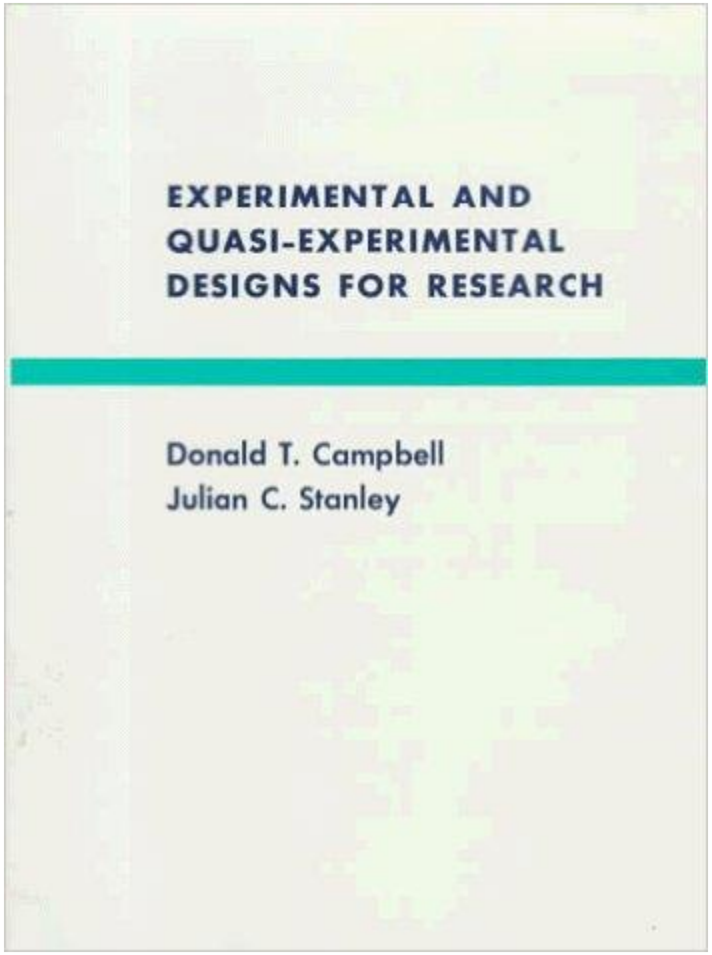
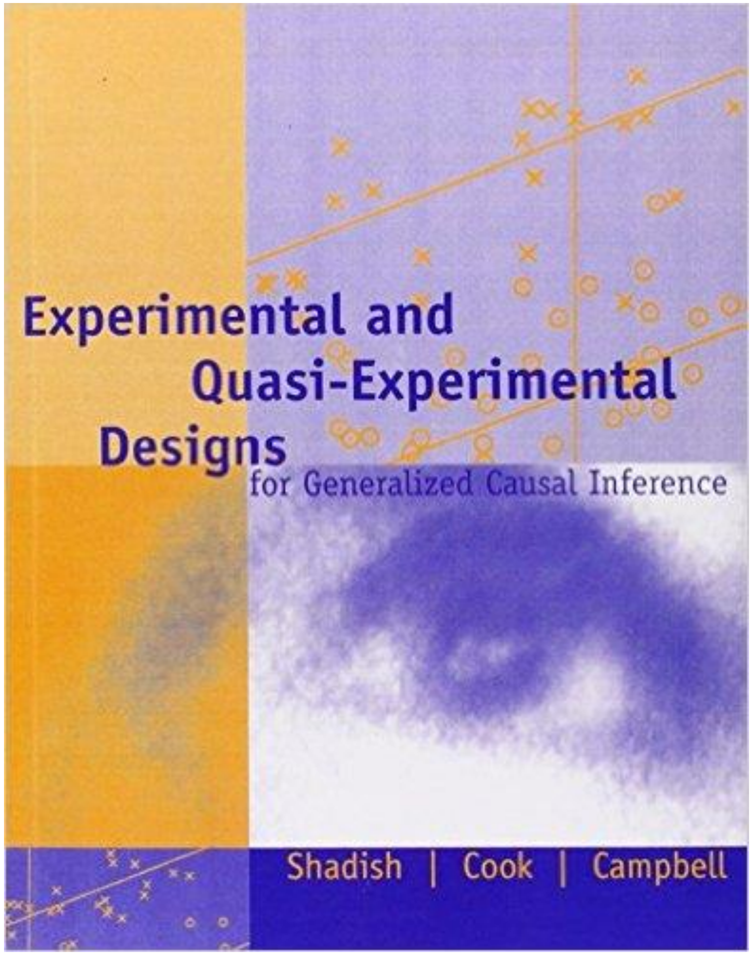
CAMPBELL SCORES: ELIMINATING COMPETING HYPOTHESES

Core Concepts

1. **Causal Analysis and the Counter-Factual**
    1. **The Competing Hypotheses Framework**
2. **Calculation of Program Effect (effect size)**
3. **Randomization Process**
4. **Control versus Comparison Group**
5. **Treatment Effect**
    1. **Average Treatment Effect (ATE)**
    2. **Intention to Treat (ITT)**
    3. **Treatment on the Treated (TOT)**

