Intro to Data: What, Where and How?

RALPH BUNCHE SUMMER INSTITUTE 2023

DUKE UNIVERSITY

METHODS LAB

MATEO VILLAMIZAR-CHAPARRO*

*Some content was taken from slides produced by Katie Webster in 2019

This Session will cover:

1. What do we mean by data?
2. Variables and types of variables
3. How and where to find data?
4. Some thoughts about measurement

BUT…first things first

We know this is getting old but …
repetition is key for learning

Tell us your name

Your research interests

Are you a Mac or a Windows user?

Have you used any statistical software before / taken classes in statistics?

Remember to access slack!

What do we mean by data? (I)

Data: Is a collection of pieces of observation gathered through a myriad of activities like filed notes, surveys, experiments, images, archives, etc.

- It can be from any topic – political behavior, economic information, war, preferences for redistribution, race, environmental factors, etc
- Recently, advances in computation analysis have allowed the use of social media, satellite images, pictures, sounds, and videos as sources of data.

What do we mean by data? (II)

Some of the most common distinctions between types of data are:

- Qualitative (interviews, focus groups, ethnography) vs quantitative (surveys, scraping data, text)
- Experimental vs observational
- We can also distinguish data with respect to time
  - Cross sectional: observations from the same period
  - Time series: observations of the same unit across time
  - Panel data: observations from multiple units across time

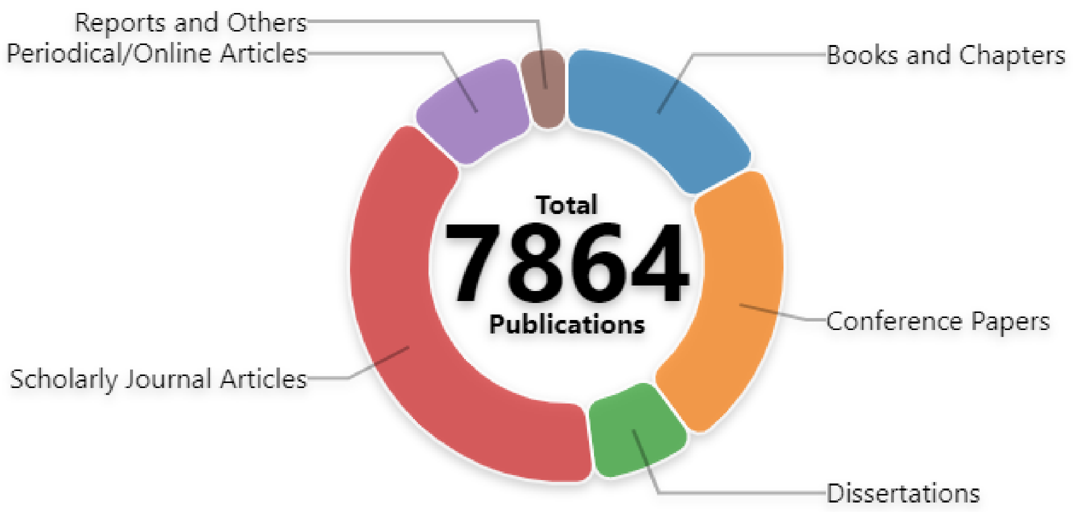
From Data to Datasets/Dataframes

A dataset is a series of organized observations. It is important to always identify what is the unit of analysis (observations) of the dataset since this will determine if the data will be useful to answer your research question:

- County level data on income -> county
- A survey like the ANES -> individual
- Census -> individual
- Data on international conflicts -> country or pairs of countries

American National Election Studies

- A series of election studies conducted during the years of every US presidential election since 1948.
- Respondents were interviewed during the two months preceding the November election (pre-election interview) and then re-interviewed during the two months following the election (post-election interview).
- Who is included? U.S. Citizens 18 or older (The unit of analysis here is the individual citizen)
- Hundreds of variables, about citizen demographics, opinions and political behavior.
- Is it a time series?



Finding data: other common datasets

- US Politics
  - General Social Survey (GSS)
  - Cooperative Congressional Election Survey (CCES)
  - Resource center for minority data
  - National Asian survey
- International relations
  - Correlates of War (COW)
  - International Crisis Behavior (ICB)
- Comparative Politics
  - World Bank Indicators
  - IPUMS international
  - Regional barometers
  - Varieties for democracy

How to find datasets?

