Three main goals of research design:

- Allow you to have confidence that the conclusions you reach are justified
- Let you extend your conclusions beyond the actual people involved in your study -- to other groups, places, and settings
- Permit you to provide the most thorough possible explanation or understanding of the phenomena that you explore in your study

There is **no** research design that will permit you to accomplish all three goals completely. It is not OK to simply ignore one of these goals of design. If you do so, your study will be seriously flawed. You need to think about these three goals and build a design that will ensure that you do everything possible, given your constraints, to achieve all three goals. Often, the best way to do this is through a multi-stage design -- to use two or even three different designs in one study. The different design groups and individual designs within them have different strengths and weaknesses. Combining them allows you to capitalize on their differences.

**Internal Validity = Confidence in Your Conclusions**

For some, internal validity refers to the degree to which you are able to establish a *direct* causal relationship between two or more phenomena. This is the positivist sense of the term. However, many research questions do not address *direct* cause and effect and some do not concern causality in any sense. From a realist epistemological perspective, *direct* cause and effect and even general "causality" are often not the only, or even the primary, goal of scientific research. For us, internal validity is much more than simply demonstrating cause and effect. Rather this concept refers to the **degree of confidence that you can have that the conclusions you draw are justified or valid.**

Sometimes, whatever our epistemological stance, we do want to show **direct cause and effect relationships**. This means that you can show that a change in one variable directly causes a change in another variable. For example, we know that there are direct causal relationships between a pregnant woman's use of drugs and all sorts of birth defects in babies. The mother's drug use causes physiological damage to the foetus -- directly.

More commonly for many researchers, we are interested in the more general idea of **causality.** For example, we know that there is a positive relationship between parents reading to young children and early cognitive skill development. The relationship is **directional** -- that is, the parents reading to the child comes **before** the cognitive skill development, a potential indicator that the relationship is **causal,** that reading to a child does stimulate cognitive skill development. However, it is not a direct cause and effect relationship. Reading to the child is one of several factors that influence cognitive development. Further, it is not a one-to-one relationship -- that every hour spent reading to the child results in a 0.05 increase in cognitive skills. Nonetheless, most experts conclude that there is a causal relationship of some sort between reading to your child and the child's cognitive skill development.

In many cases, we are not concerned much about directionality and causality. Rather, we want to understand how different phenomena interact. For example, we can show that there is a positive

relationship between socio-economic status and educational achievement. But which comes first? Is it that people with high socio-economic status tend to be able to take the time and money to achieve educationally? Or is it that people who have high educational achievement end up acquiring high socio-economic status? Or does some third factor cause both higher socio-economic status and educational achievement? We do not know. In fact, we cannot tell whether one causes the other at all. All we can say is that these two phenomena generally **co-vary.** Very consistent patterns of co-variance over time are strong indicators that causality is involved. One study at one point in time does little to establish confidence that causality is occurring. Whichever of these we are talking about -- direct cause and effect, some sort of indirect or mediated effect, or simply co-variation with no way to tell if there is a causal relationship at all - - we want to be confident that our conclusion that a relationship does exist is correct. In the broad sense, this is what internal validity means.

To establish internal validity, you need to develop a research design that allows you to do two things.

First, you want to provide positive evidence that the relationship that you claim exists does in fact exist. I poke you. You jump. That is positive evidence of direct cause and effect. I find a relationship between religiosity and attitude about same-sex marriage. That is positive evidence that these two attributes of a person are related**, but not necessarily in a causal way.**

Second, you want to eliminate other possible explanations. I put you in a soundproof room and poke you. That eliminates noise as a possible explanation for you jumping. I could put a blindfold on you, too. Then you would not jump just because you saw me start to move. These are simple examples. In research, eliminating alternative explanations is usually very difficult.