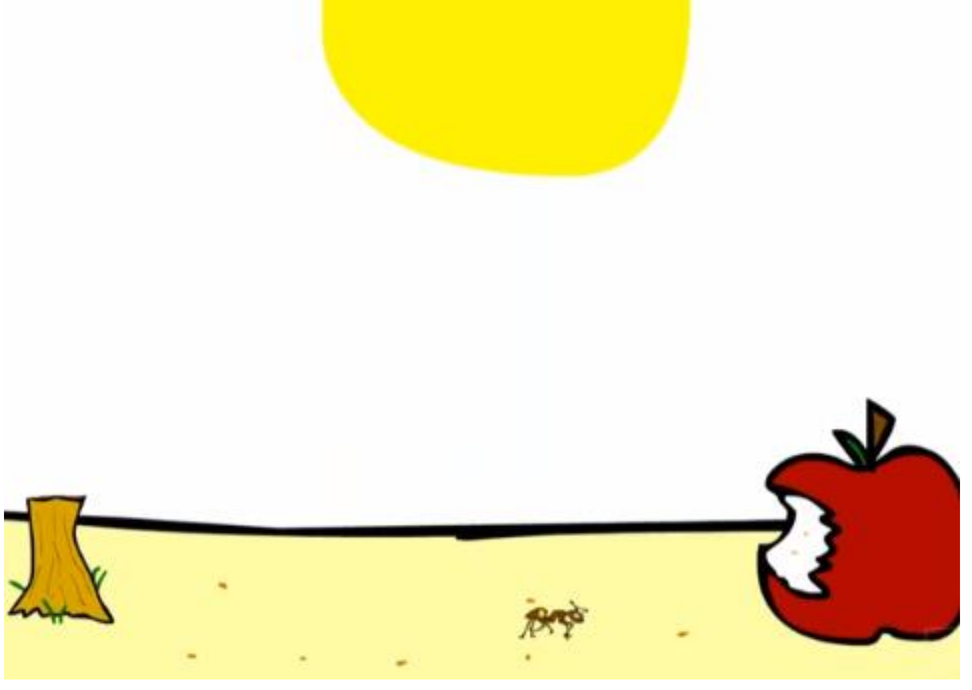




(inspiration for these assignments)

Eliminating Competing Hypotheses to Improve Internal Validity

Can Ants Count?



http://www.youtube.com/watch?v=7DDF8WZFnoU

Competing Hypotheses

The Program Hypothesis:

The change that we saw in our study group above and beyond the comparison group (the effect size) was a result of the program.

The Competing Hypothesis:

The change that we saw in our study group above and beyond the comparison group was a result of _______.

(insert any item of the Campbell Score)

*This is a general example of "**the identification problem**" in statistics. It's one thing to run a model and get significant results (the outcome definitely changed for the treatment group). Another thing to say we have properly identified the program or intervention as the cause of the change.*

The Campbell Score: Ten Competing Hypotheses

Omitted Variable Bias
Selection / Omitted Variables
Non-Random Attrition

Guilty until proven innocent

Trends in the Data
Maturation
Secular Trends
Testing
Seasonality
Regression to the Mean

Study Calibration
Measurement Error
Time-Frame of Study

Contamination Factors
Intervening Events

Innocent until proven guilty (must have evidence from the study or solid reasoning beyond simple speculation)

Scoring Items on Homework

Your job is to make a strong case. Use the definitions of Campbell Score items provided and evidence that is presented in the case studies to make your arguments.

Note that the first two items are intimately linked with omitted variable bias in program evaluation studies. Since this is the most common and most problematic issue we worry about, rigorous evaluations need to demonstrate that this problem has been addressed in order to establish a baseline of internal validity. Since most observational studies will be significantly affected by selection and attribution problems, these first two items have a “guilty until proven innocent” criteria.

The subsequent items are potential causes of concern for the internal validity of a study. Even if selection has been addressed, these other 8 things can impact our ability to generate valid causal inferences from a study, but they are less common so you cannot simply assume they will be a problem. You need to make a reasonable argument that the problem might exist in the study based upon data and evidence that is present, or sound logic and reasoning beyond speculation.

The Campbell Scores help you establish metrics for the <u>quality of evidence</u> provided in the study.

Competing Hypothesis #1

Selection Into a Program

If people have a choice to enroll in a program, those that enroll will be different than those that do not.

This is a source of omitted variable bias.

The Fix:

Randomization into treatment and control groups, or a rigorous matching process.

Randomization must be “happy”!

Test for “Happy” Randomization

TABLE 2

Background Characteristics of Students in Treatment and Control Groups
(Total numbers of cases in parentheses)

	All students in the study			*All students with scores three or four years after application*		
Characteristic	*Choice students*	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

This contrast suggests a difference in mother’s education level, but because 0.04 > 0.05/7, we do NOT reject the null that these two groups are the same. We can consider this randomization process to be “happy”.

We test for equivalence of groups by comparing their measured characteristics. The Bonferroni correction allows you to test the hypothesis that collectively multiple contrasts performed together do not suggest differences in treatment vs. control groups. Adjust the alpha by dividing 0.05 by the number of tests in a table.

Competing Hypothesis #2

Non-Random Attrition

If the people that leave a program or study are different than those that stay, the calculation of effects will be biased.

The Fix:

Examine characteristics of those that stay versus those that leave.