- Google is always your friend
- **Harvard dataverse or ICPSR**
- Scholar/Lab websites
- Published articles
- Ask around

Some important tips when looking/downloading datasets

Always look and read the codebook. It will have key information like:

- the survey specifications, sampling methods, etc
- the measures of the variables,
- The questionnaire or list of variables
- the unit of analysis and the coverage of the data (both units ad time)

When downloading the data, always take a look at its structure. For data analysis purposes some structures might be more convenient than others for you.

Sometimes not all the information you want is going to be in the same dataset, so it is always good to check if the data has some type of indicator that you could use in case you want to merge different datasets (more on this later)

How a common dataset looks like

Unit of analysis: Defines the entities at which the observations are measured

Each column contains the values for a variable in your dataset

Unit	Variable 1	Variable 2	Variable 3

Each row contains the information for one observation

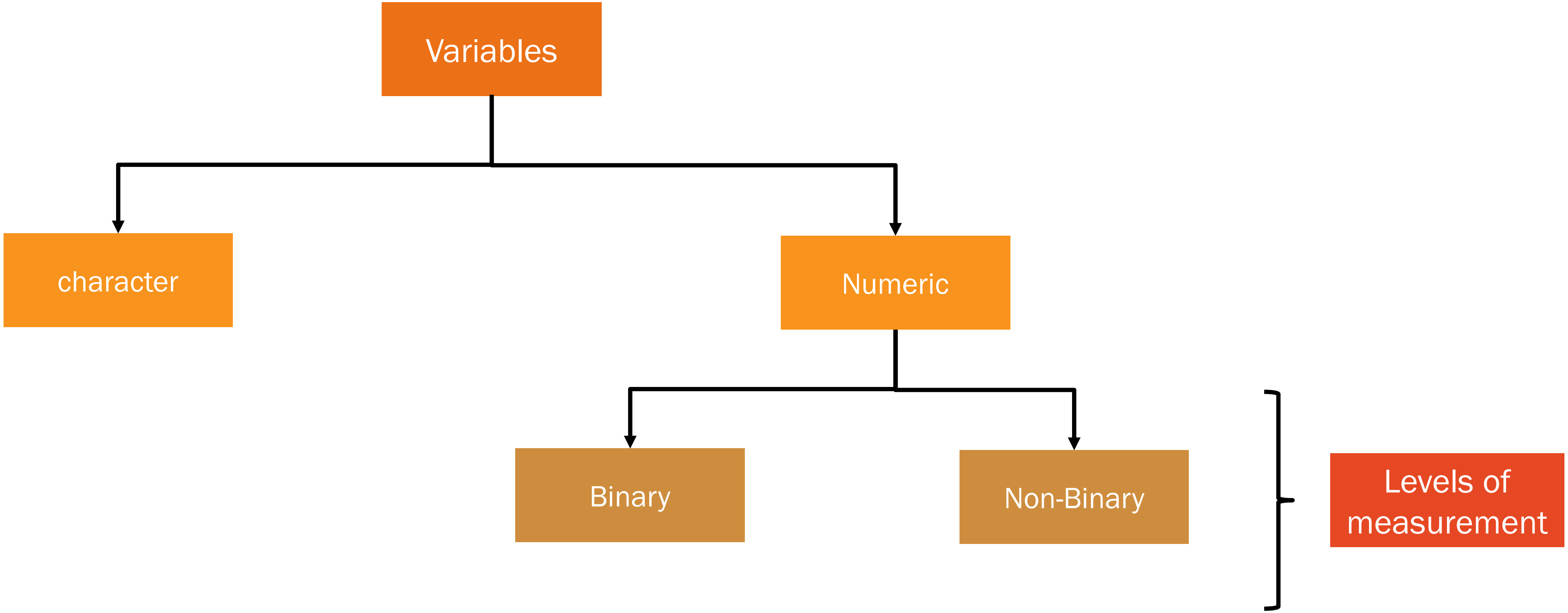
Variables

Variables are representations of properties that observations have.