One approach is to try to eliminate or account for **non-experimental, confounding, or extraneous variables.** Except for the experimental group of designs, as a practical matter this means we need to eliminate or account for relationships that we already understand or know about. For example, you want to know why adolescents get involved in gangs. We already know that early exposure to community violence has a strong positive relationship with adolescent gang membership. This is a well-established relationship and there is no reason to re-study this if your only research question is "Does exposure to community violence lead to gang membership?" Doing so will just cost you time and money and probably make it harder for you to understand the more universal processes and phenomena at work that lead kids to become a gang member. Why not study kids involved in gangs who have *not* been subjected to repeated community violence? You could go further and look at other well-established patterns and eliminate those factors from your study. You have not eliminated all non-experimental, confounding or extraneous variables, but you have removed the obvious so that you can concentrate on what we do *not* know.

Another good way to "eliminate other explanations" is to use a theoretically comparative approach in which you draw upon two or more "competing" theoretical explanations to discover which one best explains your observations (your data). Again, you cannot compare every relevant theory, but you can at least compare major contenders and move our understanding forward by determining which one seems to provide the best explanation.

*The take home lesson: do not focus solely on providing positive evidence that your explanation is the right or valid one. Think about innovative ways to rule out other explanations or to move past what we already understand and know.*

**Common Threats to Internal Validity** This is a list that almost everyone sees sooner or later about the kinds of threats to internal validity. I do **not want you to take a checklist approach to understanding and evaluating internal validity. This list is of limited utility.** You should think about all these things, but you can never eliminate all of them. Focus on the one or two that are *big, real threats* to a specific research project.

History: Outcomes are due to events and processes that occurred in the past and have a persistent influence on social processes today (or biological or physical – applies to all sciences).

Racism is a historic threat to the conclusions we reach in the social sciences. Racism is deeply embedded in many societies, including the U.S., and affects many processes today. It is almost impossible to measure the total effect it has on the conclusions we reach. In fact, we often treat "race or ethnicity" as a predictor (causal) variable in studies, and then fail to specify the mechanisms or pathways that would define the nature and extent of the threat in our individual research projects. That is, we treat "race" the variable as an explanation for outcomes without further definition.

Maturation: Outcomes are due to a natural or non-experimental change that occurs over time due to growth and/or development of individuals, groups, or organizations

You evaluate employee performance. Then the company institutes a six-month training program. You evaluate employee performance again after the program. Performance scores are much higher. You think "Great. That program worked!" But there is a problem. The improvement in performance may be due to more job experience. Perhaps most participants would have improved *more without the training program* due purely to greater experience. All sorts of studies of children and youth over time are plagued by maturation effects. Lots of changes happen just because they are maturing and changing physically and psychologically. If you are going to use an intervention -- a program, a drug treatment, something like that -- use a *treatment and control* design (an experiment). That way you can tell how much change occurred because of maturation -- that will be the average change in the control group. In many cases, however, you simply cannot eliminate maturation effects. At a minimum, you discuss them and use the literature to try to determine how important they are to your conclusions. More important*, one type of design – the longitudinal design –incorporates maturity as a component in the design*. If maturity is important and you need to understand maturation effects, this is an excellent design to select. On the other hand, if you select a cross-sectional design, you have little protection from maturation effects.

Testing Response: Outcomes are due to study participants learning from or becoming familiar with the testing procedures you use

You give people a Likert scale about their attitudes about LGBT individuals. Then they participate in a sensitivity training. You give them the same scale again. Aha! They now have a more positive attitude. You think great! My training worked. However, maybe they just figured out that you think people should not be prejudiced against LGBT populations and they're giving you the answer you wanted. This is called *social response bias* and it is especially a problem when you ask about the same thing more than once, particularly socially sensitive topics. Or you interview people about what they think our policy in China should be. Then you interview them again in three months to see if their ideas have changed. The first time you asked them your questions, many of them were formulating their ideas. They had not really systematically thought through their ideas about what we should do. Thinking through all that is just plain hard mental work, so the second time you ask your questions, lots of people repeat what they said the first time, without really reconsidering all the events that have occurred since then, new information they have received, etc. This is called *conditioning*, and we all do it when we give the least effort response. People learn in other cases. You give a test of simple

factual information. Three weeks later, you give the same test again. But people learned what questions were on the test. They may have even thought about them over the three-week period, consciously or subconsciously. The second time they score higher because they learned from the test itself.