Microfinance Example: Artificial effects in reflective (before/after) study

\$3.00	Mean = \$2.00	\$3.00	Mean = \$2.50
\$2.50		\$2.50	
\$2.00		\$2.00	
\$1.50		\$1.50	Attrit
\$1.00		\$1.00	

Test for Attrition

TABLE 2

Background Characteristics of Students in Treatment and Control Groups
(Total numbers of cases in parentheses)

	All students in the study			*All students with scores three or four years after application*		
Characteristic	*Choice students*	*Control students*	*p value*[a]	*Choice students*	*Control students*	*p value*[a]
Math scores before application	39.7 (264)	39.3 (173)	.81	40.0 (61)	40.6 (33)	.86
Reading scores before application	38.9 (266)	39.4 (176)	.74	42.1 (60)	39.2 (33)	.35
Family income	10,860 (423)	12,010 (127)	.14	10,850 (143)	11,170 (25)	.84
Mothers' education 3 = some college 4 = college degree	4.2 (423)	3.9 (127)	.04	4.1 (144)	3.8 (29)	.15
Percent married parents	24 (424)	30 (132)	.17	23 (145)	38 (29)	.11
Parents' time with children 1 = 1–2 hours/week 2 = 3–4 hours/week 3 = 5 or more	1.9 (420)	1.8 (130)	.37	1.9 (140)	1.7 (27)	.26
Parents' education expectations of children 4 = college 5 = graduate school	4.2 (422)	4.2 (129)	.85	4.2 (142)	3.7 (27)	.01

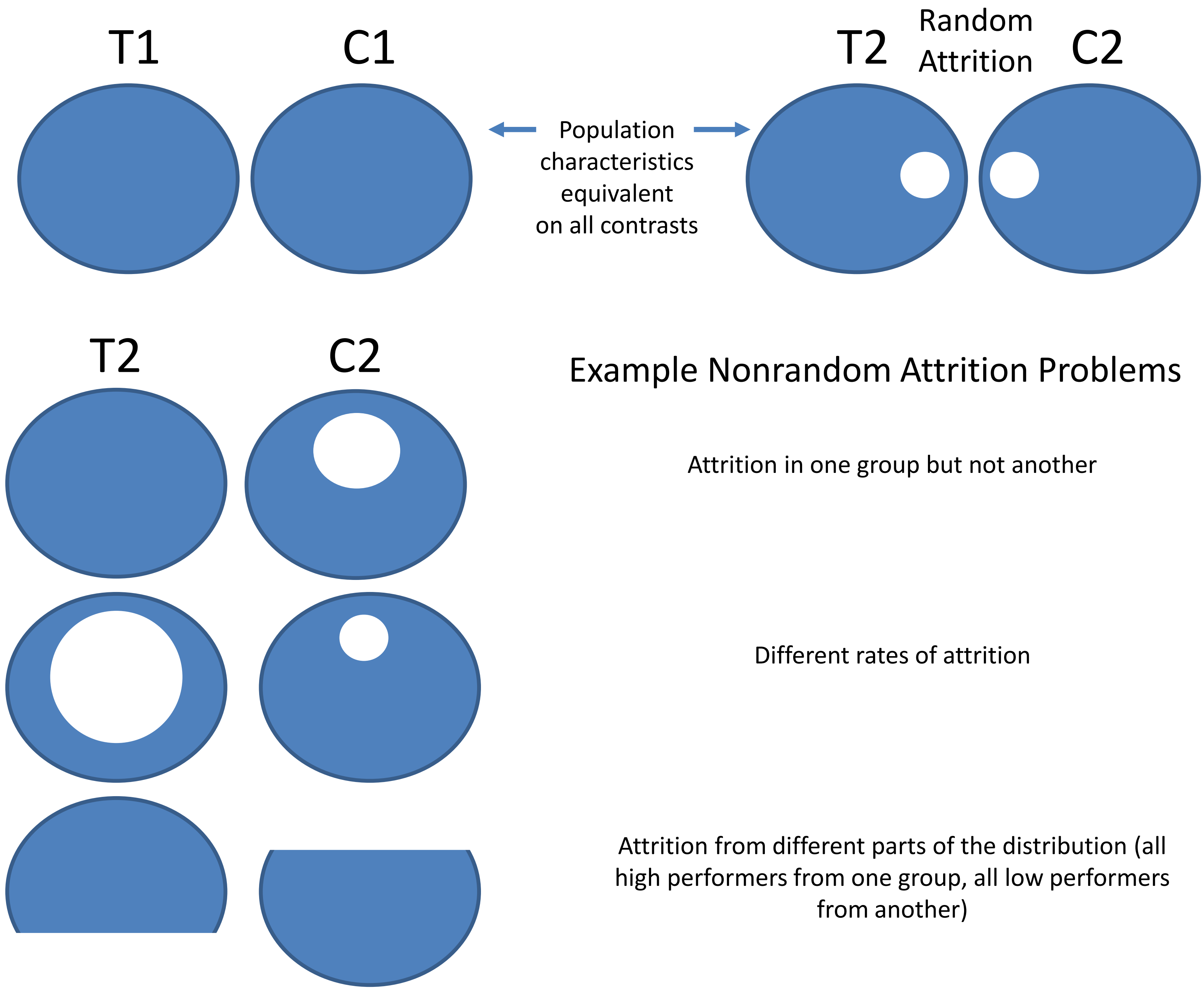
a. The tests of significance are suggestive of the equivalence of the two groups. Technically, tests of significance should be done at each point of random assignment, but the number of cases at each point is too few for such tests to be meaningful.