- They provide some information about the observations themselves.
- They usually have different values across observations (over time, space, or unit).
- You should try to match your variables to your concepts as much as possible and make sure they are measuring what you want them to measure (construct validity)

Dependent Variable	What we are trying to explain
Independent Variable	Things that explain changes in our DV
Control Variables	Things that are kept constant or are accounted for because they can affect the DV

Variables you Will probably find in R



Variables

character

Numeric

Binary

Non-Binary

Levels of measurement

Taken from Llaudet and Imai (2023) Data Analysis for social science

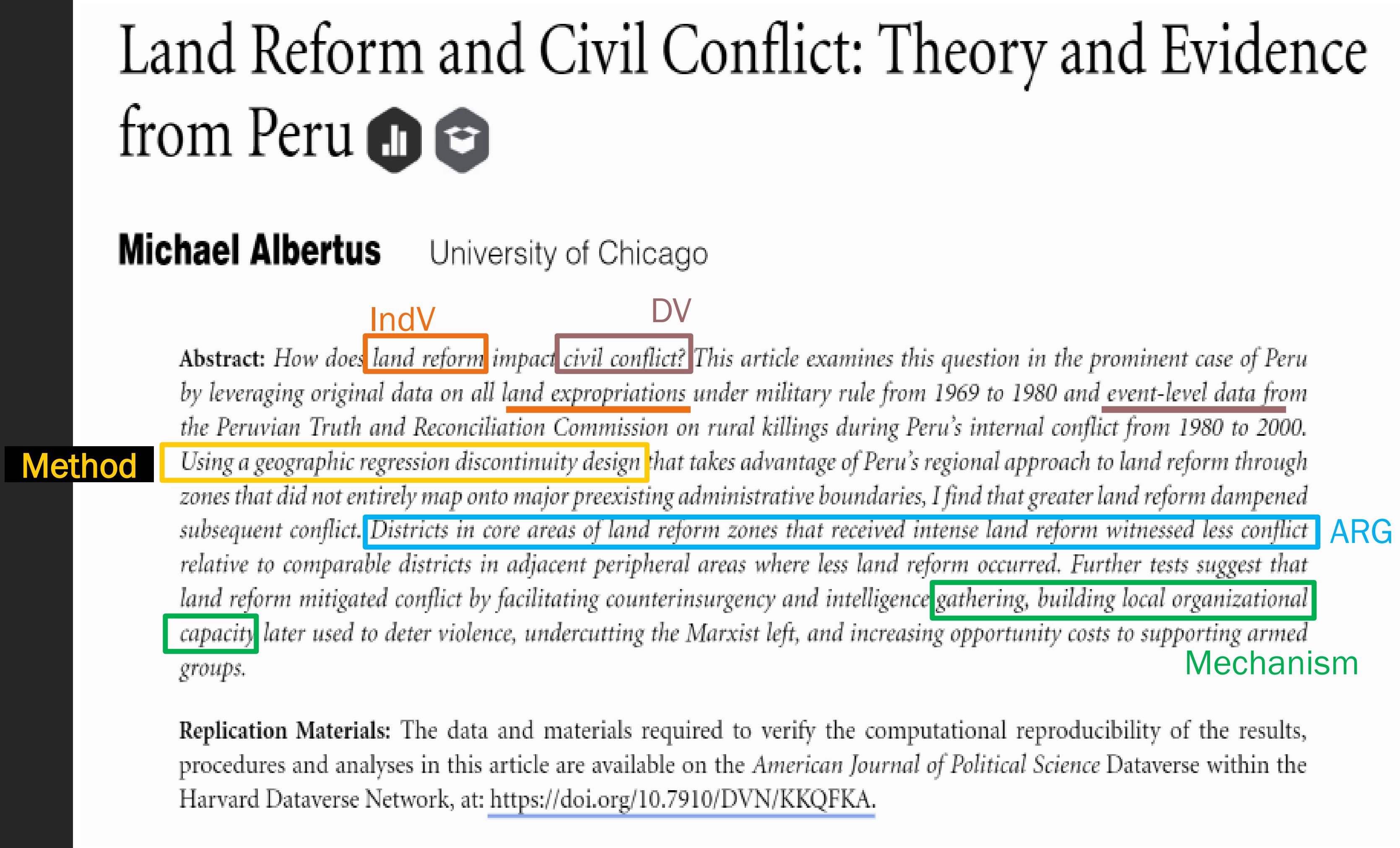
Levels of measurement

Nominal	Also known as dummy variables, indicator variables or dichotomous variables	Does not establish mathematical relationships just presence/absence	High school degree
Ordinal	Also know as categorical variables	Does establish mathematical relationships of <, ≤, ≥, >, =	Levels of education
Interval	Also know as continuous variables	Distance between values is constant and meaningful	Years of education
Ratio	Are also considered continuous variables	Interval + The zero value is a meaningful value	Municipal Budget

Dissecting and abstract

Let's put all of this together

1. Let's identify the dependent and independent variables
2. How would you measure the DV and the independent variables?
3. What is the argument?
4. What is the mechanism?
5. What is the method?



