Econ 211

Prof. Jeffrey Naecker

Wesleyan University

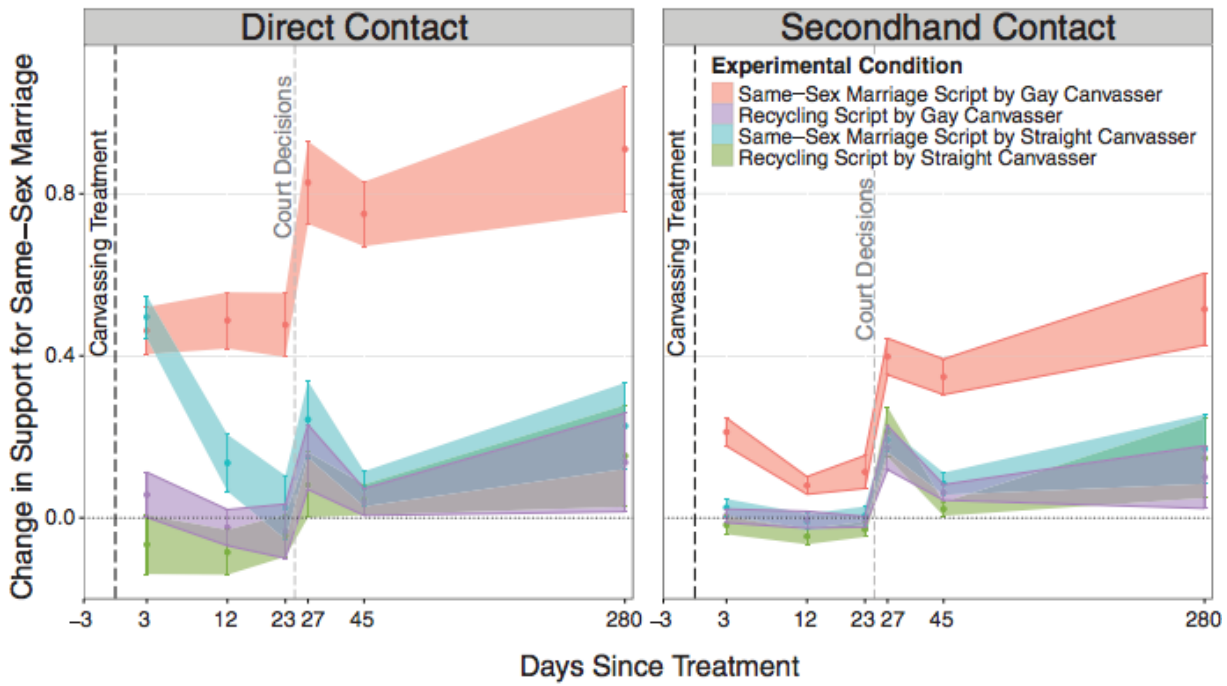
Research Transparency

LaCour and Green (2014)

- We saw that the "echo chamber effect" can make it difficult for people's opinions to change?
- But forcing "cross-cutting" interactions might sway opinions
- La Cour and Green (2014) report an experiment attempting to change opinions on gay rights via canvasing
  - Initial baseline survey of opinions of voters in Los Angeles
  - Send either gay or straight canvaser to discuss gay rights with each voter for 22 minutes on average
  - Measure opinions on gay rights again with delay of 3 weeks, 5 weeks, and 9 months
  - Also measure opinions of people in the same household who did not talk directly to canvaser
  - Outcome: response on scale of 1-100, where 1=very cold and 100=very warm to idea of gay rights (thermometer scale)

Reported Results

- Both gay and straight canvassers were able to increase support for same-sex marriage
- Effect from gay canvassers persisted (or even increased) over time
- Gay canvassers also had an effect on other members in household



