Chapter 6: Causal Inference[*]

Justin Grimmer[†] Margaret E. Roberts [‡] Brandon M. Stewart [§]

March 19, 2018

Over the last two decades, computer scientists and statisticians have made great strides in producing and developing methods to analyze natural language. When social scientists began adapting these methods to the study social phenomena, they first applied these techniques for the purposes of discovery and measurement, the tasks we focused on in the last two chapters. For example, in sociology and psychology, these methods were used to track the ways in which the U.S. government discussed arts and culture (DiMaggio, Nag and Blei, 2013) and how language reflects personality and social interaction (Schwartz et al., 2013). In political science, these methods were adapted to measure how Japanese candidates spoke about policy issues (Catalinac, 2016), the ideology of the author of a text (Slapin and Proksch, 2008; Benoit and Laver, 2003; Lowe, 2008), and the sentiment of social media posts (Hopkins and King, 2010). In economics, some of the first applications of text as data methods outside of prediction measured the ideological content of newspapers to study how media competition affects slant (Gentzkow and Shapiro, 2010).

Some of the initial work in the social sciences that used text as data for discovery and measurement explicitly incorporated metadata into the measurement model to explore variation in measurement between various subsets of data. For example, Grimmer (2010) Expressed Agenda Model explicitly models information about the author of a group of posts into a

[*] Preliminary and incomplete draft, please do not circulate.

[†] Associate Professor, Department of Political Science, University of Chicago

[‡] Assistant Professor, Department of Political Science, University of California at San Diego

[§] Assistant Professor, Department of Sociology, Princeton University topic model, estimating how the amount an issue is talked about varies by each individual author. Grimmer (2013) uses this model to explore how members of the U.S. Congress focus on different issues when talking with their constituents depending on their characteristics, such as the competitiveness of their district. Quinn et al. (2006) creates a dynamic topic model which is able to measure how topics change over time and use it to study changes to discussions in the U.S. Senate over a 7 year period.[1]

The ability to include metadata into the measurement model inspired further research that explored how text as data methods could be used in conjunction with causal inference. If metadata could be included into a measurement model of text, then could an indicator of the treatment and control condition in an experiment also be included to estimate the impact of treatment on a text outcome? Roberts et al. (2014) and Roberts, Stewart and Airoldi (2016) explore a general model, the Structural Topic Model, for including metadata into text and show how the model can be used to estimate treatment effects in an experiment where text is the outcome. Gill and Hall (2015) show how permutation tests can be used to test the impact of a randomized treatment on a text outcome. These papers took the first steps toward defining the problem and addressing the challenge of using experiments in conjunction with unstructured text.

In this chapter, we take the next step toward defining explicitly the research processes and assumptions necessary to make causal inferences where text is a measure of the treatment, outcome, or confounding variables within a causal inference design. We build off of previous work that defines the problem of making causal inferences with text to explicitly outline the opportunities of using text in causal analyses, along with challenges of making causal inferences with high-dimensional data like text.

We begin this chapter by introducing three guiding principles of using text data with

[1] These models developed in conjunction with a large literature in computer science that explored using metadata in conjunction with text in measurement models, for example Mimno and McCallum (2008); Eisenstein, Ahmed and Xing (2011).

causal inference and return to these in detail throughout the remainder of the chapter. Then we outline the basic problem of causal inference and the challenges that text data poses for causal inference. We propose a solution to these challenges, and then provide several examples of using text in experiments as an outcome and as a treatment. We also explore conditioning on text in observational studies. Our chapter draws extensively not only on the previous work on using text data in conjunction with experiments in the social sciences, but also on companion papers, Egami et al. (2018), Fong and Grimmer (2018*a*), and Roberts, Stewart and Nielsen (2018).

1 Key Principles of Causal Inference with Text

The core problems of causal inference remain, even when working with text. Nothing about using text data absolves us from the core challenges with causal inference. Even with new data like text, we run into the fundamental problem of causal inference: that we cannot observe what would have happened under a different treatment condition for an individual observation. In order to accurately estimate the counterfactual—the value of some outcome if the state of the world had been different—we require the same basic ingredients as any good causal inference. We need to have a good design, be attentive to the sample we are working with, and be cautious when relying on selection on observables. Confounding, reverse causality and dependence between observations are just as plausible in cases where we have text data as those where we do not and we be vigilant about these and other threats to inference.

g **remains a critical component of causal inference with text.** In the discovery chapter, we illustrated how the analyst could discover new concepts from text data by exploring different representations of text, or g functions. In the measurement chapter, we described how researchers can use g to measure a given concept in the text at scale. In this chapter, we show how we can leverage g to make causal inferences. However, unlike in the previous chapters, we now have to be particularly careful about how we arrived at g. We will emphasize two major risks when developing g in a causal inference framework. The first risk should feel familiar – we are concerned that we will overfit the text. That is, one risk is that we will develop a categorization scheme that captures non-systematic facets of the text, essentially uncovering ephemeral noise in the documents. When this happens with measurement or discovery the risk is that we allocate attention to categories that are exceedingly rare and unlikely to appear in other samples. Overfitting in causal inference is even more pernicious, because it can lead us to infer interventions had a systematic effect when they didn't.

The second risk is a particular kind of spillover bias, which we call an Analyst Inducted SUTVA violation (AISV). AISV occurs when the analyst creates dependence between units by developing g with reference to the outcome of treatment in the experiment. If g is developed with the same data as is used in the final analysis, then the estimated potential outcome of one unit will depend on the treatment status of other units because the g function depends on the other units. This dependence violates SUTVA, one of the necessary conditions for causal inference. In this case, the analyst creates dependence between units in the process of creating g, which is why it's essential to consider the process of creating g for making good causal inferences. We explain overfitting and AISV in more detail later in this chapter.

The challenges of making causal inferences with text underscore the need for sequential science. Social science is the process of creating generalizable knowledge that explains or predicts societal patterns. The best causal inferences, or explanations, in social science emerge from a close relationship between theory and careful empirical evidence.

This interaction is sequential – theory is updated as we learn from experiments and the experiments that we run are guided by theory. Because this interaction is sequential, the acquisition of evidence is never completely done and no one experiment is completely pivotal. Replication of experiments in different time periods or in different populations expand our evidence base from which we can draw more accurate theories. These theories in turn suggest the next experiments.

Because of the interplay between experiments and theory, many have suggested that before an experiment has been conducted, all of the parameters within the experiment should be laid out, including the variables the experiment will study, how those variables are measured, and the ultimate tests that the experiment should report. In the words of this book, the g function must be specified before the experiment is conducted. The intuition behind this recommendation is that if an experiment is truly testing a theory known beforehand, that the experiment should not change after it has been started and that reporting results outside of the pre-conceived research design may undermine the analysis. Many have suggested that researchers pre-register their experiments before conducting them, creating a *pre-analysis plan* (PAP) that ties the researchers hands so that they cannot change the data collection and analysis until after the experiment is run.

While such inflexibility in the analysis of data can be very useful in particular contexts, as we will describe below, it relies on a great deal of certainty about the about the variables we're likely to use, how those variables are measured, and the properties of the data we're examining. It relies on the experiment running smoothly, without unanticipated hiccups. Of course, it is rarely the case we have all that information in hand before running an experiment or that unexpected events during implementation change the research design. This is especially true when dealing with high-dimensional data like text, where researchers may not be able to foresee the topics and sentiments that will be expressed by the subjects. The difficulties in anticipating how the experiment will run can lead to awkward applications of the PAP and the creation of institutions that diminish their effectiveness. For example, it is now common that PAPs are amended to incorporate new data or are altered after experiences with partner organizations. And because it is impossible to anticipate all the issues with a PAP, researchers are trying to develop standard operating procedures when confronted with deviations. Once we are able to alter a bit, we open the door to many of the same abuses that PAPs, at their best, are able to defend against.

Instead of relying on perfect foresight to predict potential deviations in research design, we advocate for conducting the interplay between theory and experimentation *within each experiment*. That is, our research designs encourage users to randomly divide texts and other data into a training and test set. While this split is familiar in machine learning and used throughout Chapter 5, we use it here for different reasons. First, if we already have a g in hand, we can use the training set to rule out any deviations from our anticipated data collection efforts. We can validate that what we intended to measure is actually measured by g. Once we have done that, we can then write down a well defined PAP and analyze the test data. Second, when we do not know the g before hand, we can use the training data to discover g, exploring our data to ensure that we uncover a coding scheme that is useful to our data set. With that coding scheme in hand, we can define a clear set of rules that are easily applied test data, without unexpected hiccups. The use of the train test split, as we will show in this chapter, ensures that we avoid issues of AISV, limits our opportunity to overfit, and helps us to constrain researcher degrees of freedom credibly. We argue that it is less wasteful than conducting a pre-analysis plan before collecting any data, as it allows researchers to define the most useful measures for the data, discover unanticipated patterns, and take into account the actual implementation of the experiment, while still ensuring that final inferences made from the data are credible.

Central to the discussion of these last two principles is g – how we discover it and when we fix it to make causal inference. As a result, we start this chapter by reviewing g – why we need it, how we discover it, and what properties we would like it to have. Because how we discover g is important to the validity of our causal inferences, this provides essential background for the remainder of the chapter on how to make causal inferences with text.

2 The g function

2.1 What is g and why do we need it?

The codebook function, g, is essential because the text is typically not usable for causal inferences in its *raw* form. As we have come back to time and again throughout this book, social scientists are often interested in some emergent property of the text—such as the topic that is discussed, the sentiment expressed, or the ideological position taken. Documents are high-dimensional, complicated, and sparse. The result is that distinct blocks of text can convey similar topics and sentiment. Reducing the dimensions of the text allows us to group texts and make inferences from our data.

Suppose we are interested in understanding how candidate biographies influence the popularity of a candidate. Each biography is unique, so we cannot meaningfully estimate the effect of any individual biography on a candidate's popularity—we observe only one biography per candidate. Instead, we are interested in some latent property of the text's effect on the popularity of the candidate, such as occupational background. In this example g might compress the text of the biography into an indicator of whether the candidate is a lawyer. The analyst could define g in numerous ways including hand-coding. g could also be defined automatically from the text, by looking for the presence or absence of the word "lawyer", or a group of words or phrases that convey that someone has a legal background, such as "JD", "attorney", and "law school". Being a lawyer is just one latent feature in the text. Different g's might measure if a candidate held prior office, went to college, or served in the military.