Land Reform and Civil Conflict: Theory and Evidence from Peru

Michael Albertus University of Chicago

Abstract: *How does land reform impact civil conflict? This article examines this question in the prominent case of Peru by leveraging original data on all land expropriations under military rule from 1969 to 1980 and event-level data from the Peruvian Truth and Reconciliation Commission on rural killings during Peru's internal conflict from 1980 to 2000. Using a geographic regression discontinuity design that takes advantage of Peru's regional approach to land reform through zones that did not entirely map onto major preexisting administrative boundaries, I find that greater land reform dampened subsequent conflict. Districts in core areas of land reform zones that received intense land reform witnessed less conflict relative to comparable districts in adjacent peripheral areas where less land reform occurred. Further tests suggest that land reform mitigated conflict by facilitating counterinsurgency and intelligence gathering, building local organizational capacity later used to deter violence, undercutting the Marxist left, and increasing opportunity costs to supporting armed groups.*

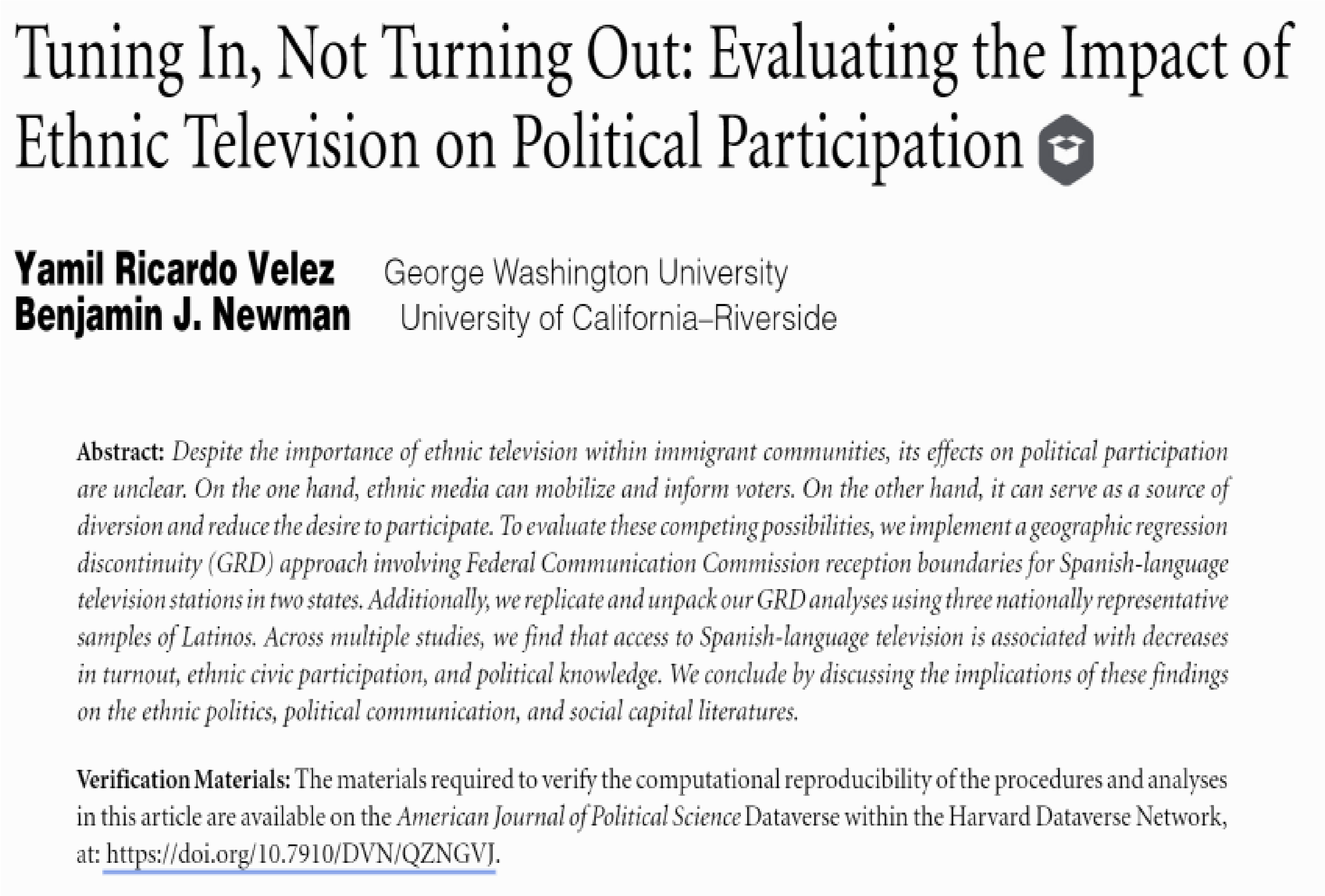
Replication Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/KKQFKA.

