Survey Research and Design

Survey Experiments

William Marble

October 5, 2023

Causation and Correlation

- Surveys are good at telling us correlations between variables: e.g., college-educated people increasing vote for Democrats; people who say crime is important tend to vote for Republicans
- Does this mean that college *causes* people to vote for Democrats? Does it mean concern about crime *causes* people to vote for Republicans?

Causation and Correlation

- Surveys are good at telling us correlations between variables: e.g., college-educated people increasing vote for Democrats; people who say crime is important tend to vote for Republicans
- Does this mean that college *causes* people to vote for Democrats? Does it mean concern about crime *causes* people to vote for Republicans?
- Maybe people's stated concerns are driven by who they support, not vice versa! Two models:
  - I think crime is a big problem $\rightsquigarrow$ Republicans are focusing on crime $\rightsquigarrow$ I vote for Republicans
  - I'm a Republican $\rightsquigarrow$ Republican politicians say crime is a big problem $\rightsquigarrow$ I say crime is a big problem
- The correlation between voting and issue attitudes isn't enough to distinguish between these two explanations

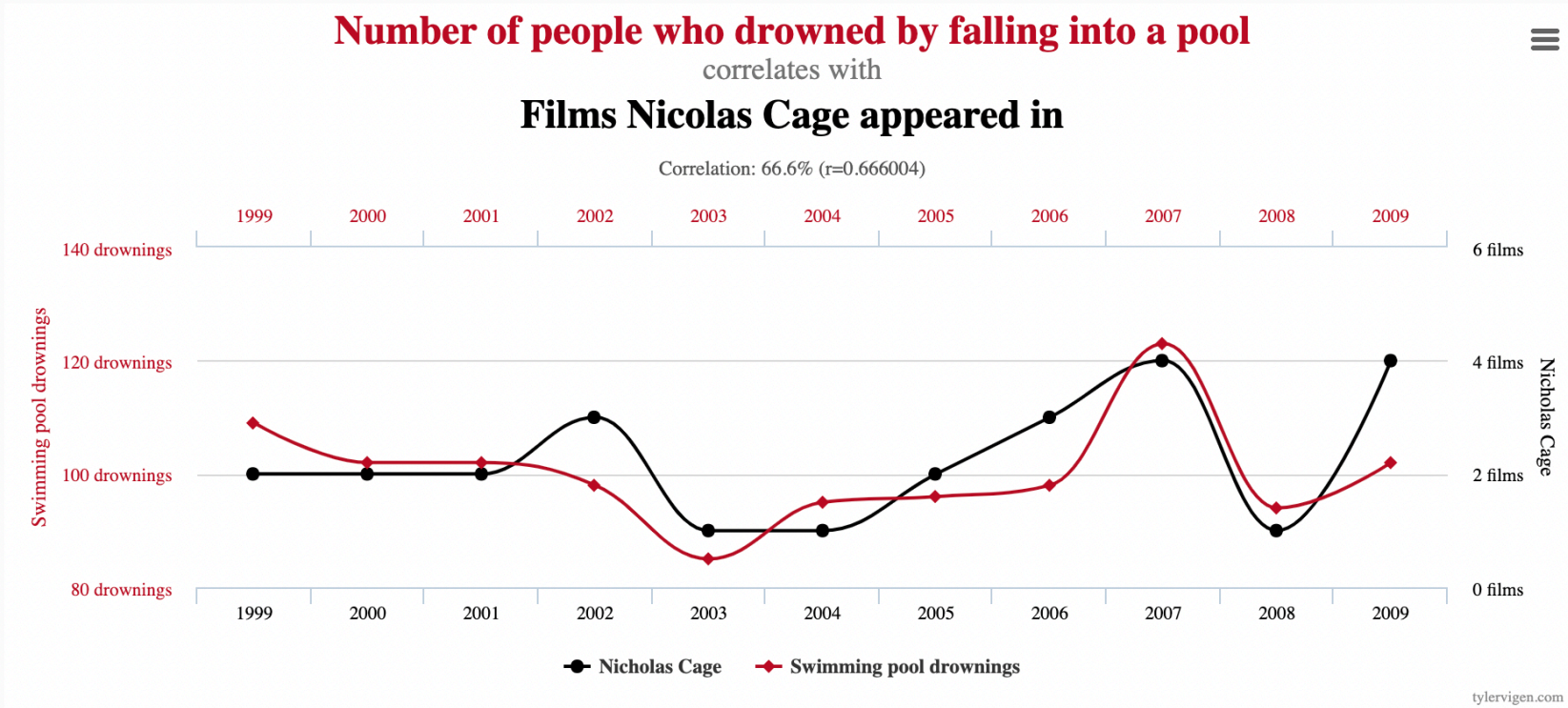
Crash Course on Causal Inference and Experiments

The Key Problem

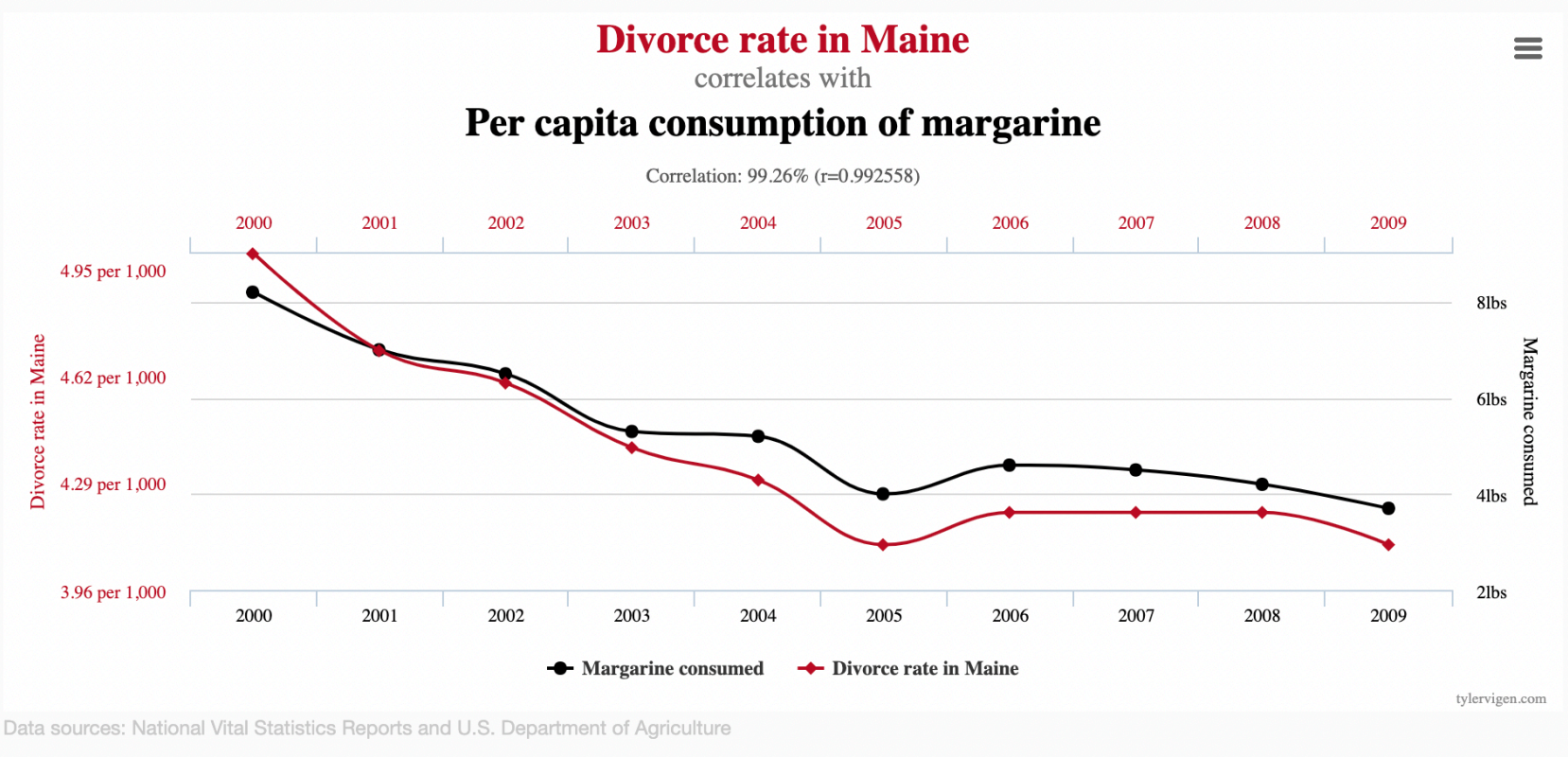
- The old mantra: correlation does not imply causation
- Just because two attitudes "go together" doesn't mean one *causes* the other

The Key Problem

- The old mantra: correlation does not imply causation
- Just because two attitudes "go together" doesn't mean one *causes* the other
- Correlations between X and Y could be caused by:
  - X affecting Y (causal relationship)
  - Y affecting X (causal relationship)
  - Z affecting both X and Y (spurious correlation, non-causal)
  - Random chance (spurious correlation, non-causal)



Data sources: Centers for Disease Control & Prevention and Internet Movie Database



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Causation as Counterfactual

Causation is about **counterfactuals**: if X had been different than it actually is — holding all else equal — how would Y be different?