Instrument Decay: Outcomes are due to changes in the instrument over time

All research uses some kind of instrument -- ranging from a physical rig of some sort to a set of questions for an interview (called the interview schedule). They can change -- usually deteriorate -- over time. For social scientists, **we** are often part of the instrumentation because we are asking the questions, recording the answers people give, etc. You get bored after you've asked 20 people the same set of questions. Unconsciously, you start to speak more quickly and people do not understand you as clearly. Or you start to look kind of bored. People give you shorter answers because they can tell you are not giving them all your attention. Or you get lazy about taking good interview notes and quit writing down things like people's facial expressions, pauses, etc. that often tell you as much as the words they speak. This is all instrument decay.

Regression to the Mean: Outcomes are due to abnormally high or low values

This is a special problem for any study where you measure something on more than one occasion. Sometimes, the conditions of the study itself cause the problem. You are studying depression. People who come to a clinic complaining of depression are diagnosed for degree of depression. Then they get counseling for eight weeks. They are re-diagnosed. Hooray!!! They are not nearly as depressed as they were before counseling! However, even for very depressed people, depression tends to go up and down over time, and they tend to seek help when they are at the nadir of depression. While not everyone will get better over time, half or so probably will, with or without treatment.

Mortality or Attrition: Outcomes are due to non-response by participants

Attrition (or mortality) can occur for many reasons. First, some people that you select for your sample will refuse to participate or you will not be able to contact them. In some cases, people inflate the required sample size to account for attrition. In other cases, researchers use **replacement**. This means that every time someone says no, or if you fail to contact a participant, you select a new participate to replace the one lost to attrition. In either case, the **respondents** are not usually **exactly the same** as the participants you selected. This can cause a big problem if it turns out that the non-respondents are **different from the ones who do respond in a way that affects your conclusions**. You want to study alcohol consumption by freshmen at UF. You get a random sample using an alphabetical list of freshmen. Of course, many of them do not want to be part of a year-long study about their alcohol use. This may not be a problem in and of itself, but you start to notice a pattern. The refusal rate among women is higher than among men. Women seem less willing than men to participate in your study. You do not know why, but it could be because women are more likely than men to think that consuming alcohol is a bad thing and do not want to tell anyone about their drinking habits. Or it could be that women consume less alcohol and are simply less interested in the whole subject. Or it could be that men are proud of drinking and are eager to tell people about it, compared to women. In any case, you have a problem due to attrition. You will not be able to draw valid comparisons about the drinking patterns of men and women freshmen.

Attrition or mortality can be a big problem for studies that extend over time because the people who "drop out" may differ from those that "stay in" the study. Take the depression example. You decide to re-contact everyone one year after the end of your counseling to see if the beneficial effects of counseling have persisted, to see if it works over the long haul. About 25% of your original

participants do not respond to your e-mails, phone calls and letters. They have been lost to **attrition**. Now you examine their original depression diagnoses. It turns out that many of the people that you cannot contact had severe depression. This is a big problem because it would seem that the non-respondents -- the folks you could not reach for the follow-up -- and the respondents -- the ones you could contact and who agreed to come in for a follow-up diagnosis -- are different **in ways that can affect your conclusions.** Maybe the really depressed people have totally given up on getting better. Maybe some committed suicide. Maybe they have concluded that counseling is worthless. At any rate, you can have little confidence in your ability to draw conclusions about the long-term benefits of counseling.

Selection Bias: Two or more groups that you want to compare differ in some important characteristic(s) that affects your conclusions

**Simple selection bias** can cause you to **conclude that two groups used for comparison differ when they do not**. You take a random sample of 500 men and 500 women and ask them to watch the first ever YouTube format presidential debate. Afterwards, they score how much they liked it on a scale of 1 to 10. Men have an average score of 8.2 and women have an average score of 6.4. The difference is statistically significant. You conclude that men like the YouTube format more than women. Then someone examines your data closely and you find that the average **age** of the men was 32 while the average age of the women was 56. You did not pick younger men on purpose. It just turned out that way. You have a selection bias problem because **age** could affect how much people like and feel comfortable with the YouTube format. Younger people might like it more because they are more accustomed to the technology, because many of the people who submitted questions were themselves younger -- a lot of reasons. The difference you found may be due to age instead of gender. You can no longer be confident of your conclusion.