Thank you! If you have any questions, slack or email us!

Data Lab Exercises

Exercise 1a

1. Let’s identify the dependent and independent variables
2. How would you measure the DV and the independent variables?
3. What is the argument?



Tuning In, Not Turning Out: Evaluating the Impact of Ethnic Television on Political Participation

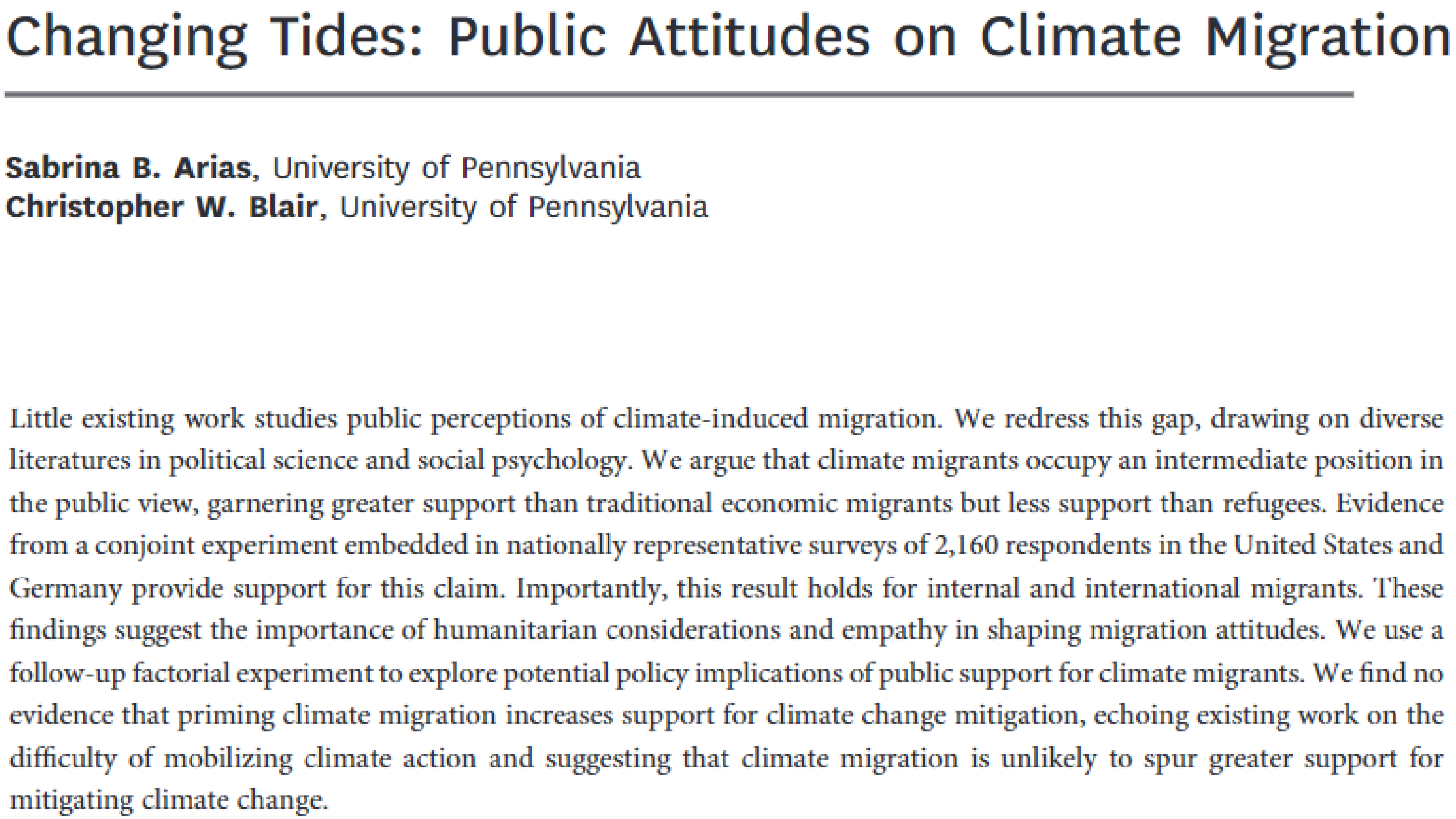
Yamil Ricardo Velez George Washington University
Benjamin J. Newman University of California–Riverside

