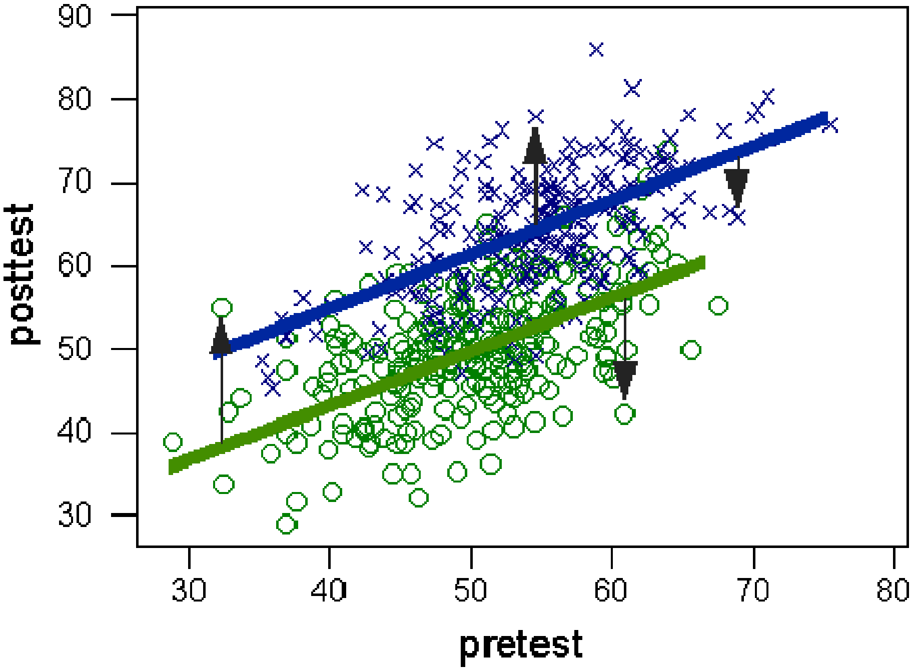
- pretest measurement error which leads to the attenuation or "flattening" of the slopes in the regression lines
- group nonequivalence

The first problem actually also occurs in randomized studies, but it doesn't lead to biased treatment effects because the groups are equivalent (at least probabilistically). It is the combination of both these conditions that causes the problem. And, understanding the problem is what leads us to a solution in this case.


Y
X
No measurement error
Y
X
Measurement error on the posttest only
Y
X
Measurement error on the pretest only

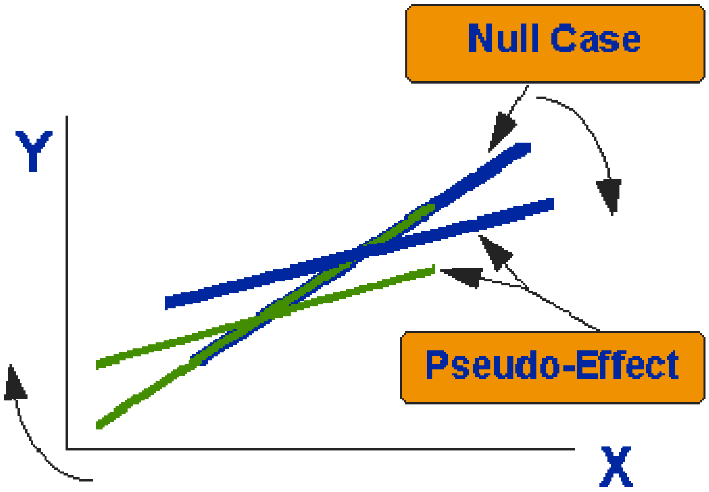
Regression and Measurement Error. We begin our attempt to understand the source of the bias by considering how error in measurement affects regression analysis. We'll consider three different measurement error scenarios to see what error does. In all three scenarios, we assume that there is no true treatment effect, that the null hypothesis is true. The first scenario is the case of no measurement error at all. In this hypothetical case, all of the points fall right on the regression lines themselves. The second scenario introduces measurement error on the posttest, but not on the pretest. The figure shows that when we have posttest error, we are disbursing the points vertically -- up and down -- from the regression lines. Imagine a specific case, one person in our study. With no measurement error the person would be expected to score on the regression line itself. With posttest measurement error, they would do better or worse on the posttest than they should. And, this would lead their score to be displaced vertically. In the third scenario we have measurement error only on the pretest. It stands to reason that in this case we would be displacing cases horizontally -- left and right -- off of the regression lines. For these three hypothetical cases, none of which would occur in reality, we can see how data points would be disbursed.

How Regression Fits Lines. Regression analysis is a **least squares** analytic procedure. The actual criterion for fitting the line is to fit it so that you minimize the sum of the squares of the residuals from the regression line. Let's deconstruct this sentence a bit. The key term is "residual." The residual is the vertical distance from the regression line to each point.

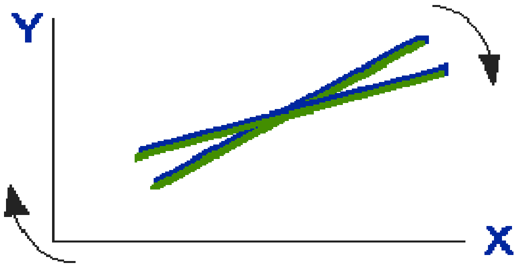


The graph shows four residuals, two for each group. Two of the residuals fall above their regression line and two fall below. What is the criterion for fitting a line through the cloud of data points? Take all of the residuals within a group (we'll fit separate lines for the program and comparison group). If they are above the line they will be positive and if they're below they'll be negative values. Square all the residuals in the group. Compute the sum of the squares of the residuals -- just add them. That's it. Regression analysis fits a line through the data that yields the smallest sum of the squared residuals. How it does this is another matter. But you should now understand what it's doing. The key thing to notice is that ***the regression line is fit in terms of the residuals and the residuals are vertical displacements from the regression line***.

How Measurement Error Affects Slope Now we're ready to put the ideas of the previous two sections together. Again, we'll consider our three measurement error scenarios described above. When there is no measurement error, the slopes of the regression lines are unaffected. The figure shown earlier shows the regression lines in this no error condition. Notice that there is no treatment effect in any of the three graphs shown in the figure (there would be a treatment effect only if there was a vertical displacement between the two lines). Now, consider the case where there is measurement error on the posttest. Will the slopes be affected? The answer is no. Why? Because in regression analysis we fit the line relative to the vertical displacements of the points. Posttest measurement error affects the vertical dimension, and, if the errors are random, we would get as many residuals pushing up as down and the slope of the line would, on average, remain the same as in the null case. There would, in this posttest measurement error case, be more variability of data around the regression line, but the line would be located in the same place as in the no error case.



Now, let's consider the case of measurement error on the pretest. In this scenario, errors are added along the horizontal dimension. But regression analysis fits the lines relative to vertical displacements. So how will this affect the slope? The figure illustrates what happens. If there was no error, the lines would overlap as indicated for the null case in the figure. When we add in pretest measurement error, we are in effect elongating the horizontal dimension without changing the vertical. Since regression analysis fits to the vertical, this would force the regression line to stretch to fit the horizontally elongated distribution. The only way it can do this is by rotating around its center point. The result is that the line has been "flattened" or "attenuated" -- the slope of the line will be lower when there is pretest measurement error than it should actually be. You should be able to see that if we flatten the line in each group by rotating it around its own center that this introduces a displacement between the two lines that was not there originally. Although there was no treatment effect in the original case, we have introduced a false or "pseudo" effect. The biased estimate of the slope that results from pretest measurement error introduces a phony treatment effect. In this example, it introduced an effect where there was none. In the simulated example shown earlier, it exaggerated the actual effect that we had constructed for the simulation.



Why Doesn't the Problem Occur in Randomized Designs? So, why doesn't this pseudo-effect occur in the randomized Analysis of Covariance design? The next figure shows that even in the randomized design, pretest measurement error *does* cause the slopes of the lines to be flattened. But, we don't get a pseudo-effect in the randomized case even though the attenuation occurs. Why? Because in the randomized case the two groups are equivalent on the pretest -- there is no horizontal difference between the lines. The lines for the two groups overlap perfectly in the null case. So, when the attenuation occurs, it occurs the same way in both lines and there is no vertical displacement introduced between the lines. Compare this figure to the one above. You should now see that the difference is that in the NEGD case above we have the attenuation of slopes and the initial nonequivalence between the groups. Under these circumstances the flattening of the lines introduces a displacement. In the randomized case we also get the flattening, but there is no displacement because there is no nonequivalence between the groups initially.

Summary of the Problem. So where does this leave us? The ANCOVA statistical model seemed at first glance to have all of the right components to correctly model data from the NEGD. But we found that it didn't work correctly -- the estimate of the treatment effect was biased. When we examined why, we saw that the bias was due to two major factors: the attenuation of slope that results from pretest measurement error coupled with the initial nonequivalence between the groups. The problem is not caused by posttest measurement error because of the criterion that is used in regression analysis to fit the line. It does not occur in randomized experiments because there is no pretest nonequivalence. We might also guess from these arguments that the bias will be greater with greater nonequivalence between groups -- the less similar the groups the bigger the problem. In real-life research, as opposed to simulations, you can count on measurement error on all measurements -- we never measure perfectly. So, in nonequivalent groups designs we now see that the ANCOVA analysis that seemed intuitively sensible can be expected to yield incorrect results!

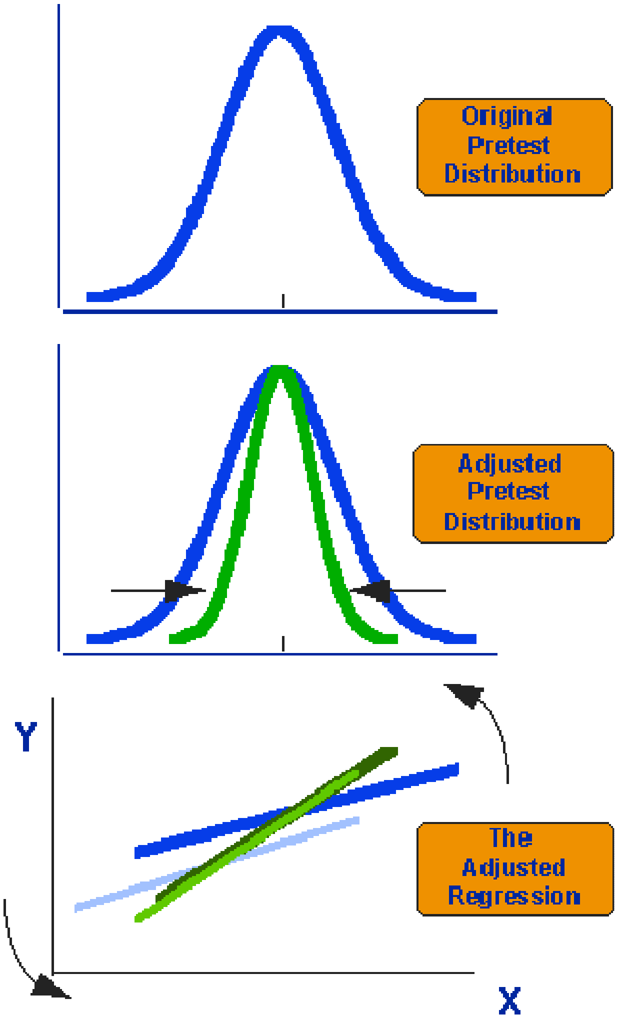
The Solution

Now that we understand the problem in the analysis of the NEGD, we can go about trying to fix it. Since the problem is caused in part by measurement error on the pretest, one way to deal with it would be to address the measurement error issue. If we could remove the pretest measurement error and approximate the no pretest error case, there would be no attenuation or flattening of the regression lines and no pseudo-effect introduced. To see how we might adjust for pretest measurement error, we need to recall what we know about measurement error and its relation to reliability of measurement.

Recall from reliability theory and the idea of true score theory that reliability can be defined as the ratio:

$$\frac{\mathbf{var(T)}}{\mathbf{var(T) + var(e)}}$$

where T is the true ability or level on the measure and e is measurement error. It follows that the reliability of the pretest is directly related to the amount of measurement error. If there is no measurement error on the pretest, the var(e) term in the denominator is zero and reliability = 1. If the pretest is nothing but measurement error, the Var(T) term is zero and the reliability is 0. That is, if the measure is nothing but measurement error, it is totally unreliable. If half of the measure is true score and half is measurement error, the reliability is.5. This shows that there is a direct relationship between measurement error and reliability -- reliability reflects the proportion of measurement error in your measure. Since measurement error on the pretest is a necessary condition for bias in the NEGD (if there is no pretest measurement error there is no bias even in the NEGD), if we correct for the measurement error we correct for the bias. But, we can't see measurement error directly in our data (remember, only God can see how much of a score is True Score and how much is error). However, we can estimate the reliability. Since reliability is directly related to measurement error, we can use the reliability estimate as a proxy for how much measurement error is present. And, we can adjust pretest scores using the reliability estimate to correct for the attenuation of slopes and remove the bias in the NEGD.