Our most consequential decision about g is the space we compress the text into. Options for this space could include discrete categories, proportions, or continuous variables (like ideal point estimates). We will call the lower-dimensional space $\mathcal{Z}$. Typically these low-dimensional representations are then given a label for interpretation. For example, we might use g to bin social media posts into "positive," "negative," or "neutral," or, put portions of documents into topics that we label "Sports," "Weather," or "Politics."

Social scientists working on text as data have adopted this compression approach, although the low-dimensional representation is often only implicit (Laver, Benoit and Garry, 2003; Grimmer, Messing and Westwood, 2012; Spirling, 2012; Catalinac, 2016). We can also think of g as the *codebook function* because it plays the role of a codebook in a manual content analysis, describing a procedure for *organizing* the researcher's texts in some systematic way. g takes on a central role because it connects the raw text to the underlying property that the researcher cares about. While applied work on measurement often describes the categories under study, discussion of the implications of g as an object of interest is rare. Nevertheless, g is always implicitly present in any systematic analysis of text—any instance where a set of documents is placed into a common set of categories or is assigned a common set of properties. Once a researcher decides on and estimates g, then text is usually ready to be used in statistical analysis.

2.2 Discovering g

While g is necessary to make causal inference, rarely is it determined exactly from theory or prior research. Even in manual content analysis (Krippendorff, 2004; Neuendorf, 2016), researchers typically read at least a portion of the documents to write a codebook that determines how coders should put documents into the categories of interest. More recently, a wide array of machine learning methods are used to discover g from the data, like those that we covered in the discovery chapter (Blei, Ng and Jordan, 2003; Hopkins and King, 2010). These newly discovered categories can help shape research questions, identify powerful textual interventions, and capture text-based outcomes.

Manual content analysts have been creating text-based g functions for years. Content analysis handbooks such as Krippendorff (2004) or Neuendorf (2016) advocate the use of codebooks which are essentially codifications of manual g functions. A manual content analysis scheme has many benefits. For example, there are many features of text which humans are quite adept at detecting, but computers find more challenging (e.g. humor and sarcasm) and when these are the quantity of interest a manual content analysis scheme might be preferable. Because the investigator has more freedom to choose the component of text that interests them, they arguable have greater freedom to choose features of theoretical interest (although they are still bound by the capacity of human coders).[2]

In spite of its central role across forms of text analysis, social scientists rarely discuss the process of discovery that lead to a particular codebook that is then subsequently used for causal inference. In practice, these coding schemes are developed through iteration between coding rules and the documents to be coded. We raise two main points about the discovery of g when making causal inferences with texts that apply regardless of the methodology applied.

1) We can (and often do) learn g from the data. There are three strategies for learning g from the data. First, we could read a sample of text. In manual content analysis, g often relies on some familiarity with the text or reading a sample of documents to decide how the text should map into categories. Second, we could use a method to classify texts into categories using hand coded examples for training. Supervised methods, which are conceptually similar to manual content analysis, use statistical and algorithmic methods

[2] Indeed the aspirations of investigators exceeding the capacity of human coders has led to trenchant critiques of the content analysis enterprise by those concerned that the coding is not able to capture the nuance of the underlying text and thus inadvertently 'invents facts' (Biernacki, 2012).

attempting to estimate the best g from hand coded or otherwise labeled documents. Last, unsupervised learning discovers a low-dimensional representation and assigns documents to that representation.

2) There is no single correct g. As we discussed in Chapter 4, regardless of the methods used in discovery, the analyst chooses a g on the basis of their theoretical question of interest. Different theories imply different organizations of the text and, therefore, different g's. However, we can and *should* evaluate g once we have defined a question of interest. Given a particular function and a particular purpose, we can label the identified latent features, the scales measured, and the classification accuracy, as we did in the measurement chapter. The *post hoc* validation of g provides clarity for both the researcher and the reader to correctly interpret the underlying latent features (Grimmer and Stewart, 2013). Our goal in the validation is to ensure that the interpretation implicit in our theoretical argument arises from and corresponds with the mapping in our chosen g.

2.3 Desirable Properties of g

Although there is no application-independent correct g for causal inference, once we have a question of interest, there are properties of g that are useful: interpretability, theoretical interest, label fidelity, and tractability.

Property 1: Interpretability First, g should be *interpretable.* To claim that a measure is theoretically interesting, we have to interpret it. Interpretability is research and text specific, but our articles must communicate to the reader what the measure in a specific study is capturing. This is particularly important for g's discovered from text data, which are based on underlying covariances in the data and thus will not necessarily be interpretable.

Property 2: Theoretical Interest The codebook function should also create measures of *theoretical interest*. We want to find low-dimensional representations of text that operationalize concepts from a theory and identify causal effects that test observable implications of the theory. Ideally, we would like to focus on large magnitude causal effects. All else equal, larger effects help us to explain more of the behavior of theoretical interest.

Property 3: Fidelity We also want to choose functions with high *fidelity* between the label we give to the components of g and the text it is compressing. Establishing fidelity involves producing evidence that the latent variable $\boldsymbol{z}$ accurately captures the property implied by the label. This is a common exercise in the social sciences; there is always an implicit mapping between the labels we use for our variables and the reality of what our labels measure. For text analysis, we think of maximizing label fidelity as minimizing the surprise that a reader would have in going from the label to reading the text. Fidelity is closely connected to the literature on validity in measurement and manual content analysis (see e.g., Grimmer and Stewart, 2013; Quinn et al., 2010; Krippendorff, 2004).

Property 4: Tractable Finally, we want the development and deployment of g to be *tractable*. In the context of manual content analysis this means the codebook can be applied accurately and reliably by human coders and that the number of documents to be coded is feasible for the resources available. In the case of learning g statistically, tractability implies that we have a model which can be estimated using reasonable computational resources and that it is able to learn a useful representation with the number of documents we possess.

There is an inherent tension between the four properties. This is most acute with the tension between theoretical interest and label fidelity. It is often tempting to assign a very general label even though g is more specific. This increases theoretical relevance, but lowers fidelity. The consequence can be research that is more difficult to replicate. Alternatively, we might have a g that coincides with a label because it increases the chances that our result can be replicated. But this could reduce the theoretical interest.

The analog of g lurks in every research design, including those that use standard data. Invariably when making an argument the researcher needs to find empirical surrogates or operationalized the concepts in her theoretical argument. For example, every time a researcher uses gross domestic product (GDP) as a stand-in for the size of the economy, she is projecting a high-dimensional and complicated phenomenon—the economy—into a lower-dimensional and more tractable variable—GDP. The causal estimand is defined in terms of its effect on GDP, but the theoretical argument is made about the size of the economy. While there is no correct measure to use for the economy, the reader can and should still interrogate the degree to which the chosen measure appropriately captures the broader theoretical concept that the researcher wants to speak to.

3 The Problem of Causal Inference with g

Text is high-dimensional, so we use the codebook function, g, to learn a low-dimensional representation to make inferences. But using g to compress text introduces new problems for causal inference. In this section we explain how g facilitates causal inference with text and then characterize the problems it creates.[3] In Section 3.1 we place g in the traditional causal inference setting. Section 3.2 explains how the use of g leads to the problems of an *analyst induced SUTVA violation* and *overfitting.*

[3] At a technical level we can think of an experiment with the process of discovery as a form of data-adaptive estimation (van der Laan, Hubbard and Pajouh, 2013), a framework which originates from biostatistics and describes circumstances where our target estimation is not fixed in advance.

3.1 Causal inference with g

To begin, we review potential outcomes notation and assumptions used when there is no text or dimensionality reduction and we are analyzing a unidimensional treatment and outcome (Imbens and Rubin, 2015). Denote our dependent variable for each unit i ($i \in 1, 2, \ldots, N$) with Y_i, the treatment condition for unit i will be T_i. We define the space of all possible outcomes as $\boldsymbol{\mathcal{Y}}$ and the space of all possible treatments as $\boldsymbol{\mathcal{T}}$. When the treatment is binary we refer to $Y_i(1)$ as the potential outcome for unit i under treatment and $Y_i(0)$ as the potential outcome under control and the individual causal effect (ICE) for unit i is given by $\text{ICE}_i = Y_i(1) - Y_i(0)$. Our typical estimand is some function of the individual causal effects such as the average treatment effect (ATE), $E[Y_i(1) - Y_i(0)]$.

To identify the average treatment effect using a randomized experiment we make three key assumptions. First, we assume that the response depends only on the assigned treatment, often called the Stable Unit Treatment Value Assumption (SUTVA). Specifically:

Assumption 1 (SUTVA). *For all individuals i, $Y_i(T) = Y_i(T_i)$.*

Second, we will assume that our treatment is randomly assigned:

Assumption 2 (Ignorability). $Y_i(t) \perp\!\!\!\perp T_i$

Third,we will assume that every treatment has a chance of being seen:

Assumption 3 (Positivity). *$Pr(T_i = t) > 0$ for $t \in \mathcal{T}$.*

The second and third assumptions are guaranteed by proper randomization of the experiment whereas the first is an assumption that is generally understood to mean that there is no interference between units and no hidden values of treatment. For each observation we observe only a single potential outcome corresponding to the realized treatment.

Building off of this notation, we can introduce mathematical notation to cover high-dimensional text and the low-dimensional representation of texts derived from g that we will use for our inferences. We start by extending our notation to cover multi-dimensional outcomes, $\boldsymbol{Y}_i$, and multi-dimensional treatments, $\boldsymbol{T}_i$. We will suppose, for now, that we have already determined g, the codebook function. Recall g is applicable regardless of whether the coding is done by a machine learning algorithm, a team of undergraduate research assistants or an expert with decades of experience.