- If Jane Smith had been more concerned about crime, would she have voted for a Republican?
- If inflation were lower, would Biden's approval rate be higher?
- Do natives prefer immigrants with more education, holding other characteristics equal?
- If fewer people got divorced in Maine, would margarine consumption be lower?

Causation as Counterfactual

Causation is about **counterfactuals**: if X had been different than it actually is — holding all else equal — how would Y be different?

- If Jane Smith had been more concerned about crime, would she have voted for a Republican?
- If inflation were lower, would Biden's approval rate be higher?
- Do natives prefer immigrants with more education, holding other characteristics equal?
- If fewer people got divorced in Maine, would margarine consumption be lower?

Correlations cannot give insight into these counterfactual "ifs." Need another method to investigate these questions.

Potential Outcomes

Potential outcomes framework for studying causality is ubiquitous in social science and medicine

We are interested in some outcome Y which is a function of a **treatment** D

Potential Outcomes

Potential outcomes framework for studying causality is ubiquitous in social science and medicine

We are interested in some outcome Y which is a function of a **treatment** D

Treatment:

- a medicine, information, an incentive, etc. etc.
- refer to anything we want to know the causal effect of as a "treatment"
- If a unit receives treatment then $D_i = 1$; if they are in the *control group* then $D_i = 0$

Outcome:

- health, longevity, opinions, vote choice, etc.

Potential Outcomes

Every unit has two **potential outcomes** — one under treatment and one under control: $Y(D = 1)$ and $Y(D = 0)$

Causal effect is the difference in outcome under each treatment condition: $\tau_i = Y_i(1) - Y_i(0)$

We can't ever observe both $Y_i(1)$ and $Y_i(0)$ for a single unit: "fundamental problem of causal inference"

Potential Outcomes

Every unit has two **potential outcomes** — one under treatment and one under control: $Y(D = 1)$ and $Y(D = 0)$

Causal effect is the difference in outcome under each treatment condition: $\tau_i = Y_i(1) - Y_i(0)$

We can't ever observe both $Y_i(1)$ and $Y_i(0)$ for a single unit: "fundamental problem of causal inference"

So we're usually interested in the **average treatment effect**: $E[\tau_i] = E[Y_i(1) - Y_i(0)]$.

Potential Outcomes

Every unit has two **potential outcomes** — one under treatment and one under control: $Y(D = 1)$ and $Y(D = 0)$

Causal effect is the difference in outcome under each treatment condition: $\tau_i = Y_i(1) - Y_i(0)$

We can't ever observe both $Y_i(1)$ and $Y_i(0)$ for a single unit: "fundamental problem of causal inference"

So we're usually interested in the **average treatment effect**: $E[\tau_i] = E[Y_i(1) - Y_i(0)]$.

For a sample,

$$E[\tau_i] = \frac{1}{n}\sum_{i=1}^{n} Y_i(1) - Y_i(0) = \frac{1}{n}\sum Y_i(1) - \frac{1}{n}\sum Y_i(0)$$

Experiments

To estimate the average treatment effect (ATE), we need a good estimate of the average outcome under treatment $E[Y(1)]$ and the average outcome under control $E[Y(0)]$.

The central insight of experiments is **random assignment**:

- The researcher randomly chooses some units to assign to "treatment" and other units to assign to "control"
- Because assignment is random, "treatment" and "control" units should be exactly the same, on average, on all other characteristics
- Any differences in outcomes between groups can therefore be attributed to the treatment

Under random assignment, mean of the observed treated units is an unbiased estimator for $E[Y(1)]$, and likewise mean of observed control units are unbiased for $E[Y(0)]$

Analogy to Survey Sampling?

Is there an analogy to survey sampling?

Analogy to Survey Sampling?

Is there an analogy to survey sampling?

True random sampling: sample can stand in for the population

Random assignment: sample of potential outcomes can stand in for the population of potential outcomes

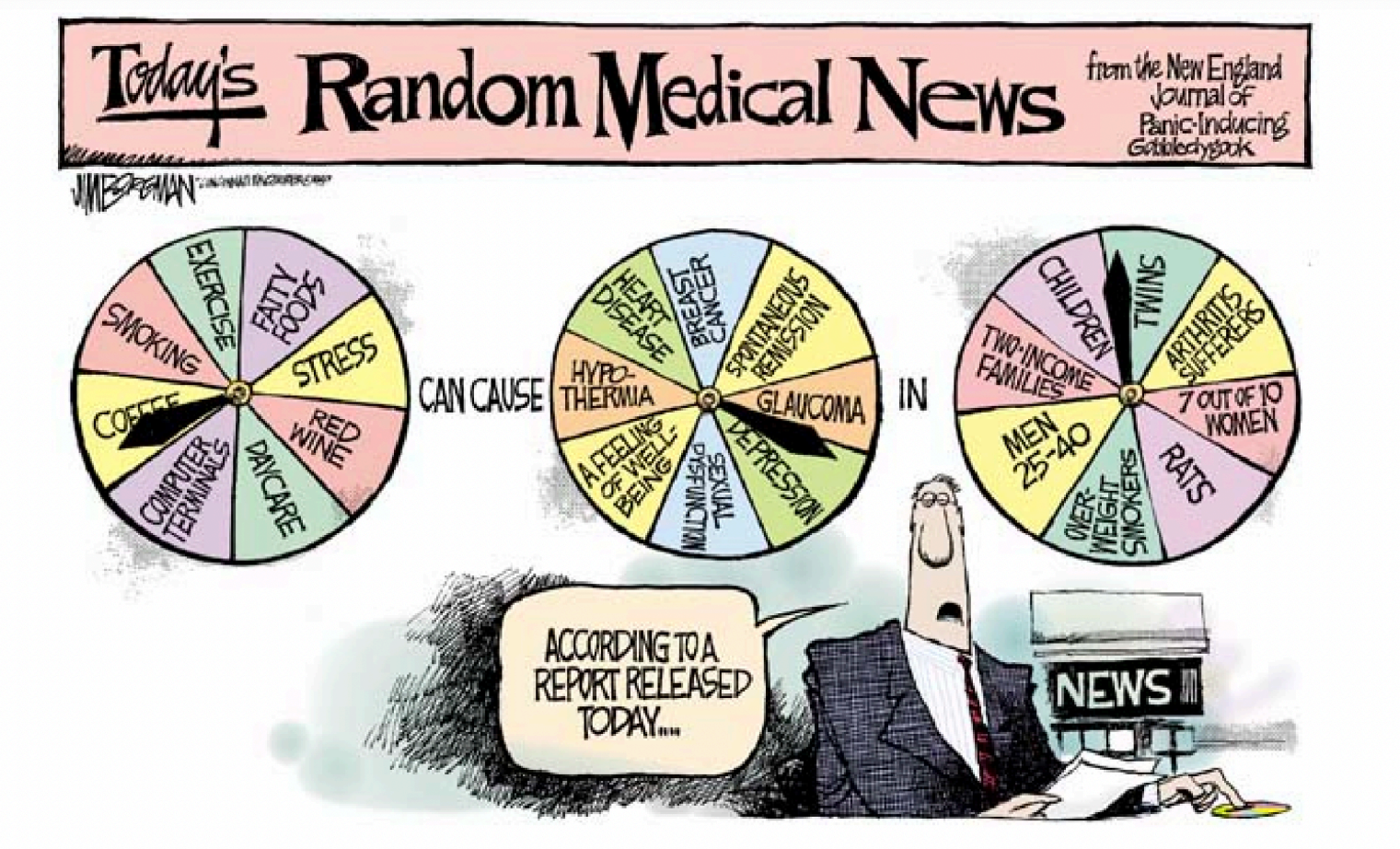
Analogy to Survey Sampling?

Is there an analogy to survey sampling?