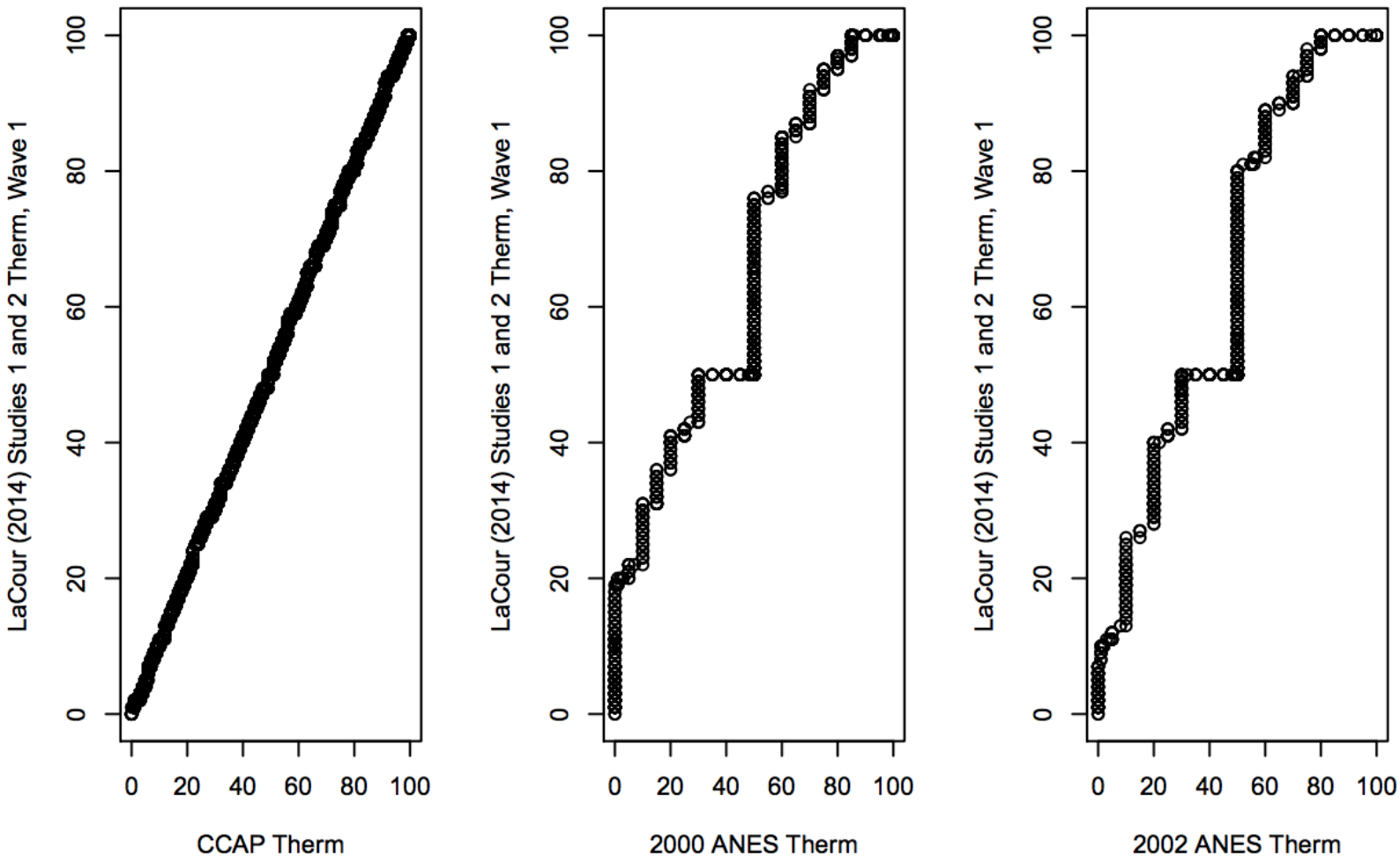
Just One Problem

- All the results reported by LaCour and Green (2014) were fabricated
- The deception appears to have been perpetrated entirely by LaCour (a graduate student at the time)
  - Canvassing was actually carried out as described by a non-profit (at great expense of time and money)
  - However, pre- and post-canvasing responses (allegedly collected via online surveys send to the canvassed households) were entirely made up by LaCour
  - LaCour even fabricated the research grants that he supposedly used to fund the surveys