We write the set of possible values for the mapped text as $\mathcal{Z}$ with a subscript to indicate if it is the dependent variable or treatment. We denote the realized values of the low-dimensional representation for unit i as $\boldsymbol{z}_i$ $(i = 1, \ldots, N)$. We suppose that when the outcome is text $g : \mathcal{Y} \to \mathcal{Z}_Y$ and $g(\boldsymbol{Y}_i) = \boldsymbol{z}_i$, and when the treatment is text $g : \mathcal{T} \to \mathcal{Z}_T$ and $g(\boldsymbol{T}_i) = \boldsymbol{z}_i$. The set $\mathcal{Z}$ is a lower-dimensional representation of the text and can take on a variety of forms depending upon the study of interest. For example, if we are hand coding our documents into two mutually-exclusive and exhaustive categories, then $\mathcal{Z}$ is $\{0, 1\}$. If we are using a mixed-membership topic model to measure the prevalence of K topics as our dependent variable, then $\mathcal{Z}$ is a $K - 1$ dimensional simplex. And if we are using texts as a treatment, we might suppose that $\mathcal{Z}$ is the set of K binary feature vectors, representing the presence or absence of an underlying treatment (see Egami et al. (2018) for the reason we prefer binary treatments, though continuous treatments also fit within our framework). There are numerous other types of g that we might use—including latent scales, dictionary-based counts of terms, or crowd-sourced measures of content. The only requirement for g is that it is a function.

We next use g to write our causal quantity of interest in terms of the low-dimensional representation. To make this concrete, consider a case where we have a binary non-text treatment and a text-based outcome (we consider other causal estimands below). Suppose we hand code each document into one of K categories such that for unit i we can write the coded text under treatment as $g(\boldsymbol{Y}_i(1)) = \boldsymbol{z_i}(1)$. We can then define the average treatment effect for category k to be:

$$\begin{aligned} \text{ATE}_k &= E[g(\boldsymbol{Y}_i(1))_k - g(\boldsymbol{Y}_i(0))_k] \qquad (3.1) \\ &= E[z_{i,k}(1) - z_{i,k}(0)] \end{aligned}$$

where $z_{i,k}(1)$ indicates the value of the k-th category, for unit i, under treatment.

3.2 The Problems: Identification and Overfitting

Equation 3.1 supposes that we already have a g in hand. As we mentioned above, g is often discovered by interacting with some of the data, either by reading or through machine learning. To describe this problem more clearly, we denote the set of documents considered in development of g as $\boldsymbol{I}$ and write $g_{\boldsymbol{I}}$ to indicate the dependence of g on the documents. Problems of identification and estimation arise where the set of documents used to develop g, $\boldsymbol{I}$, overlaps with the set of documents used in estimation which we will call $\boldsymbol{O}$. There are two broad concerns: an identification problem arising from an *Analyst Induced SUTVA Violation* (AISV) and an estimation problem with overfitting.

3.2.1 Identification concerns: Analyst Induced SUTVA Violations

If Assumption 1 holds then each observation's response does not depend on other units' treatment status. But even when Assumption 1 holds, when we discover $g_{\boldsymbol{I}}$, we can create a dependence across observations in $\boldsymbol{I}$ because the particular randomization may affect the $g_{\boldsymbol{I}}$ we estimate. This violation occurs because the treatment vector $\boldsymbol{T}_{\boldsymbol{I}}$ – the treatment assignments for all documents $\boldsymbol{I}$– affects the g that we obtain, inducing dependence across *all* observations in $\boldsymbol{I}$. If we then try to use the documents in $\boldsymbol{I}$ for estimation of the effect, we have violated SUTVA. This violation is induced by the analyst in the process of discovering g, which is why we call it an *Analyst* induced violation. Egami et al. (2018) provides a technical definition of the AISV.

To see how the AISV works in practice, consider a stylized experiment on four units with a dichotomous intervention defining a treatment and control condition and a text-based outcome. We might imagine potential outcomes that have a simple relationship between treatment and the text-based outcome such as the one shown in Table 1. Treated units talk about Candidate Morals and Polarization and control units talk about Taxes and Immigration.

	Treated	Control
Person 1	Candidate Morals	Taxes
Person 2	Candidate Morals	Taxes
Person 3	Polarization	Immigration
Person 4	Polarization	Immigration

Table 1: A stylized experiment indicating the potential outcomes of a textual response.

Using Table 1 we can imagine the properties of an estimator applied to this text-based experiment as we rerandomize. Suppose that for each randomization we decide on both the form of g and estimate the treatment effect given g. For example, if randomization results in the treatment vector (1,1,0,0), then we would observe only two of the four categories: morals and immigration. A reasonable g might compress the text based responses to two variables: an indicator variable for discussing morals and an indicator variable for discussing immigration. If we randomize again and obtain (1,0, 1, 0) we observe all four categories. In this case, g might map the text based responses to a four-element long vector, with an indicator for whether each distinct category is discussed in the response. Under a third randomization (0,0,1,1) we are back to only two categories: taxes and polarization; so g might be two bivariate indicator variables, with the categories corresponding to whether someone discussed taxes or not or polarization or not.

As we randomize we estimate new g's with different categories. This lack of *category stability* complicates our ability to analyze our estimators as we traditionally do, using a framework based on re-randomization. We take this category and classification stability for granted in standard experiments because categories are supposed to be fixed before the experiment. But when we estimate categories from data the discovered g depends on the randomization and therefore dependence between units is induced by the analyst. And even if we fix the categories, as we might do with a supervised model, different randomizations may lead to different rules for assigning documents to categories, leading to a lack of *classification stability*. Or, unanticipated developments in the implementation of the experiment may require us to respecify g after looking at the data to truly capture the quantity we are interest in. If, however, we fix g before estimating the effects, the problem is solved.

3.2.2 Estimation concerns: Overfitting

Even if we assume away the AISV, estimating g means that researchers might *overfit*: discover effects that are present in a particular sample but not in the population. This is a particular risk when researchers are searching over different g's to find those that best meet the criteria of interpretability, interest, fidelity and tractability. The overfitting problem is particularly acute when a researcher is fishing — searching over g's to obtain statistical significance or estimates that satisfy a related criterion. But overfitting can occur even if researchers are conducting data analysis without ill-intentions. This happens because following best practice with almost all available text as data methods requires some iteration. With hand coding iteration refines the codebook, with supervised models it occurs when we refine a classifier, and with unsupervised methods it happens as we adjust parameters to examine new organizations.

Fishing and overfitting are a problem in all experimental designs and not just those with text. The problem of respecifying g until finding a significant result is analogous to the problem of researchers recoding variables or ignoring conditions in an experiment, which can lead to false-positive results. (Simmons, Nelson and Simonsohn, 2011).The problem with text-based inferences is heightened because texts are much more flexible than other types of variables, creating a much wider range of potential g's. This wider range increases the risk of overfitting, even amongst well-intentioned analysts. Overfitting is also likely in texts because it is so easy to justify a particular g after the fact – the human brain is well-equipped to identify and justify a pattern in a low-dimensional representation of text, even if that pattern emerges merely out of randomness. This means that validation steps alone may be insufficient safeguard against overfitting, even though texts provide a rich set of material to validate the content.

4 A Train/Test Split Procedure for Valid Causal Inference with Text

To address the identification issues caused by the AISV and the estimation challenges of overfitting, we must break the dependence between the discovery of g and the estimation of the causal effect. The most straightforward approach is to define g before looking at the documents. Defining the categories beforehand, however, limits our coding scheme, excluding information about the language used in the experiment's interventions or what units said in response to a treatment. If we define our codebook before seeing text we will miss important concepts and have a poorer measure of key theoretical concepts.

We could also assume the problem away. Specifically, to eliminate the AISV it is sufficient to assume that the codebook that we obtain is invariant to randomization. Take for example the text as outcome case; if over different randomizations of the treatment the g we learned does not change, we don't have an AISV.

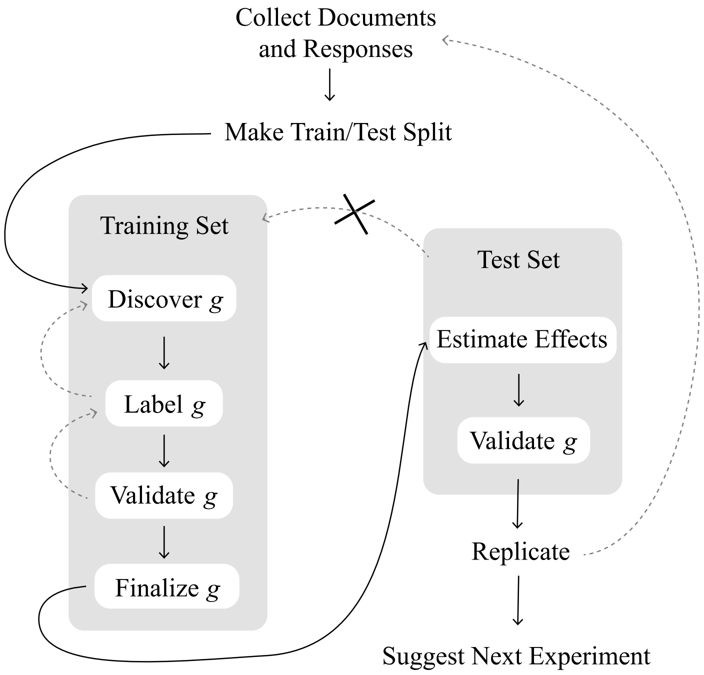
Our preferred procedure is to explicitly separate the creation of g and the estimation of treatment effects. This procedure avoids the AISV and provides a natural check against overfitting. To explicitly separate the creation of the codebook and its application to estimate effects, we randomly divide our data into a training set and a test set. Specifically, we randomly create a set of units in a training set denoted by the indices $\boldsymbol{I}$ and a non-overlapping test set denoted by the indices $\boldsymbol{O}$. We use only the training set to estimate the $g_{\boldsymbol{I}}$ function and then discard it. We then use the test set exclusively to estimate the causal effect on the documents in $\boldsymbol{O}$.

This division between the training and test set addresses both the identification and estimation problems. It avoids the AISV in the test set because the function g does not depend on the randomization in the test set, so that each test set unit's response depends only on its assigned treatment status. There is still a dependence on the training set observations and their treatment assignment. This, however, is analogous to the analyst shaping the object of inquiry or creating a codebook after a pre-test. With the AISV addressed, it is now possible to define key properties of the estimator, like bias or consistency.