The Reliability-Corrected ANCOVA. We're going to solve the bias in ANCOVA treatment effect estimates for the NEGD using a "reliability" correction that will adjust the pretest for measurement error. The figure shows what a reliability correction looks like. The top graph shows the pretest distribution as we observe it, with measurement error included in it. Remember that I said above that adding measurement error widens or elongates the horizontal dimension in the bivariate distribution. In the frequency distribution shown in the top graph, we know that the distribution is wider than it would be if there was no error in measurement. The second graph shows that what we really want to do in adjusting the pretest scores is to squeeze the pretest distribution inwards by an amount proportionate to the amount that measurement error elongated widened it. We will do this adjustment separately for the program and comparisons groups. The third graph shows what effect "squeezing" the pretest would have on the regression lines -- It would increase their slopes rotating them back to where they truly belong and removing the bias that was introduced by the measurement error. In effect, we are doing the opposite of what measurement error did so that we can correct for the measurement error.

All we need to know is how much to squeeze the pretest distribution in to correctly adjust for measurement error. The answer is in the reliability coefficient. Since reliability is an estimate of the proportion of your measure that is true score relative to error, it should tell us how much we have to "squeeze." In fact, the formula for the adjustment is very simple:

$$X_{adj} = \bar{X} + r(X - \bar{X})$$

where:

X_{adj} = adjusted pretest value

$\bar{X}$ = original pretest value

r = reliability

The idea in this formula is that we are going to construct new pretest scores for each person. These new scores will be "adjusted" for pretest unreliability by an amount proportional to the reliability. Each person's score will be closer to the pretest mean for that group. The formula tells us how much closer. Let's look at a few examples. First, let's look at the case where there is no pretest measurement error. Here, reliability would be 1. In this case, we actually don't want to adjust the data at all. Imagine that we have a person with a pretest score of 40, where the mean of the pretest for the group is 50. We would get an adjusted score of:

$\mathbf{X_{adj} = 50 + 1(40-50)}$
$\mathbf{X_{adj} = 50 + 1(-10)}$
$\mathbf{X_{adj} = 50 -10}$
$\mathbf{X_{adj} = 40}$

Or, in other words, we wouldn't make any adjustment at all. That's what we want in the no measurement error case.

Now, let's assume that reliability was relatively low, say .5. For a person with a pretest score of 40 where the group mean is 50, we would get:

$\mathbf{X_{adj} = 50 + .5(40-50)}$
$\mathbf{X_{adj} = 50 + .5(-10)}$
$\mathbf{X_{adj} = 50 - 5}$
$\mathbf{X_{adj} = 45}$

Or, when reliability is .5, we would move the pretest score halfway in towards the mean (halfway from its original value of 40 towards the mean of 50, or to 45).

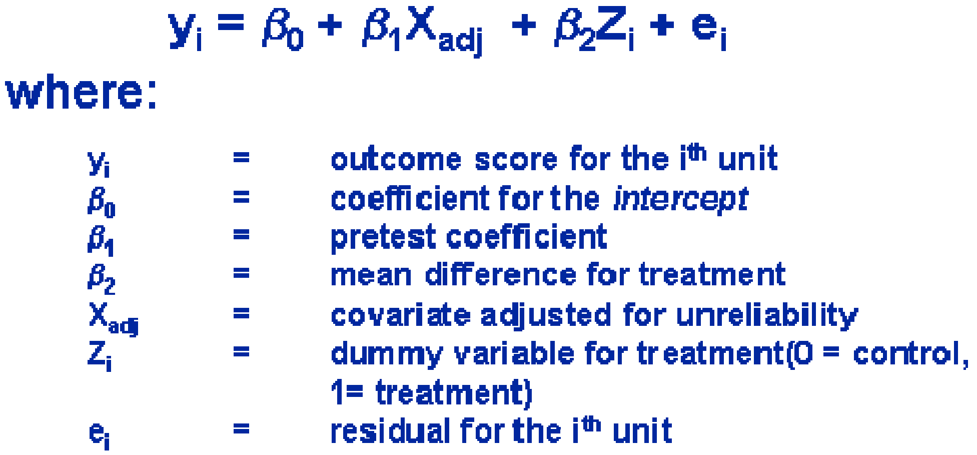
Finally, let's assume that for the same case the reliability was stronger at .8. The reliability adjustment would be:

X_{adj} = 50 + .8(40-50)
X_{adj} = 50 + .8(-10)
X_{adj} = 50 - 8
X_{adj} = 42

That is, with reliability of .8 we would want to move the score in 20% towards its mean (because if reliability is .8, the amount of the score due to error is 1 -.8 = .2).

You should be able to see that if we make this adjustment to all of the pretest scores in a group, we would be "squeezing" the pretest distribution in by an amount proportionate to the measurement error (1 - reliability). It's important to note that we need to make this correction separately for our program and comparison groups.

We're now ready to take this adjusted pretest score and substitute it for the original pretest score in our ANCOVA model:



$$y_i = \beta_0 + \beta_1 X_{adj} + \beta_2 Z_i + e_i$$

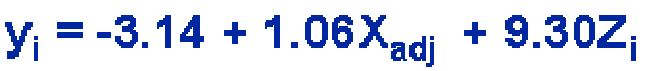
where:

y_i	=	outcome score for the ith unit
β_0	=	coefficient for the *intercept*
β_1	=	pretest coefficient
β_2	=	mean difference for treatment
X_{adj}	=	covariate adjusted for unreliability
Z_i	=	dummy variable for treatment(0 = control, 1= treatment)
e_i	=	residual for the ith unit

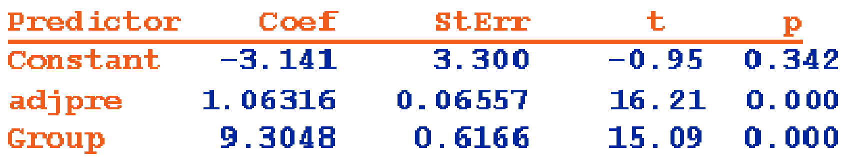
Notice that the only difference is that we've changed the X in the original ANCOVA to the term X_{adj}.

The Simulation Revisited.

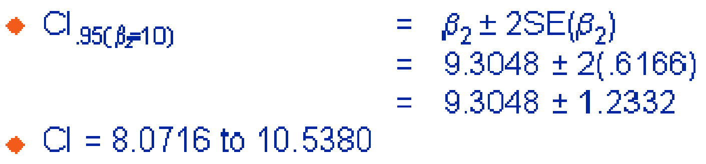
So, let's go see how well our adjustment works. We'll use the same simulated data that we used earlier. The results are:



$$y_i = -3.14 + 1.06X_{adj} + 9.30Z_i$$



```
Predictor      Coef      StErr        t       p
Constant     -3.141      3.300     -0.95   0.342
adjpre      1.06316    0.06557     16.21   0.000
Group        9.3048     0.6166     15.09   0.000
```



◆ $CI_{.95(\beta_2=10)} = \beta_2 \pm 2SE(\beta_2)$

$= 9.3048 \pm 2(.6166)$

$= 9.3048 \pm 1.2332$

◆ CI = 8.0716 to 10.5380

This time we get an estimate of the treatment effect of 9.3048 (instead of 11.2818). This estimate is closer to the true value of 10 points that we put into the simulated data. And, when we construct a 95% confidence interval for our adjusted estimate, we see that the true value of 10 falls within the interval. That is, the analysis estimated a treatment effect that is not statistically different from the true effect -- it is an unbiased estimate.

You should also compare the slope of the lines in this adjusted model with the original slope. Now, the slope is nearly 1 at 1.06316, whereas before it was .626 -- considerably lower or "flatter." The slope in our adjusted model approximates the expected true slope of the line (which is 1). The original slope showed the attenuation that the pretest measurement error caused.

So, the reliability-corrected ANCOVA model is used in the statistical analysis of the NEGD to correct for the bias that would occur as a result of measurement error on the pretest.

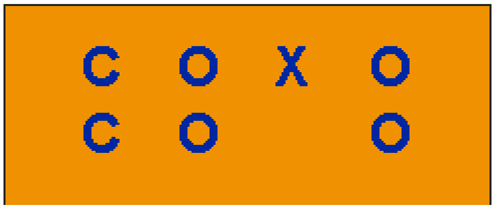
Which Reliability To Use?

There's really only one more major issue to settle in order to finish the story. We know from reliability theory that we can't calculate the true reliability, we can only estimate it. There a variety of reliability estimates and they're likely to give you different values. Cronbach's Alpha tends to be a high estimate of reliability. The test-retest reliability tends to be a lower-bound estimate of reliability. So which do we use in our correction formula? The answer is: both! When analyzing data from the NEGD it's safest to do two analyses, one with an upper-bound estimate of reliability and one with a lower-bound one. If we find a significant treatment effect estimate with both, we can be fairly confident that we would have found a significant effect in data that had no pretest measurement error.

This certainly doesn't feel like a very satisfying conclusion to our rather convoluted story about the analysis of the NEGD, and it's not. In some ways, I look at this as the price we pay when we give up random assignment and use intact groups in a NEGD -- our analysis becomes more complicated as we deal with adjustments that are needed, in part, because of the nonequivalence between the groups. Nevertheless, there are also benefits in using nonequivalent groups instead of randomly assigning. You have to decide whether the tradeoff is worth it.

- **Regression-Discontinuity Analysis**

Analysis Requirements



The basic RD Design is a two-group pretest-posttest model as indicated in the design notation. As in other versions of this design structure (e.g., the Analysis of Covariance Randomized Experiment, the Nonequivalent Groups Design), we will need a statistical model that includes a term for the pretest, one for the posttest, and a dummy-coded variable to represent the program.

Assumptions in the Analysis

It is important before discussing the specific analytic model to understand the assumptions which must be met. This presentation assumes that we are dealing with the basic RD design as described earlier. Variations in the design will be discussed later. There are five central assumptions which must be made in order for the analytic model which is presented to be appropriate, each of which is discussed in turn:

1. **The Cutoff Criterion.** The cutoff criterion must be followed without exception. When there is misassignment relative to the cutoff value (unless it is known to be random), a selection threat arises and estimates of the effect of the program are likely to be biased. Misassignment relative to the cutoff, often termed a "fuzzy" RD design, introduces analytic complexities that are outside the scope of this discussion.
2. **The Pre-Post Distribution.** It is assumed that the pre-post distribution is describable as a polynomial function. If the true pre-post relationship is logarithmic, exponential or some other function, the model given below is misspecified and estimates of the effect of the program are likely to be biased. Of course, if the data can be transformed to create a polynomial distribution prior to analysis the model below may be appropriate although it is likely to be more problematic to interpret. It is also sometimes the case that even if the true relationship is not polynomial, a sufficiently high-order polynomial will adequately account for whatever function exists. However, the analyst is not likely to know whether this is the case.
3. **Comparison Group Pretest Variance.** There must be a sufficient number of pretest values in the comparison group to enable adequate estimation of the true relationship (i.e., pre-post regression line) for that group. It is usually desirable to have variability in the program group as well although this is not strictly required because one can project the comparison group line to a single point for the program group.
4. **Continuous Pretest Distribution.** Both groups must come from a single continuous pretest distribution with the division between groups determined by the cutoff. In some cases one might be able to find intact groups (e.g., two groups of patients from two different geographic locations) which serendipitously divide on some measure so as to imply some cutoff. Such naturally discontinuous groups must be used with caution because of the greater likelihood that if they differed naturally at the cutoff prior to the program such a difference could reflect a selection bias which could introduce natural pre-post discontinuities at that point.
5. **Program Implementation.** It is assumed that the program is uniformly delivered to all recipients, that is, that they all receive the same dosage, length of stay, amount of training, or whatever. If

this is not the case, it is necessary to model explicitly the program as implemented, thus complicating the analysis somewhat.

The Curvilinearity Problem

The major problem in analyzing data from the RD design is model misspecification. As will be shown below, when you misspecify the statistical model, you are likely to get biased estimates of the treatment effect. To introduce this idea, let's begin by considering what happens if the data (i.e., the bivariate pre-post relationship) are curvilinear and we fit a straight-line model to the data.

Figure 1. A curvilinear relationship.