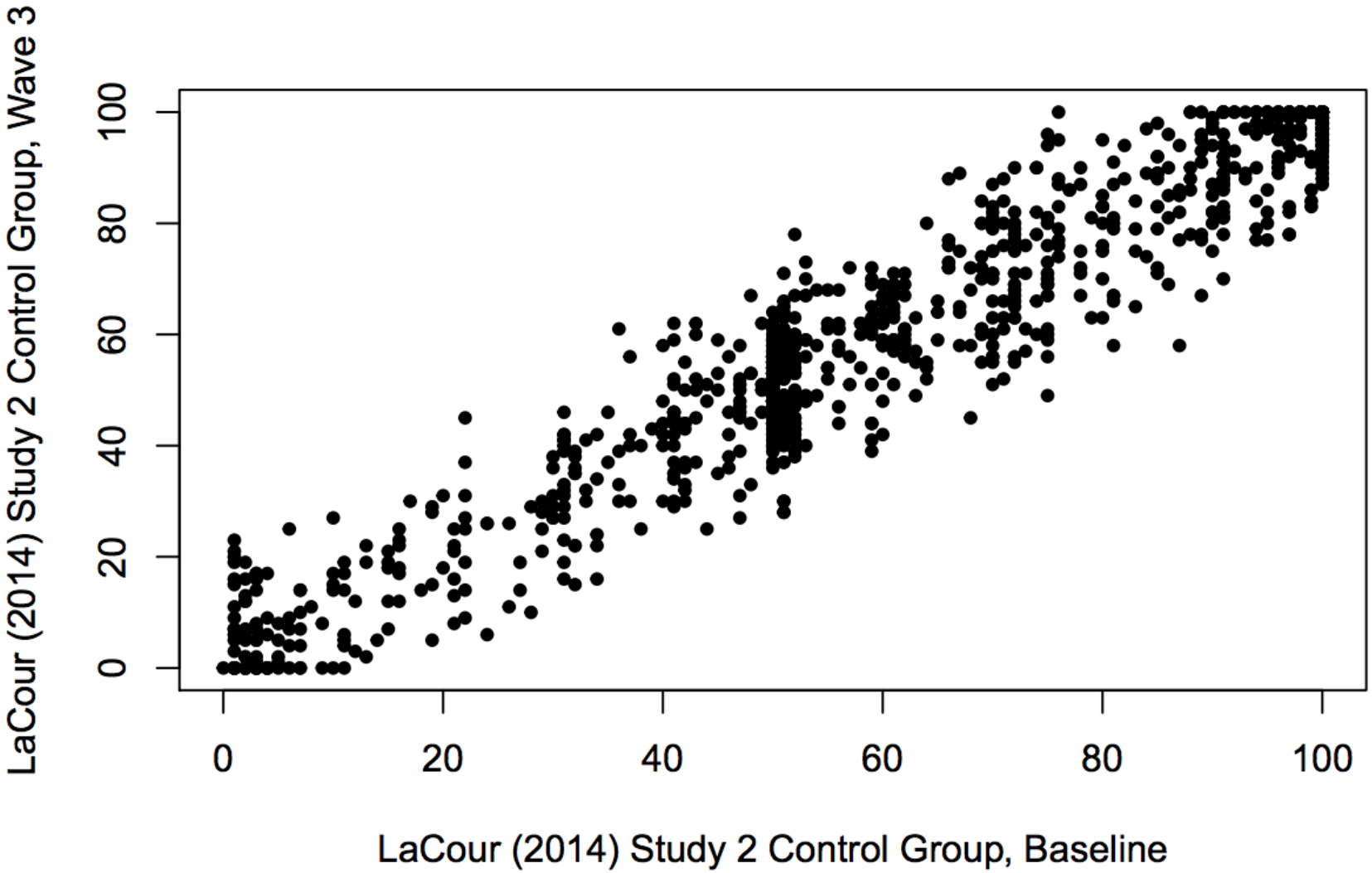
How Was This Discovered?

- Two researchers, Josh Kalla and David Broockman, attempted to replicated LaCour and Green's methods, but with the goal of reducing transphobia
- However, did not get responses rates to follow-up surveys that were similar to LaCour
- Suspicious, they investigated individual response data from LaCour (which was published along with paper)
- They found several suspicious trends in data:
  - Initial survey responses were remarkably similar to responses from another well-known paper that used same thermometer scale
  - Follow-up responses were much more highly correlated with initial responses than usually seen in literature
  - Follow-up responses seemed to be created by taking initial responses and adding positive random numbers

LaCour Data Nearly Identical to Other Study



LaCour Baseline vs Follow-up



This Has Happened Before

- This is not the only time such fabrication has happened, unfortunately
  - One social psychology researcher in the Netherlands believed to have fabricated data in over 50 published papers
  - Not just social science: A Japanese anesthesiologist believed to have fabricated data in at least 172 papers
  - Hundreds of examples across all major research fields