A different form of selection bias is called **selection interaction bias.** In this case, the problem is that **you cannot detect a real difference that does exist**. A school district has some grant money to provide a new computer-based math program to its students. The program is new and has not been well tested. The grant funds will run out after one year. Subscribing to the computer- based program is expensive. You decide to test program efficacy before you commit scarce budget dollars to the program. You pay for using the program on a per participant basis to keep the cost of testing low. You want to use your grant money to help the students who need it the most, so you give students who have a D or E in math access to the program. These people are in the treatment group. After 8 weeks, you give everyone a standardized math test. The average score of the treatment group is 73. The average score of the control group is 77. There is **no significant difference** between the treatment and control groups on the test. What can you conclude? It may be true that the program worked and really helped the poor performers, but you cannot be sure because there is **no significant difference** between treatment and control groups on the test. This happened because you did not randomly assign students to treatment and control -- this would eliminate the problem. Or you could have given everyone the standardized test before starting the intervention (a pre-test) and then tested again at 8 weeks (a post-test) and measured **change in score.** Given that you failed to do either of these things, you cannot really know whether program "works" or not. This is selection interaction bias because the way that you selected the participants for the treatment or intervention interfered with being able to draw firm conclusions.

**External Validity = Ability to Generalize**

We conduct research with a specific set of people, in a specific place, under a specific set of conditions. But we almost always want to be able to **generalize** our conclusions to other people, places and conditions. Being "sure" of what happened with 150 students at the University of Florida in

fall semester of 2018 is good, but we want to know if it is true for all students, at all major public universities, every year. ***External validity*** refers to the degree to which you can extend your conclusions to other people, other places, and other conditions.

There are two aspects to generalization -- statistical and theoretical generalization. We are usually concerned with both, and always with theoretical generalization.

***Statistical generalization*** refers to the degree to which: (1) any characteristics of the sample of people in our study (descriptive statistics) can be applied to the population as a whole from which they were selected and (2) the degree to which the results of statistical tests (inferential statistics) can be applied to the population of interest.

Generalizing Characteristics of the Sample to the Population. You want to know whether freshmen gain weight during the first year away from home at college (too much pizza, hamburgers and KFC). You get an alphabetical list of every incoming freshman at UF at the beginning of fall semester, 2018 You take a completely random sample of 500 students from that list. You weigh them at the beginning of fall semester. You weigh them again at the end of spring semester, 2019. On average, they gained 4.8 pounds. You ***can*** generalize that the average freshman at UF gains 4.8 pounds during their freshman year because there is no reason to believe that a person's surname somehow affects their eating habits and weight gain. This is fine. In fact, you can probably generalize to freshmen at major public universities in the United States because there's no reason to think that freshmen at UF somehow differ in the eating habits they adopt as freshmen from students at Ohio State or Berkeley. However, if you took the same sample of 500 freshmen at UF and measured their average GPA at the end of the freshman year, you would have to be much more cautious. You ***could generalize*** to UF freshmen as a whole. **But** UF attracts very good students. We differ from "large public universities" as a whole in this respect. So you probably could not generalize that the average GPA for your sample -- say 3.8 -- is the average GPA for freshmen at major public universities in the U.S. as a whole.

Generalizing the Results of Statistical Tests from a Sample to a Population. The same general rules hold for generalizing the results of statistical tests. Assume that in your weight gain study you select ***two separate samples*** of 500 women and 500 men. On average, freshmen women at UF gain 5.4 pounds and men gain 4.2 pounds. You run a student t-test and it turns out that the difference is statistically significant. Women did statistically gain more than men. Just like before, you could generalize to UF as a whole -- women freshmen at UF on average gain more than men freshmen ***and you could generalize to freshmen at major public universities as a whole*** because there is no reason to believe that differences between men and women freshmen would differ simply due to where they go to school.

