Dataset Generation

The CSV files and python scripts to generate the datasets for each study are included.

Study 1: *dataset_formative1.csv* and *data_generator_formative1.py*

Study 2: *dataset_formative2.csv* and *data_generator_formative2.py*

Main Study: *dataset_main_study.csv* and *data_generator_main_study.py*

Each script takes a number *n* representing the number of rows (politicians) to generate. The script then generates one politician at a time, appending it to a list, then writing the complete list to a file line-by-line.

An individual politician is generated by sampling each attribute value from the distributions defined in the script. The distributions for each attribute vary per study and can be found in the respective scripts. For Study 1 and the Main Study, we sought for the dataset as a whole to adhere to specific controlled distributions -- hence, for these studies, we generated additional rows (politicians) for underrepresented attributes and pruned rows for overrepresented attributes until the desired distribution was achieved.

The distributions of attributes sampled from for each of the three studies are found in the following table.

	Attribute	Formative Study 1 (X)	Formative Study 2 (Y)	Main Study (Z)
Biographical Attributes	**Name**	Sampled randomly by gender		
	Party	50% Democrat; 50% Republican	46% Democrat ; 54% Republican	
	Gender	50% Female; 50% Male	Female (28% if Democrat; 12% if Republican); Male (72% if Democrat; 88% if Republican)	
	Occupation	25% each: Career Politician, Doctor, Lawyer, Business	26% Career Politician; 24% Business Person; 17% Lawyer; 11% Educator; 7% Judge; 3% Financier; 3% Doctor; 3% Farmer; 2% Military; 2% Engineer; 1% Minister; 1% Scientist	38% Lawyer; 23% Career Politician; 21% Business Person; 9% Educator; 5% Scientist; 4% Doctor
	Education	-	4% High School; 2% Associate's; 25% Bachelor's; 22% Master's; 5% PhD; 38% Law; 4% Medical, constrained by Occupation	-
	Religion	-	88% Christian; 6% Jewish; 2% Mormon; 1% Muslim; 1% Hindu; 2% Unaffiliated	
	Age (Years)	-	Sampled from normal distr. with μ = 58 years, σ = 10 years	
	Experience (Years)	33% each: Low, Medium, High	Sampled from normal distr. with μ = 9 years, σ = 3 years	
Policy Attributes	**Ban Abortion After 6 Weeks**	33% each: In Favor, Neutral, Opposed	-/+ 3, constrained by party: D (-) R (+)	33% each: In Favor, Neutral, Opposed
	Legalize Medical Marijuana	-	+/- 3, constrained by party: D (+) R (-)	-
	Budget for Free School Lunch	-	+/- 3, constrained by party: D (+) R (-)	-
	Increase Gun Control Legislation	-	+/- 3, constrained by party: D (+) R (-)	-
	Ban Alcohol Sales on Sundays	-	-/+ 3, constrained by party: D (-) R (+)	-
	Increase Budget for Medicare	-	+/- 3, constrained by party: D (+) R (-)	-
	Increase Budget for VA	-	+/- 3, constrained by party: D (+) R (-)	-

Formative Study 1

We created a dataset of 144 politicians containing one politician with each combination of *Gender* (Male, Female), *Party* (Republican, Democrat), *Occupation* (Doctor, Lawyer, Business, Career Politician), *Experience* (Low, Medium, High), and the policy view *Ban Abortion After 6 Weeks* (Opposed, Neutral, In Favor). Names for each politician were generated based on U.S. census data.

Formative Study 2

We sought to increase the realism in this study by increasing the dimensionality of the dataset (i.e., people have more features to keep in mind during their decision), and deriving the dataset of fictitious politicians based on distributions found in the 115th US House of Representatives. In this version of the dataset, each of 100 fictitious politicians is described by biographical attributes (e.g., *Occupation*, *Religion*, *Experience*, etc) and policy attributes (e.g., each politician's view on issues like *Legalize Medical Marijuana*, etc). Policy attributes take on one of seven discrete values ranging from *strongly opposed* to *strongly in favor* with a *neutral* option. Politicians are assumed to primarily vote along party lines, with a 1% chance of voting against their party and a 5% chance of a neutral policy (defined arbitrarily). For non-neutral policy positions, values were sampled from a distribution of 30% *somewhat* {*opposed*, *in favor*}, 50% {*opposed*, *in favor*}, and 20% *strongly* {*opposed*, *in favor*}, representing our general view that more neutral policies are somewhat more likely than more extreme policies, with party-dependent policies being most likely.

Main Study

Compared to Study 2, we reduced the cardinality of *Occupation* to 6 options (reduced from 12). We removed *Education* (given that *Occupation* is often highly correlated). We removed all policy-related attributes, except for *Ban Abortion After 6 Weeks*.