Abstract: *Despite the importance of ethnic television within immigrant communities, its effects on political participation are unclear. On the one hand, ethnic media can mobilize and inform voters. On the other hand, it can serve as a source of diversion and reduce the desire to participate. To evaluate these competing possibilities, we implement a geographic regression discontinuity (GRD) approach involving Federal Communication Commission reception boundaries for Spanish-language television stations in two states. Additionally, we replicate and unpack our GRD analyses using three nationally representative samples of Latinos. Across multiple studies, we find that access to Spanish-language television is associated with decreases in turnout, ethnic civic participation, and political knowledge. We conclude by discussing the implications of these findings on the ethnic politics, political communication, and social capital literatures.*

Verification Materials: The materials required to verify the computational reproducibility of the procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/QZNGVJ.

Exercise 1b

1. Let's identify the dependent and independent variables
2. How would you measure the DV and the independent variables?
3. What is the argument?



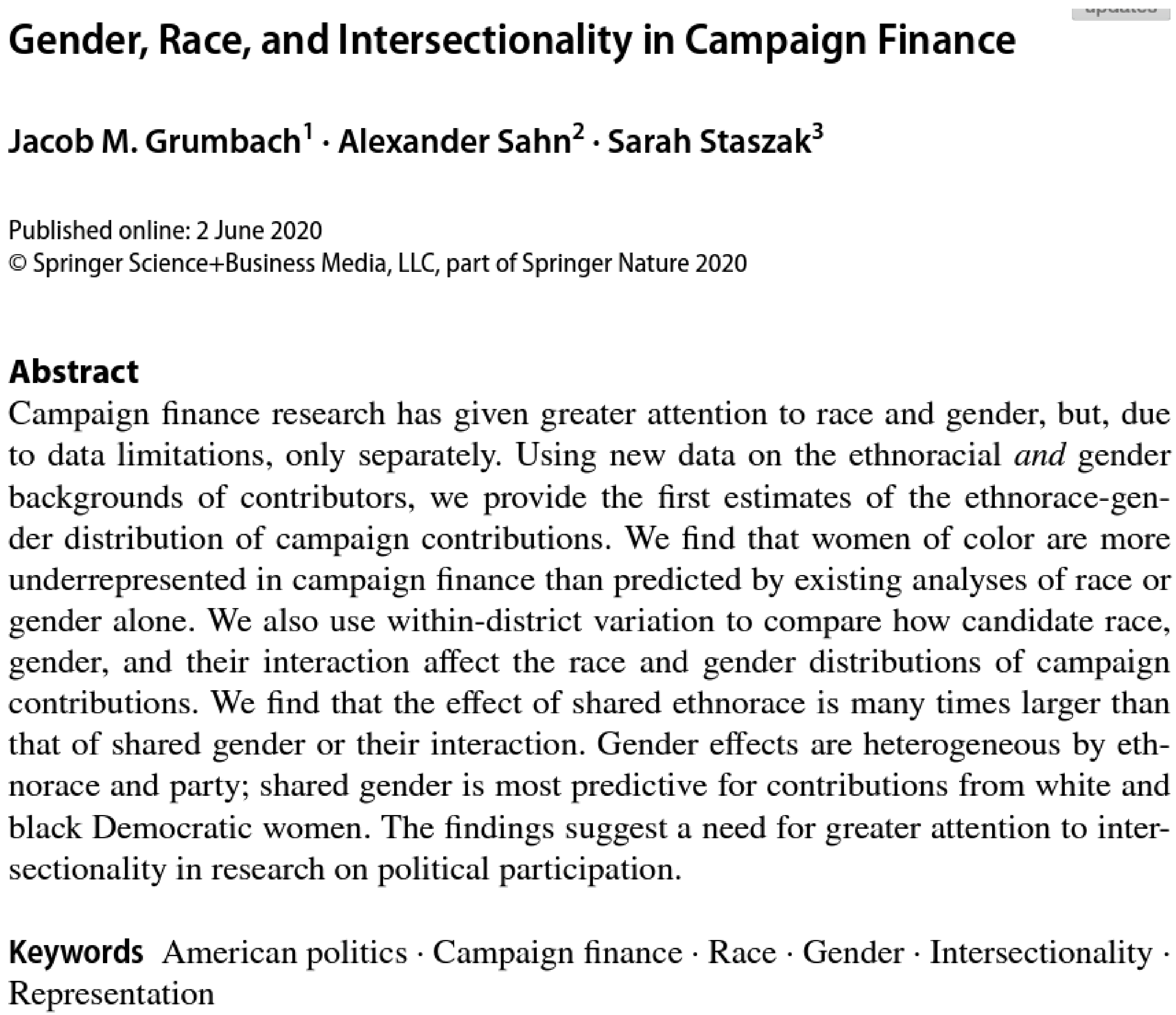
Changing Tides: Public Attitudes on Climate Migration