Statistical generalization ***depends*** on having a ***statistically representative sample*** of the population you want to talk about. We will discuss this more under sampling, but much of this is just common sense. For example, in your weight gain study you would probably want to ask potential participants if they have any eating disorder and ***eliminate those who do***. This is called screening.

A statistically representative sample ***does NOT mean a sample that is just like the population as a whole in every way.*** It has to be "like the population as a whole" in terms of characteristics that matter to what you are studying. It is common for novices to believe that research findings cannot be generalized because the sample differed from the population as a whole in terms of age, or race, or ethnicity, or gender. These characteristics of the sample are only important to your conclusions and your ability to generalize ***if they are likely to affect the attribute you want to study.***

Consider IQ. You have the same sample of 500 freshmen at UF, chosen by alphabetical listing. The average IQ is 120. As you look at some other ***descriptive statistics*** about your sample, you see that the sample differs from freshmen at UF as a whole. Just by accident, 41% of your sample consists of Latinx people whereas 23% of UF's incoming freshman class are Latinx. We say that Latinx people "over-represented" in the sample, compared to the population as a whole. But there is ***zero valid evidence*** to suggest that ethnicity is related to IQ. You can conclude that the average IQ of incoming freshmen at UF is 120 because ethnicity and IQ are unrelated characteristics.

On the other hand, if 38% of your sample speaks two or more languages, you probably cannot conclude that 38% of incoming freshmen at UF speak two or more languages. People of Latinx heritage are probably more likely to speak a second language than lots of other people who come to UF based on exposure to Spanish in the home. Now the ethnic difference between your sample and the UF freshman class as a whole does limit your ability to generalize.

***Theoretical generalization*** refers to the degree to which you can use your findings to gain a broadly applicable understanding or explanation of the relationships between different phenomena. We base research on theories because theories provide our basis for figuring out how things work not just with one set of people, or in one place, or at one time, but in general. By basing research on a theory, we move beyond ***descriptive research*** to ***explanatory research.*** Descriptive research just tells us "what happened with these participants in this study." Explanatory research tells us "what happened with these participants in this study -- ***and*** how can I explain what happened and understand why it happened. Therefore, we always want to be able to theoretically generalize. In order to be able to theoretically generalize, you need to ensure ***the theoretical adequacy*** of your study. This involves four things, and failure to ensure any of the four is a threat to your ability to theoretically generalize your findings:

• Make sure that the theory applies to what you are studying
• Make sure that you use the constructs in the theory in your research
• Make your study as "reproducible" as possible
• Select a sample that is theoretically sound

Domain of the Theory. There is no theory that explains every possible human behavior. For example, the theory of reasoned action explains purposeful decision making. If you want to understand how people behave in mob situations, you should not use the theory of reasoned action for your study. Mob behavior is not a part of the domain or set of behaviors covered by this theory.

Content (sometimes called construct) Validity. The building blocks or components of theories are constructs, but constructs are abstract ideas -- not "real" things. We turn constructs into variables, the major concern of research ***methodology***. Suffice it here to say that you must make sure that your variables actually represent the concepts or constructs in the theory you are using. Attitude is an important construct in the theory of reasoned action. We usually use a Likert scale to measure attitudes -- where you indicate how much you agree or disagree with a set of statements. A common beginner's error is to include a statement of fact -- rather than opinion -- in the set of statements. You want to know about people's attitudes about dieting as a way to control weight. You include the statement "Most people regain weight once the dieting period is over." This is true. it is a fact. I could mark that I strongly agree with this statement because I think it is true, and still think that dieting is a very good way to control weight. In this case, by including a statement of fact instead of opinion, your variable did not adequately represent the construct of "attitude or opinion."