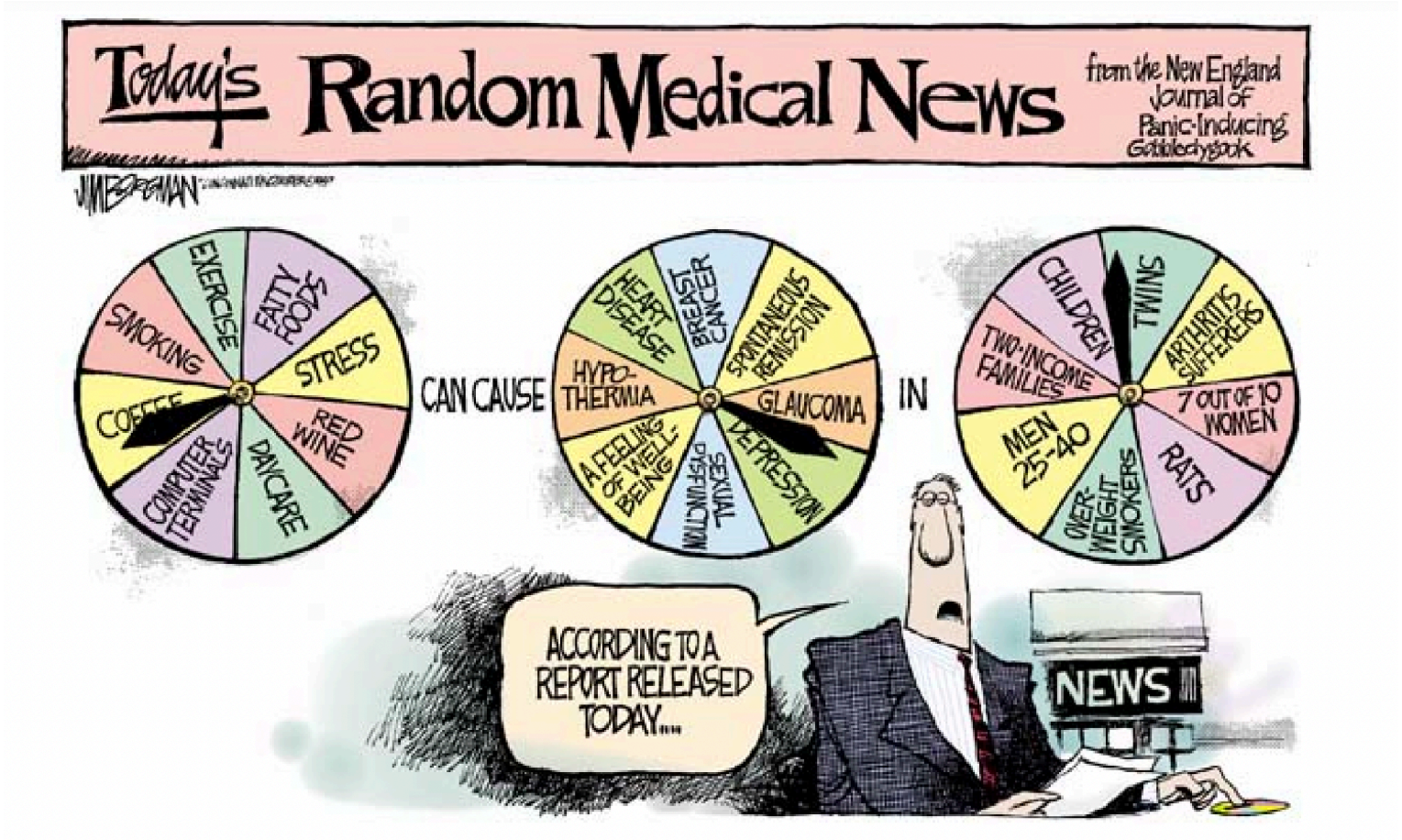
True random sampling: sample can stand in for the population

Random assignment: sample of potential outcomes can stand in for the population of potential outcomes

Analogy to models of nonresponse?



Lancet 2001: negative correlation between coronary heart disease mortality and level of vitamin C in bloodstream (controlling for age, gender, blood pressure, diabetes, and smoking)



Lancet 2002: no effect of vitamin C on mortality in controlled placebo trial (controlling for nothing)

Survey Experiments

- Survey experiments embed an experiment inside a survey
- The "treatment" could be:
  - The wording of a question
  - Information given to respondents
  - Video, pictures, or text shown to respondents
  - The answer choices
  - Etc. etc.
- Outcomes measured *after* the treatment is administered

Vignette Experiments

Vignette Experiments

- Have respondents read a "vignette" — a paragraph or so — where parts of the text is randomly varied
- Manipulation of text elements gives different information to different survey respondents
- Differences in outcomes across groups can be attributed to the experimental manipulation (aka the treatment)

Vignette Experiment: Democratic Peace

- Tomz and Weeks (2013) study the "democratic peace": Why do democracies rarely fight each other?
- One possibility: democracy gives leaders an incentive to avoid wars because they're unpopular
- Another possibility: public is especially averse to war against *other democracies*
- To investigate, researchers present survey respondents with hypothetical scenarios that might prompt use of force, then measure public support for using force in each scenario
- They vary the scenarios to investigate whether respondents are more supportive of force against non-democracies

Democratic Peace: Introduction

There is much concern these days about the spread of nuclear weapons. We are going to describe a situation the United States could face in the future. For scientific validity the situation is general, and is not about a specific country in the news today. Some parts of the description may strike you as important; other parts may seem unimportant. Please read the details very carefully. After describing the situation, we will ask your opinion about a policy option.

Democratic Peace: Experimental Treatment

Here is the situation:

- *A country is developing nuclear weapons and will have its first nuclear bomb within six months. The country could then use its missiles to launch nuclear attacks against any country in the world.*
- *The country [has / has not] signed a military alliance with the United States.*
- *The country [has / does not have] high levels of trade with the United States.*
- *The country [is a democracy, and shows every sign that it will remain a democracy / is not a democracy, and shows no sign of becoming a democracy].*
- *The country's nonnuclear military forces are half as strong as U.S. nonnuclear forces.*
- *The country's motives remain unclear, but if it builds nuclear weapons, it will have the power to blackmail or destroy other countries.*
- *The country has refused all requests to stop its nuclear weapons program.*

By attacking the country's nuclear development sites now, the United States could prevent the country from making any nuclear weapons. Would you favor or oppose using the U.S. military to attack the country's nuclear development sites? (Favor strongly, favor somewhat, neither favor nor oppose, oppose somewhat, oppose strongly)

Democratic Peace Experiment: Logic

- Are people less supportive of using force against a democracy than an autocracy?
- Authors want to answer this question *in general* so they present a hypothetical scenario
- The scenario is exactly the same for all respondents, **except** for whether the country is a democracy/non-democracy, whether it has an alliance with the U.S., and whether it has a high level of trade with the U.S.
- Differences in support for use of force across conditions $\rightsquigarrow$ the **causal effect** of the condition
- Each bullet point (democracy, trade, and alliances) was manipulated independently: thus there were $2 \times 2 \times 2 = 8$ total treatments

Democratic Peace Experiment: Results

TABLE 1. The Effect of Democracy on Willingness to Strike

	United Kingdom (between)	United States (between)	United States (within)
Not a democracy	34.2	53.3	50.0
Democracy	20.9	41.9	38.5
Effect of democracy	−13.3	−11.4	−11.5
95% C.I.	(−19.6 to −6.9)	(−17.0 to −5.9)	(−14.7 to −8.3)

Note: The table gives the percentage of respondents who supported military strikes when the target was a democracy and when it was not. The difference is the estimated effect of democracy. In the United Kingdom, we obtained between-subject estimates by comparing 364 cases in which the target was a democracy, versus 398 cases in which it was not a democracy. In the United States, we obtained between-subject estimates by comparing 639 cases in which the target was a democracy, versus 634 in which it was not. The United States within-subject estimates were based on 972 respondents, each of whom assessed two scenarios, one in which the target was a democracy and another in which the target was not a democracy. 95% confidence intervals appear in parentheses.

Further Explaining the Effect

Do perceptions about likely outcomes differ according to whether the target is a democracy or not?

TABLE 3. The Effect of Democracy on Perceptions of Threat

If the U.S. did not attack, the country would . . .	Belief if Autocracy	Effect of Democracy
Build nuclear weapons	75	−3*
Threaten to use nukes vs. another country	52	−14*
Threaten to use nukes vs. U.S. or U.S. ally	45	−11*
Launch a nuclear attack vs. another country	34	−8*
Launch a nuclear attack vs. U.S. or U.S. ally	30	−6*
Average	47	−9*

Note: The first column gives the percentage of respondents who thought the event had more than a 50% chance of happening when the country was an autocracy. The second column shows how much that percentage changed when the same respondents considered an identical scenario involving a democracy. Each row is based on 972 respondents. Asterisks denote effects that were significant at the .05 level.