Sabrina B. Arias, University of Pennsylvania
Christopher W. Blair, University of Pennsylvania

Little existing work studies public perceptions of climate-induced migration. We redress this gap, drawing on diverse literatures in political science and social psychology. We argue that climate migrants occupy an intermediate position in the public view, garnering greater support than traditional economic migrants but less support than refugees. Evidence from a conjoint experiment embedded in nationally representative surveys of 2,160 respondents in the United States and Germany provide support for this claim. Importantly, this result holds for internal and international migrants. These findings suggest the importance of humanitarian considerations and empathy in shaping migration attitudes. We use a follow-up factorial experiment to explore potential policy implications of public support for climate migrants. We find no evidence that priming climate migration increases support for climate change mitigation, echoing existing work on the difficulty of mobilizing climate action and suggesting that climate migration is unlikely to spur greater support for mitigating climate change.

Exercise 1c

1. Let's identify the dependent and independent variables
2. How would you measure the DV and the independent variables?
3. What is the argument?



Gender, Race, and Intersectionality in Campaign Finance

Jacob M. Grumbach[1] · Alexander Sahn[2] · Sarah Staszak[3]

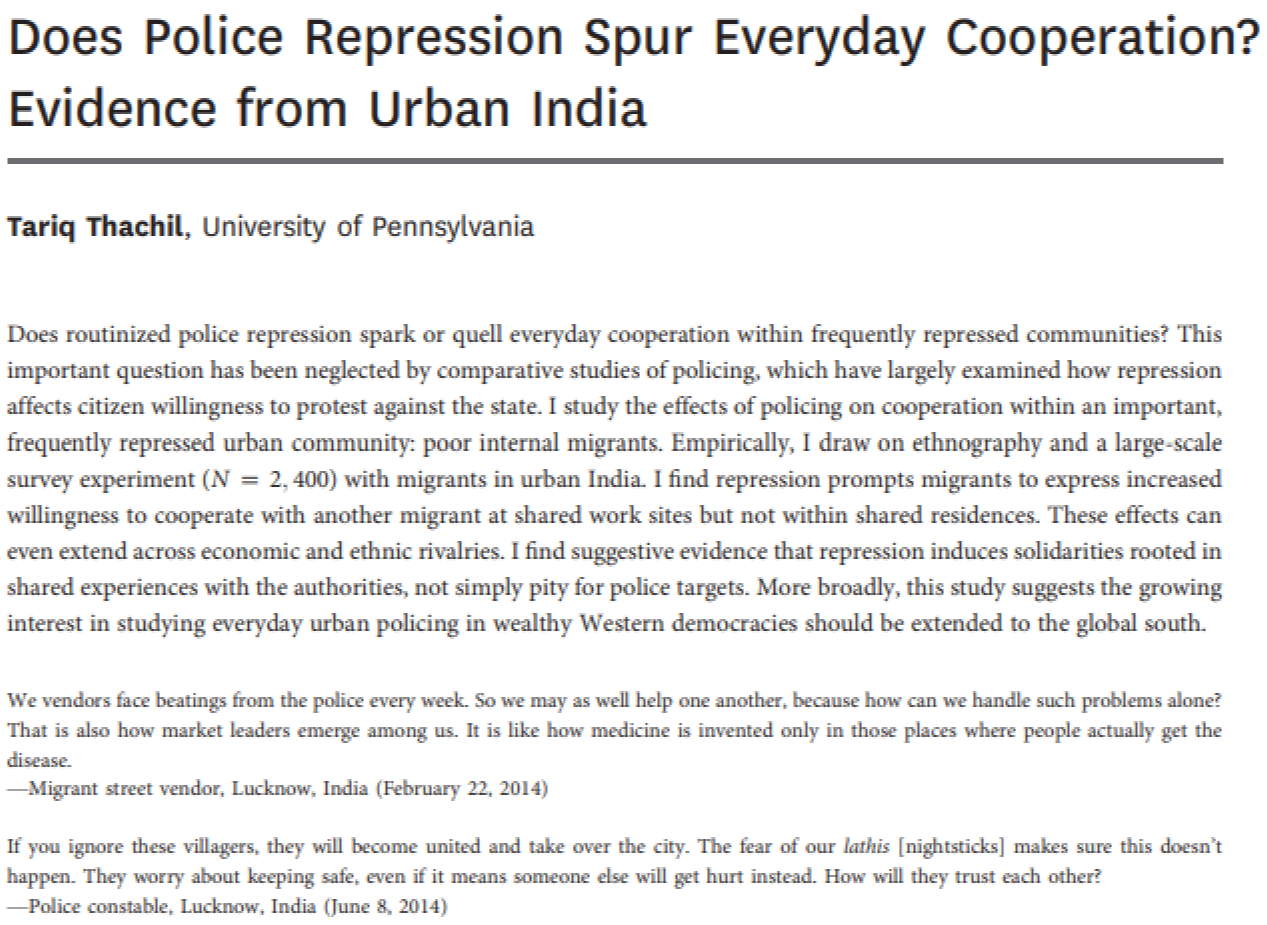
Published online: 2 June 2020
© Springer Science+Business Media, LLC, part of Springer Nature 2020