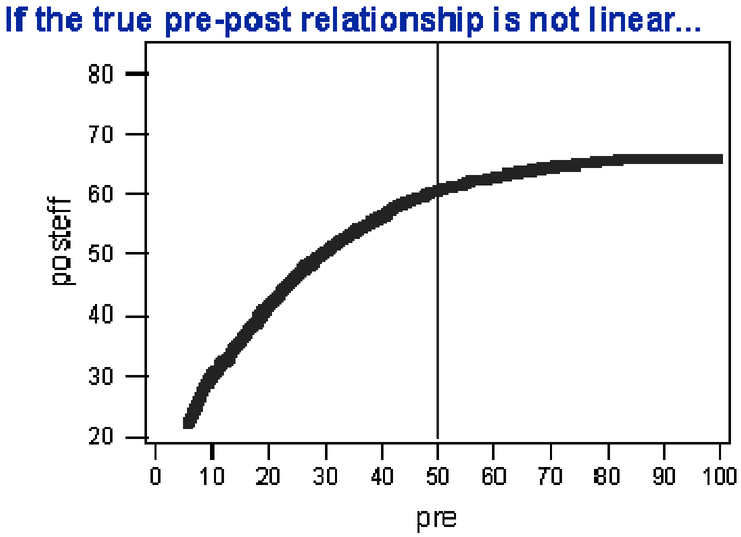
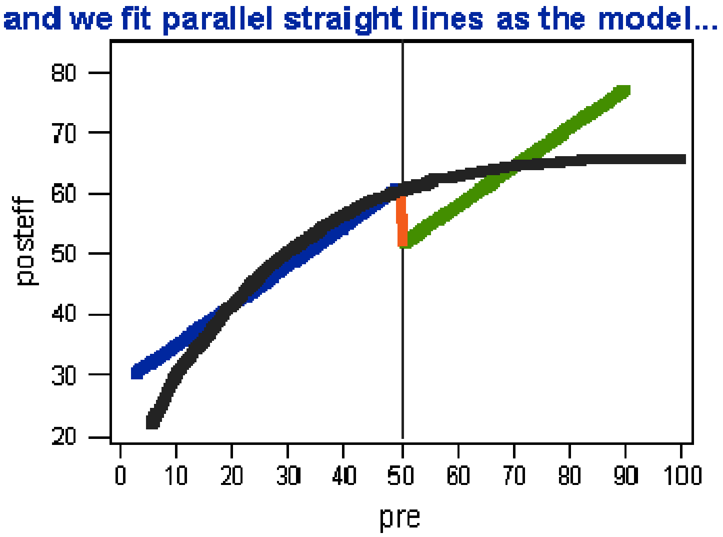


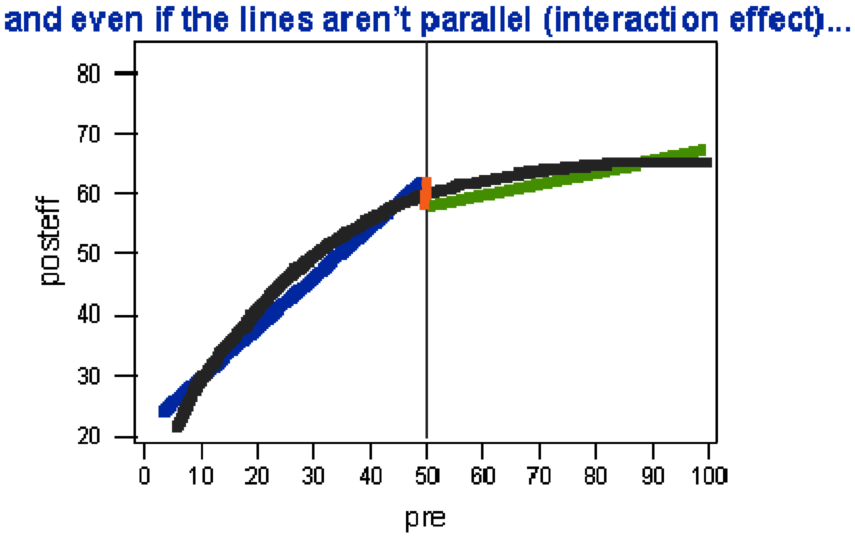
Figure 1 shows a simple curvilinear relationship. If the curved line in Figure 1 describes the pre-post relationship, then we need to take this into account in our statistical model. Notice that, although there is a cutoff value at 50 in the figure, there is no jump or discontinuity in the line at the cutoff. This indicates that there is no effect of the treatment.

Figure 2. A curvilinear relationship fit with a straight-line model.



Now, look at Figure 2. The figure shows what happens when we fit a straight-line model to the curvilinear relationship of Figure 1. In the model, we restricted the slopes of both straight lines to be the same (i.e., we did not allow for any interaction between the program and the pretest). You can see that the straight line model suggests that there is a jump at the cutoff, even though we can see that in the true function there is no discontinuity.

Figure 3. A curvilinear relationship fit with a straight-line model with different slopes for each line (an interaction effect).



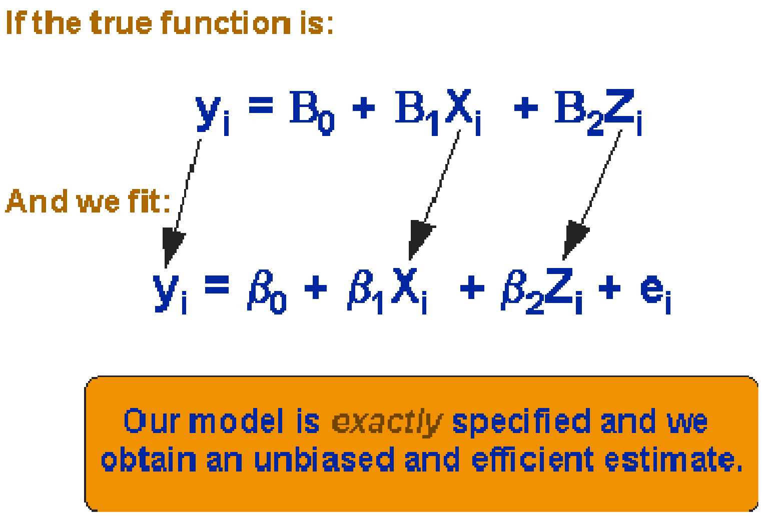
Even allowing the straight line slopes to differ doesn't solve the problem. Figure 3 shows what happens in this case. Although the pseudo-effect in this case is smaller than when the slopes are forced to be equal, we still obtain a pseudo-effect.

The conclusion is a simple one. If the true model is curved and we fit only straight-lines, we are likely to conclude wrongly that the treatment made a difference when it did not. This is a specific instance of the more general problem of model specification.

Model Specification

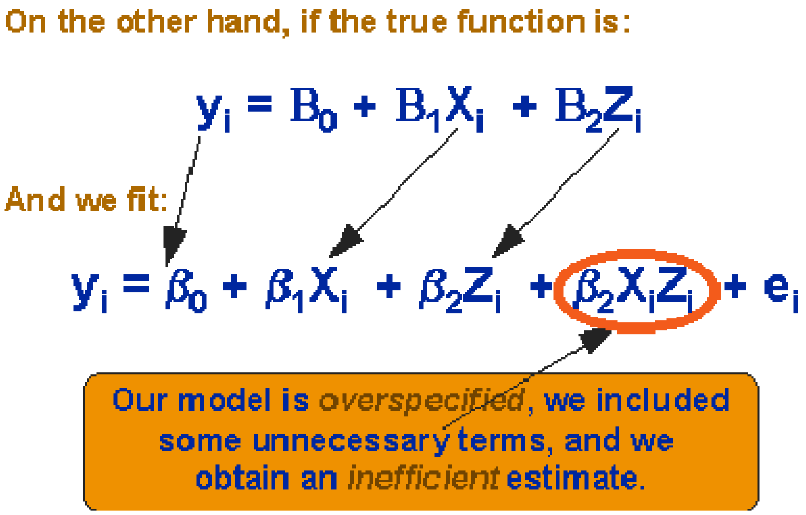
To understand the model specification issue and how it relates to the RD design, we must distinguish three types of specifications. Figure 4 shows the case where we **exactly specify** the true model. What does "exactly specify" mean? The top equation describes the "truth" for the data. It describes a simple straight-line pre-post relationship with a treatment effect. Notice that it includes terms for the posttest Y, the pretest X, and the dummy-coded treatment variable Z. The bottom equation shows the model that we specify in the analysis. It too includes a term for the posttest Y, the pretest X, and the dummy-coded treatment variable Z. And that's all it includes -- there are no unnecessary terms in the model that we specify. When we exactly specify the true model, we get unbiased and efficient estimates of the treatment effect.

Figure 4. An exactly specified model.


If the true function is:
$y_i = B_0 + B_1X_i + B_2Z_i$
And we fit:
$y_i = \beta_0 + \beta_1X_i + \beta_2Z_i + e_i$
Our model is *exactly* specified and we obtain an unbiased and efficient estimate.

Now, let's look at the situation in Figure 5. The true model is the same as in Figure 4. However, this time we specify an analytic model that includes an extra and unnecessary term. In this case, because we included all of the necessary terms, our estimate of the treatment effect will be unbiased. However, we pay a price for including unneeded terms in our analysis -- the treatment effect estimate will not be efficient. What does this mean? It means that the chance that we will conclude our treatment doesn't work when it in fact does is increased. Including an unnecessary term in the analysis is like adding unnecessary noise to the data -- it makes it harder for us to see the effect of the treatment even if it's there.

Figure 5. An overspecified model.



On the other hand, if the true function is:

$$y_i = B_0 + B_1X_i + B_2Z_i$$

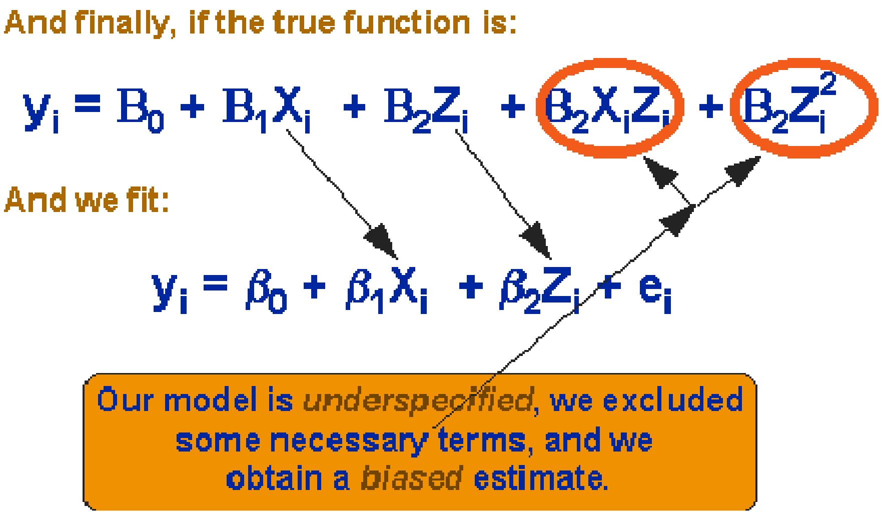
And we fit:

$$y_i = \beta_0 + \beta_1X_i + \beta_2Z_i + \beta_2X_iZ_i + e_i$$

Our model is *overspecified*, we included some unnecessary terms, and we obtain an *inefficient* estimate.

Finally, consider the example described in Figure 6. Here, the truth is more complicated than our model. In reality, there are two terms that we did not include in our analysis. In this case, we will get a treatment effect estimate that is both biased and inefficient.

Figure 6. An underspecified model.



And finally, if the true function is:

$$y_i = B_0 + B_1X_i + B_2Z_i + B_2X_iZ_i + B_2Z_i^2$$

And we fit:

$$y_i = \beta_0 + \beta_1X_i + \beta_2Z_i + e_i$$

Our model is *underspecified*, we excluded some necessary terms, and we obtain a *biased* estimate.

Analysis Strategy

Given the discussion of model misspecification, we can develop a modeling strategy that is designed, first, to guard against biased estimates and, second, to assure maximum efficiency of estimates. The best option would obviously be to specify the true model exactly. But this is often difficult to achieve in practice because the true model is often obscured by the error in the data. If we have to make a mistake -- if we must misspecify the model -- we would generally prefer to overspecify the true model rather than underspecify. Overspecification assures that we have included all necessary terms even at the expense of unnecessary ones. It will yield an unbiased estimate of the effect, even though it will be inefficient. Underspecification is the situation we would most like to avoid because it yields both biased and inefficient estimates.

Given this preference sequence, our general analysis strategy will be to begin by specifying a model that we are fairly certain is overspecified. The treatment effect estimate for this model is likely to be unbiased although it will be inefficient. Then, in successive analyses, gradually remove higher-order terms until the treatment effect estimate appears to differ from the initial one or until the model diagnostics (e.g., residual plots) indicate that the model fits poorly.

Steps in the Analysis

The basic RD analysis involves five steps:

1. **Transform the Pretest.**

2. The analysis begins by subtracting the cutoff value from each pretest score, creating the modified pretest term shown in Figure 7. This is done in order to set the intercept equal to the cutoff value. How does this work? If we subtract the cutoff from every pretest value, the modified pretest will be equal to 0 where it was originally at the cutoff value. Since the intercept is by definition the y-value when x=0, what we have done is set X to 0 at the cutoff, making the cutoff the intercept point.