TABLE 4. The Effect of Democracy on Perceptions of Cost, Success, and Morality

If the U.S. did attack . . .	Belief if Autocracy	Effect of Democracy
Cost		
The country would attack U.S. or U.S. ally	39	0
The U.S. military would suffer many casualties	32	1
The U.S. economy would suffer	31	0
U.S. relations with other countries would suffer	49	4*
Average	38	1
Success		
It would prevent nukes in the near future	66	−5*
It would prevent nukes in the long run	30	−5*
Average	48	−5*
Morality		
It would be immoral	31	7*

Note: For our measures of cost and success, the first column gives the percentage of respondents who thought the event had more than a 50% chance of happening when the country was an autocracy. The second column shows how much that percentage changed when the same respondents considered an identical scenario involving a democracy. Each row is based on 972 respondents. Asterisks denote effects that were significant at the .05 level.

Democratic Peace Experiment: Takeaways

- Less public support for use of force against democracies
- Reasons:
  - More likely to think autocracies will threaten and/or attack other countries
  - Slightly less likely to think attack against democracy will be successful
  - More likely to think attacking a democracy is immoral
- Experimental design enables these *causal* conclusions
- Enables measuring opinion in a hypothetical scenario not directly tied to current events

Democratic Peace Experiment: Drawbacks and Critiques

Democratic Peace Experiment: Drawbacks and Critiques

Attitudes might differ in a real-world situation:

- Specific circumstances might overwhelm general "democracy effect"
- Real-world scenario: real costs to military action

Democratic Peace Experiment: Drawbacks and Critiques

Attitudes might differ in a real-world situation:

- Specific circumstances might overwhelm general "democracy effect"
- Real-world scenario: real costs to military action

Respondents might be thinking of a particular scenario

- Maybe attitudes are shaped by particular political scenarios (e.g. Iran or North Korea)
- Researchers say it's hypothetical scenario
- They conduct a replication study specifying the country is in Africa

Democratic Peace Experiment: Drawbacks and Critiques

Attitudes might differ in a real-world situation:

- Specific circumstances might overwhelm general "democracy effect"
- Real-world scenario: real costs to military action

Respondents might be thinking of a particular scenario

- Maybe attitudes are shaped by particular political scenarios (e.g. Iran or North Korea)
- Researchers say it's hypothetical scenario
- They conduct a replication study specifying the country is in Africa

Other limitations

- Views towards use of force may differ in the U.S. and U.K. compared to other countries
- High-stakes nuclear proliferation scenario may differ from lower-stakes scenarios

Question Wording Experiments

Question Wording Experiments

In political communication, terms are chosen carefully:

- "Estate tax" vs. "death tax"
- "Illegal alien" vs. "undocumented immigrant"
- "Global warming" vs. "climate change"

Question Wording Experiments

In political communication, terms are chosen carefully:

- "Estate tax" vs. "death tax"
- "Illegal alien" vs. "undocumented immigrant"
- "Global warming" vs. "climate change"

What are the effects of this language on public attitudes?

Question Wording Experiments

In political communication, terms are chosen carefully:

- "Estate tax" vs. "death tax"
- "Illegal alien" vs. "undocumented immigrant"
- "Global warming" vs. "climate change"

What are the effects of this language on public attitudes?

One way to investigate is to use a **question-wording experiment**:

- Ask two versions of the same question, where the specific wording is assigned randomly
- Differences in responses across groups are due to differences in question wording

Global Warming vs. Climate Change

- Stark divisions in the U.S. over whether human activity is affecting the global climate
- Politicians and interest groups use different language to talk about the issue depending on their position
- "Global warming" focuses attention on *increases* in temperature; "climate change" focuses attention on all types of changes to weather patterns
- "Global warming" used more by climate skeptics, "climate change" used by groups who support more aggressive action
- How do these frames affect (measured) public opinion?

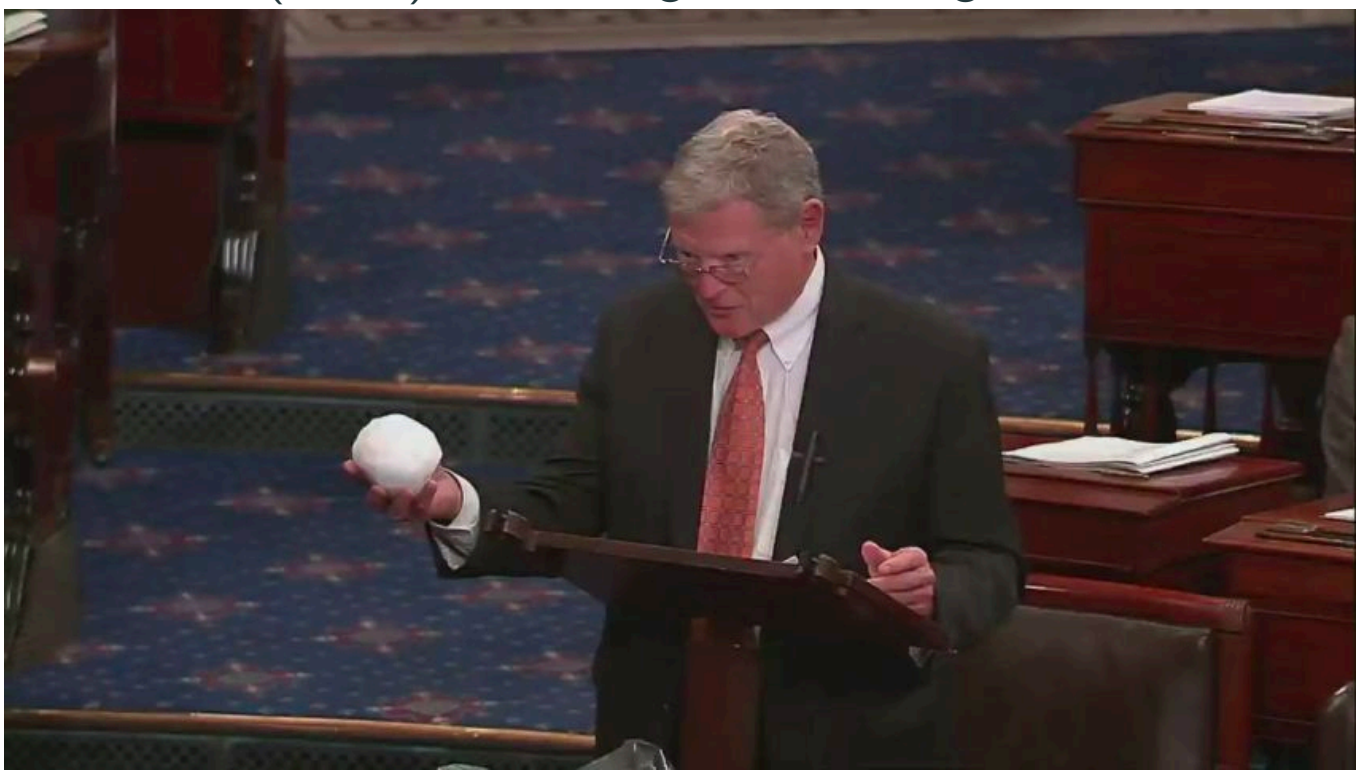
Use of “Global Warming” by Climate Skeptics

Drudge report headline from 2004: “Gore to warn of ‘global warming’ on New York City’s coldest day in decades!”

Use of "Global Warming" by Climate Skeptics

Drudge report headline from 2004: "Gore to warn of 'global warming' on New York City's coldest day in decades!"

Sen. Inhofe (R-OK) debunks "global warming":



Global Warming Experiment: Treatment

Schuldt et al. (2011) use a question-wording experiment to assess the effects of these different frames:

> *You may have heard about the idea that the world's temperature may have been [going up / changing] over the past 100 years, a phenomenon sometimes called ['global warming' / 'climate change']. What is your personal opinion regarding whether or not this has been happening? (Definitely has not been happening; Probably has not been happening; Unsure, but leaning toward it has not been happening; Not sure either way; Unsure, but leaning toward it has been happening; Probably has been happening; Definitely has been happening)*

Global Warming Experiment: Results

Percent saying it has been happening, by question wording:

Group	Global Warming	Climate Change	Difference
All respondents	67.7%	74.0%	6.3
Democrats	86.9	86.4	-0.2
Republicans	44.0	60.2	16.2

Global Warming Experiment: Conclusion

Takeaways

- Republicans more likely to agree that "climate change" is happening than "global warming"
- Rhetorical frames can influence measurement of public opinion

Global Warming Experiment: Conclusion

Takeaways

- Republicans more likely to agree that "climate change" is happening than "global warming"
- Rhetorical frames can influence measurement of public opinion

Critiques/Follow-Up Questions?