The sample split also addresses the concerns about overfitting. The analyst can explore in the training set as much as she likes, but, because findings are verified in a test set that is only accessed once, she is incentivized to find a robust underlying pattern. Patterns in the training set which are due to idiosyncratic noise are highly unlikely to also arise in the test set which helps assure the analyst that patterns which are confirmed by the separate test set will be replicable in further experiments. By locking g into place in the training set, the properties of the tests in the test set do not depend upon the number of different g's considered in the training set. In practice, we find splitting the sample ensures that we are able to consider several models to find the g that best captures the data and aligns with our theoretical quantity of interest without worrying about accidentally p-hacking.

The training/test split in our research design brings the quantitative analysis of text in line with many best practices used by manual content analysts when they hand code documents. For example, a common practice is to run pilot-tests on a subset of data in order to develop a codebook or to refine particular text-based treatments. In other settings

Figure 1: Our Procedure for Text-Based Causal Inferences



scholars doing handcoding in experiments will develop a codebook on a subset of data and discard those data, using the codebook to classify the remaining documents for analysis. Our approach extends the logic of these well-established practices to instances where g is learned using machine learning methods.

4.1 Procedure

In this section we discuss the general procedure for implementing the train/test split to estimate the above quantities of interest. This procedure follows the schematic in Figure 1. Considerations specific to treatment or outcome are deferred to Sections 5 and 6 below.

4.1.1 Splitting the sample

The first major choice that the analyst faces is how to split the sample into two pieces: the training set and the test set. A default recommendation is to split 50% of the documents in training and 50% in the test set. But this depends on how the researcher evaluates the tradeoff between discovery of g and testing. Additional documents in the training set enables learning a more complicated g or more precise coding rules. Additional documents in the test set enable estimation of a more precise effect. While the test set should be representative of the population that you want to make inference about, the training set can draw on additional non-representative documents as long as they are similar enough to the test set to aid in learning a useful g. Finally, when taking the sample the analyst can stratify on characteristics of interest to ensure that the split has appropriate balance between the train and test set on those characteristics.

Once the test set is decided, the single most important rule is that the test set is used once, solely for estimation. If the analyst revises g after looking at the test set data, she reintroduces the AISV and risks overfitting. Setting aside test data must be true for all features of the analysis: even preliminary steps like preprocessing must not include the test data set. Third parties, such as survey firms and research agencies, could be helpful in credibly setting the data aside.

4.1.2 Discover g

We use the training set and text as data methods to find a g that is interpretable, of theoretical interest, has high label fidelity and is tractable. In this chapter we use the Structural Topic Model (Roberts, Stewart and Airoldi, 2016) and the Supervised Indian Buffet Process (Fong and Grimmer, 2018*a*) but there are numerous other methods that are applicable.

4.1.3 Validation in the training set

Validation is an important part of the text analysis process and researchers should apply the normal process of validation to establish label fidelity. These validations are often application-specific and draw on close reading of the texts. In Chapters 4 and 5 we have described the types of validation and the `stm` package (Roberts, Stewart and Tingley, 2017) provides tools designed to assist with validation. These validations should be completed in the training set as part of the process of discovering and labeling g, before the test set is opened.

4.1.4 Before opening the test set

While obtaining g in the training set, we can refit g as often as it is useful for our analysis. But once applied to the test set we cannot alter g further. Therefore, we advise two final steps.

1) Take One More Look at g

   Be sure g is capturing the aspect of the texts that you want to capture, assign labels and then validate to ensure that the conceptual gap between those labels and the representation g produces is as small as possible. While validation approaches may vary- this necessarily involves reading documents (Krippendorff, 2004; Quinn et al., 2010; Grimmer and Stewart, 2013). It is helpful to fix a set of human assigned labels, example documents and automated keyword labels in advance to avoid subtle influence from the test set.

2) Fix Your Evaluation Plan

   While we focus on inference challenges with g, standard experimental challenges remain. Here we can draw from the established literature on best practices in experiments (Gerber and Green, 2012) potentially including a pre-analysis plan (Humphreys, Sanchez de la Sierra and Van der Windt, 2013). This can include multiple-testing and false-discovery rate corrections.

4.1.5 Applying g and estimating causal effects

Mechanically, applying g in the test set is straightforward and is essentially the process of making a prediction for a new document. After calculating the quantities $g_{\boldsymbol{I}}(\mathbf{Y}_{\boldsymbol{O}})$ we can use standard estimators appropriate to our estimand, such as the difference of means to estimate the average treatment effect.

4.1.6 Validation in the test set

It is also necessary to ensure that the model fits nearly as well on the test set as it did on the training set. When both the training and test sets are random draws from the same population this will generally be true. But overfitting or a small sample size can result in different model fit. The techniques used to validate the original model can be used in the test set as well as common measures of model fit such as log likelihood. Unlike the validation in the training set, during the validation in the test set the analyst cannot return to make changes to the model. Nevertheless, validation in the test set helps the analyst understand the substantive meaning of what is being estimated and provides guidance for future experiments. Formally, our estimand is defined in terms of the empirically discovered g in the training set. However, invariably the analyst making a broader argument indicated by the label. Validation in the test set verifies that *label fidelity* holds and that g represents the concept in the test set of documents.

4.2 Tradeoffs

We have used splits throughout the book to simulate the sequential nature of social science inferences. And like other settings, splitting our data addresses several research problems, but it does cause some new concerns about the inferences we can make. Efficiency loss is the biggest concern. In a 50/50 train-test split, half the data is used in each phase of analysis, implying half the data is excluded from each step. At the outset, it is difficult to assess how much data is necessary for either the training or the test set. The challenge in setting the size of the test set is that the analyst does not yet know what the outcome (or treatment) will be when the decision is made on the size of the split. The problem in setting the size of the training set is that *we don't know the power we need for discovery.* Alternatively, we could focus first on determining the power needed for estimation of an effect and then allocate the remaining data for discovery. This can be effective, but it requires that we are able to anticipate characteristics of our discovered treatment or outcome.

5 Text as a Dependent Variable

In the next two sections, we go over two cases – when text is the dependent variable in a causal analysis and when text is an independent variable in a causal analysis. We provide experimental examples of each of these cases and discuss how the train-test split avoids overfitting and AISV in each case. From there, we move on to considering text in observational studies, in particular when text is a confounder.

The text as outcome setting is analytically straightforward. The particular g that the analyst chooses defines the categories of the outcome from which the estimand will be defined. Our goal is to obtain a consistent (and preferably unbiased) estimator for the ATE (or other causal quantities of interest) assuming a particular g. Using Assumptions 1-3, a consistent estimator will be:

$$\widehat{ATE} \quad = \quad \sum_{i \in \boldsymbol{O}} \frac{I(T_i = 1) g_{\boldsymbol{I}}(\boldsymbol{Y}_i(1))}{\sum_{i \in \boldsymbol{O}} I(T_i = 1)} - \sum_{i \in \boldsymbol{O}} \frac{I(T_i = 0) g_{\boldsymbol{I}}(\boldsymbol{Y}_i(0))}{\sum_{i \in \boldsymbol{O}} I(T_i = 0)}$$

When g is fixed before documents $\boldsymbol{O}$ are examined, we can essentially treat the mapped outcome $g_{\boldsymbol{I}}(\boldsymbol{Y_O})$ as an observed variable.[4] See Egami et al. (2018) for the identification proof.

5.1 An experiment on immigration

To first demonstrate how to use text as a response in a causal inference framework, we apply the structural topic model to open-ended responses from a survey experiment on immigration (Roberts et al., 2014). Specifically, we build on an experiment first introduced in Cohen, Rust and Steen (2004) to assess how knowledge about an individual's criminal history affects respondent's preference for punishment and deportation. These experimental results contribute to a large literature about Americans' preferences about immigrants and immigration policy (see Hainmueller and Hopkins 2014 for a review) and a literature on the punishments people view as appropriate for crimes (Carlsmith, Darley and Robinson, 2002). Critically, in both conditions of our experiment an individual has broken the same law, entering the country illegally, but differs solely on past criminal history. We therefore ask how someone's past criminal behavior affects the public's preference for future punishment and use the open-ended responses to gather a stated reason for that preference.

To address this question we report the results from three iterations of a similar experiment. With each experiment we report our procedure for choosing g and the treatment effects in order to provide clarity and to demonstrate how the process described in Figure 1 works in practice. The first results are based on responses initially recorded in Cohen, Rust and Steen (2004). We use this initial set of responses to estimate an initial g and to provide

[4] It is still important to verify that the mapped variable is capturing what you care about the underlying text. Ultimately this is not any different than ensuring that a chosen outcome for an experiment captures the phenomenon of interest to the researcher.

baseline categories for the considerations respondents raise when explaining why someone deserves punishment. In a second experiment we build on Cohen, Rust and Steen (2004), but address issues in the wording of questions, expand the set of respondents who are asked to provide an open ended response, and update the results with contemporary data. We then run a third experiment because we discovered our g performed poorly in the test set of the second experiment. We also used that opportunity to improve small features of the design of the experiment.

We report the results of each experiment in order to be transparent about our research process, something we suggest that researchers do in order to avoid selective reporting based on an experiment's results. The three sets of experimental results show that there has been surprising stability in the considerations Americans raise when explaining their punishment preferences, though there are some new categories that emerge. There is also a consistent inclination to punish individuals who have previously committed a crime, even though they committed the same crime as someone without a criminal history.

5.1.1 Experiment 1

As a starting point, we conduct an analysis of the results of an experiment reported in Cohen, Rust and Steen (2004). The survey experiment was administered in the context of a larger study of public perceptions of the criminal justice system. The survey was conducted in 2000 by telephone random-digit dial and includes 1,300 respondents.[5]

In the experiment, respondents were given two scenarios of a criminal offense. In both the treatment and control conditions, the same crime was committed: illegal entry to the United States. In the treatment condition, respondents were told that the person had previously committed a violent crime and had been deported. In the control condition, respondents

[5] More details about the survey are available in Cohen, Rust and Steen (2002).

were told that the person had never been imprisoned before.