Reproducibility. You almost always want to make your study as reproducible as possible so that other

researchers in the future can repeat it or build on it. Sometimes we do want to study very rare or unusual things, but usually not, so you do not want to create a study that depends on some very specific conditions or set of participants that other researchers cannot hope to reproduce. Reproducibility does *NOT* mean that someone else has to be able to find exactly the same participants, or even necessarily a very similar set of participants. It does mean that you want other people to be able to find or create similar *conditions* for a study. For example, if your theory is supposed to help us understand the factors that affect how people perceive of government services generally, conducting a study with people from New Orleans right after Hurricane Katrina is a bad idea. The conditions were extreme and hopefully not reproducible.

Theoretically Representative Sample. This is very different from a statistically representative sample. It means that the participants in your study adequately represent the population whose behavior the theory is supposed to help us understand or explain. For example, the theory of reasoned action deals with purposeful behavior. If you select a sample of people who are all under treatment for psychotic behavior, your sample is not valid. This theory does not deal with psychotic behavior. On the other hand, if you want to apply the theory of reasoned action to understanding how people decide to smoke (or not), it is fine if your sample consists of middle-class college students. There is no reason to think that their reasoning process somehow differs from that of poor people, older people, etc. They are a theoretically adequate sample and you could compare your findings directly to those of another researcher whose sample consisted of middle-age bankers.

Other Threats to External Validity. There are two other major threats to our ability to generalize our findings beyond our specific study group, place, etc.

One is *sensitization.* Sometimes just including people in a study makes them act differently than they normally do. Let's say I'm a volunteer in a study about eating habits. I'm not supposed to change my habits during the study -- just write down what I eat. But I cannot help but start paying more attention to my diet when I write it down. "Wow! I'm drinking five soft drinks a day! That's too much." So I reduce my soft drink consumption.

The other is *artificiality.* Some think this is a big problem for experiments and quasi- experiments. Some argue that a social psychology experiment where you ask a randomly selected group of students who do not know each other to develop a set of rules for punishing students who are late to class can never generate reproducible outcomes because the study involves atypical conditions in a limited time during their first meeting while being observed by researchers. Some argue that this makes it impossible to generalize the findings of these studies. I disagree. *IF* the experiment meets the conditions for theoretical adequacy, the findings can be generalized.

There was a fairly well-known social psychology study done many years ago. A researcher wanted to try to understand why guards become abusive toward prisoners and thought it had to do with group identity. He assigned student volunteers randomly to either a "guard" or "prisoner" role and isolated them in a facility where they played out their roles. They were continuously observed. These participants were certainly under "artificial" conditions. The study was supposed to last several weeks. He had to discontinue it after just a couple of weeks because the "guards" started to become extremely abusive of the prisoners -- even when they were acquaintances or friends -- and the researcher became fearful of actual serious physical abuse. People thought this was one of those "artificial" studies -- until the experiences of American servicemen guarding prisoners in Iraq. This study was directly cited as a good example of the group dynamics that occur when some people are placed in control of other people to whom they are "superior" in some sense. In fact, the power of his study to explain these abusive patterns of behavior is probably even stronger precisely because his student participants did not in fact differ culturally, often knew each other, were strongly conditioned to

treat each other as equals, etc.

I personally do not know how much to treat artificiality as a threat. You need to think about it. But it is also true that with the very strict informed consent procedures that we must use, almost every study is subject to both sensitization and artificiality threats. Try to minimize them, but do not be thwarted completely by them.

**Explanatory Power = Thorough Explanation or Understanding**

Very few theories in any discipline explain everything about what we want to understand. Very few studies can explore every possible relationship. In general, however, one of the goals of research is to provide as full or complete an understanding as possible -- not to settle for understanding just one relationship, or how something works under just one very restrictive set of conditions. You want a powerful explanation or understanding. There are three basic approaches that you can use to make your explanations or understanding more powerful.