TEST:

T2 = C2
on all contrasts
(only use measures from before the treatment occurred)

If attrition was non-random but occurred equally across groups then it will typically not bias results. Not helpful in reflexive designs.

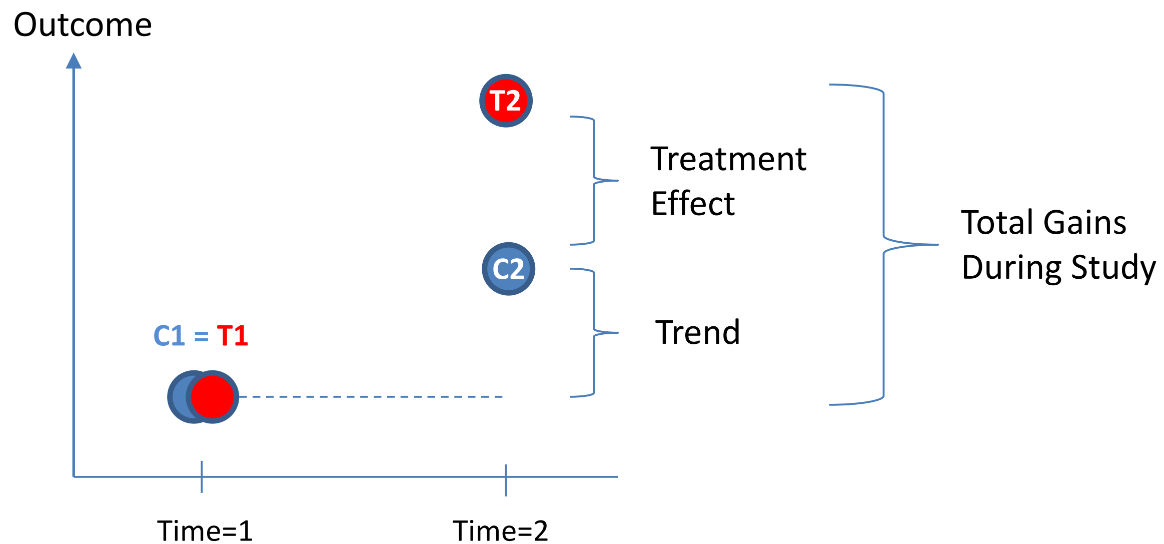
Can also be tested in another way:
If T1 = T2 then attrition was random
(useful for reflexive studies)



T1
C1
Population
characteristics
equivalent
on all contrasts
T2
Random
Attrition
C2
T2
C2
Example Nonrandom Attrition Problems
Attrition in one group but not another
Different rates of attrition
Attrition from different parts of the distribution (all high performers from one group, all low performers from another)