Reinhart and Rogoff

- Reinhart and Rogoff (2009) reported that countries with debt above 90% of GDP have lower growth
- Paper was influential for policy during great recession and financial crisis
  - Used to justify austerity measures in Europe, for example
- However, other economists could not replicate results
- Turns out Reinhart and Rogoff used Excel for data analysis, and didn't select the right cells of the spreadsheet when crunching their numbers
- After correcting their errors, no apparent threshold at 90%

Research Integrity More Broadly

- The above are extreme and (hopefully) rare example
- However, even seemingly benign choices by researcher can call results into question
  - Choice of which data to use: throw out outliers, focus on subsample analysis, pilot several designs of experiment
  - Choice of which regressions to run and which variables to include
  - Choice of which statistical tests to use
- These issues put under the general umbrella of *p-hacking*
- Also called *research degrees of freedom* or *the garden of forking paths*

Motivating Example

- Suppose you are running a simple experiment
  - Randomly assign people to either hot or cold room
  - Ask whether they would like \$10 now (impatient) or \$11 tomorrow (patient)
- Suppose your sample size is $N = 2$ individuals, one to each treatment
- Suppose you find that the person in the hot room takes the patient option and the person in the cold room takes the impatient option
- Can you conclude that warmer rooms cause people to act more patient?
  - No; even if temperature has no effect on patience, there is a 50% chance of getting the result we did
  - This is because there is 50% chance that we just happened to select the more patient person for the hot treatment
  - Thus in this example, the p-value is 0.5

Hypothesis Testing

- More generally, are testing whether we can accept or reject a certain hypothesis
- Typically, the *null hypothesis* predicts that there will be no difference between our treatments, while the *alternate hypothesis* predicts there will be a difference
- In temperature example:
  - Null hypothesis: temperature has no effect on patience
  - Alternate hypothesis: temperature causes people to act more patient

