Threats to Internal Validity

The true experiment is considered to offer the greatest protection against threats to ***internal validity.*** Note in this discussion that pre- and post-tests are the same test, although question order is normally changed.

History: some event occurs, beyond the researcher's control, that affects the outcome of the study
Example: A study is underway in 2001 in which the DOD (Department of Defense) is comparing three recruiting approaches, (1) the standard "wait for the recruit to walk into the office" approach, (2) cash payment for enlisting and bringing in one additional recruit, and (3) cash payments for enlistment. The attacks on the World Trade Center occur. Enlistment goes way up. The true experiment protects against this threat because of random assignment to the three treatments. The increase will be greater for ***all*** treatments, which will still permit the DOD to determine which produces more recruits.

Maturation: scores on the post-test go up compared to pre-test scores just because the participants mature in some way (get older, gain more experience, attend a class, etc. – any form of maturing)
Example: new employees in a corporation are randomly assigned to treatment (mentoring) and control (one new employee seminar) groups to test which is a more effective way to teach them about corporate procedures. The study runs for 6 months. All of them should improve their scores on how to complete various procedures as they gain experience, but some employees will learn from experience better than others. Random assignment to treatment and control allows the corporation to evaluate the value of mentoring because there should be a random distribution of employees who learn well and who do not learn well from experience in each group.

Testing response: participants do better on the post-test than the pre-test just because they already have experience with the questions (know what to expect, thought about the questions between tests, etc.)
Example: a study is underway to evaluate the effect of traditional (lecture) versus active (exercise-based) learning. Participants take a pre-test, participate in a 3-week training program, and then take a post-test. All scores should go up, but some participants will have more knowledge initially than others and may learn more or less as a result of their prior knowledge. Random assignment should distribute more and less knowledgeable participants equally in the two groups.

Instrument decay: reuse, especially several times, of an instrument literally causes it to wear and become less accurate. This includes human instruments conducting interviews and such. The interviewer is typically "fresher," more attentive, and more interactive on the 1st interview than on the 20th interview. If you don't believe this, try it yourself.
Example: the researcher conducts pre- and post-tests of participants in an intervention program by interview. There are the normal comparison groups (like improved vs. traditional type of intervention). The researcher is "fresh" during the pre-tests and takes better notes, asks more follow-up and probe questions, etc., than during the post-test. The overall change may be lower than the researcher anticipated because of this "instrument decay" on the part of the interviewer. If the decay is too serious, the study will fail because the post-test result will be invalid. The researcher will conclude that the improved intervention yields no better result than the traditional one. However, if the decay is not too serious, the difference may be less than one

would expect, but will still be detectable. The true experiment does not provide complete protection against this threat – but no design does.

Regression to the mean: a study fails to show significant results between pre- and post-test because random differences in the performance of participants causes the mean for each test to be the same.

Example: middle school students with below-grade math skills are randomly assigned to treatment (computer tutoring) and control (classroom instruction only) groups. A few students in the treatment group do really well on the pre-test – they were just really "on" that day. At the time of the post-test, these students actually did ***worse*** than on the pre-test. As a result, there was little change in mean score for the treatment group. Meanwhile, none of the students were "on" the day of the pre-test. Mean score for the control group does increase modestly on the post-test. Random assignment should help prevent this. Presumably the "on" students would be equally present in the treatment and control groups on the pre-test date. However, there is really no "gold" defense against this threat in any design.

Mortality: this does not mean necessarily (although it can, especially in medical trials) that participants died during the study. It means that there was some sort of differential failure to stick it out throughout the study. This is particularly problematic with studies that extend over time or that require consistent effort on the part of participants.

Example: obese adolescents are randomly assigned to treatment (weekly mentoring sessions) and comparison (e-mail reminders of good eating habits) groups in a study about approaches to weight loss. Both groups record what they ate each day for one week before the treatments start. They are supposed to record what they eat every day during the study. Both groups have the same scores on a "healthy diet" measure at the beginning. The study goes on for 6 weeks. At the end of the study, drop-out rate is higher in the treatment group. Further analysis shows that twice as many participants in the treatment group with the ***lowest*** scores on the index dropped out than participants with higher initial scores (better diets). This difference in drop out rate makes the results inconclusive. One would expect the greatest improvement (change in pre- and post-test scores) to occur for those with the worst initial diets, but they're the ones who had the high drop-out rate in the treatment group. There is no real defense against this, but random assignment helps. Presumably the "low motivation" and "high motivation" participants would be equally distributed across both groups. In fact, however, the weekly mentoring sessions may be the problem – too much effort required.

Selection bias: some aspect of how participants were selected affects the outcome of the study.

Example: participants are asked to volunteer to participate in a study about parenting skills for first-time parents. They are randomly assigned to treatment (one month of active-learning training with "baby props" and exposure to the kinds of "mini crises" that new parents face) and comparison (one month of traditional training sessions about general parenting skills) groups. At the end of the study there is no significant difference between groups – in fact, there was very little improvement at all for either group. The real problem here is probably in selection. Motivated people probably volunteered and they may have been actively seeking out all sorts of advice and information about parenting. However, random assignment could help because we would anticipate that highly motivated learners would be equally distributed in both groups. If there is any effect due to treatment, even a small one, random assignment would make it much more likely for the researcher to detect it. Random assignment is particularly important in studies where the effect is apt to be small or subtle.