Global Warming Experiment: Conclusion

Takeaways

- Republicans more likely to agree that "climate change" is happening than "global warming"
- Rhetorical frames can influence measurement of public opinion

Critiques/Follow-Up Questions?

- Do these differences in belief that the phenomenon is occurring translate to differences in policy preferences?
- Does this question-wording experiment generalize to the way politicians talk about the issue?

Question-Wording Experiments

- Useful for validating questionnaire design: if small changes in question wording affect results, suggests measure may have problems with validity and/or reliability
- Simple way to test different messages, as in global warming experiment
- However, may have limited generalizability due to their simplicity

Conjoint Experiments

Measuring the Effects of Multiple Attributes

- Vignette and question wording experiments allow us to estimate the effect of a small number of attributes: democracy vs. non-democracy or global warming vs. climate change
- Other scenarios have more relevant dimensions:
  - Candidate platforms: taxes, immigration, LGBT+ rights, education policy, etc.
  - Where to live: price of home, urban vs. suburban vs. rural, proximity to family, size of house/apartment, etc.
  - Type of car to buy: size of car, price, MPG, stereo system, safety features
- **Conjoint experiments** are a method to measure the importance of many attributes simultaneously

Conjoint Experiments: Set Up

- Researcher decides which attributes (or *factors*) they want to investigate and comes up with alternatives (*levels*) for each attribute.
- For example, might randomize candidate across several issues:
  - Tax Policy: large tax increase on people making over $400,000; small tax increase on people making over $400,000; no change; small decrease; large decrease
  - Immigration Policy: Grant legal status to all undocumented immigrants; Grant legal status to all illegal immigrants who have paid taxes for at least 3 years; Don't grant legal status but don't deport undocumented immigrants unless they have been convicted of a violent felony; Deport all undocumented immigrants
  - and so on for Education Policy, Abortion Policy, Foreign Policy, etc.

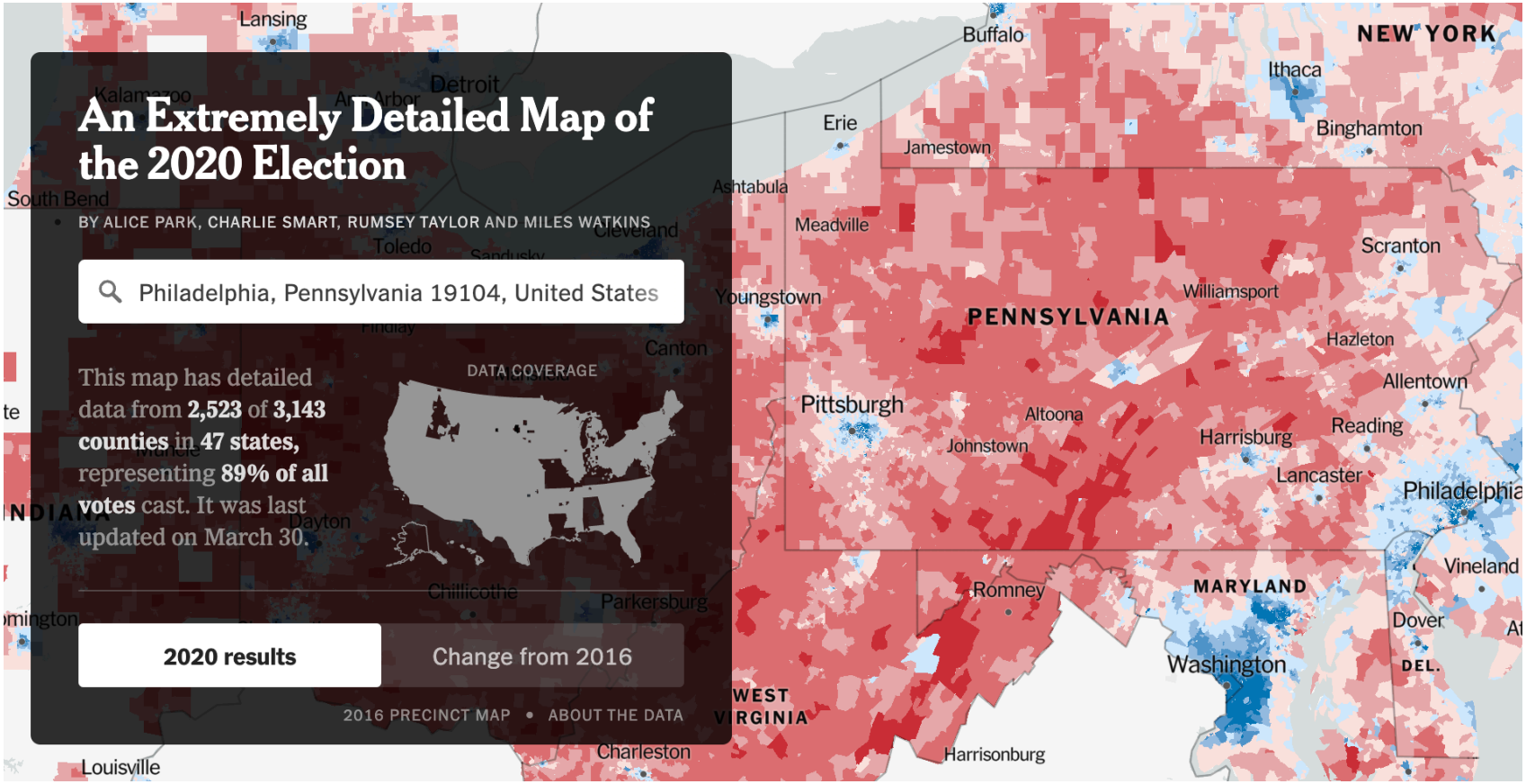
Conjoint Experiments: Set Up

- Researcher decides which attributes (or *factors*) they want to investigate and comes up with alternatives (*levels*) for each attribute.
- For example, might randomize candidate across several issues:
  - Tax Policy: large tax increase on people making over $400,000; small tax increase on people making over $400,000; no change; small decrease; large decrease
  - Immigration Policy: Grant legal status to all undocumented immigrants; Grant legal status to all illegal immigrants who have paid taxes for at least 3 years; Don't grant legal status but don't deport undocumented immigrants unless they have been convicted of a violent felony; Deport all undocumented immigrants
  - and so on for Education Policy, Abortion Policy, Foreign Policy, etc.
- Generate pairs of "profiles" that contain randomly generated levels for each attribute
- Ask respondents to compare two randomly generated profiles and select which one they prefer (**forced choice conjoint**)
- Ask respondents to rate

Conjoint Experiments: Analysis

- Quantity of interest is called the **average marginal component effect**
- Answers the question: how much more likely are respondents to select a profile containing a particular level, relative to a baseline level?
- E.g., how much more likely are people to prefer a policy platform that features a large tax increase on the rich, relative to a policy platform that features no change in taxes on the rich?
- Estimate this by running a regression, where the outcome variable is a $0/1$ for selecting the profile and the independent variables are $0/1$ indicator variables for the levels in that profile