In the treatment condition, respondents were told:

> “A 28-year-old single man, a citizen of another country, was convicted of illegally entering the United States. Prior to this offense, he had served two previous prison sentences each more than a year. One of these previous sentences was for a violent crime and he had been deported back to his home country.”

In the control condition, respondents were told:

> “A 28-year-old single man, a citizen of another country, was convicted of illegally entering the United States. Prior to this offense, he had never been imprisoned before.”

Respondents were then asked a close-ended question about whether or not the person should go to jail. If they responded that the person should not go to jail, they were asked to respond to an open-ended question, “Why?” The key inferential goal of the initial study was determining if a respondent believed a person should be deported, jailed, or given some other punishment.

5.1.2 Experiment 2

After analyzing the results of Experiment 1, we ran a second experiment using the same treatment and control conditions, but with slight design differences to build upon and improve the original experimental protocol. First, all respondents were asked the open-ended question, not just those who advocated for not sending the individual to jail. Second, we redesigned the survey to avoid order effects. Third, we asked a more specific open-ended question. We still asked 'Should this offender be sent to prison?' (responses: yes, no, don't know) but followed by asking "Why or why not? Please describe in **at least two sentences** what actions if any the U.S. government should take with respect to this person and why?"[6] Experiment 2 was run on Mechanical Turk on July 16, 2017 with 1000 respondents.

5.1.3 Experiment 3

We expected Experiment 2 to be our last experiment, but we encountered a design problem. After we estimated g in the training set using STM and fit it to the test data, we realized that some of our topic labels were inaccurate. In particular, we had attempted to label topics using three pre-determined categories: prison, deport, and allow to stay. But the data in the second experiment suggested some new categories. We could not simply relabel the topics in the test set, because this would eliminate the value of the train/test split. Instead we verified the results of experiment 2 with an additional experiment.[7] Experiment 3 was run on Mechanical Turk on September 10, 2017 with 1000 respondents. To avoid labeling mistakes, two members of our team labeled the topics independently using the training data and then compared labels with one another to create a final set of congruent labels before

[6] Per our IRB we added the statement "(Please **do not** include any identifying information such as your name or other information about you in this open-ended response.)"

[7] We also took the opportunity to make a few design changes. We had previously included an attention check which appeared after the treatment question. We moved the attention check to before the treatment. We also had not previously used the MTurk qualification enforcing the location to be in the U.S. although we did in Experiment 3. Finally, we blocked workers who had taken the survey in Experiment 2 using the `MTurkR` package (Leeper, 2017).

applying the g to the test set.

5.1.4 Results

In each experiment, we used equal proportions of the sample in the train and test sets. In each experiment we fit several models in the training set before choosing a single model that we then applied to the test set.

We include the results from all three experiments below.For Experiment 3, Table 2 shows the words with the highest probability in each of 11 topics and the documents most representative of each topic, respectively. Topics range from advocating for rehabilitation or assistance for remaining in the country to suggesting that the person should receive maximal punishment.

	Label	Highest Probability Words
Topic 1	Limited punishment with help to stay in country, complaints about immigration system	legal, way, immigr, danger, peopl, allow, come, countri, can, enter
Topic 2	Deport	deport, think, prison, crime, alreadi, imprison, illeg, sinc, serv, time
Topic 3	Deport because of money	just, send, back, countri, jail, come, prison, let, harm, money
Topic 4	Depends on the circumstances	first, countri, time, came, jail, man, think, reason, govern, put
Topic 5	More information needed	state, unit, prison, crime, immigr, illeg, take, crimin, simpli, put
Topic 6	Crime, small amount of jail time, then deportation	enter, countri, illeg, person, jail, deport, time, proper, imprison, determin
Topic 7	Punish to full extent of the law	crime, violent, person, law, convict, commit, deport, illeg, punish, offend
Topic 8	Allow to stay, no prison, rehabilitate, probably another explanation	dont, crimin, think, tri, hes, offens, better, case, know, make
Topic 9	No prison, deportation	deport, prison, will, person, countri, man, illeg, serv, time, sentenc
Topic 10	Should be sent back	sent, back, countri, prison, home, think, pay, origin, illeg, time
Topic 11	Repeat offender, danger to society	believ, countri, violat, offend, person, law, deport, prison, citizen, individu

Table 2: Experiment 3: Topics and highest probability words

After discovering, labeling, and finalizing g in the training set, we estimated the effect of treatment on the topics in the test set. In Figure 2 we show large impacts of treatment on topics. Treatment (indicating that the person had a previous criminal history) increased the amount of writing about maximal punishment, deportation, and sending the person back to

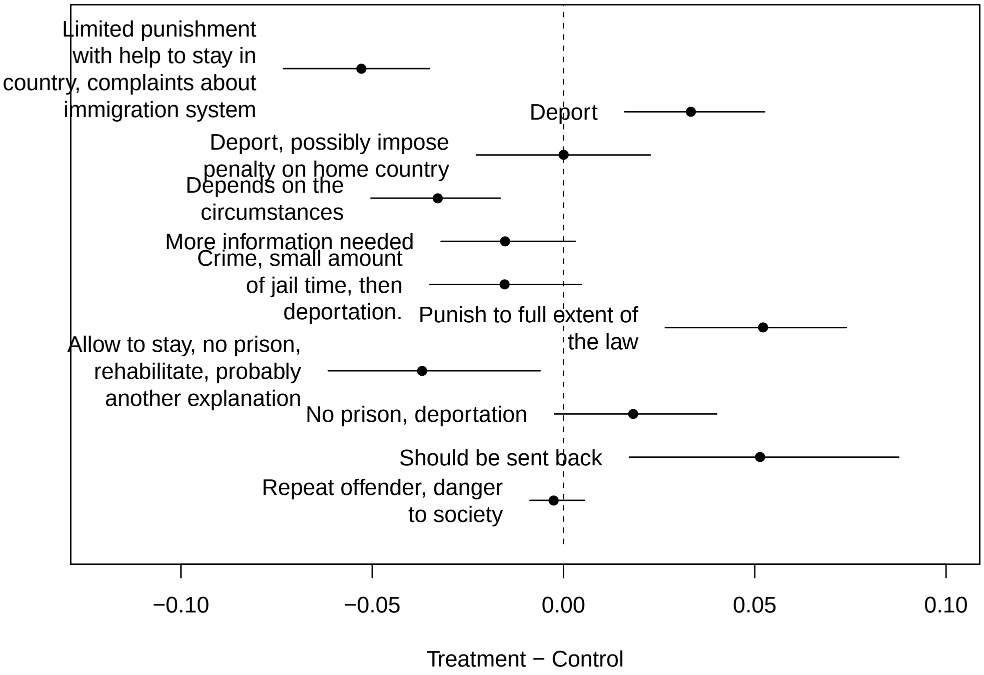


Figure 2: Test Set results for Immigration Experiment 3. Point estimates and 95% confidence intervals.

their country of origin. The control group was more likely to advocate that the person should be able to stay in the country or that the punishment should depend on the circumstances of the crime.

We found qualitatively similar results in Experiments 1 and 2 (Figure 3), even though g is different in both cases and the set of people who were asked to provide a reason is different. In each case, the description of a criminal history significantly increases the likelihood that the respondent advocates for more severe punishment or deportation.

5.2 The Effect of Presidential Public Appeals

In our second example or using text as a dependent variable, Franco, Grimmer and Lim (2018) examine the effect of presidents making public appeals during prime time television

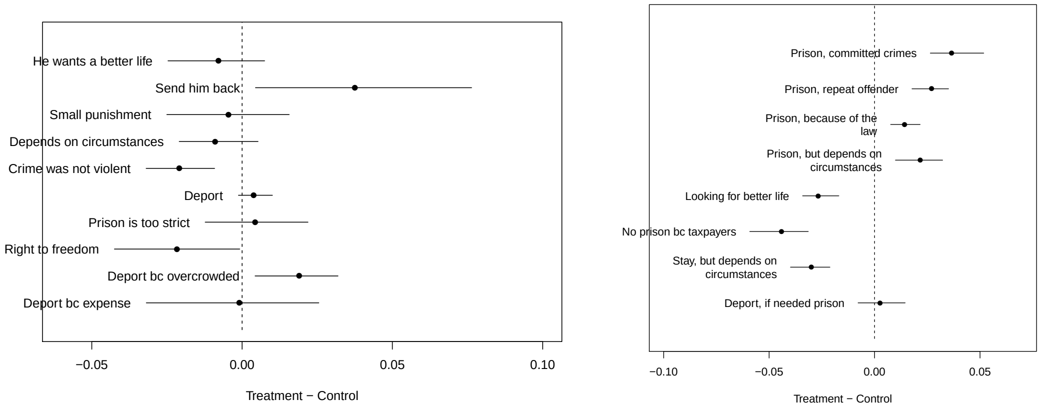


Figure 3: Test Set results for Experiment 1 (left) and Experiment 2 (right). Point estimates and 95% confidence intervals.