   Figure 7. Transforming the pretest by subtracting the cutoff value.

   $$\tilde{X}_i = X_i - X_c$$

3. **Examine Relationship Visually.**

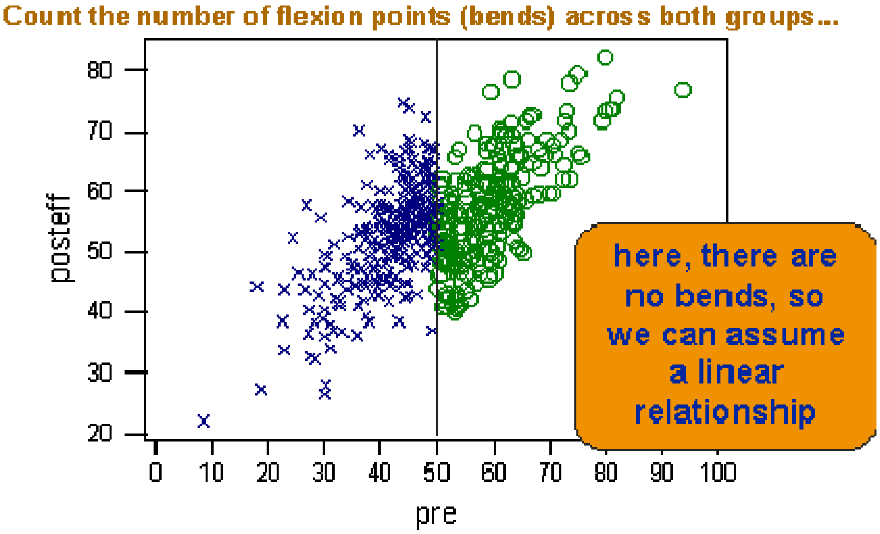
   There are two major things to look for in a graph of the pre-post relationship. First it is important to determine whether there is any visually discernable discontinuity in the relationship at the cutoff. The discontinuity could be a change in level vertically (main effect), a change in slope (interaction effect), or both. If it is visually clear that there is a discontinuity at the cutoff then one should not be satisfied with analytic results which indicate no program effect. However, if no discontinuity is visually apparent, it may be that variability in the data is masking an effect and one must attend carefully to the analytic results.

   The second thing to look for in the bivariate relationship is the degree of polynomial which may be required as indicated by the bivariate slope of the distribution, particularly in the comparison group. A good approach is to count the number of flexion points (i.e., number of times the distribution "flexes" or "bends") which are apparent in the distribution. If the distribution appears linear, there are no flexion points. A single flexion point could be indicative of a second (quadratic) order polynomial. This information will be used to determine the initial model which will be specified.

4. **Specify Higher-Order Terms and Interactions.**

Depending on the number of flexion points detected in step 2, one next creates transformations of the modified assignment variable, X. The rule of thumb here is that you go two orders of polynomial higher than was indicated by the number of flexion points. Thus, if the bivariate relationship appeared linear (i.e., there were no flexion points), one would want to create transformations up to a second-order (0 + 2) polynomial. This is shown in Figure 8. There do not appear to be any inflexion points or "bends" in the bivariate distribution of Figure 8.

Figure 8. Bivariate distribution with no flexion points.



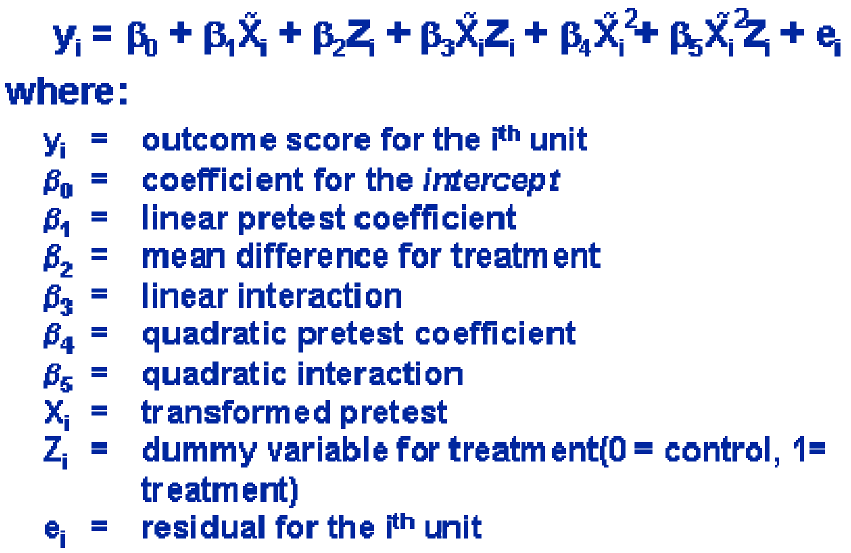
The first order polynomial already exists in the model (X) and so one would only have to create the second-order polynomial by squaring X to obtain X^2. For each transformation of X one also creates the interaction term by multiplying the polynomial by Z. In this example there would be two interaction terms: X_iZ_i and $X_i^2Z_i$. Each transformation can be easily accomplished through straightforward multiplication on the computer. If there appeared to be two flexion points in the bivariate distribution, one would create transformations up to the fourth (2 + 2) power and their interactions.

Visual inspection need not be the only basis for the initial determination of the degree of polynomial which is needed. Certainly, prior experience modeling similar data should be taken into account. The rule of thumb given here implies that one should err on the side of overestimating the true polynomial function which is needed for reasons outlined above in discussing model specification. For whatever power is initially estimated from visual inspection one should construct all transformations and their interactions up to that power. Thus if the fourth power is chosen, one should construct all four terms X to X^4 and their interactions.

5. **Estimate Initial Model.**

At this point, one is ready to begin the analysis. Any acceptable multiple regression program can be used to accomplish this on the computer. One simply regresses the posttest scores, Y, on the modified pretest X, the treatment variable Z, and all higher-order transformations and interactions created in step 3 above. The regression coefficient associated with the Z term (i.e., the group membership variable) is the estimate of the main effect of the program. If there is a vertical discontinuity at the cutoff it will be estimated by this coefficient. One can test the significance of the coefficient (or any other) by constructing a standard t-test using the standard error of the coefficient which is invariably supplied in the computer program output.

Figure 9. The initial model for the case of no flexion points (full quadratic model specification).



$$y_i = \beta_0 + \beta_1\tilde{X}_i + \beta_2 Z_i + \beta_3\tilde{X}_i Z_i + \beta_4\tilde{X}_i^2 + \beta_5\tilde{X}_i^2 Z_i + e_i$$

where:

y_i = outcome score for the ith unit

β_0 = coefficient for the *intercept*

β_1 = linear pretest coefficient

β_2 = mean difference for treatment

β_3 = linear interaction

β_4 = quadratic pretest coefficient

β_5 = quadratic interaction

X_i = transformed pretest

Z_i = dummy variable for treatment(0 = control, 1= treatment)

e_i = residual for the ith unit

If the analyst at step 3 correctly overestimated the polynomial function required to model the distribution then the estimate of the program effect will at least be unbiased. However, by including terms which may not be needed in the true model, the estimate is likely to be inefficient, that is, standard error terms will be inflated and hence the significance of the program effect may be underestimated. Nevertheless, if at this point in the analysis the coefficient is highly significant, it would be reasonable to conclude that there is a program effect. The direction of the effect is interpreted based on the sign of the coefficient and the direction of scale of the posttest. Interaction effects can also be examined. For instance, a linear interaction would be implied by a significant regression coefficient for the XZ term.

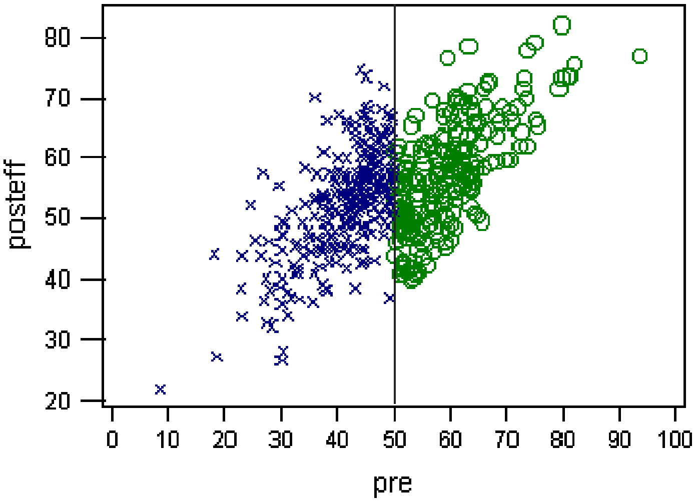
6. **Refining the Model.**

On the basis of the results of step 4 one might wish to attempt to remove apparently unnecessary terms and reestimate the treatment effect with greater efficiency. This is a tricky procedure and should be approached cautiously if one wishes to minimize the possibility of bias. To accomplish this one should certainly examine the output of the regression analysis in step 4 noting the degree to which the overall model fits the data, the presence of any insignificant coefficients and the pattern of residuals. A conservative way to decide how to refine the model would be to begin by examining the highest-order term in the current model and its interaction. If both coefficients are nonsignificant, and the goodness-of-fit measures and pattern of residuals indicate a good fit one might drop these two terms and reestimate the resulting model. Thus, if one estimated up to a fourth-order polynomial, and found the coefficients for X^4 and X^4Z were nonsignificant, these terms can be dropped and the third-order model respecified. One would repeat this procedure until: 1) either of the coefficients is significant; b) the goodness-of-fit measure drops appreciably; or, c) the pattern of residuals indicates a poorly fitting model. The final model may still include unnecessary terms but there are likely to be fewer of these and, consequently, efficiency should be greater. Model specification procedures which involve dropping any term at any stage of the analysis are more dangerous and more likely to yield biased estimates because of the considerable multicolinearity which will exist between the terms in the model.

Example Analysis

It's easier to understand how data from a RD Design is analyzed by showing an example. The data for this example are shown in Figure 10.

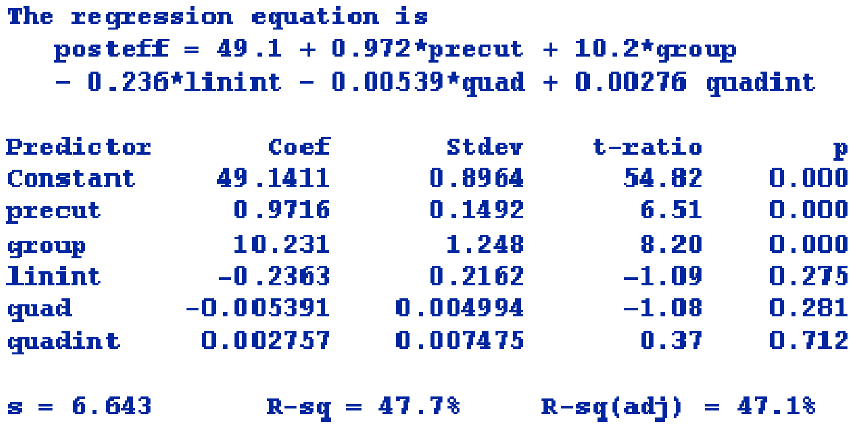
Figure 10. Bivariate distribution for example RD analysis.



Several things are apparent visually. First, there is a whopping treatment effect. In fact, Figure 10 shows simulated data where the true treatment effect is 10 points. Second, both groups are well described by straight lines -- there are no flexion points apparent. Thus, the initial model we'll specify is the full quadratic one shown above in Figure 9.

The results of our initial specification are shown in Figure 11. The treatment effect estimate is the one next to the "group" variable. This initial estimate is 10.231 (SE = 1.248) -- very close to the true value of 10 points. But notice that there is evidence that several of the higher-order terms are not statistically significant and may not be needed in the model. Specifically, the linear interaction term "linint" (XZ), and both the quadratic (X^2) and quadratic interaction (X^2Z) terms are not significant.

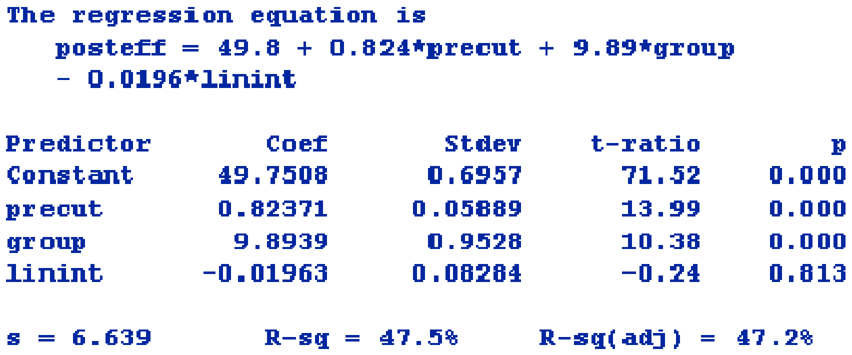
Figure 11. Regression results for the full quadratic model.



```
The regression equation is
   posteff = 49.1 + 0.972*precut + 10.2*group
   - 0.236*linint - 0.00539*quad + 0.00276 quadint

Predictor        Coef        Stdev    t-ratio        p
Constant      49.1411       0.8964      54.82    0.000
precut         0.9716       0.1492       6.51    0.000
group          10.231        1.248       8.20    0.000
linint        -0.2363       0.2162      -1.09    0.275
quad        -0.005391     0.004994      -1.08    0.281
quadint      0.002757     0.007475       0.37    0.712

s = 6.643       R-sq = 47.7%     R-sq(adj) = 47.1%
```

Although we might be tempted (and perhaps even justified) to drop all three terms from the model, if we follow the guidelines given above in Step 5 we will begin by dropping only the two quadratic terms "quad" and "quadint". The results for this model are shown in Figure 12.