Selection Interaction Bias: some aspect of how participants were selected causes they to respond ***differently*** to treatment and control

Example: middle school students with below-grade reading skills are randomly assigned to treatment 1 (computer tutoring) and treatment 2 (in-person tutoring) groups. After the study is completed, performance goes up more for the treatment 1 group than the treatment 2 group – contrary to expectations. Selection interaction bias can occur for many reasons, but one is particularly common and I will use that here. Researchers must inform participants in any study about the amount of time or effort participation will require. If there are significant differences in effort required of participants between treatments, the researcher must tell participants about these differences. ***Only those participants who agree to the higher effort treatment are eligible for assignment to this treatment.*** Thus, the pool for people to assign to the "high demand" group will be limited and probably will consist of people who have more motivation, even if the total pool was originally randomly selected. Even if the limited pool of students who say they're willing to participate in the more intensive (personal tutoring) treatment are randomly assigned to treatments, none of the "lower motivation" students can be included in the intensive treatment. There is an interaction between selection (in this case willingness to engage in the intensive treatment) and treatments that nullifies the results. Selection interaction bias can occur in many designs, not just true experiments. In fact, true experiments offer the greatest (and still limited) protection against this threat. It's not perfect, but it's the best we have. Quasi-experiments, as we will see, are particularly prone to this threat.

Threats to External Validity

True experiments do not offer as much protection against threats to external validity. In fact, some argue that they increase some of these threats (see Mark reading). Nonetheless, remember that the purpose of the true experiment is to find out if there is a direct causal relationship between treatment(s) and outcome (internal validity). We are only trying to generalize that the intervention "works." The experiment is set up to control variance – to see if the intervention has any effect at all. That's why we take a homogeneous sample (screen) from a homogeneous population. Other designs (cross-sectionals, case studies, longitudinal) address other kinds of research questions where generalization has to do with the degree to which an intervention "works" over a wide range of variance. Review "Types of Research Designs" in module 5 if you are unclear about this difference in purpose for designs.

Selection Interaction Bias: some aspect of how participants were selected causes they to respond ***differently*** to treatment and control – which presents a threat to external as well as internal validity
Example: take the same example of the middle school students with below-grade reading skills. Based on the results of that study, we would have to conclude that the results could only be extended to highly motivated students. In simple terms, our initial theoretical population was "middle school students with below-grade reading skills" and we have to change that to be "highly motivated middle school students with below-grade reading skills." We can no longer generalize to the original theoretical population.

Sensitization: participants start to behave differently, learn more, try harder, etc. just because they're participating in a study – they are sensitized to the behavior, knowledge, etc.
Example: all overweight adolescents in a school district are invited to participate in a study about weight loss. They are randomly assigned to treatment (counseling) and control (nothing) groups. Everyone records what they eat every day for 6 weeks. To the researcher's surprise, everyone loses weight, although weight loss is somewhat (not statistically significant) higher in the treatment group. The control group may have been sensitized to what they eat. Even though nobody is asking them to change eating habits or counseling them about how to lose weight, they are more aware of all those chips, ice cream, doughnuts, etc. that they regularly eat. They

lose weight. If we can generalize anything at all from this study, it would be something like "people tend to lose weight when they become aware of what they eat" – not a very useful thing to generalize and it tells us nothing about whether counseling would help. I know of no defense to this threat to external validity. It is a potential problem in ***any*** study. Even if you are just interviewing people or something like that (no treatment as in a cross-sectional), people may change what they think and do just because you asked them about it.

Artificial Response: participants' behavior is not indicative of what they would do in the "real world" because of the conditions under which an experiment is conducted.
Example: a true study – students in a university who knew each other, sometimes every well, were assigned to two roles, "prisoner" and "guard." The assignment to role was random. The objective of the experiment was to see whether role assignment would alter their behavior toward each other. This experiment actually had to be discontinued mid-stream because the guards became very abusive and the researcher was afraid that the abuse would cause psychological damage to prisoners and even the guards and that physical abuse might start. The question, of course, is "Does this behavioral change indicate what people would do in the 'real world,' or did these students just become abusive because they got a chance to 'play act' a few times a week, under observation, in a totally contrived 'laboratory' setting?" As it turns out, in this case, we later learned that artificial environment was apparently not a factor. All sorts of people who gain power over others, especially 'total' power become abusive. That's why we train prison guards and MP (military police). We've learned that failure to do so can lead to some really abusive situations. Apparently we lost sight of this lesson learned in Iraq. The number of prisoners taken was much higher than anticipated and there was a shortage of trained MPs. The abuse at AbuGraid prison resulted. I must admit, I was pleasantly surprised when one news network cited the research done many years before in the "artificial" university study as evidence that we should not have been surprised at what happened. Regular soldiers are not trained to resist the psychological processes that lead to abuse by those in power. DeVaus makes a very big deal about this (too many studies with psychology students in laboratories or something to that effect). I think he is wrong. Remember, the purpose of true experiments is to establish (or not) direct cause and effect. It may be true that in a "real world" setting the effect would be lessened or increased. But that's a different research question, addressed by other designs. With the experiment, we just want to generalize that there is or is not a causal relationship.

Explanatory Power: the limited number of constructs that can be tested in any experiment limits the degree to which we gain a full understanding of the phenomenon of interest
Any example shows this. Most experiments have at most two or three treatments, sometimes with multiple levels of each. I think the biggest I ever conducted had two treatments with four levels per treatment and four replications. That was ***thirty-two*** groups (in my case, test plots). It was a bear just to handle! Experiments do allow us to eliminate some treatments (no effect), which is just as important to explanatory power as knowing what does have an effect. However, no design tells us everything we need to know. That's why, as a good scientific realist, I argue that we should use all the design groups. I also expect you to assess explanatory power based, in part, on the degree to which the body of knowledge has been built on one design. In my book, that's a flawed body of knowledge. Different designs answer different questions. We need them all to have good explanatory power. In fact, your greatest contribution as a student researcher may be to use a design that has been "underutilized" in constructing a body of knowledge.