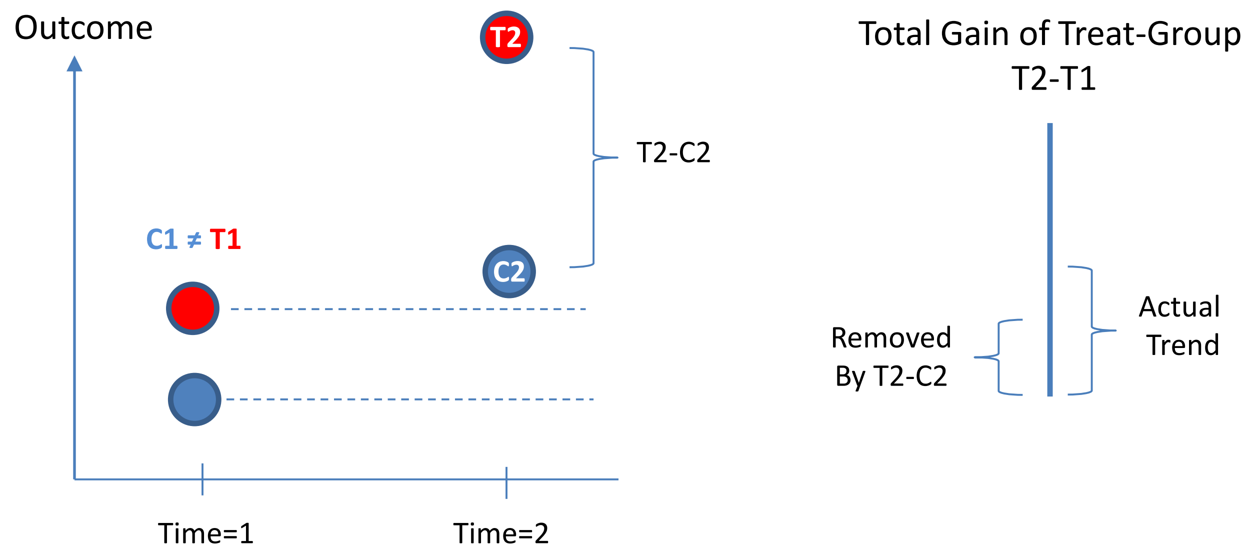
Separating Trend from Effects

T2-C2 removes trend



Separating Trend from Effects

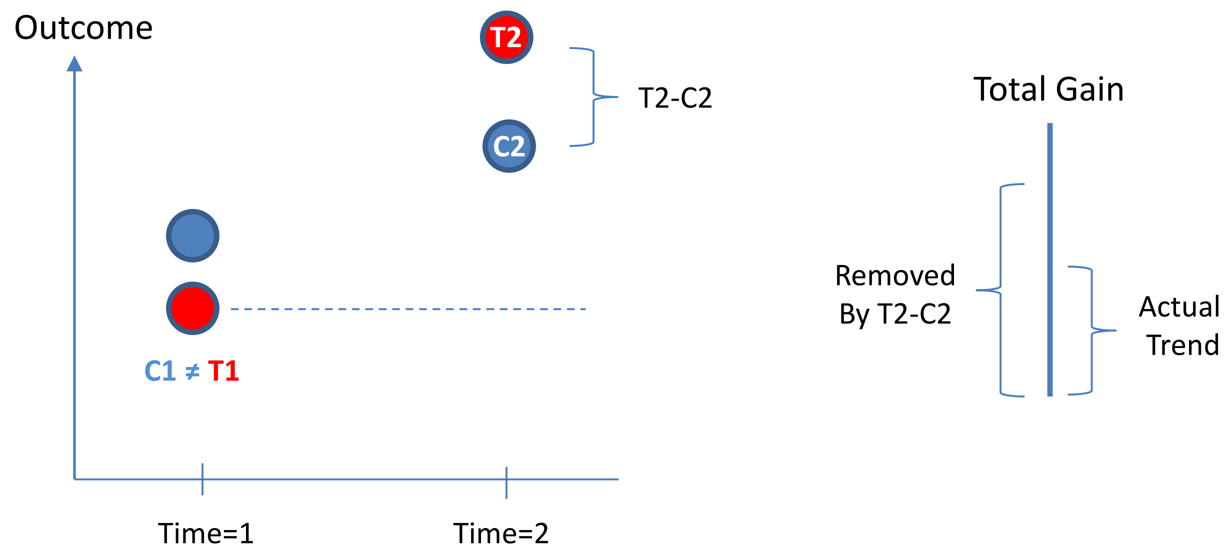
T2-C2 does NOT fully remove trend



NOTE, diff-in-diff separates trends even when groups are not equivalent.

Separating Trend from Effects

T2-C2 removes too much trend



NOTE, diff-in-diff separates trends even when groups are not equivalent.

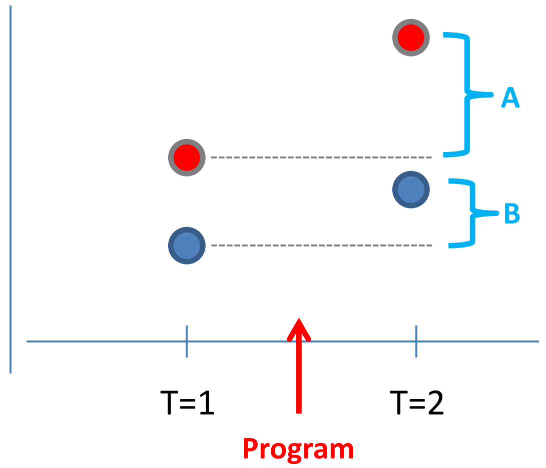
Competing Hypothesis #3

Maturation

Occurs when growth is expected naturally, such as increase in cognitive ability of children because of natural development independent of program effects.

The Fix:

Use a comparison group to remove the trend.



Pre-Post
With Control

Effect:
A-B

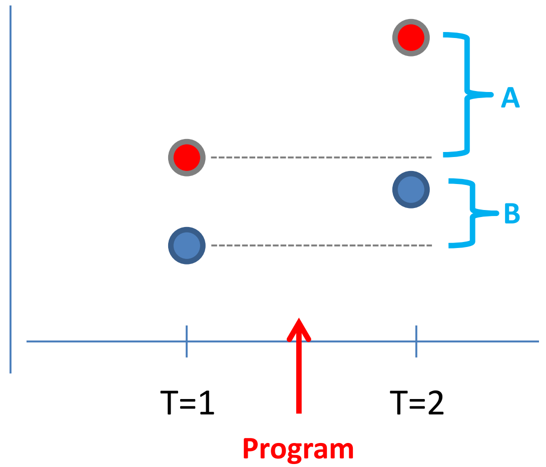
Competing Hypothesis #4

Secular Trends

Very similar to maturation, except the trend in the data is caused by a global process outside of individuals, such as economic or cultural trends.

The Fix:

Use a comparison group to remove the trend.