Figure 12. Regression results for initial model without quadratic terms.

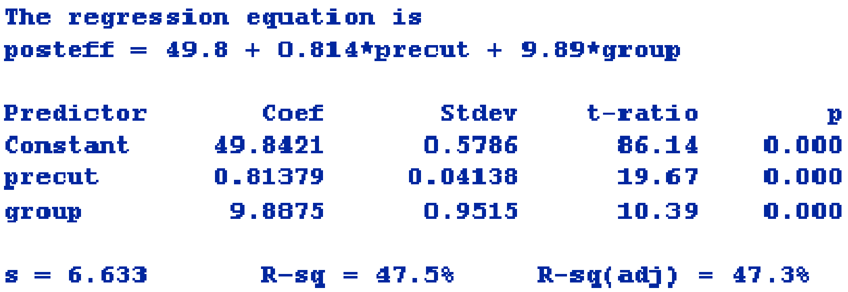


```
The regression equation is
   posteff = 49.8 + 0.824*precut + 9.89*group
   - 0.0196*linint

Predictor        Coef        Stdev    t-ratio        p
Constant      49.7508       0.6957      71.52    0.000
precut        0.82371      0.05889      13.99    0.000
group          9.8939       0.9528      10.38    0.000
linint       -0.01963      0.08284      -0.24    0.813

s = 6.639       R-sq = 47.5%     R-sq(adj) = 47.2%
```

We can see that in this model the treatment effect estimate is now 9.89 (SE = .95). Again, this estimate is very close to the true 10-point treatment effect. Notice, however, that the standard error (SE) is smaller than it was in the original model. This is the gain in efficiency we get when we eliminate the two unneeded quadratic terms. We can also see that the linear interaction term "linint" is still nonsignificant. This term would be significant if the slopes of the lines for the two groups were different. Visual inspection shows that the slopes are the same and so it makes sense that this term is not significant.

Finally, let's drop out the nonsignificant linear interaction term and respecify the model. These results are shown in Figure 13.

Figure 13. Regression results for final model.

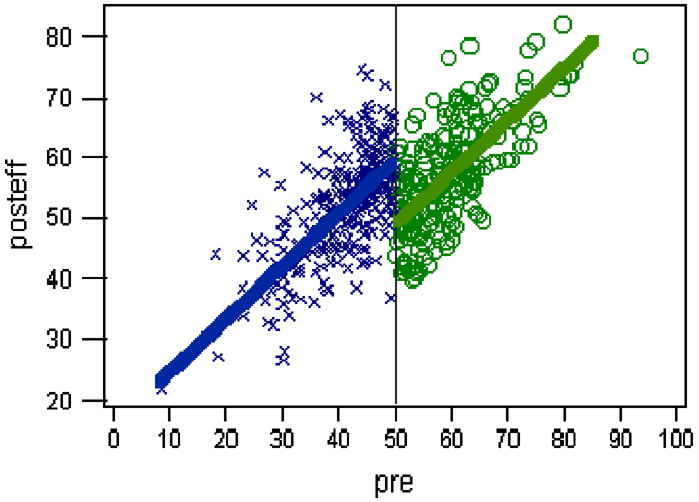


```
The regression equation is
posteff = 49.8 + 0.814*precut + 9.89*group

Predictor        Coef        Stdev     t-ratio        p
Constant      49.8421       0.5786       86.14    0.000
precut        0.81379      0.04138       19.67    0.000
group          9.8875       0.9515       10.39    0.000

s = 6.633        R-sq = 47.5%      R-sq(adj) = 47.3%
```

We see in these results that the treatment effect and SE are almost identical to the previous model and that the treatment effect estimate is an unbiased estimate of the true effect of 10 points. We can also see that all of the terms in the final model are statistically significant, suggesting that they are needed to model the data and should not be eliminated.

So, what does our model look like visually? Figure 14 shows the original bivariate distribution with the fitted regression model.

Figure 14. Bivariate distribution with final regression model.



Clearly, the model fits well, both statistically and visually.

- **Regression Point Displacement Analysis**

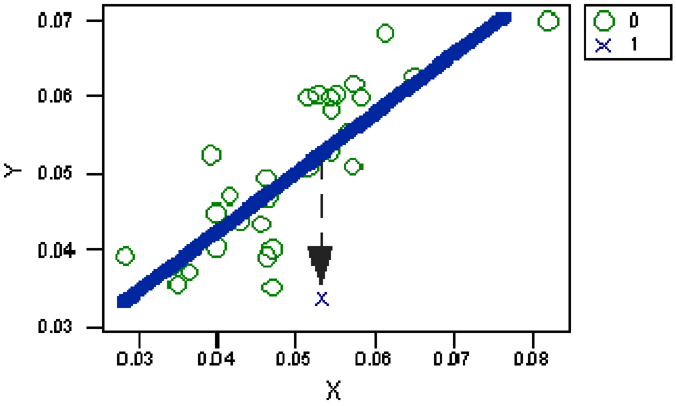
Statistical Requirements

$N_{(n=1)}$ O X O
N O O

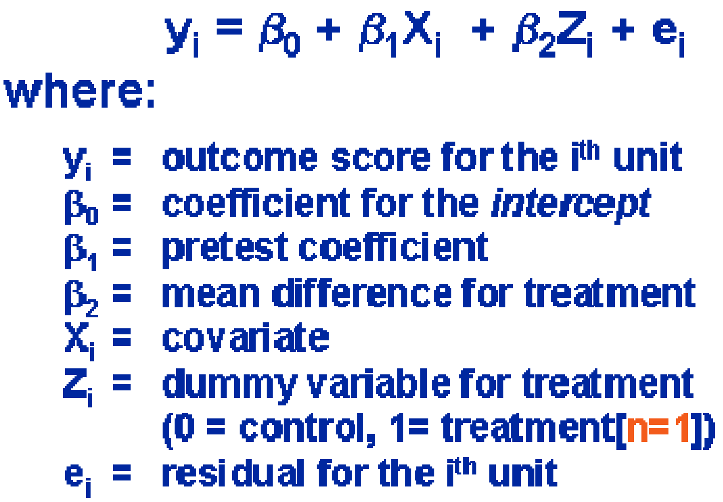
The notation for the Regression Point Displacement (RPD) design shows that the statistical analysis requires:

- a posttest score
- a pretest score
- a variable to represent the treatment group (where 0=comparison and 1=program)

These requirements are identical to the requirements for the Analysis of Covariance model. The only difference is that the RPD design only has a single treated group score.



The figure shows a bivariate (pre-post) distribution for a hypothetical RPD design of a community-based AIDS education program. The new AIDS education program is piloted in one particular county in a state, with the remaining counties acting as controls. The state routinely publishes annual HIV positive rates by county for the entire state. The x-values show the HIV-positive rates per 1000 people for the year preceding the program while the y-values show the rates for the year following it. Our goal is to estimate the size of the vertical displacement of the treated unit from the regression line of all of the control units, indicated on the graph by the dashed arrow. The model we'll use is the Analysis of Covariance (ANCOVA) model stated in regression model form:

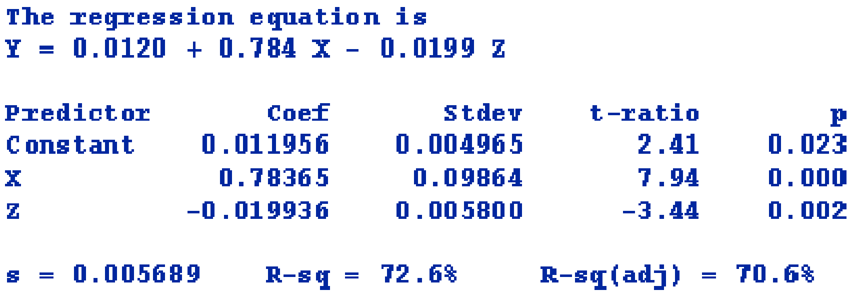


$$y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + e_i$$

where:

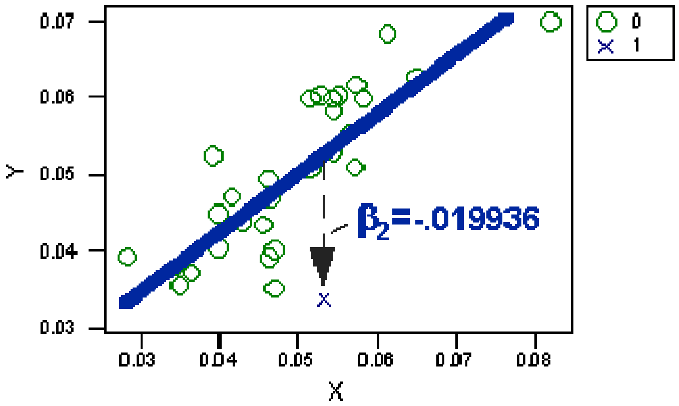
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the *intercept*
β_1 = pretest coefficient
β_2 = mean difference for treatment
X_i = covariate
Z_i = dummy variable for treatment
(0 = control, 1= treatment[n=1])
e_i = residual for the i^{th} unit

When we fit the model to our simulated data, we obtain the regression table shown below:



```
The regression equation is
Y = 0.0120 + 0.784 X - 0.0199 Z

Predictor        Coef       Stdev    t-ratio       p
Constant     0.011956    0.004965       2.41   0.023
X             0.78365     0.09864       7.94   0.000
Z           -0.019936    0.005800      -3.44   0.002

s = 0.005689     R-sq = 72.6%     R-sq(adj) = 70.6%
```

The coefficient associated with the dichotomous treatment variable is the estimate of the vertical displacement from the line. In this example, the results show that the program lowers HIV positive rates by .019 and that this amount is statistically significant. This displacement is shown in the results graph:



For more details on the statistical analysis of the RPD design, you can view an entire paper on the subject entitled " The Regression Point Displacement Design for Evaluating Community-Based Pilot Programs and Demonstration Projects."

Write-Up

So now that you've completed the research project, what do you do? I know you won't want to hear this, but your work is still far from done. In fact, this final stage -- writing up your research -- may be one of the most difficult. Developing a good, effective and concise report is an art form in itself. And, in many research projects you will need to write multiple reports that present the results at different levels of detail for different audiences.

There are several general considerations to keep in mind when generating a report:

- The Audience

  Who is going to read the report? Reports will differ considerably depending on whether the audience will want or require technical detail, whether they are looking for a summary of results, or whether they are about to examine your research in a Ph.D. exam.

- The Story

  I believe that every research project has at least one major "story" in it. Sometimes the story centers around a specific research finding. Sometimes it is based on a methodological problem or challenge. When you write your report, you should attempt to tell the "story" to your reader. Even in very formal journal articles where you will be required to be concise and detailed at the same time, a good "storyline" can help make an otherwise very dull report interesting to the reader.

  The hardest part of telling the story in your research is finding the story in the first place. Usually when you come to writing up your research you have been steeped in the details for weeks or months (and sometimes even for years). You've been worrying about sampling response, struggling with operationalizing your measures, dealing with the details of design, and wrestling with the data analysis. You're a bit like the ostrich that has its head in the sand. To find the story in your research, you have to pull your head out of the sand and look at the big picture. You have to try to view your research from your audience's perspective. You may have to let go of some of the details that you obsessed so much about and leave them out of the write up or bury them in technical appendices or tables.

- Formatting Considerations

  Are you writing a research report that you will submit for publication in a journal? If so, you should be aware that every journal requires articles that you follow specific

> formatting guidelines. Thinking of writing a book. Again, every publisher will require specific formatting. Writing a term paper? Most faculty will require that you follow specific guidelines. Doing your thesis or dissertation? Every university I know of has very strict policies about formatting and style. There are legendary stories that circulate among graduate students about the dissertation that was rejected because the page margins were a quarter inch off or the figures weren't labeled correctly.

To illustrate what a set of research report specifications might include, I present in this section general guidelines for the formatting of a research write-up for a class term paper. These guidelines are very similar to the types of specifications you might be required to follow for a journal article. However, you need to check the specific formatting guidelines for the report you are writing -- the ones presented here are likely to differ in some ways from any other guidelines that may be required in other contexts.

I've also included a sample research paper write-up that illustrates these guidelines. This sample paper is for a "make-believe" research project. But it illustrates how a final research report might look using the guidelines given here.

Key Elements

This page describes the elements or criteria that you must typically address in a research paper. The assumption here is that you are addressing a causal hypothesis in your paper.

I. Introduction

1. **Statement of the problem:** The general problem area is stated clearly and unambiguously. The importance and significance of the problem area is discussed.
2. **Statement of causal relationship:** The cause-effect relationship to be studied is stated clearly and is sensibly related to the problem area.
3. **Statement of constructs:** Each key construct in the research/evaluation project is explained (minimally, both the cause and effect). The explanations are readily understandable (i.e., jargon-free) to an intelligent reader.
4. **Literature citations and review:** The literature cited is from reputable and appropriate sources (e.g., professional journals, books and not Time, Newsweek, etc.) and you have a minimum of five references. The literature is condensed in an intelligent fashion with only the most relevant information included. Citations are in the correct format (see APA format sheets).
5. **Statement of hypothesis:** The hypothesis (or hypotheses) is clearly stated and is specific about what is predicted. The relationship of the hypothesis to both the problem statement and literature review is readily understood from reading the text.