Work Hard to Put Your Ideas to a Tough Test. Science is not about "proving" something is true, nor is it just about "disproving" a hypothesis. it is about taking your favorite explanation and putting it to the toughest possible test. Conduct your study under conditions, with a sample, in a situation where your favorite explanation will be put to a test. One of my favorites on this (a **negative** example) is that we have some "great" program designed to improve say self- esteem. We love it. We think it is wonderful. We want to "prove to the world" how well it works. We conduct a pre- and post-test of the self-esteem of our participants. Sure enough – their self- esteem increased. We conclude that our "great" program really **IS** great. This is essentially a bogus study. There is no comparison group here. There is no way to conclude that our program had any effect. In fact, all of our participants **chose to participate voluntarily** in our program. Maybe they were all low in self-esteem. Of course, their self-esteem increased – because somebody paid attention to them. **Any** attention would have worked. Or perhaps they all had high self-esteem. They were, in some sense self-esteem seekers and reinforcers. Again, our program worked because these are folks who spend a lot of time working on their self-esteem, not because we have a great program. Be tough on your own ideas.

Gain Breadth. Try to understand how several different phenomena fit together, to expand the number of linkages that we can explain or understand, or sometimes to show that a theory applies well beyond the conditions (place, people, setting) in which it was originally proposed. Theory-comparative studies help us do this. Working with a population or under a set of conditions that stretches the explanation (like a study of kids from perfectly nice homes and communities involved in gangs) can help with this. Using comparison groups that differ greatly in ways that may affect the outcomes can help. What does **not** help is one more study exploring what previous research has already established.

Explore the full variance. For example, I participated in a study about the effects of grape juice on cognitive function. As a condition of this study, I had to **NOT** eat or drink the juice of any red or blue "berry" -- things like cranberry juice, strawberries, blueberries, raspberry jam, etc. Except for raspberry jam, I eat a LOT of these things. But these researchers needed to isolate the effect of grape juice so they had to eliminate these confounding, non-experimental variables from the study -- get the study participants to be alike in the sense that our only source of the beneficial compounds in red and blue berries in our diets was grape juice. We took five tests of cognitive ability every week. Assume that the people in the treatment group (got real grape juice) showed a 30% increase in cognitive function while the control group (phony grape juice) showed no change in cognitive function. The researchers have shown direct cause and effect -- grape juice works. However, the magnitude of the effect – the 30% increase in cognitive function – may not be generalizable. Outside the experimental setting when

people eat strawberries, drink cranberry juice and have raspberry preserves on their toast every day, the effects of the grape juice might be minimal, not even measurable. Is grape juice "good for" your cognitive functions? Yes, it is, but you cannot conclude that "on average, people who drink 12 oz per day of grape juice will show a 30% increase in cognitive ability." That depends on what else they are eating and drinking. A person who loves strawberries, blueberries, cranberry juice and raspberries might add 12 oz of grape juice to their diet and get no improvement in cognitive function.

**Summary**

Research design has three main goals:
    Provide confidence in the conclusions we reach
    Allow us to generalize beyond our specific study
    Give us a robust explanation or understanding of what we study

There are three major design groups
    Experiments and quasi-experiments
    Longitudinal and cross-sectional designs
    Case studies

None of these design groups is "better" than the others in any overall sense. They have different strengths -- and weaknesses. Experiments, for example, are the only group that can demonstrate direct cause and effect. Longitudinal and Cross-sectional designs allow us to explore differences in patterns and processes among existing groups of people. Case studies allow us to understand complex interactions in detail.

**Keys to success**

***Keep the three goals in mind*** -- do not just ignore them and remember that ***the research question drives the design***.

| Research Question: | What happens when I? |
| --- | --- |
| Research Objective: | Demonstrate DIRECT cause and effect. |
| Design Group: | Experiments |

| Research Question: | How does that work? |
| --- | --- |
| Research Objective: | Understand the relationships among several phenomena and processes |
| Design Group: | Observational |

| Research Question: | Why did that happen? |
| --- | --- |
| Research Objective: | Explain an outcome state |
| Design Group: | Case Study |

***Used a multi-stage design if possible***. Take advantage of the strengths of different designs. You will gain in all three areas.

Every study has some limitations in terms of internal validity, external validity, and explanatory power. Discuss them openly in your thesis or research article and let the reader decide for him/herself how applicable your results are to their situation. It ***is*** your responsibility to inform the reader of the limitations of your study. It is ***the reader's*** responsibility to decide whether those limitations restrict the applicability of your conclusions to his/her intended application.