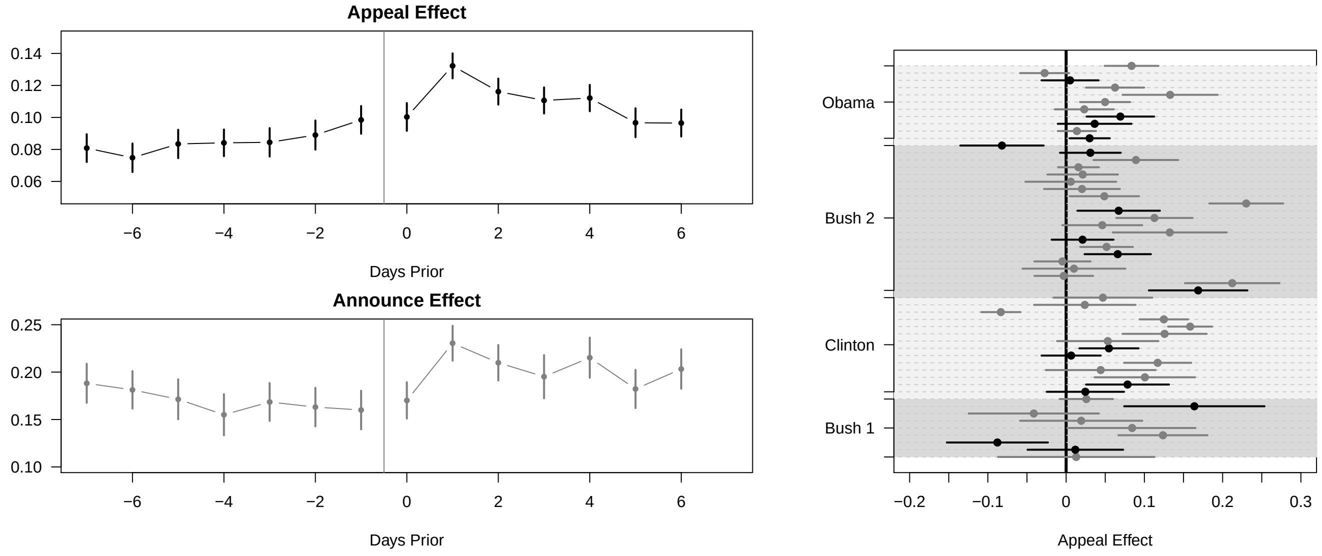
on subsequent media coverage. A large literature presents mixed results about the ability of presidents to affect public opinion and how they are covered in the news. One reason for the mixed results is that prior studies use different research designs and shifting types of evidence to asses the effect of presidential appeals. To isolate the effect of the appeals, Franco, Grimmer and Lim (2018) restrict attention to presidential public statements made during prime time television coverage. They then collect coverage of the president from the top ten newspapers in the week before and week after the speech, along with broadcast news transcripts from the same time period. The result is a collection of nearly 160,000 stories that mention the president, collected across the president's public appeals.

Franco, Grimmer and Lim (2018) use these data to ask the effect presidents have on the type of coverage they receive from the media. To make this assessment Franco, Grimmer and Lim (2018) use a structural topic model to estimate a set of topics, using 10% of the news stories and transcripts as a training set and the remaining 90% as a test set. This split ensures there is enough information in the training set to discover the salient topics and enough content in the test set to infer if their is a treatment effect.

Given this topic fit Franco, Grimmer and Lim (2018) then identify the salient topic of each presidential public appeal. Using this identified salient topic, they then estimate the effect of the president's speech on this salient topic within the subsequent news media.

The top plot in Figure 4 shows the effect of the president's speech on the topic coverage in two different situations: 1) when the president makes an appeal to Congress (the black top lines) and 2) when the president makes an announcement that conveys that some major world event happens (the grey-bottom line). Using these percentage, Franco, Grimmer and Lim (2018) infer the effect of the speech by comparing the share of coverage in the day immediately before and after the speech was given.

Figure 4: The Effect of Presidential Public Appeals



This top plot shows that when presidents make public appeals they substantially increase coverage. Following an appeal to Congress, presidents receive 3.4 percentage points more coverage about the topic, compared to the rate immediately before the speech. The bottom plot shows why it can be problematic to assess the effect of presidential speech after he announces a major world event. After the president announces an event, there is a seven percentage point increase in the coverage of that topic in articles that mention the president. Of course, some of that increase is attributable to the event that occurred, rather than the president's speech, so the estimate conflates the president's speech with the additional coverage of that event in the world. Perhaps not surprisingly, we see that the increased coverage of announcements lasts longer than the appeals.

Franco, Grimmer and Lim (2018) also provide speech-by-speech estimates of the effect of appeals on subsequent news coverage. The right plot in Figure 4 shows the increase in coverage on a topic for announcements (grey) and appeals (black). This figure shows that while news coverage rises more after the president makes an announcement related to a world event, again, this may be because the major news story provides additional coverage outside of the president's speech. Further, this figure shows that presidents do not appear to get better at influencing media coverage over their presidency, later speeches, which are closer to the top of the Figure are not more effective than earlier speeches at the bottom of the figure.

6 Text as Intervention

Text can also be used as the treatment in an experiment. For example, we may ask individuals to read a candidate's biography and then evaluate how the candidate's favorability on a scale of 0 to 100. The treatment, $\boldsymbol{T}_i$, is the text description of the candidate assigned to the respondents. The potential outcomes $Y_i(\boldsymbol{T}_i)$ describes respondent i's rating of the candidate under the treatment assigned to respondent i.

While we could compare two completely separate candidate descriptions, social scientists are almost always interested in how some underlying feature of a document affects responses—that is the researcher is interested in estimating how an *aspect* or *latent* value of the text influences the outcome.[8] For example, the researcher might be interested in whether including military service in the description has an impact on the respondents' ratings of

[8] This distinguishes our framework from A/B tests commonly found in industry settings which evaluate different blocks of text without attempting to understand why there are differences across the texts.

the candidate. Military service is a latent variable – there are many ways that the text could describe military service that all would count as the inclusion of military service and many ways that the text could omit military service that all would count as the absence of the latent variable. The researcher might assign 100 different candidate descriptions, some which mention the candidate's military service and some which do not. In this case, the treatment of interest is $Z_i = g(T_i)$ which maps the treatment text to an indicator variable that indicates whether or not the text contains a description of the candidate's military service. To estimate the impact of a binary treatment, we could use the estimator:

$$\widehat{ATE} = \sum_{i \in \boldsymbol{O}} \frac{I(Z_i = g_{\boldsymbol{I}}(\boldsymbol{T}_i) = 1) Y_i(1))}{\sum_{i \in \boldsymbol{O}} I(Z_i = g_{\boldsymbol{I}}(\boldsymbol{T}_i) = 1)} - \sum_{i \in \boldsymbol{O}} \frac{I(Z_i = g_{\boldsymbol{I}}(\boldsymbol{T}_i) = 0) Y_i(0))}{\sum_{i \in \boldsymbol{O}} I(Z_i = g_{\boldsymbol{I}}(\boldsymbol{T}_i) = 0)}$$

With text as treatment, we may be interested in more than just one latent treatment. The presence of multiple latent treatments requires different causal estimands and enables us to ask different questions about how features of the text affect responses. For example, we can learn the marginal effect of military service and how military service interacts with other features of the candidate's background—such as occupation or family life. Typically with multidimensional treatments we are interested in the effect of one treatment holding all others constant. This complicates the use of topic models which suppose $\mathcal{Z}$ is a simplex (all topic proportions are non-negative and sum to one) because there is no straightforward way to change one topic holding others constant (see Fong and Grimmer 2016*a* and Egami et al. (2018)). Instead we will work with g that compress the text $\boldsymbol{T}$ to a vector of K binary treatments $\boldsymbol{Z}_j \in \mathcal{Z}$ where $\mathcal{Z}$ represents all 2^K possible combinations of the treatments. We could also, of course, suppose that g maps $\boldsymbol{T}$ to a set of continuous underlying treatments, but this requires additional functional form assumptions.

The use of binary features leads naturally to the *Average Marginal Component Effect* (AMCE), the causal estimand commonly used in conjoint experiments (Hainmueller, Hopkins and Yamamoto, 2013*a*). The AMCE estimates the marginal effect of one component k, averaging over the values of the other components:

$$AMCE_k = \sum_{\boldsymbol{Z}_{-k}} E[Y(Z_k = 1, \boldsymbol{Z}_{-k}) - Y(Z_k = 0, \boldsymbol{Z}_{-k})] m(\boldsymbol{Z}_{-k})$$

The $AMCE_k$ describes the average effect of component k, summed over all other values of k, weighted by $m(Z_{-k})$, or an analyst determined distribution of Z_{-k}. The AMCE can be thought of as an estimate of the effect of component k, averaging over the distribution of other components in the population—therefore providing a sense of how an intervention will matter averaging over other characteristics.

In order to discover the mapping from text to latent treatments we an additional assumption than in the text as outcome case. This is because analysts are usually only able to randomize at the text level, but we are interested in identifying the effect of latent treatments we are unable to manipulate directly. Consequently, we need to make an additional assumption beyond the three mentioned above in Section 3.1 (SUTVA, Ignorability and Positivity[9]).The Sufficiency Assumption states that our g captures all the information relevant to the response in $\boldsymbol{T}$ is contained in $\boldsymbol{Z}$

Fong and Grimmer (2018*b*) shows that for sufficiency to hold for any individual the response to two documents with the same latent feature representation might differ, but on average over individuals the responses are the same. Mathematically, it is written as:[10]

[9] To address the multidimensional treatments, the positivity assumption becomes the common support assumption which states that all combinations of treatments have non-zero probability $f(\boldsymbol{Z}_i) > 0$ for all $\boldsymbol{Z}_i \in \text{Range } g(\cdot)$.

[10] Fong and Grimmer (2016*a*) present a stronger and more intuitive version. Fong and Grimmer (2016*a*)

Assumption 4 (Sufficiency). *For all* $\boldsymbol{T}$ *and* $\boldsymbol{T}^{'}$ *such that* $g(\boldsymbol{T}) = g(\boldsymbol{T}^{'})$ *then* $E[Y_i(g(\boldsymbol{T}))] = E[Y_i(g(\boldsymbol{T}^{'}))]$.

Fong and Grimmer (2018*b*) shows that this assumption is equivalent to supposing that the components of the document that affect the response and are not included in the latent feature representation are orthogonal to the latent feature representation. Technically, we can define $\epsilon_i(T) = Y_i(T) - Y_i(g(T))$ and then this more general assumption is equivalent to assuming that $E_i[\epsilon_i(T)] = 0$ for all T. Fong and Grimmer (2016*a*) and Fong and Grimmer (2018*b*) provide an identification proof.

6.1 Text as treatment: Consumer Financial Protection Bureau

We now provide an applied example of how to use text as treatment. We examine the features of a complaint that causes the Consumer Financial Protection Bureau (CFPB) to reach a timely resolution of the issue. The CFPB is a product of Dodd-Frank legislation and is (in part) charged with offering protections to consumers. The CFPB solicits complaints from consumers across a variety of financial products and then addresses those complaints. It also has the power to secure payments for consumers from companies, impose fines on

show that sufficiency holds if $Y_i(\boldsymbol{T}_i) = Y_i(g(\boldsymbol{T}_i))$ for all documents and for all respondents. In words, this assumption requires that the potential outcome response to the text be identical to the potential outcome response to all documents with the same latent feature representation. This assumption is strong because it requires that there is no other information contained in the text that matters for the response beyond what is contained in the latent feature representation. In our running example about military service, this would mean that the inclusion or exclusion of military service is the only aspect relevant to the effect of the document on the individual's rating. Particularly for text, we could imagine that this assumption could easily be violated. If both versions of the treatment contain "The candidate served in the military", but one also adds "The candidate was dishonorably discharged" we might expect that this additional text added in addition to $\boldsymbol{Z}$ may be relevant to the responses.

firms found to have acted illegally, or both.