Pre-Post
With Control

Effect:
A-B

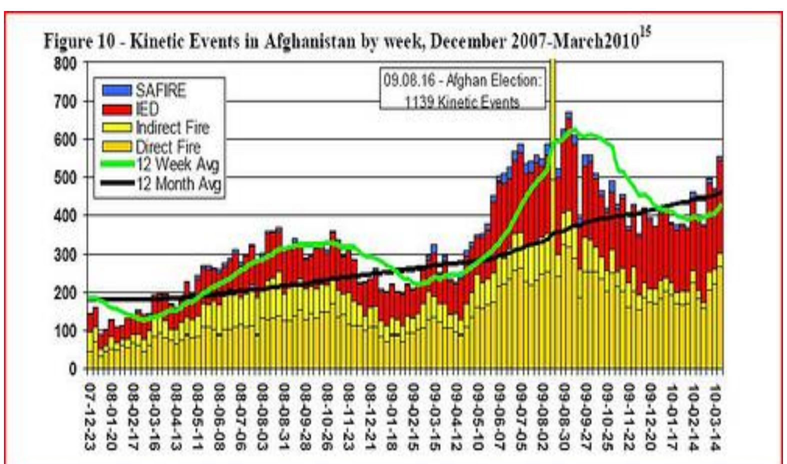
Competing Hypothesis #5

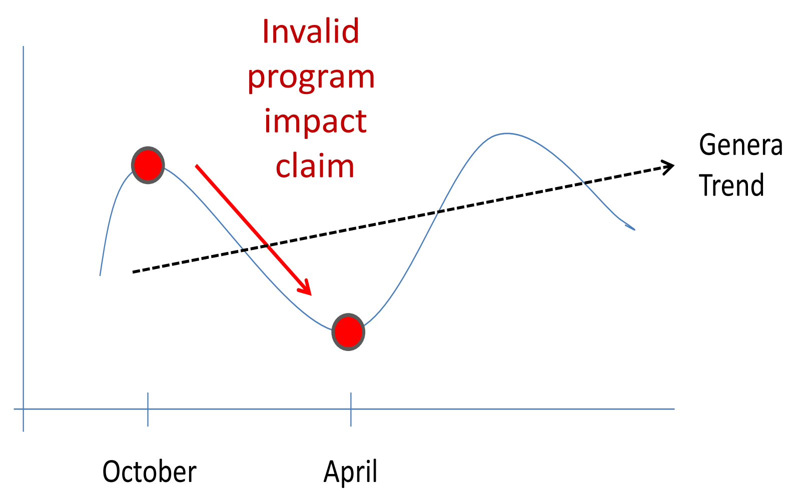
Seasonality

Data with seasonal trends or other cycles will have natural highs and lows.

The Fix:

Only compare observations from the same time period, or average observations over an entire year (or cycle period).





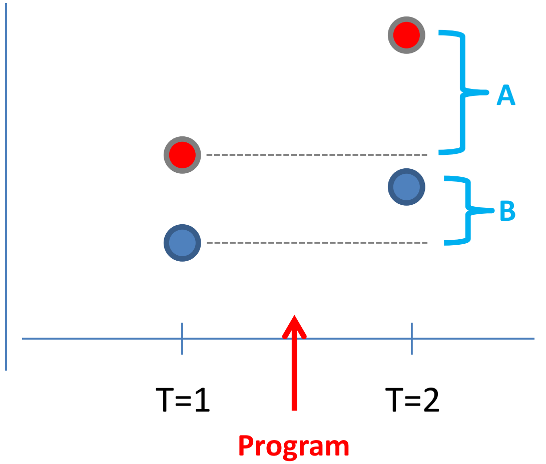
Competing Hypothesis #6

Testing

When the same group is exposed repeatedly to the same set of questions or tasks they can improve independent of any training.

The Fix:

This problem only applies to a small set of programs. Change tests, use post-test only designs, or use a control group that receives the test.



Pre-Post
With Control

Effect:
A-B

Competing Hypothesis #7

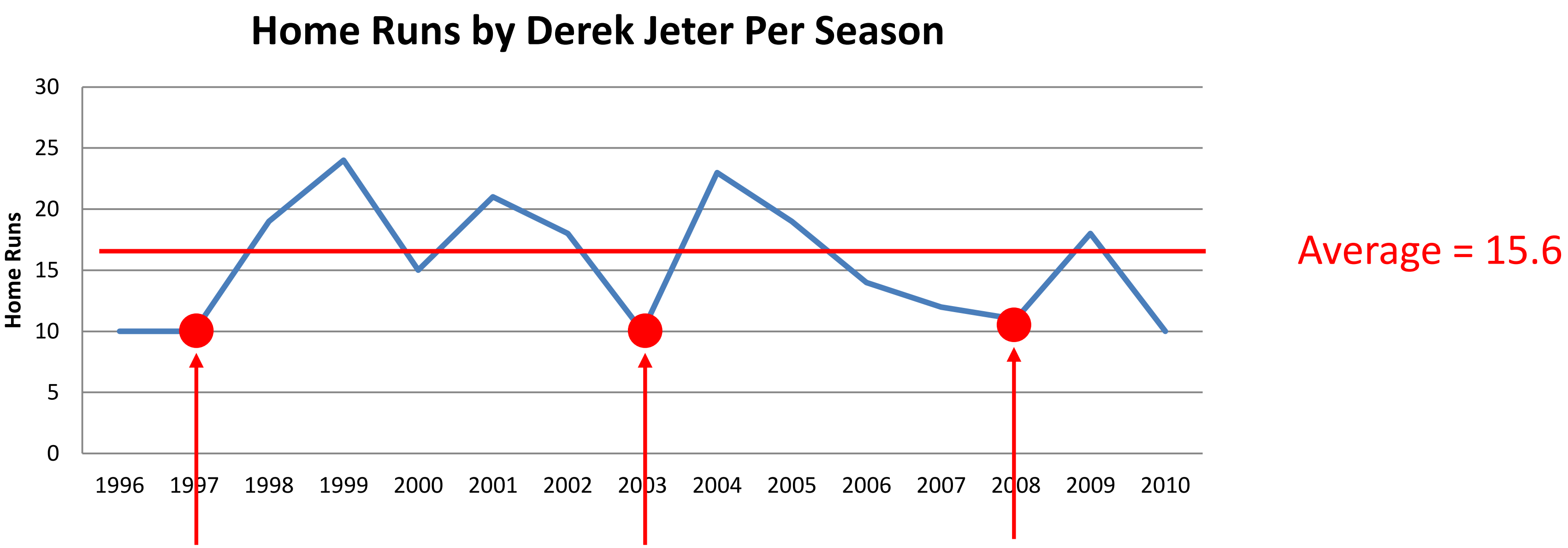
Regression to the Mean

Every time period that you observe an outcome, during the next time period the outcome naturally has a higher probability of being closer to the mean than it does of staying the same or being more extreme. As a result, quality improvement programs for low-performing units often have a built-in improvement bias regardless of program effects.