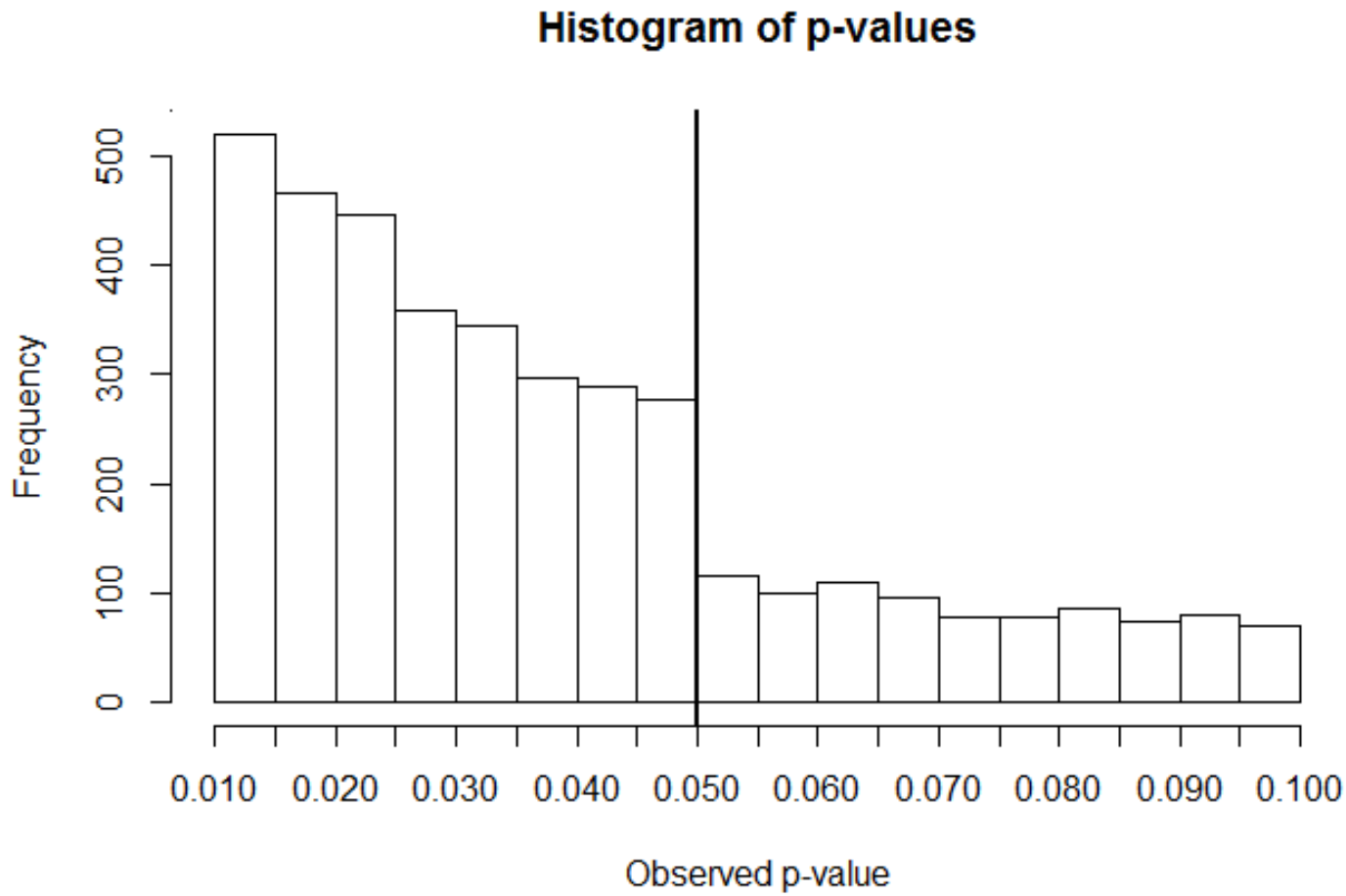
p-values

- The *p*-value measures the probability of getting the observed result *under the null hypothesis*
  - A *p*-value close to 0 means that there is only a small likelihood that results are due to chance
  - A *p*-value close to 1 means that there is a high likelihood that results are due to chance
- For historical and largely arbitrary reasons, a *p*-value of 0.05 or less is considered "statistically significant"
- If we look at *p*-values across an entire field, distribution should be smooth

Research Degrees of Freedom

- Consider all the choices we made when running the temperature experiment:
  - What temperature to make the rooms
  - What size prizes to use
- And choices made when analyzing the data:
  - Throw out responses from that one subject that fell asleep
  - Maybe we should control for gender, or GPA, or income, or ...
- If we make these choices in an attempt to get $p = 0.05$ (even subconsciously), then these are all ways of p-hacking

Visualization of p-hacking



Data: 3627 p-values reported in 3 different psychology journals, from Masicampo and LaLande (2012)