II. Methods

Sample section:

1. **Sampling procedure specifications:** The procedure for selecting units (e.g., subjects, records) for the study is described and is appropriate. The author state which sampling method is used and why. The population and sampling frame are described. In an evaluation, the program participants are frequently self-selected (i.e., volunteers) and, if so, should be described as such.
2. **Sample description:** The sample is described accurately and is appropriate. Problems in contacting and measuring the sample are anticipated.
3. **External validity considerations:** Generalizability from the sample to the sampling frame and population is considered.

Measurement section:

1. **Measures:** Each outcome measurement construct is described briefly (a minimum of two outcome constructs is required). For each construct, the measure or measures are described briefly and an appropriate citation and reference is included (unless you created the measure). You describe briefly the measure you constructed and provide the entire measure in an Appendix. The measures which are used are relevant to the hypotheses of the study and are included in those hypotheses. Wherever possible, multiple measures of the same construct are used.
2. **Construction of measures:** For questionnaires, tests and interviews: questions are clearly worded, specific, appropriate for the population, and follow in a logical fashion. The standards for good questions are followed. For archival data: original data collection procedures are adequately described and indices (i.e., combinations of individual measures) are constructed correctly. For scales, you must describe briefly which scaling procedure you used and how you implemented it. For qualitative measures, the procedures for collecting the measures are described in detail.
3. **Reliability and validity:** You must address both the reliability and validity of all of your measures. For reliability, you must specify what estimation procedure(s) you used. For validity, you must explain how you assessed construct validity. Wherever possible, you should minimally address both convergent and discriminant validity. The procedures which are used to examine reliability and validity are appropriate for the measures.

Design and Procedures section:

1. **Design:** The design is clearly presented in both notational and text form. The design is appropriate for the problem and addresses the hypothesis.
2. **Internal validity:** Threats to internal validity and how they are addressed by the design are discussed. Any threats to internal validity which are not well controlled are also considered.
3. **Description of procedures:** An overview of how the study will be conducted is included. The sequence of events is described and is appropriate to the design. Sufficient information is included so that the essential features of the study could be replicated by a reader.

III. Results

1. **Statement of Results:** The results are stated concisely and are plausible for the research described.
2. **Tables:** The table(s) is correctly formatted and accurately and concisely presents part of the analysis.
3. **Figures:** The figure(s) is clearly designed and accurately describes a relevant aspect of the results.

IV. Conclusions, Abstract and Reference Sections

1. **Implications of the study:** Assuming the expected results are obtained, the implications of these results are discussed. The author mentions briefly any remaining problems which are anticipated in the study.
2. **Abstract:** The Abstract is 125 words or less and presents a concise picture of the proposed research. Major constructs and hypotheses are included. The Abstract is the first section of the paper. See the format sheet for more details.
3. **References:** All citations are included in the correct format and are appropriate for the study described.

Stylistic Elements

I. Professional Writing

First person and sex-stereotyped forms are avoided. Material is presented in an unbiased and unemotional (e.g., no "feelings" about things), but not necessarily uninteresting, fashion.

II. Parallel Construction

Tense is kept parallel within and between sentences (as appropriate).

III. Sentence Structure

Sentence structure and punctuation are correct. Incomplete and run-on sentences are avoided.

IV. Spelling and Word Usage

Spelling and use of words are appropriate. Words are capitalized and abbreviated correctly.

V. General Style.

The document is neatly produced and reads well. The format for the document has been correctly followed.

Formatting

Overview

The instructions provided here are for a research article or a research report (generally these guidelines follow the formatting guidelines of the American Psychological Association documented in Publication Manual of the American Psychological Association, 4th Edition). Please consult the specific guidelines that are required by the publisher for the type of document you are producing.

All sections of the paper should be typed, double-spaced on white 8 1/2 x 11 inch paper with 12 pitch typeface with all margins set to 1 inch. REMEMBER TO CONSULT THE APA PUBLICATION MANUAL, FOURTH EDITION, PAGES 258 - 264 TO SEE HOW TEXT SHOULD APPEAR. Every page must have a header in the upper right corner with the running header right-justified on the top line and the page number right-justified and double-spaced on the line below it. The paper must have all the sections in the order given below, following the specifications outlined for each section (all pages numbers are approximate):

- Title Page
- Abstract (on a separate single page)
- The Body (no page breaks between sections in the body)
    - Introduction (2-3 pages)
    - Methods (7-10 pages)
        - Sample (1 page)
        - Measures (2-3 pages)
        - Design (2-3 pages)
        - Procedures (2-3 pages)
    - Results (2-3 pages)
    - Conclusions (1-2 pages)
- References
- Tables (one to a page)
- Figures (one to a page)
- Appendices

Title Page

On separate lines and centered, the title page has the title of the study, the author's name, and the institutional affiliation. At the bottom of the title page you should have the words (in caps) RUNNING HEADER: followed by a short identifying title (2-4 words) for the study. This running header should also appear on the top right of every page of the paper.

Abstract

The abstract is limited to one page, double-spaced. At the top of the page, centered, you should have the word '**Abstract**'. The abstract itself should be written in paragraph form and should be a concise summary of the entire paper including: the problem; major hypotheses; sample and population; a brief description of the measures; the name of the design or a short description (no design notation here); the major results; and, the major conclusions. Obviously, to fit this all on one page you will have to be very concise.

Body

The first page of the body of the paper should have, centered, the complete title of the study.

Introduction

The first section in the body is the introduction. There is no heading that says 'Introduction,' you simply begin the paper in paragraph form following the title. Every introduction will have the following (roughly in this order): a statement of the problem being addressed; a statement of the cause-effect relationship being studied; a description of the major constructs involved; a brief review of relevant literature (including citations); and a statement of hypotheses. The entire section should be in paragraph form with the possible exception of the hypotheses, which may be indented.

Methods

The next section of the paper has four subsections: Sample; Measures; Design; and, Procedure. The Methods section should begin immediately after the introduction (no page break) and should have the centered title 'Methods'. Each of the four subsections should have an underlined left justified section heading.

Sampling

This section should describe the population of interest, the sampling frame, the method for selecting the sample, and the sample itself. A brief discussion of external validity is appropriate here, that is, you should state the degree to which you believe results will be generalizable from your sample to the population. (Link to Knowledge Base on sampling).

Measures

This section should include a brief description of your constructs and all measures that will be used to operationalize them. You may present short instruments in their entirety in this section. If you have more lengthy instruments you may present some "typical" questions to give the reader a sense of what you will be doing (and include the full measure in an Appendix). You may include any instruments in full in appendices rather than in the body. Appendices should be labeled by letter. (e.g., 'Appendix A') and cited appropriately in the body of the text. For pre-existing instruments you should cite any relevant information about reliability and validity if it is available. For all instruments, you should briefly state how you will determine reliability and validity, report the results and discuss. For reliability, you must describe the methods you used and report results. A brief discussion of how you have addressed construct validity is essential. In general, you should try to demonstrate both convergent and discriminant validity. You must discuss the evidence in support of the validity of your measures. (Link to Knowledge Base on measurement).

Design

You should state the name of the design that is used and tell whether it is a true or quasi-experiment, nonequivalent group design, and so on. You should also present the design structure in X and O notation (this should be indented and centered, not put into a sentence). You should also include a discussion of internal validity that describes the major likely threats in your study and how the design accounts for them, if at all. (Be your own study critic here and provide enough information to show that you understand the threats to validity, whether you've been able to account for them all in the design or not.) (Link to Knowledge Base on design).

Procedures

Generally, this section ties together the sampling, measurement, and research design. In this section you should briefly describe the overall plan of the research, the sequence of events from beginning to end (including sampling, measurement, and use of groups in designs), how participants will be notified, and how their confidentiality will be protected (where relevant). An essential part of this subsection is a description of the program or independent variable that you are studying. (Link to Knowledge Base discussion of validity).

Results

The heading for this section is centered with upper and lower case letters. You should indicate concisely what results you found in this research. Your results don't have to confirm your hypotheses. In fact, the common experience in social research is the finding of no effect.

Conclusions

Here you should describe the conclusions you reach (assuming you got the results described in the Results section above). You should relate these conclusions back to the level of the construct and the general problem area which you described in the Introduction section. You should also discuss the overall strength of the research proposed (e.g. general discussion of the strong and weak validity areas) and should present some suggestions for possible future research which would be sensible based on the results of this work.

References

There are really two parts to a reference citation. First, there is the way you cite the item in the text when you are discussing it. Second, there is the way you list the complete reference in the reference section in the back of the report.

Reference Citations in the Text of Your Paper

Cited references appear in the text of your paper and are a way of giving credit to the source of the information or quote you have used in your paper. They generally consist of the following bits of information:

The author's last name, unless first initials are needed to distinguish between two authors with the same last name. If there are six or more authors, the first author is listed followed by the term, et al., and then the year of the publication is given in parenthesis. Year of publication in parenthesis. Page numbers are given with a quotation or when only a specific part of a source was used.

"To be or not to be" (Shakespeare, 1660, p. 241)

One Work by One Author:

Rogers (1994) compared reaction times...

One Work by Multiple Authors:

Wasserstein, Zappulla, Rosen, Gerstman, and Rock (1994) [first time you cite in text]

Wasserstein et al. (1994) found [subsequent times you cite in text]

Reference List in Reference Section

There are a wide variety of reference citation formats. Before submitting any research report you should check to see which type of format is considered acceptable for that context. If there is no official format requirement then the most sensible thing is for you to select one approach and implement it consistently (there's nothing worse than a reference list with a variety of formats). Here, I'll illustrate by example some of the major reference items and how they might be cited in the reference section.

The References lists all the articles, books, and other sources used in the research and preparation of the paper and cited with a parenthetical (textual) citation in the text. These items are entered in alphabetical order according to the authors' last names; if a source does not have an author, alphabetize according to the first word of the title, disregarding the articles "a", "an", and "the" if they are the first word in the title.

EXAMPLES BOOK BY ONE AUTHOR:

Jones, T. (1940). My life on the road. New York: Doubleday.

BOOK BY TWO AUTHORS:

Williams, A., & Wilson, J. (1962). New ways with chicken. New York: Harcourt.

BOOK BY THREE OR MORE AUTHORS:

Smith, J., Jones, J., & Williams, S. (1976). Common names. Chicago: University of Chicago Press.

BOOK WITH NO GIVEN AUTHOR OR EDITOR:

Handbook of Korea (4th ed.). (1982). Seoul: Korean Overseas Information, Ministry of Culture & Information.

TWO OR MORE BOOKS BY THE SAME AUTHOR:

Oates, J.C. (1990). Because it is bitter, and because it is my heart. New York: Dutton.

Oates, J.C. (1993). Foxfire: Confessions of a girl gang. New York: Dutton.

Note: Entries by the same author are arranged chronologically by the year of publication, the earliest first. References with the same first author and different second and subsequent authors are listed alphabetically by the surname of the second author, then by the surname of the third author. References with the same authors in the same order are entered chronologically by year of publication, the earliest first. References by the same author (or by the same two or more authors in identical order) with the same publication date are listed alphabetically by the first word of the title following the date; lower case letters (a, b, c, etc.) are included after the year, within the parentheses.

BOOK BY A CORPORATE (GROUP) AUTHOR:

President's Commission on Higher Education. (1977). Higher education for American democracy . Washington, D.C.: U.S. Government Printing Office.

BOOK WITH AN EDITOR:

Bloom, H. (Ed.). (1988). James Joyce's Dubliners. New York: Chelsea House.

A TRANSLATION:

Dostoevsky, F. (1964). Crime and punishment (J. Coulson Trans.). New York: Norton. (Original work published 1866)

AN ARTICLE OR READING IN A COLLECTION OF PIECES BY SEVERAL AUTHORS (ANTHOLOGY):

O'Connor, M.F. (1975). Everything that rises must converge. In J.R. Knott, Jr. & C.R. Raeske (Eds.), Mirrors: An introduction to literature (2nd ed., pp. 58-67). San Francisco: Canfield.

EDITION OF A BOOK:

Tortora, G.J., Funke, B.R., & Case, C.L. (1989). Microbiology: An introduction (3rd ed.). Redwood City, CA: Benjamin/Cummings.

DIAGNOSTIC AND STATISTICAL MANUAL OF MENTAL DISORDERS:

American Psychiatric Association. (1994). Diagnostic and statistical manual of mental disorders (4th ed.). Washington, D.C.: Author.

A WORK IN SEVERAL VOLUMES:

Churchill, W.S. (1957). A history of the English speaking peoples: Vol. 3. The Age of Revolution. New York: Dodd, Mead.

ENCYCLOPEDIA OR DICTIONARY:

Cockrell, D. (1980). Beatles. In The new Grove dictionary of music and musicians (6th ed., Vol. 2, pp. 321-322). London: Macmillan.

ARTICLE FROM A WEEKLY MAGAZINE:

Jones, W. (1970, August 14). Todays's kids. Newseek, 76, 10-15.