The Fix:

Take care not to select a study group from the top or bottom of the distribution in a single time period (only high or low performers).

Regression to the Mean Example



Only sent to batting coach when having a slump. Which direction does the trend go after a slump? Is it because of the batting coach? (NO – reg to mean)

Surgery Is One Hell Of A Placebo

Weirdly enough, surgery's invasiveness may explain some of its potency. Studies have shown that invasive procedures produce a stronger placebo effect than non-invasive ones, said researcher Jonas Bloch Thorlund of the University of Southern Denmark. A pill can provoke a placebo effect, but an injection produces an even stronger one. Cutting into someone appears to be more powerful still.

Even without a robust placebo effect, an ineffective surgery may *seem* helpful. Chronic pain often peaks and wanes, which means that if a patient sought treatment when the pain was at its worst, the improvement of symptoms after surgery could be the result of a condition's natural course, rather than the treatment. That softening of symptoms from an extreme measure of pain is an example of the statistical concept of regression to the mean.

https://fivethirtyeight.com/features/surgery-is-one-hell-of-a-placebo

Competing Hypothesis #8

Measurement Error

If there is significant measurement error in the dependent variables, it will bias the effects towards zero and make programs look less effective.

The Fix:

Use better measures of dependent variables.

Competing Hypothesis #9

Study Time-Frame

If the study is not long enough it make look like the program had no impact when in fact it did. If the study is too long then attrition becomes a problem.

The Fix:

Use prior knowledge or research from the study domain to pick an appropriate study period.

Examples:

- Michigan Affirmative Action Study
- Iowa liquor law change

Time frame: one-year post policy change: conclusion is POLICY CHANGE BAD

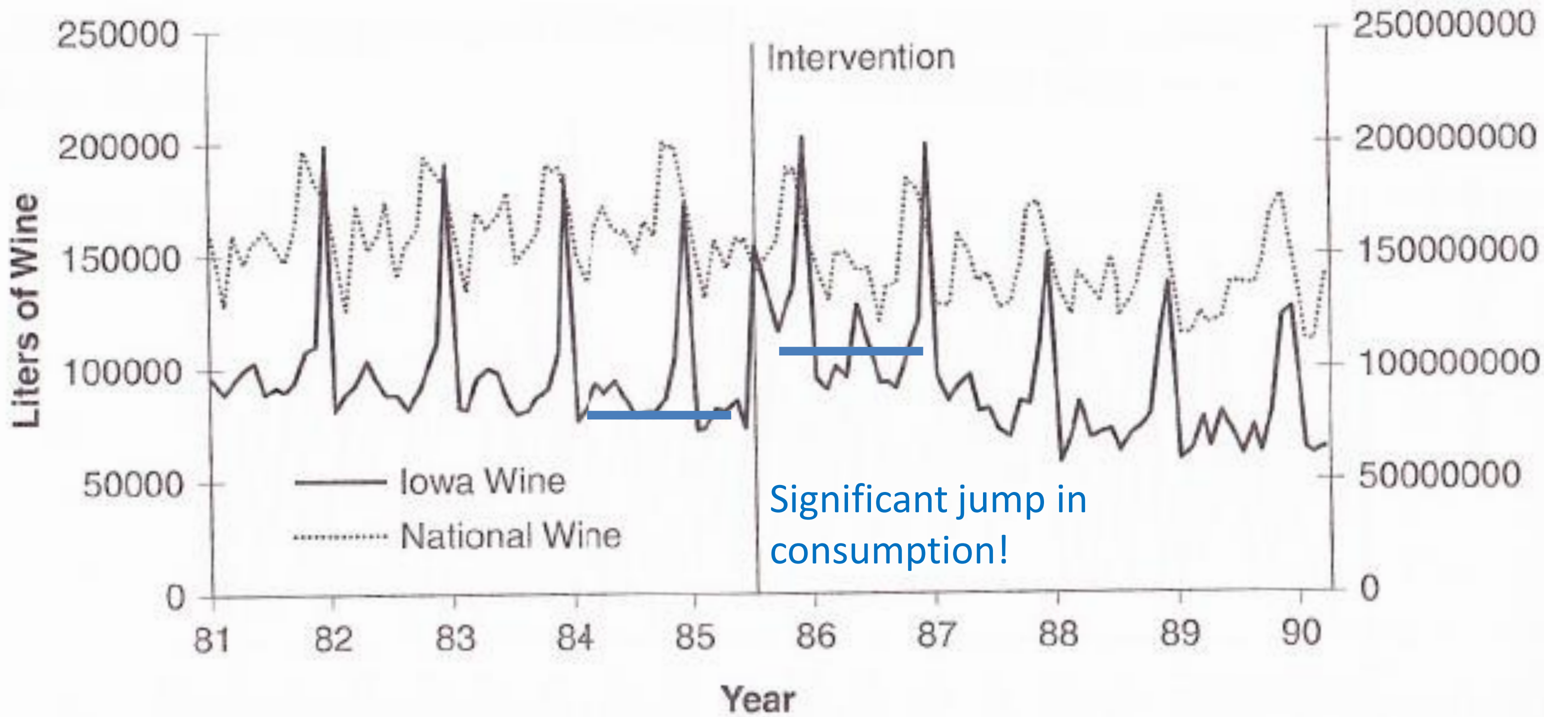


FIGURE 6.5 The effects of legislation in Iowa allowing private sector liquor stores on wine sales, using national data as a control

From "Alcohol Availability and Consumption: Iowa Sales Data Revisited," by H. A. Mulford, J. Ledolter, and J. L. Fitzgerald, 1992, *Journal of Studies on Alcohol, 53*, pp. 487–494.