Partisan Residential Preferences



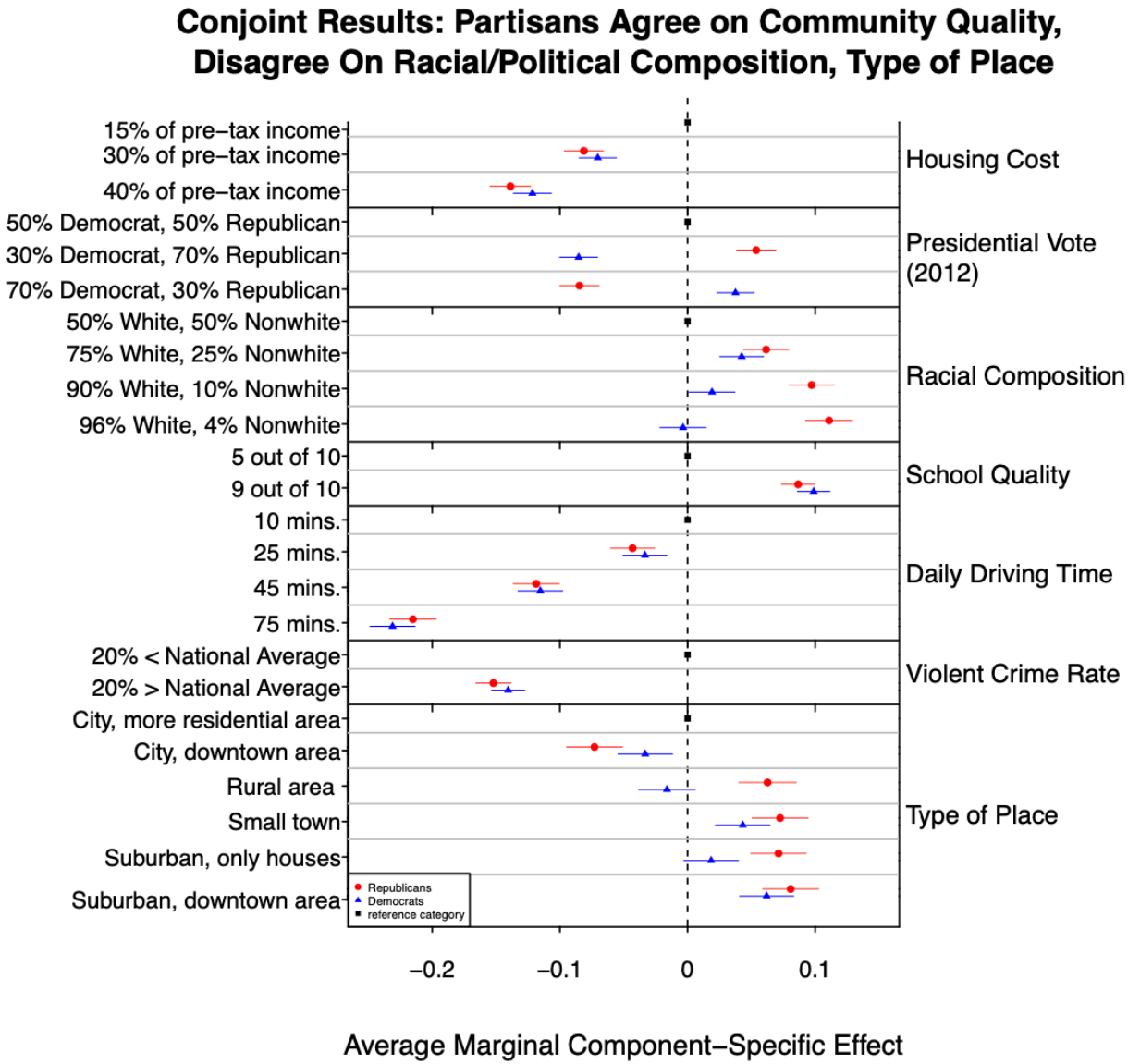
Partisan Residential Preferences

- What drives residential segregation between Democrats and Republicans?
- Do partisans prefer to live in like-minded communities? Or have different preferences on other dimensions?
- Nall and Mummolo (2017) use a conjoint experiment to investigate what features are most important for selecting where to live
- Factors: commute time, type of place, housing cost, racial composition of community, violent crime rate, presidential voting patterns, school quality

Partisan Residential Preferences: Profiles

COMMUNITY TRAIT	COMMUNITY A	COMMUNITY B
Total Daily Driving Time for Commuting and Errands	25 min	45 min
Type of Place	Suburban neighborhood with mix of shops, houses, businesses	Small town
Housing Cost	15 percent of pre-tax income	30 percent of pre-tax income
Race	90% White, 10% Nonwhite	50% White, 50% Nonwhite
Violent Crime Rate (Vs National Rate)	20% More Crime Than National Average	20% Less Crime Than National Average
Presidential Vote in 2012	30% Democrat, 70% Republican	70% Democrat, 30% Republican
School Quality Rating (1=Worst, 10=Best)	9	9

Partisan Residential Preferences: Results



Partisan Residential Preferences: Takeaways

- On "community quality" factors like school quality and crime rates, Democrats and Republicans have similar preferences
- On "home qualities" factors like cost and commute time, they have similar preferences
- Partisan differences over ideological and racial composition of community, as well as type of community
- Other results in paper: these ideological preferences aren't very important in practice

Conjoint Experiments

- Enable comparison of many attributes at once: useful for studying multidimensional choices
- Often set up to resemble real-world decision tasks: move to place A or place B? vote for candidate A or candidate B?
- Research suggests conjoints provide good estimates of real-world behavior

Taking Stock

- Survey experiments enable researchers to measure **causal effects** in surveys
- Intuition: randomly assign some respondents to a "treatment" and measure differences in outcomes across groups
- Because respondents are randomly assigned, they are the same on average except for their exposure to the treatment
- A few types of survey experiments:
  - vignette experiments
  - question-wording experiments
  - conjoint experiments

Some Additional Issues

1. Explaining "null" findings
2. Validity of survey experiments

Null Findings and Manipulation Checks

Example: Secrecy in Foreign Affairs

Myrick (2020): Does the public prefer transparency in foreign affairs? Does the public disapprove of covert military action?

Example: Secrecy in Foreign Affairs

Myrick (2020): Does the public prefer transparency in foreign affairs? Does the public disapprove of covert military action?

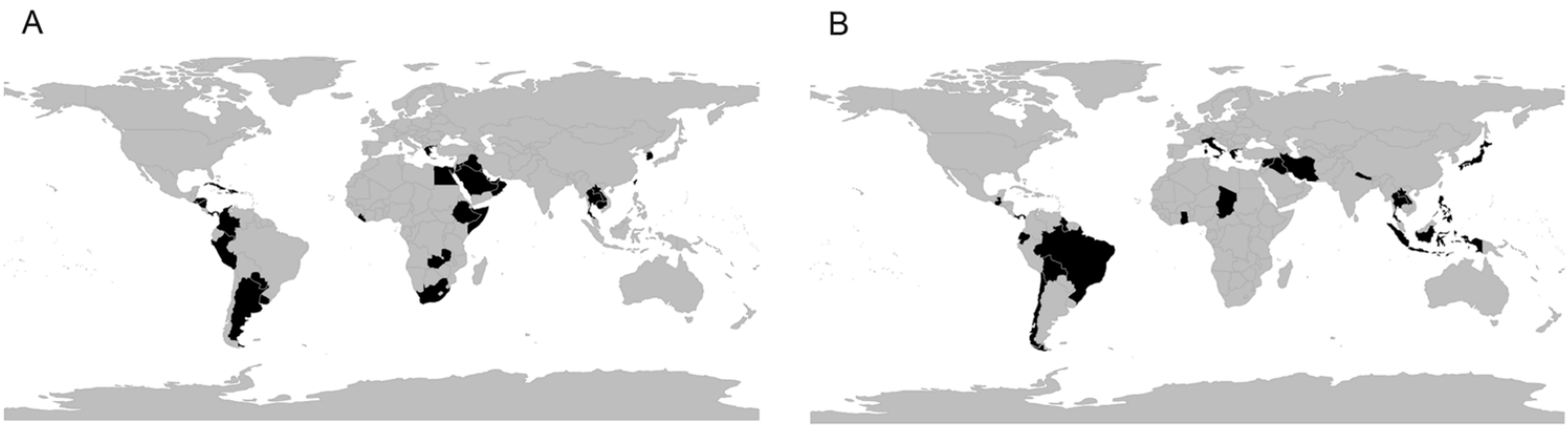


Figure 1. CIA interventions to (*A*) support existing leaders and (*B*) install new leaders, 1947–89

Treatment Vignette: Secrecy in Foreign Affairs

Consider the following situation:

- A dictator in [Asia / Africa / Latin America / the Middle East] is widely known for torturing and repressing his people and threatening stability in the region.
- Rebels within the country are attempting to overthrow the current government but have been unsuccessful so far.
- After debating different policies, the U.S. government decided to [T1: send money and weapons to assist the rebels / send in a small military force to assist the rebels].
- [T2: The government informed the American public about the operation / The government kept the operation completely secret from the American public].

How much do you approve or disapprove of the actions taken by the U.S. government in this situation?

Explaining Null Findings

What would we conclude about the public's attitudes if we found:

- the "secret" scenario received lower approval than the "informed" scenario?
- the "secret" scenario received higher approval than the "informed" scenario?

Explaining Null Findings

What would we conclude about the public's attitudes if we found:

- the "secret" scenario received lower approval than the "informed" scenario?
- the "secret" scenario received higher approval than the "informed" scenario?
- **the "secret" scenario received the same amount of approval as the "informed" scenario?** (failing to reject the *null hypothesis* of no treatment effect)

Explaining Null Findings

What would we conclude about the public's attitudes if we found:

- the "secret" scenario received lower approval than the "informed" scenario?
- the "secret" scenario received higher approval than the "informed" scenario?
- **the "secret" scenario received the same amount of approval as the "informed" scenario?** (failing to reject the *null hypothesis* of no treatment effect)

At least two explanations for a null finding:

- On average, respondents no more or less likely to support a secret operation than a transparent operation

Explaining Null Findings

What would we conclude about the public's attitudes if we found:

- the "secret" scenario received lower approval than the "informed" scenario?
- the "secret" scenario received higher approval than the "informed" scenario?
- **the "secret" scenario received the same amount of approval as the "informed" scenario?** (failing to reject the *null hypothesis* of no treatment effect)

At least two explanations for a null finding:

- On average, respondents no more or less likely to support a secret operation than a transparent operation
- Respondents did not notice or internalize information about secrecy/transparency — perhaps they are skimming (satisficing)

Manipulation Checks: Secrecy in Foreign Affairs

Good practice to include a **manipulation check** to verify that respondents noticed and internalized the treatment. Ask a simple, easy-to-answer question that verifies the respondents were paying attention.