ARTICLE FROM A MONTHLY MAGAZINE:

Howe, I. (1968, September). James Baldwin: At ease in apocalypse. Harper's, 237, 92-100.

ARTICLE FROM A NEWSPAPER:

Brody, J.E. (1976, October 10). Multiple cancers termed on increase. New York Times (national ed.). p. A37.

ARTICLE FROM A SCHOLARLY ACADEMIC OR PROFESSIONAL JOURNAL:

Barber, B.K. (1994). Cultural, family, and personal contexts of parent-adolescent conflict. Journal of Marriage and the Family, 56, 375-386.

GOVERNMENT PUBLICATION:

U.S. Department of Labor. Bureau of Labor Statistics. (1980). Productivity. Washington, D.C.: U.S. Government Printing Office.

PAMPHLET OR BROCHURE:

Research and Training Center on Independent Living. (1993). Guidelines for reporting and writing about people with disabilities. (4th ed.) [Brochure]. Lawrence, KS: Author.

Tables

Any Tables should have a heading with 'Table #' (where # is the table number), followed by the title for the heading that describes concisely what is contained in the table. Tables and Figures are typed on separate sheets at the end of the paper after the References and before the Appendices. In the text you should put a reference where each Table or Figure should be inserted using this form:

Insert Table 1 about here

Figures

Figures are drawn on separate sheets at the end of the paper after the References and and Tables, and before the Appendices. In the text you should put a reference where each Figure will be inserted using this form:

Insert Figure 1 about here

Appendices

Appendices should be used only when absolutely necessary. Generally, you will only use them for presentation of extensive measurement instruments, for detailed descriptions of the program or independent variable and for any relevant supporting documents which you don't include in the body. Even if you include such appendices, you should briefly describe the relevant material in the body and give an accurate citation to the appropriate appendix (e.g., 'see Appendix A').

Sample Paper

This paper should be used only as an example of a research paper write-up. Horizontal rules signify the top and bottom edges of pages. For sample references which are not included with this paper, you should consult the **Publication Manual of the American Psychological Association, 4th Edition.**

This paper is provided only to give you an idea of what a research paper might look like. You are not allowed to copy any of the text of this paper in writing your own report.

Because word processor copies of papers don't translate well into web pages, you should note that an actual paper should be formatted according to the formatting rules for your context. Note especially that there are three formatting rules you will see in this sample paper which you should NOT follow. First, except for the title page, the running header should appear in the upper right corner of every page with the page number below it. Second, paragraphs and text should be double spaced and the start of each paragraph should be indented. Third, horizontal lines are used to indicate a mandatory page break and should not be used in your paper.

The Effects of a Supported Employment Program on Psychosocial Indicators for Persons with Severe Mental Illness

William M.K. Trochim

Cornell University

Running Head: SUPPORTED EMPLOYMENT

Abstract

This paper describes the psychosocial effects of a program of supported employment (SE) for persons with severe mental illness. The SE program involves extended individualized supported employment for clients through a Mobile Job Support Worker (MJSW) who maintains contact with the client after job placement and supports the client in a variety of ways. A 50% simple random sample was taken of all persons who entered the Thresholds Agency between 3/1/93 and 2/28/95 and who met study criteria. The resulting 484 cases were randomly assigned to either the

SE condition (treatment group) or the usual protocol (control group) which consisted of life skills training and employment in an in-house sheltered workshop setting. All participants were measured at intake and at 3 months after beginning employment, on two measures of psychological functioning (the BPRS and GAS) and two measures of self esteem (RSE and ESE). Significant treatment effects were found on all four measures, but they were in the opposite direction from what was hypothesized. Instead of functioning better and having more self esteem, persons in SE had lower functioning levels and lower self esteem. The most likely explanation is that people who work in low-paying service jobs in real world settings generally do not like them and experience significant job stress, whether they have severe mental illness or not. The implications for theory in psychosocial rehabilitation are considered.

The Effects of a Supported Employment Program on Psychosocial Indicators for Persons with Severe Mental Illness

Over the past quarter century a shift has occurred from traditional institution-based models of care for persons with severe mental illness (SMI) to more individualized community-based treatments. Along with this, there has been a significant shift in thought about the potential for persons with SMI to be "rehabilitated" toward lifestyles that more closely approximate those of persons without such illness. A central issue is the ability of a person to hold a regular full-time job for a sustained period of time. There have been several attempts to develop novel and radical models for program interventions designed to assist persons with SMI to sustain full-time employment while living in the community. The most promising of these have emerged from the tradition of psychiatric rehabilitation with its emphases on individual consumer goal setting, skills training, job preparation and employment support (Cook, Jonikas and Solomon, 1992). These are relatively new and field evaluations are rare or have only recently been initiated (Cook and Razzano, 1992; Cook, 1992). Most of the early attempts to evaluate such programs have naturally focused almost exclusively on employment outcomes. However, theory suggests that sustained employment and living in the community may have important therapeutic benefits in addition to the obvious economic ones. To date, there have been no formal studies of the effects of psychiatric rehabilitation programs on key illness-related outcomes. To address this issue, this study seeks to examine the effects of a new program of supported employment on psychosocial outcomes for persons with SMI.

Over the past several decades, the theory of vocational rehabilitation has experienced two major stages of evolution. Original models of vocational rehabilitation were based on the idea of sheltered workshop employment. Clients were paid a piece rate and worked only with other individuals who were disabled. Sheltered workshops tended to be "end points" for persons with severe and profound mental retardation since few ever moved from sheltered to competitive employment (Woest, Klein & Atkins, 1986). Controlled studies of sheltered workshop performance of persons with mental illness suggested only minimal success (Griffiths, 1974) and other research indicated that persons with mental illness earned lower wages, presented more behavior problems, and showed poorer workshop attendance than workers with other disabilities (Whitehead, 1977; Ciardiello, 1981).

In the 1980s, a new model of services called Supported Employment (SE) was proposed as less expensive and more normalizing for persons undergoing rehabilitation (Wehman, 1985). The SE model emphasizes first locating a job in an integrated setting for minimum wage or above, and then placing the person on the job and providing the training and support services needed to remain employed (Wehman, 1985). Services such as individualized job development, one-on-one job coaching, advocacy with co-workers and employers, and "fading" support were found to be effective in maintaining employment for individuals with severe and profound mental retardation (Revell, Wehman & Arnold, 1984). The idea that this model could be generalized to persons with all types of severe disabilities, including severe mental illness, became commonly accepted (Chadsey-Rusch & Rusch, 1986).

One of the more notable SE programs was developed at Thresholds, the site for the present study, which created a new staff position called the mobile job support worker (MJSW) and removed the common six month time limit for many placements. MJSWs provide ongoing, mobile support and intervention at or near the work site, even for jobs with high degrees of independence (Cook & Hoffschmidt, 1993). Time limits for many placements were removed so that clients could stay on as permanent employees if they and their employers wished. The suspension of time limits on job placements, along with MJSW support, became the basis of SE services delivered at Thresholds.

There are two key psychosocial outcome constructs of interest in this study. The first is the overall *psychological functioning* of the person with SMI. This would include the specification of severity of cognitive and affective symptomotology as well as the overall level of psychological functioning. The second is the level of self-reported *self esteem* of the person. This was measured both generally and with specific reference to employment.

The key hypothesis of this study is:

H_O: A program of supported employment will result in either *no change or negative effects* on psychological functioning and self esteem.

which will be tested against the alternative:

H_A: A program of supported employment will lead to *positive effects* on psychological functioning and self esteem.

Method

Sample

The population of interest for this study is all adults with SMI residing in the U.S. in the early 1990s. The population that is accessible to this study consists of all persons who were clients of the Thresholds Agency in Chicago, Illinois between the dates of March 1, 1993 and February 28, 1995 who met the following criteria: 1) a history of severe mental illness (e.g., either schizophrenia, severe depression or manic-depression); 2) a willingness to achieve paid employment; 3) their primary diagnosis must not include chronic alcoholism or hard drug use;

and 4) they must be 18 years of age or older. The sampling frame was obtained from records of the agency. Because of the large number of clients who pass through the agency each year (e.g., approximately 500 who meet the criteria) a simple random sample of 50% was chosen for inclusion in the study. This resulted in a sample size of 484 persons over the two-year course of the study.

On average, study participants were 30 years old and high school graduates (average education level = 13 years). The majority of participants (70%) were male. Most had never married (85%), few (2%) were currently married, and the remainder had been formerly married (13%). Just over half (51%) are African American, with the remainder Caucasian (43%) or other minority groups (6%). In terms of illness history, the members in the sample averaged 4 prior psychiatric hospitalizations and spent a lifetime average of 9 months as patients in psychiatric hospitals. The primary diagnoses were schizophrenia (42%) and severe chronic depression (37%). Participants had spent an average of almost two and one-half years (29 months) at the longest job they ever held.

While the study sample cannot be considered representative of the original population of interest, generalizability was not a primary goal -- the major purpose of this study was to determine whether a specific SE program *could* work in an accessible context. Any effects of SE evident in this study can be generalized to urban psychiatric agencies that are similar to Thresholds, have a similar clientele, and implement a similar program.

Measures

All but one of the measures used in this study are well-known instruments in the research literature on psychosocial functioning. All of the instruments were administered as part of a structured interview that an evaluation social worker had with study participants at regular intervals.

Two measures of psychological functioning were used. The Brief Psychiatric Rating Scale (BPRS)(Overall and Gorham, 1962) is an 18-item scale that measures perceived severity of symptoms ranging from "somatic concern" and "anxiety" to "depressive mood" and "disorientation." Ratings are given on a 0-to-6 Likert-type response scale where 0="not present" and 6="extremely severe" and the scale score is simply the sum of the 18 items. The Global Assessment Scale (GAS)(Endicott et al, 1976) is a single 1-to-100 rating on a scale where each ten-point increment has a detailed description of functioning (higher scores indicate better functioning). For instance, one would give a rating between 91-100 if the person showed "no symptoms, superior functioning..." and a value between 1-10 if the person "needs constant supervision..."

Two measures of self esteem were used. The first is the Rosenberg Self Esteem (RSE) Scale (Rosenberg, 1965), a 10-item scale rated on a 6-point response format where 1="strongly disagree" and 6="strongly agree" and there is no neutral point. The total score is simply the sum across the ten items, with five of the items being reversals. The second measure was developed explicitly for this study and was designed to measure the Employment Self Esteem (ESE) of a person with SMI. This is a 10-item scale that uses a 4-point response format where 1="strongly disagree" and 4="strongly agree" and there is no neutral point. The final ten items were selected from a pool of 97 original candidate items, based upon high item-total score correlations and a judgment of face validity by a panel of three psychologists. This instrument was deliberately kept simple -- a shorter response scale and no reversal items -- because of the difficulties associated with measuring a population with SMI. The entire instrument is provided in Appendix A.

All four of the measures evidenced strong reliability and validity. Internal consistency reliability estimates using Cronbach's alpha ranged from .76 for ESE to .88 for SE. Test-retest reliabilities were nearly as high, ranging from .72 for ESE to .83 for the BPRS. Convergent validity was evidenced by the correlations within construct. For the two psychological functioning scales the correlation was .68 while for the self esteem measures it was somewhat lower at .57. Discriminant validity was examined by looking at the cross-construct correlations which ranged from .18 (BPRS-ESE) to .41 (GAS-SE).

Design

A pretest-posttest two-group randomized experimental design was used in this study. In notational form, the design can be depicted as:

R O X O

R O O

where:

R = the groups were randomly assigned

O = the four measures (i.e., BPRS, GAS, RSE, and ESE)

X = supported employment

The comparison group received the standard Thresholds protocol which emphasized in-house training in life skills and employment in an in-house sheltered workshop. All participants were measured at intake (pretest) and at three months after intake (posttest).

This type of randomized experimental design is generally strong in internal validity. It rules out threats of history, maturation, testing, instrumentation, mortality and selection interactions. Its primary weaknesses are in the potential for treatment-related mortality (i.e., a type of selection-mortality) and for problems that result from the reactions of participants and administrators to knowledge of the varying experimental conditions. In this study, the drop-out rate was 4% (N=9) for the control group and 5% (N=13) in the treatment group. Because these rates are low and are approximately equal in each group, it is not plausible that there is differential mortality. There is a possibility that there were some deleterious effects due to participant knowledge of the other group's existence (e.g., compensatory rivalry, resentful demoralization). Staff were debriefed at several points throughout the study and were explicitly asked about such issues. There were no reports of any apparent negative feelings from the participants in this regard. Nor is it plausible that staff might have equalized conditions between the two groups. Staff were given extensive training and were monitored throughout the course of the study. Overall, this study can be considered strong with respect to internal validity.

Procedure

Between 3/1/93 and 2/28/95 each person admitted to Thresholds who met the study inclusion criteria was immediately assigned a random number that gave them a 50/50 chance of being selected into the study sample. For those selected, the purpose of the study was explained, including the nature of the two treatments, and the need for and use of random assignment. Participants were assured confidentiality and were given an opportunity to decline to participate in the study. Only 7 people (out of 491) refused to participate. At intake, each selected sample member was assigned a random number giving them a 50/50 chance of being assigned to either the Supported Employment condition or the standard in-agency sheltered workshop. In addition, all study participants were given the four measures at intake.