The CFPB is particularly compelling for our analysis because it provides a massive database on the text of the complaint from the consumer and how the company responded. If the person filing the complaint consents, the CFPB posts the text of the complaint in their database, along with a variety of other data about the nature of the complaint. For example, one person filed a complaint stating that

> the service representative was harsh and not listening to my questions. Attempting to collect on a debt I thought was in a grace period ...They were aggressive and unwilling to hear it

and asked for remedy. The CFPB also records whether a business offers a timely response once the CFPB raises the complaint to the business. In total, we use a collection of 113,424 total complaints downloaded from the CFPB's public website.

The texts are not randomly assigned to the CFPB, but we view the use of CFPB data as still useful for demonstrating our framework. Much of the information available to bureaucrats at the CFPB will be available in the complaint, because of the way complaints are recorded in the CFPB data. To be clear, for the effect of the text to be identified, we would need to assume that the texts provide all the information for the outcome and that any remaining information is orthogonal to the latent features of the text. We view the example of the CFPB as useful, because it provides us a clear way to think through how this assumption could be violated. If there are other non-textual factors that correlate with the text content, then our estimated treatment effects will be biased. For example, if working with the CFPB directly to resolve the complaint were important and individuals who submitted certain kinds of complaints were less well equipped to assist the CFPB, then we would be concerned about whether selection on observables holds. Or, there could be demographic factors that confound the analysis. For example, minorities may receive a slower response from CFPB bureaucrats or a more adversarial response from financial institutions (Butler, 2014; Costa, 2017) and minorities may be more likely to write about particular topics. While this is certainly plausible, many of the effects that we estimate of the text are large, so they would be difficult to explain solely through this confounding.

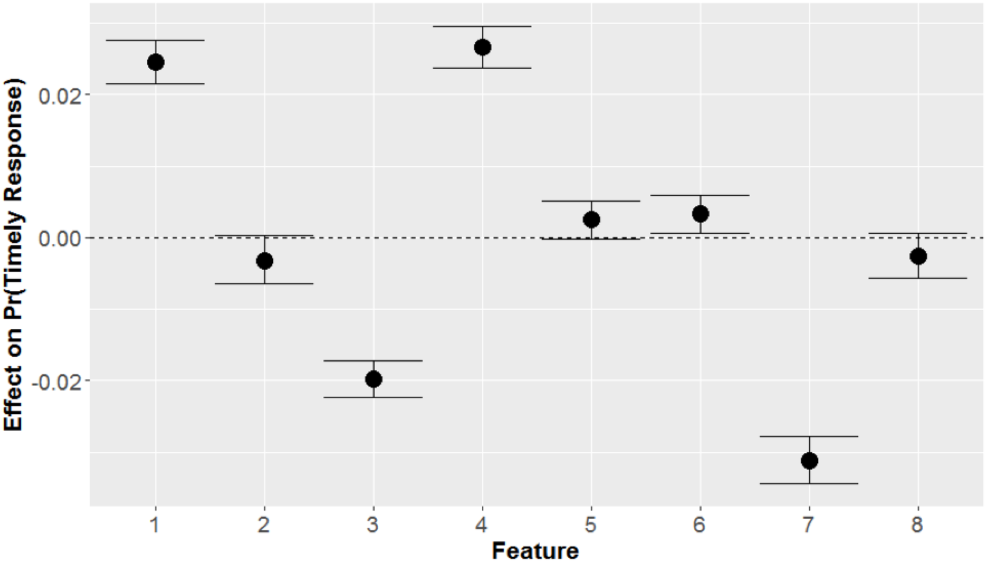
Our goal is to discover the treatments and estimate their effect on the probability of a response. We discover g using the supervised Indian Buffet Process developed for this setting in Fong and Grimmer (2016a) and implemented in the `texteffect` package in `R` (Fong, 2017). The model learns a set of latent binary features which are predictive of both the text and the outcome. To do this, we first randomly divide the data, placing 10% in the training set and 90% of the data in the test set. We place more data in the test set because our large sample ($\approx$ 11K) provides ample opportunity to discover the latent-treatments in the training set and to provide greater power when estimating effects in the test set. In the training set we apply the sIBP to the text of the complaints and whether there was a timely response. We use an extensive search to determine the number of features to include and the particular model run to use. The sIBP is a nonparametric Bayesian method; based on a user-set hyperparameter, it estimates the number of features to include in the model, though the number estimated from a nonparametric method rarely corresponds to the optimal number for a particular application. To select a final model we then evaluate the candidate model fits utilizing a model fit statistic introduced in Fong and Grimmer (2016a) that provides a quantitative measure of model fit. The train/test split ensures that we can refit the model several times choosing the estimate that provides the features that provide the best substantive insights.

Once we have fit the model in the training set, we use it to the infer the treatments in the test set. Table 3 provides the inferred latent treatments from the CFPB complaint data. The *Automatic Keywords* are the words with the largest values in the estimated latent factors for

Table 3: Consumer Financial Protection Bureau Latent Treatments

No.	Automatic Keywords	Manual Keyword
1	payment, payments, amount, interest, balance, paid, month	loan
2	card, called, call, branch, money, deposit, credit_card, told	bank
3	debt, debt_collection, account, number, validation, dispute, collection	debt collection
4	xxxx, account, time xxxx_xxxx, request, copy, received, letter	detailed complaint
5	payment, payments, pay, told, amount, month, called	disputed payment
6	loan, mortgage, modification, house foreclosure, payments	mortgage
7	debt, debt_collection, collection, credit_reporting, proof, credit_report	threat
8	fcra, credit_report, credit_reporting, reporting, debt, violation, law	credit report

Figure 5: The Effect of Complaint Features on a Prompt Response



each treatment, and the manual keyword is a phrase that we assign to each category after assessing the categories. Using these features we can then infer their presence or absence in the treated documents and then estimate their effect. To do this we use the regression procedure from Fong and Grimmer (2016*a*) and then use a bootstrap to capture uncertainty from estimation.

Figure 5 shows the effects of each latent feature on the probability of a timely response. The black dots are point estimates and the lines are 95-percent confidence intervals. Figure 5 reveals that when consumers offer more detailed feedback (Treatment 4) and when complaints are made about payments to repay a loan (Treatment 1), the probability of a prompt response increases. In contrast, the CFPB is much less successful at obtaining prompt responses from debt collectors—either when those collectors are explicitly attempting to collect a debt (Treatment 3) or when the debt collectors are threatening credit reports (Treatment 7). The inability to obtain a prompt response from debt collectors is perhaps not surprising—debt collection companies exist to successfully recover funds and are likely less concerned with their perceived reputation with debtors. It also demonstrates that it can be harder to remedy consumer complaints in some areas than others, even if the CFPB is generally able to assist complaints.

6.2 A Candidate Biography Experiment

In another example, Fong and Grimmer (2016*b*) use a text as treatment framework to study the features of candidates biographies that affect constituent evaluations. A large literature in political science asks how voters evaluate the characteristics of people who run for office (Canon, 1990; Popkin, 1994; Bartels, 2002; Carnes, 2012; Campbell and Cowley, 2014). For example, recent work studies the overrepresentation of lawyers in elected office (Bonica and Sen, Forthcoming). One way scholars study how a candidate's personal history affects their electoral support is with survey experiments. In these experiments, researchers manipulate some feature of the text, hoping to identify how individuals react to a particular piece of information. For example, we might construct a biography that alters an individual's occupation. Or, we might run a conjoint study that varies several features of the candidate's background.

Fong and Grimmer (2016*b*) take a different approach to assessing the effect of candidate biographies. They use a collection of candidate biographies to discover features of candidates' backgrounds that voters find appealing. To uncover the features of candidate biographies that voters are responsive to they acquired a collection of 1,246 Congressional candidate biographies from Wikipedia. Fong and Grimmer (2016*b*) then anonymize the biographies—replacing names and removing other identifiable information—to ensure that the only information available to the respondent was explicitly present in the text.

A condition for this experiment to uncover latent treatments is that each vector of treatments has non-zero probability of occurring. This is equivalent to assuming that none of the treatments are *aliased*, or perfectly correlated (Hainmueller, Hopkins and Yamamoto, 2013*b*). Aliasing would be more likely if there are only a few distinct texts that are provided to participants in our experiment. Therefore, Fong and Grimmer (2016*b*) assign each respondent in each evaluation round a distinct candidate biography. To bolster statistical power, they ask respondents to evaluate up to four distinct candidate biographies, resulting in each respondent evaluating 2.8 biographies on average.[11] After presenting the respondents with a candidate's biography, they ask each respondent to rate the candidate using a *feeling thermometer*: a well-established social science scale that goes from 0 when a respondent is "cold" to a candidate to 100 when a respondent is "warm" to the candidate.

Fong and Grimmer (2016*b*) recruited a sample of 1,886 participants using Survey Sampling International (SSI), an online survey platform. The sample is census matched to reflect US demographics on sex, age, race, and education. Using the sample there are 5,303 total observations. They assign 2,651 responses to the training set and 2,652 to the test set. Fong and Grimmer (2016*b*) then apply the sIBP process to the training data. To apply the model, they standardize the feeling thermometer to have mean zero and standard deviation 1. They set K to a relatively low value ($K = 10$) reflecting a quantitative and qualitative search over K. They select the final model varying the parameters and evaluating the CE score.

Table 4 provides the top words for each of the ten treatments the sIBP discovered in the training set. Fong and Grimmer (2016*b*) selected ten treatments using a combination of guidance from the sIBP, assessment using CE scores, and their own qualitative assessment

[11] The multiple evaluations of candidate biographies is problematic if there is spillover across rounds of our experiment. We have little reason to believe observing one candidate biography would systematically affect the response in subsequent rounds.