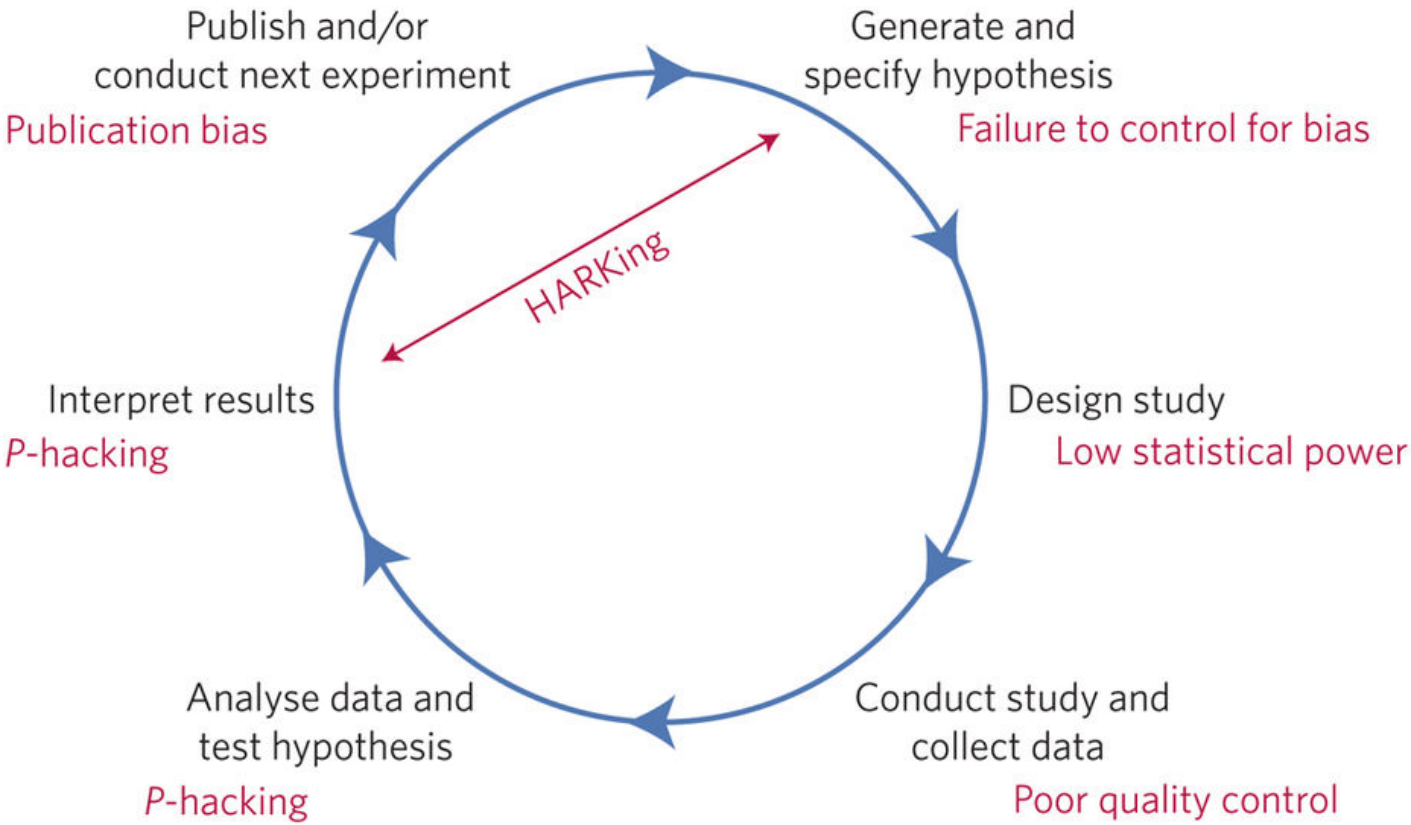
Returning to Example

- Now suppose sample size was $N = 100$, with 50 people in each treatment
- Suppose you find that all 50 people in the hot room take the patient option and all 50 people in the cold room take the impatient option
- Now can you conclude that temperature has an effect on patience?
  - Almost certainly yes: getting this result by chance if the null was true is extremely unlikely
  - If we assume that people are equally likely to be patient or impatient under null, then getting this result is like flipping 50 heads in a row on a fair coin
  - Thus the p-value is essentially 0

Other Transparency Issues

- Publication bias
  - Journals have a tendency to favor studies with statistically significant results
  - This leads to *publication bias*: significant results are published more quickly and in higher-status journals
  - Also causes *file drawer effect*: researchers don't even try to publish null (non-significant) results
- HARKing (Hypothesizing After the Results are Known)
  - Ideally, hypothesis should be generated *before* experiment is run or data are analyzed
  - However, researchers often generate hypothesis/theory *after* data are analyzed to make it seem like they predicted the results all along

Overview



Source: Munafo et al (2017)

Reproducibility vs Replication

- A study is *reproducible* if the exact same results can be re-generated using the exact same data set and (intended) methods
  - LaCour fabrication is not reproducible since data don't exist
  - Reinhart and Rogoff also not reproducible since methods not executed as intended
- A study is *replicable* if the results can be re-generated using similar data and methods
  - Replications attempt to verify the underlying theory and/or methods
  - Studies that are fully reproducible may still not replicate
  - In recent replication projects, only about 40% of pysch studies and about 60% of econ studies replicated

What To Do?

1. Make the research process transparent and reproducible
   - Make all researcher publish raw data and code
   - Issue: what about proprietary/sensitive data?
2. Encourage replication
   - Don't put too much credence in results until they have been replicated independently
   - Issue: how to incentivize more replications?
3. Encourage pre-analysis plans (also know as pre-registration)
   - Force researchers to register experimental designs and analysis plans (eg which regressions to run) before running experiment
   - Would alleviate p-hacking and file-drawer effect

Coda

- Recall that Broockman and Kalla were attempting to replicate LaCour and Green's canvassing methods to reduce transphobia
- Their paper was recently published in Science (same journal that publish now-discredited LaCour and Green paper)
- Data: 1825 voters in Florida
- What they found:
  - Both transgender and non-transgender canvassers effective at changing opinions
  - These changes lasted at least 3 months
  - Key seems to be forcing respondents to do "perspective-taking" rather than logical or legal arguments