All participants spent the initial two weeks in the program in training and orientation. This consisted of life skill training (e.g., handling money, getting around, cooking and nutrition) and job preparation (employee roles, coping strategies). At the end of that period, each participant was assigned to a job site -- at the agency sheltered workshop for those in the control condition, and to an outside employer if in the Supported Employment group. Control participants were expected to work full-time at the sheltered workshop for a three-month period, at which point they were posttested and given an opportunity to obtain outside employment (either Supported Employment or not). The Supported Employment participants were each assigned a case worker -- called a Mobile Job Support Worker (MJSW) -- who met with the person at the job site two times per week for an hour each time. The MJSW could provide any support or assistance deemed necessary to help the person cope with job stress, including counseling or working beside the person for short periods of time. In addition, the MJSW was always accessible by cellular telephone, and could be called by the participant or the employer at any time. At the end of three months, each participant was post-tested and given the option of staying with their current job (with or without Supported Employment) or moving to the sheltered workshop.

Results

There were 484 participants in the final sample for this study, 242 in each treatment. There were 9 drop-outs from the control group and 13 from the treatment group, leaving a total of 233 and 229 in each group respectively from whom both pretest and posttest were obtained. Due to unexpected difficulties in coping with job stress, 19 Supported Employment participants had to be transferred into the sheltered workshop prior to the posttest. In all 19 cases, no one was transferred prior to week 6 of employment, and 15 were transferred after week 8. In all analyses, these cases were included with the Supported Employment group (intent-to-treat analysis) yielding treatment effect estimates that are likely to be conservative.

The major results for the four outcome measures are shown in Figure 1.

Insert Figure 1 about here

It is immediately apparent that in all four cases the null hypothesis has to be accepted -- contrary to expectations, Supported Employment cases did significantly *worse* on all four outcomes than did control participants.

The mean gains, standard deviations, sample sizes and t-values (t-test for differences in average gain) are shown for the four outcome measures in Table 1.

Insert Table 1 about here

The results in the table confirm the impressions in the figures. Note that all t-values are negative except for the BPRS where high scores indicate greater severity of illness. For all four outcomes, the t-values were statistically significant (p<.05).

Conclusions

The results of this study were clearly contrary to initial expectations. The alternative hypothesis suggested that SE participants would show improved psychological functioning and self esteem after three months of employment. Exactly the reverse happened -- SE participants showed significantly worse psychological functioning and self esteem.

There are two major possible explanations for this outcome pattern. First, it seems reasonable that there might be a delayed positive or "boomerang" effect of employment outside of a sheltered setting. SE cases may have to go through an initial difficult period of adjustment (longer than three months) before positive effects become apparent. This "you have to get worse before you get better" theory is commonly held in other treatment-contexts like drug addiction and alcoholism. But a second explanation seems more plausible -- that people working full-time jobs in real-world settings are almost certainly going to be under greater stress and experience more negative outcomes than those who work in the relatively safe confines of an in-agency sheltered workshop. Put more succinctly, the lesson here might very well be that work is hard. Sheltered workshops are generally very nurturing work environments where virtually all employees share similar illness histories and where expectations about productivity are relatively low. In contrast, getting a job at a local hamburger shop or as a shipping clerk puts the person in contact with co-workers who may not be sympathetic to their histories or forgiving with respect to low productivity. This second explanation seems even more plausible in the wake of informal debriefing sessions held as focus groups with the staff and selected research participants. It was clear in the discussion that SE persons experienced significantly higher job stress levels and more negative consequences. However, most of them also felt that the experience was a good one overall and that even their "normal" co-workers "hated their jobs" most of the time.

One lesson we might take from this study is that much of our contemporary theory in psychiatric rehabilitation is naive at best and, in some cases, may be seriously misleading. Theory led us to believe that outside work was a "good" thing that would naturally lead to "good" outcomes like increased psychological functioning and self esteem. But for most people (SMI or not) work is at best tolerable, especially for the types of low-paying service jobs available to study participants. While people with SMI may not function as well or have high self esteem, we should balance this with the desire they may have to "be like other people" including struggling with the vagaries of life and work that others struggle with.

Future research in this are needs to address the theoretical assumptions about employment outcomes for persons with SMI. It is especially important that attempts to replicate this study also try to measure how SE participants feel about the decision to work, even if traditional outcome indicators suffer. It may very well be that negative outcomes on traditional indicators can be associated with a "positive" impact for the participants and for the society as a whole.

References

Chadsey-Rusch, J. and Rusch, F.R. (1986). The ecology of the workplace. In J. Chadsey-Rusch, C. Haney-Maxwell, L. A. Phelps and F. R. Rusch (Eds.), School-to-Work Transition Issues and Models. (pp. 59-94), Champaign IL: Transition Institute at Illinois.

Ciardiello, J.A. (1981). Job placement success of schizophrenic clients in sheltered workshop programs. Vocational Evaluation and Work Adjustment Bulletin, 14, 125-128, 140.

Cook, J.A. (1992). Job ending among youth and adults with severe mental illness. Journal of Mental Health Administration, 19(2), 158-169.

Cook, J.A. & Hoffschmidt, S. (1993). Psychosocial rehabilitation programming: A comprehensive model for the 1990's. In R.W. Flexer and P. Solomon (Eds.), Social and Community Support for People with Severe Mental Disabilities: Service Integration in Rehabilitation and Mental Health. Andover, MA: Andover Publishing.

Cook, J.A., Jonikas, J., & Solomon, M. (1992). Models of vocational rehabilitation for youth and adults with severe mental illness. American Rehabilitation, 18, 3, 6-32.

Cook, J.A. & Razzano, L. (1992). Natural vocational supports for persons with severe mental illness: Thresholds Supported Competitive Employment Program, in L. Stein (ed.), New Directions for Mental Health Services, San Francisco: Jossey-Bass, 56, 23-41.

Endicott, J.R., Spitzer, J.L. Fleiss, J.L. and Cohen, J. (1976). The Global Assessment Scale: A procedure for measuring overall severity of psychiatric disturbance. Archives of General Psychiatry, 33, 766-771.

Griffiths, R.D. (1974). Rehabilitation of chronic psychotic patients. Psychological Medicine, 4, 316-325.

Overall, J. E. and Gorham, D. R. (1962). The Brief Psychiatric Rating Scale. Psychological Reports, 10, 799-812.

Rosenberg, M. (1965). Society and Adolescent Self Image. Princeton, NJ, Princeton University Press.

Wehman, P. (1985). Supported competitive employment for persons with severe disabilities. In P. McCarthy, J. Everson, S. Monn & M. Barcus (Eds.), School-to-Work Transition for Youth with Severe Disabilities, (pp. 167-182), Richmond VA: Virginia Commonwealth University.

Whitehead, C.W. (1977). Sheltered Workshop Study: A Nationwide Report on Sheltered Workshops and their Employment of Handicapped Individuals. (Workshop Survey, Volume 1), U.S. Department of Labor Service Publication. Washington, DC: U.S. Government Printing Office.

Woest, J., Klein, M. and Atkins, B.J. (1986). An overview of supported employment strategies. Journal of Rehabilitation Administration, 10(4), 130-135.

Table 1. Means, standard deviations and Ns for the pretest, posttest and gain scores for the four outcome variables and t-test for difference between average gains.

BPRS		**Pretest**	**Posttest**	**Gain**
Treatment	Mean	3.2	5.1	1.9
	sd	2.4	2.7	2.55
	N	229	229	229
Control	Mean	3.4	3.0	-0.4
	sd	2.3	2.5	2.4
	N	233	233	233

t =	**9.979625**	**p<.05**		

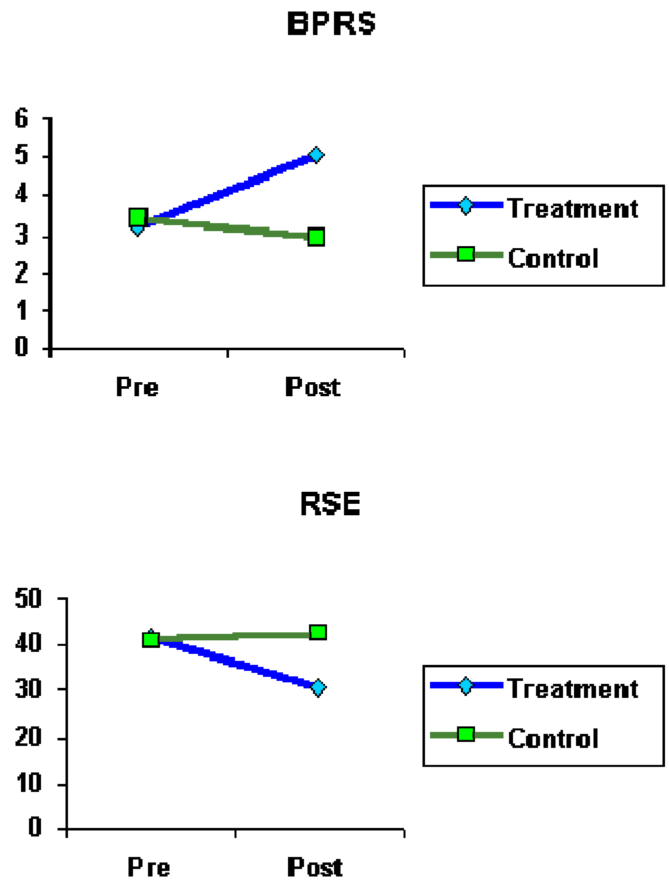
GAS		**Pretest**	**Posttest**	**Gain**
Treatment	Mean	59	43	-16
	sd	25.2	24.3	24.75
	N	229	229	229
Control	Mean	61	63	2
	sd	26.7	22.1	24.4
	N	233	233	233
t =	**-7.87075**	**p<.05**		

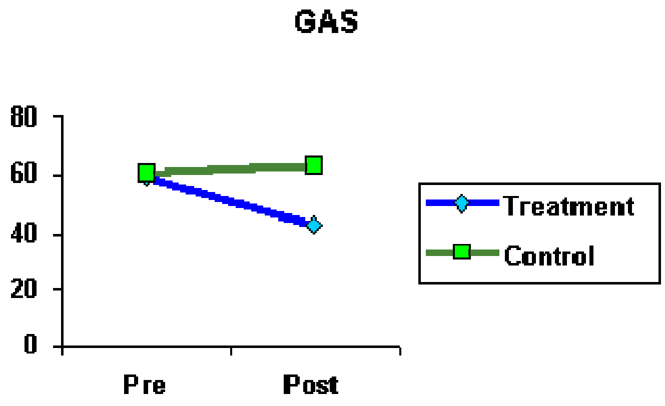
RSE		**Pretest**	**Posttest**	**Gain**
Treatment	Mean	42	31	-11
	sd	27.1	26.5	26.8
	N	229	229	229
Control	Mean	41	43	2
	sd	28.2	25.9	27.05
	N	233	233	233
t =	**-5.1889**	**p<.05**		

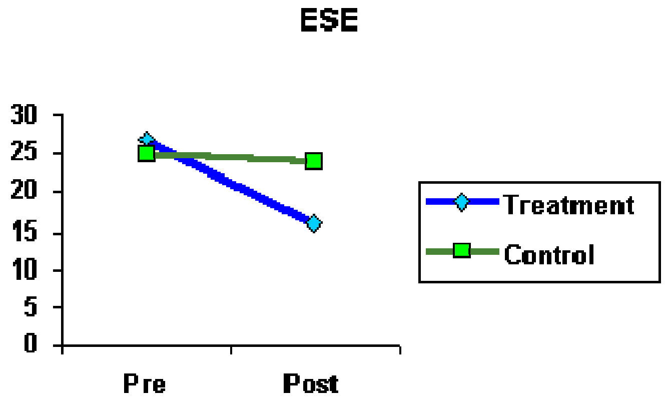
ESE		**Pretest**	**Posttest**	**Gain**
Treatment	Mean	27	16	-11
	sd	19.3	21.2	20.25
	N	229	229	229

Control	Mean	25	24	-1
	sd	18.6	20.3	19.45
	N	233	233	233
t =	**-5.41191**	**p<.05**		

Figure 1. Pretest and posttest means for treatment (SE) and control groups for the four outcome measures.







Appendix A

The Employment Self Esteem Scale

Please rate how strongly you agree or disagree with each of the following statements.

Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	1. I feel good about my work on the job.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	2. On the whole, I get along well with others at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	3. I am proud of my ability to cope with difficulties at work.
				4. When I feel uncomfortable at

Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	work, I know how to handle it.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	5. I can tell that other people at work are glad to have me there.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	6. I know I'll be able to cope with work for as long as I want.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	7. I am proud of my relationship with my supervisor at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	8. I am confident that I can handle my job without constant assistance.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	9. I feel like I make a useful contribution at work.
Strongly Disagree	Somewhat Disagree	Somewhat Agree	Strongly Agree	10. I can tell that my co-workers respect me.