Abstract
Campaign finance research has given greater attention to race and gender, but, due to data limitations, only separately. Using new data on the ethnoracial *and* gender backgrounds of contributors, we provide the first estimates of the ethnorace-gender distribution of campaign contributions. We find that women of color are more underrepresented in campaign finance than predicted by existing analyses of race or gender alone. We also use within-district variation to compare how candidate race, gender, and their interaction affect the race and gender distributions of campaign contributions. We find that the effect of shared ethnorace is many times larger than that of shared gender or their interaction. Gender effects are heterogeneous by ethnorace and party; shared gender is most predictive for contributions from white and black Democratic women. The findings suggest a need for greater attention to intersectionality in research on political participation.

Keywords American politics · Campaign finance · Race · Gender · Intersectionality · Representation

Exercise 1d

1. Let’s identify the dependent and independent variables
2. How would you measure the DV and the independent variables?
3. What is the argument?



Does Police Repression Spur Everyday Cooperation? Evidence from Urban India

Tariq Thachil, University of Pennsylvania

Does routinized police repression spark or quell everyday cooperation within frequently repressed communities? This important question has been neglected by comparative studies of policing, which have largely examined how repression affects citizen willingness to protest against the state. I study the effects of policing on cooperation within an important, frequently repressed urban community: poor internal migrants. Empirically, I draw on ethnography and a large-scale survey experiment ($N = 2,400$) with migrants in urban India. I find repression prompts migrants to express increased willingness to cooperate with another migrant at shared work sites but not within shared residences. These effects can even extend across economic and ethnic rivalries. I find suggestive evidence that repression induces solidarities rooted in shared experiences with the authorities, not simply pity for police targets. More broadly, this study suggests the growing interest in studying everyday urban policing in wealthy Western democracies should be extended to the global south.

We vendors face beatings from the police every week. So we may as well help one another, because how can we handle such problems alone? That is also how market leaders emerge among us. It is like how medicine is invented only in those places where people actually get the disease.
—Migrant street vendor, Lucknow, India (February 22, 2014)

If you ignore these villagers, they will become united and take over the city. The fear of our *lathis* [nightsticks] makes sure this doesn’t happen. They worry about keeping safe, even if it means someone else will get hurt instead. How will they trust each other?
—Police constable, Lucknow, India (June 8, 2014)

Exercise 2: Write down your ideas

In the next five minutes, think about three different ideas you might want to explore for your final RBSI paper.

- This is a first try, your final topic might change but is good to start thinking about it.

Exercise 3: Pair and Share

From the three ideas of the previous exercise, choose the one you like the most and tell your partner:

- What is the research question you might want to explore during RBSI?
  - Don’t think on sources right now
- Identify your dependent and independent variables?
- What would your unit of analysis be?