Manipulation Checks: Secrecy in Foreign Affairs

Good practice to include a **manipulation check** to verify that respondents noticed and internalized the treatment. Ask a simple, easy-to-answer question that verifies the respondents were paying attention.

> *We are checking to make sure you read carefully. In the last scenario, what did the U.S. government do?*
>
> - *the government informed the American public about the operation.*
> - *the government kept the operation completely secret from the American public.*
> - *I don't know.*

If treatment is working $\rightsquigarrow$ more people in "secret" condition will answer "secret", more people in "informed" condition will answer "informed".

Manipulation Checks: Secrecy in Foreign Affairs

Good practice to include a **manipulation check** to verify that respondents noticed and internalized the treatment. Ask a simple, easy-to-answer question that verifies the respondents were paying attention.

> *We are checking to make sure you read carefully. In the last scenario, what did the U.S. government do?*
>
> - *the government informed the American public about the operation.*
> - *the government kept the operation completely secret from the American public.*
> - *I don't know.*

If treatment is working $\rightsquigarrow$ more people in "secret" condition will answer "secret", more people in "informed" condition will answer "informed".

If people pass manipulation check but you find no treatment effects, easier to argue that the treatment really doesn't have an effect.

Results: Secrecy in Foreign Affairs

- Outcome: answer to a 7-point approve/disapprove scale
- Results: Respondents express lower approval of secret operations: 0.604 points lower on the 7-point scale (this result is statistically significant)
- However: After given additional information about whether the military action was successful or not, this effect shrinks in half
- Conclusion: Normative preference for transparency, but outcomes matter more than process

Lessons for Survey Design

- Before the study: think through inference if there is/isn't an effect
- Want respondents to NOTICE the treatment – don't be subtle!
- When possible: include manipulation check

Validity in Survey Experiments

Survey Experiments

- Survey experiments can be powerful tools for identifying causal effects
- But they are artificial and may not perfectly mimic real-world "treatments"
- Under what assumptions can we *generalize* from a survey experiment to real-world scenarios?

Internal and External Validity

Two more notions of "validity" for causal inference.

Internal and External Validity

Two more notions of "validity" for causal inference.

Internal validity: how confident are we that the observed correlation is *causal*?

Internal and External Validity

Two more notions of "validity" for causal inference.

Internal validity: how confident are we that the observed correlation is *causal*?

- Randomized experiments have high internal validity because the researcher controls the assignment process: randomization ensures no "confounding" variables
- Observational studies often have lower internal validity

Internal and External Validity

Two more notions of "validity" for causal inference.

Internal validity: how confident are we that the observed correlation is *causal*?

- Randomized experiments have high internal validity because the researcher controls the assignment process: randomization ensures no "confounding" variables
- Observational studies often have lower internal validity

External validity: do the results generalize to the population of interest and to real-world settings?

Internal and External Validity

Two more notions of "validity" for causal inference.

Internal validity: how confident are we that the observed correlation is *causal*?

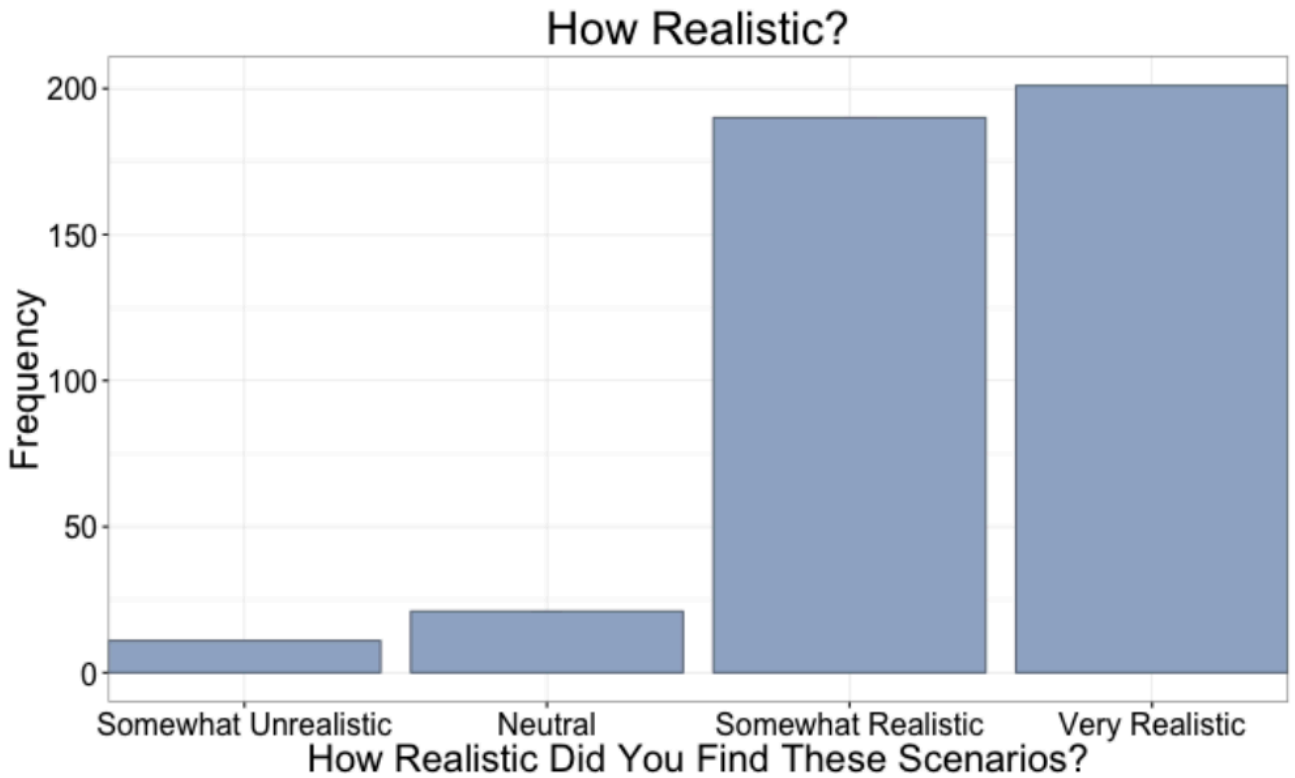
- Randomized experiments have high internal validity because the researcher controls the assignment process: randomization ensures no "confounding" variables
- Observational studies often have lower internal validity

External validity: do the results generalize to the population of interest and to real-world settings?

- Survey experiments often present slightly unrealistic scenarios $\rightsquigarrow$ would real-world scenarios have the same effects?
- Survey experiments might not be conducted on a representative sample $\rightsquigarrow$ are the effects in one sample the same as effects in a another?

External Validity: How Realistic is the Treatment?

In a pre-test (a small-scale survey on a cheaper sample), Myrick asked respondents to rate how realistic the scenario seemed:



Does this assuage concerns about external validity?

External Validity: Sample Selection

Survey experiments can have high internal validity even with highly non-representative samples (e.g. college students, opt-in survey respondents, etc.).

When will these results have external validity?

External Validity: Sample Selection

Survey experiments can have high internal validity even with highly non-representative samples (e.g. college students, opt-in survey respondents, etc.).

When will these results have external validity?

- Similar assumptions as in weighting: the average treatment effect in the sample is equal to the average treatment effect in the population

External Validity: Sample Selection

Survey experiments can have high internal validity even with highly non-representative samples (e.g. college students, opt-in survey respondents, etc.).

When will these results have external validity?

- Similar assumptions as in weighting: the average treatment effect in the sample is equal to the average treatment effect in the population
- Assumption satisfied under either of the following scenarios:
  1. There is no heterogeneity in the treatment effect across the population $\rightsquigarrow$ e.g., all people have the same difference in approval of secret vs. informed operation
  2. There is heterogeneity in the treatment effect, but it is not correlated with response propensity $\rightsquigarrow$ e.g., under true random sample

External Validity: Sample Selection

Survey experiments can have high internal validity even with highly non-representative samples (e.g. college students, opt-in survey respondents, etc.).