Evaluation two or more years after policy change: conclusion POLICY GOOD

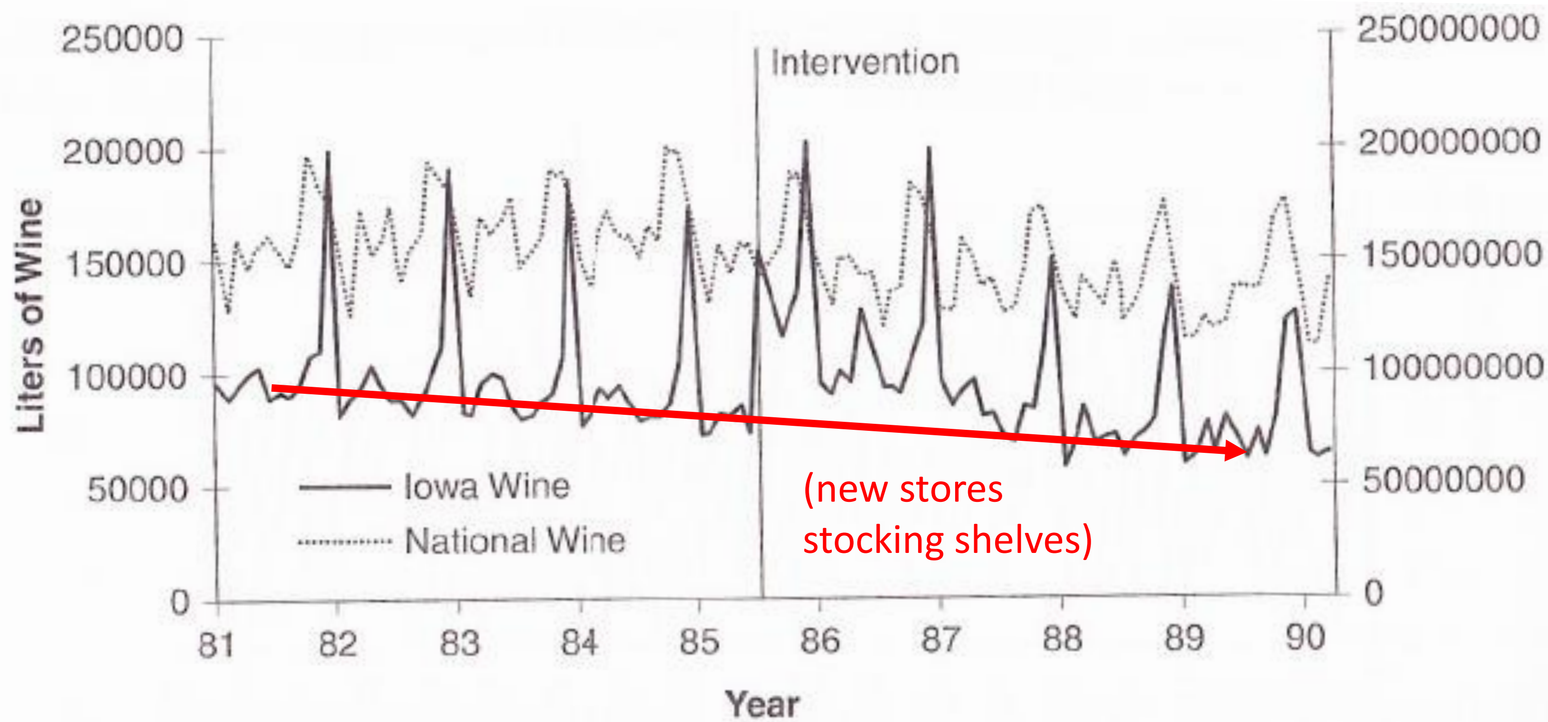


FIGURE 6.5 The effects of legislation in Iowa allowing private sector liquor stores on wine sales, using national data as a control

From "Alcohol Availability and Consumption: Iowa Sales Data Revisited," by H. A. Mulford, J. Ledolter, and J. L. Fitzgerald, 1992, *Journal of Studies on Alcohol, 53*, pp. 487–494. Copyright 1992 by Alcohol Research Documentation, Inc., Rutgers Center of Alcohol Studies, Piscataway NJ 08855.

Note "consumption" is measured as wholesale volume, not consumer consumption, so there is a serious measurement problem as well!

Competing Hypothesis #10

Intervening Events

Has something happened during the study that affects one of the groups (treatment or control) but not the other?

Example, treatment group school burns down. Prices change for substitute goods for control group.

The Fix:

If there is an intervening event, it may be hard to remove the effects from the study.