When will these results have external validity?

- Similar assumptions as in weighting: the average treatment effect in the sample is equal to the average treatment effect in the population
- Assumption satisfied under either of the following scenarios:
  1. There is no heterogeneity in the treatment effect across the population $\rightsquigarrow$ e.g., all people have the same difference in approval of secret vs. informed operation
  2. There is heterogeneity in the treatment effect, but it is not correlated with response propensity $\rightsquigarrow$ e.g., under true random sample
- Easiest solution: conduct the experiment on a high-quality, representative sample — a *population-based survey experiment* (Mutz 2013)

External Validity: Sample Selection

Survey experiments can have high internal validity even with highly non-representative samples (e.g. college students, opt-in survey respondents, etc.).

When will these results have external validity?

- Similar assumptions as in weighting: the average treatment effect in the sample is equal to the average treatment effect in the population
- Assumption satisfied under either of the following scenarios:
  1. There is no heterogeneity in the treatment effect across the population $\rightsquigarrow$ e.g., all people have the same difference in approval of secret vs. informed operation
  2. There is heterogeneity in the treatment effect, but it is not correlated with response propensity $\rightsquigarrow$ e.g., under true random sample
- Easiest solution: conduct the experiment on a high-quality, representative sample — a *population-based survey experiment* (Mutz 2013)
- Otherwise, consider the selection process and look for evidence of possible bias
- Weighting methods are also available but not used as often as standard survey weights

External Validity: Convergence Between Survey and Naturalistic Evidence

Observational analysis often better reflects real-world events, but there are limitations:

- May not have exactly the data necessary to test the substantive hypothesis (e.g. about mechanisms or subgroups)
- May have lower internal validity $\rightsquigarrow$ assumptions necessary for causal inference may be less plausible

External Validity: Convergence Between Survey and Naturalistic Evidence

Observational analysis often better reflects real-world events, but there are limitations:

- May not have exactly the data necessary to test the substantive hypothesis (e.g. about mechanisms or subgroups)
- May have lower internal validity $\rightsquigarrow$ assumptions necessary for causal inference may be less plausible

Researchers often supplement observation analysis with carefully designed survey experiments.

- Survey experiments (often) have higher internal validity due to random assignment, but lower external validity
- If results of observational analysis and survey experiment converge, more evidence for the findings

Validation of Survey Experiments

Hainmueller, Hangartner, and Yamamoto (2015) validate survey experimental estimates against real-world behavior.

Research question: Which features of immigrants are seen as preferable by natives? Which immigrants would natives like to give citizenship to?

Validation of Survey Experiments

Hainmueller, Hangartner, and Yamamoto (2015) validate survey experimental estimates against real-world behavior.

Research question: Which features of immigrants are seen as preferable by natives? Which immigrants would natives like to give citizenship to?

Observational research design: In Swiss municipalities, referendum on whether to give individual migrants citizenship. Using a regression model, authors estimate the causal effect of immigrant characteristics on voters' choices (e.g. gender, country of origin, education, etc.).

Validation of Survey Experiments

Hainmueller, Hangartner, and Yamamoto (2015) validate survey experimental estimates against real-world behavior.

Research question: Which features of immigrants are seen as preferable by natives? Which immigrants would natives like to give citizenship to?

Observational research design: In Swiss municipalities, referendum on whether to give individual migrants citizenship. Using a regression model, authors estimate the causal effect of immigrant characteristics on voters' choices (e.g. gender, country of origin, education, etc.).

Survey experiment design: Conjoint and vignette experiments randomly assigning the same information about migrants as in the real-world referendums.

Validation of Survey Experiments

Hainmueller, Hangartner, and Yamamoto (2015) validate survey experimental estimates against real-world behavior.

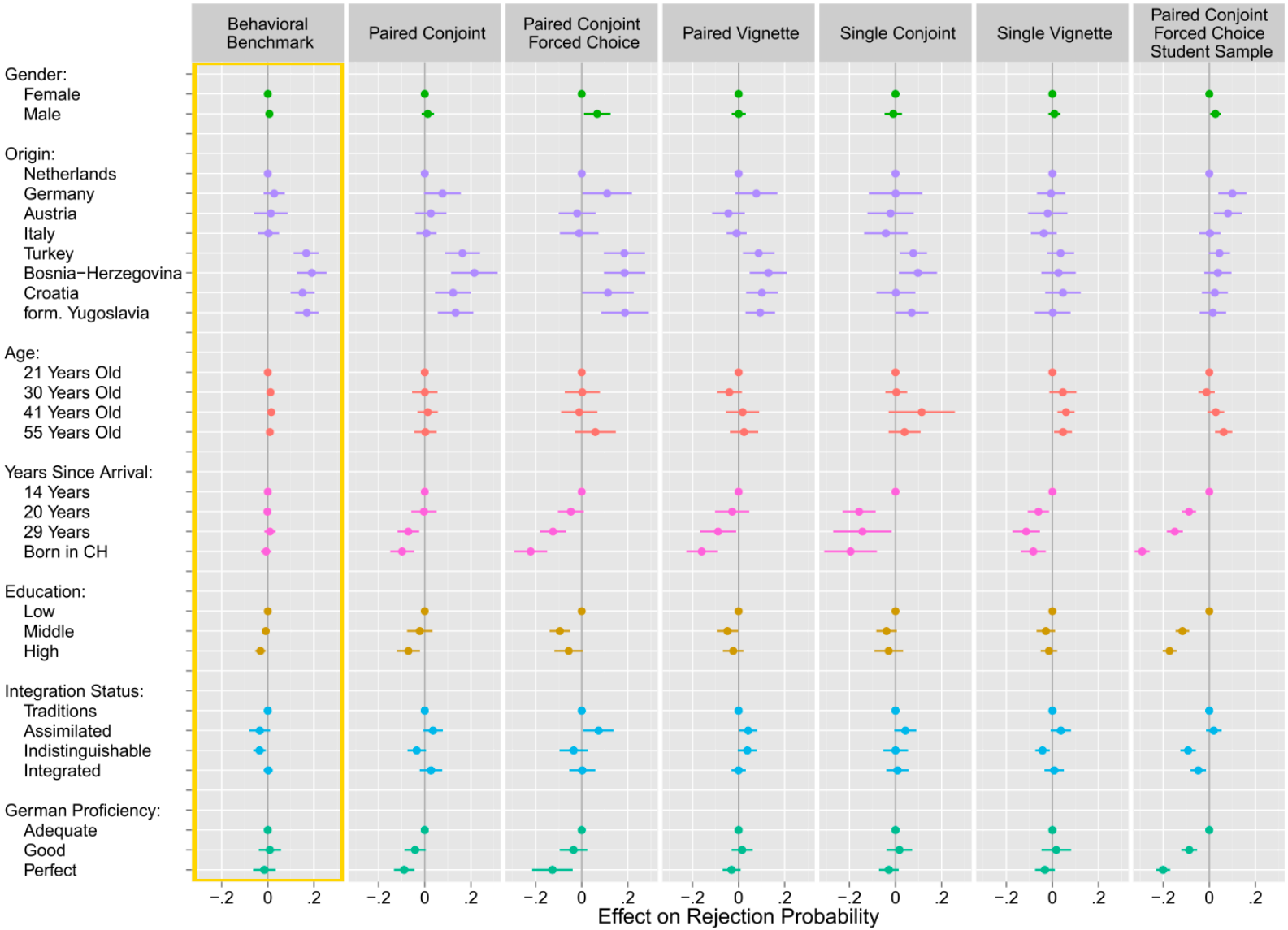
Research question: Which features of immigrants are seen as preferable by natives? Which immigrants would natives like to give citizenship to?

Observational research design: In Swiss municipalities, referendum on whether to give individual migrants citizenship. Using a regression model, authors estimate the causal effect of immigrant characteristics on voters' choices (e.g. gender, country of origin, education, etc.).

Survey experiment design: Conjoint and vignette experiments randomly assigning the same information about migrants as in the real-world referendums.

If results are the same $\rightsquigarrow$ provides real-world validation of the estimates derived from the survey experiment.

Naturalization Preferences: Real-World vs. Survey Experiment



Takeaways from Validation

Does this study convince you that conjoint experiments results reflect real-world behavior?

Takeaways from Validation

Does this study convince you that conjoint experiments results reflect real-world behavior?

- Rare case where the actual decision and information environment can be closely replicated in a survey.
- Provides evidence against surveys *per se* mismeasuring preferences/behavior.
- External validity of this study?

Next Time

- Political polling: problems, solutions, and philosophy
- Read Morris, *Strength in Numbers* for next Thursday
- Begin programming your survey in Qualtrics

Have a great fall break!