Table 4: Top Words for 10 Treatments sIBP Discovered

Treatment 1	**Treatment 2**	**Treatment 3**	**Treatment 4**	**Treatment 5**
appointed	fraternity	director	received	elected
school_graduated	distinguished	university	washington_university	house
governor	war_ii	received	years	democratic
worked	chapter	president	death	seat
older	air_force	master_arts	company	republican
law_firm	phi	phd	training	served
elected	reserve	policy	military	committee
grandfather	delta	public	including	appointed
office	air	master	george_washington	defeated
legal	states_air	affairs	earned_bachelors	office
Treatment 6	**Treatment 7**	**Treatment 8**	**Treatment 9**	**Treatment 10**
united_states	republican	star	law	war
military	democratic	bronze	school_law	enlisted
combat	elected	germany	law_school	united_states
rank	appointed	master_arts	juris_doctor	assigned
marine_corps	member	awarded	student	army
medal	incumbent	played	earned_juris	air
distinguished	political	yale	earned_law	states_army
air_force	father	football	law_firm	year
states_air	served	maternal	university_school	service
air	state	division	body_president	officer

of the models (Grimmer and Stewart, 2013). The treatments cover salient features of Congressional biographies from the time period that we analyze. For example, treatments 6 and 10 capture a candidate's military experience. Treatment 5 and 7 are about previous political experience and Treatment 3 and 9 refer to a candidate's education experience. Obviously, there are many features of a candidate's background missing here, but the treatments discovered provide a useful set of dimensions to assess how voters respond to a candidate's background.

After training the model on the training set, Fong and Grimmer (2016*b*) apply it to the test set to infer the treatments in the biographies. They assume there are no interactions between the discovered treatments in order to estimate their effects.[12] Figure 6 shows the

[12] This assumption is not necessary for the framework we propose here. Interaction effects could be mod-

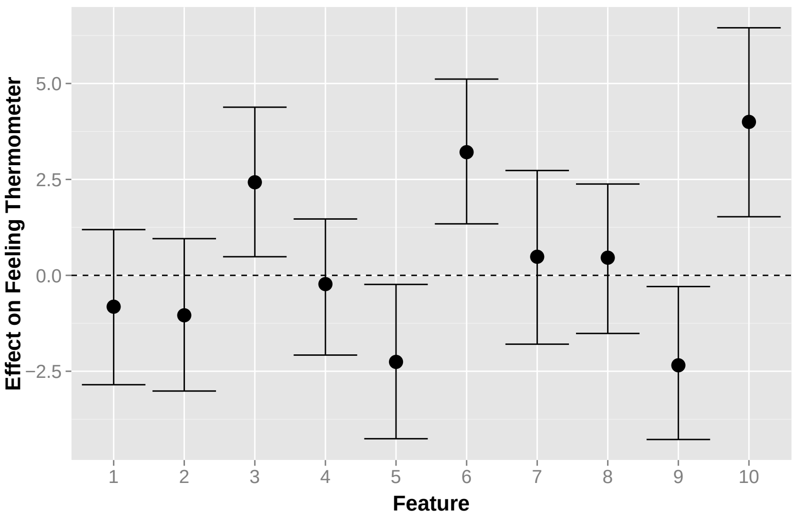


Figure 6: 95% Confidence Intervals for Effects of Discovered Treatments: The mean value of the feeling thermometer is 62.3

point estimate and 95-percent confidence intervals, which take into account uncertainty in inferring the treatments from the texts and the relationship between those treatments and the response.

The treatment effects reveal intuitive, though interesting, features of candidate biographies that affect respondent's evaluations. For example, Figure 6 reveals a distaste for political and legal experience—even though a large share of Congressional candidates have previous political experience and a law degree. Treatment 5, which describes a candidate's previous political experience, causes an 2.26 point reduction in feeling thermometer evaluation (95 percent confidence interval, [-4.26,-0.24]). Likewise, Treatment 9 shows that respondents dislike lawyers, with the presence of legal experience causing a 2.34 point reduction in feeling thermometer (95-percent confidence interval, [-4.28,-0.29]). The aversion to lawyers is not, however, an aversion to education. Treatment 3, a treatment that describes advanced degrees, causes a 2.43 point increase in feeling thermometer evaluations (95-percent

eled, but it would require us to make much stronger parametric assumptions using a method for heterogeneous treatments such as Imai and Ratkovic (2013).

confidence interval, [0.49,4.38]).

In contrast, Figure 6 shows that there is a consistent bonus for military experience. This is consistent with intuition from political observers that the public supports veterans. For example, treatment 6, which describes a candidate's military record, causes a 3.21 point increase in feeling thermometer rating (95-percent confidence interval, [1.34,5.12]) and treatment 10 causes a 4.00 point increase (95-percent confidence interval, [1.53,6.45]).

This experiment reveals how varying the texts can lead to intuitive, though still useful insights into the effect of texts on outcomes. In our next section, we consider how text can be used to eliminate confounding between a numeric treatment and outcome.

7 Text as a Confounder

Not only can text be used as an outcome or an intervention in experiments, text can also be used to control for variables in observational analyses. Here we draw on work by Roberts, Stewart and Nielsen (2018) to describe how text can be a confounder in causal analyses and how we can use text analysis to control for this type of confounding.

Text as a confounding variable occurs frequently within social science applications. To take one example, Maliniak, Powers and Walter (2013) provide evidence that articles written by women in International Relations are less likely to be cited than articles written by men. The authors find this relationship, even when accounting for scholarly credentials, such as tenure and rank of school, and for article-level covariates such as publication venue, topic, perspective, and article age.

Is an article with a female author name less likely to be cited than had the same article had a male author name? In this case, the gender of the author's name is treatment and the outcome are the number of citations. However, because gender is not randomly assigned, we can not simply take the mean difference in citations between men and women as our estimate of the causal effect of gender on citations. Most importantly, women write about different topics in International Relations than men do. The textual content of the article may be related to both treatment (perceived gender) and the outcome (citations), therefore the text of the article confounds the relationship between gender and citations. As a result, Maliniak, Powers and Walter (2013) spend much of the paper accounting for this by using hand coded categories that describe the article content as control variables.

Text as a confounder appears surprisingly often in the analysis of social data. The content of legislative bills might confound the relationship between veto threats and repositioning in Congress. Students college admissions profiles or recommendation letters may be associated both with their race or gender and their probability of achieving admissions to college. The content of international agreements might confound the relationship between trade and international cooperation.

Controlling for numeric variables is relatively straight forward – the researcher can use basic linear regression or matching to condition on the confounder. However, controlling for the content of a text is much more difficult. The researcher cannot include every unique word in the article in the match or the regression because there are probably many more words than observations. In addition, the researcher does not want to completely control for *all of* the text – ideally they just want to adjust for the words or clusters of words that confound the treatment and the outcome.

Adjusting for confounding using a high-dimensional covariate like text poses the challenge of the "curse of dimensionality." The difficulty with conditioning on text is to figure out which aspects of the text should be included and which elements of the text should *not* be conditioned on. Oftentimes, the researcher does not know of hand what aspects of the text are related both to the treatment and to the outcome. We might not know how women and men discuss topics differently in International Relations and which aspects of these differences might be related to citations counts. Therefore, in these cases researchers must *estimate* the aspects of the text that are confounding the relationship.

Mathematically, the structure of the confounding problem is as follows. As before in this chapter, we start with a data set of N units. Each unit i is assigned treatment T_i, which takes a value of 1 for treated units and 0 for control. Under the potential outcomes framework, the outcome variable Y_i takes on the value $Y_i(1)$ when unit i is treated and $Y_i(0)$ when unit i is a control.

Because we have observational data, T_i is not randomly assigned and treated and control groups may not be comparable. A common practice is to match on P pre-treatment covariates $\boldsymbol{X} = (X_1, X_2, \dots X_P)$ to improve similarity in the distribution of covariates within treatment and control groups, a condition called balance. When $\boldsymbol{X}$ includes all sources of confounding, we can use matching to achieve balance ($\boldsymbol{X} \perp\!\!\!\perp T$) which allows unbiased estimation of the population average treatment effect on the treated.

In most matching applications, $\boldsymbol{X}$ is low-dimensional, with $P \ll N$. Under selection on observables, we assume conditional ignorability: $T_i \perp\!\!\!\perp Y_i(0), Y_i(1) | \boldsymbol{X}$. Under this assumption, balancing the distribution of observed covariates $\boldsymbol{X}$ across treatment groups provides us a way to estimate the causal effect of interest. In some settings, the P variables of $\boldsymbol{X}$ under which selection on observables would hold are known to the researcher because the treatment assignment mechanism is transparent.

However, in the cases we consider where we want to condition not only on $\boldsymbol{X}$, but also on the document term matrix $\boldsymbol{W}$ the dimensionality of these two matrices combined, $P + J$ is very large. In this case, J, the vocabulary of the corpus, may be much larger than N, the number of observations. Further, we do not necessarily know which aspects of $\boldsymbol{W}$ are related to the treatment and outcome. We believe that some words in the text affect treatment, but are unsure which ones. If we attempt to match on all words, we will not identify any matches unless two texts have *identical* word frequencies, a possibility that becomes vanishingly small as the dimension of $\boldsymbol{W}$ grows large.

Our approach is to estimate a low-dimensional summary of the variables in $\boldsymbol{W}$ that we can use to address confounding, a g function for the confounders instead of the treatment and the outcome as we had before. Assuming positivity and Stable Unit Treatment Values (SUTVA), text matching requires the following:

Assumption 5 (Conditional Ignorability). $T_i \perp\!\!\!\perp Y_i(0), Y_i(1) | g(\boldsymbol{W})$

Assumption 6 (SUTVA). *For all individuals i, $Y_i(T) = Y_i(T_i)$.*

Assumption 7 (Positivity). *For all individuals i $Pr(T_i = t) > 0$ for all $t \in \mathcal{T}$.*[13]

In this section, we explore a wide variety of different methods for finding the right g to adjust for confounding using text data – both regression-based and matching-based adjustments. To do this, throughout the chapter we use the example from Maliniak, Powers and Walter (2013) studying gendered citation bias in the discipline of International Relations (IR). Evidence of citation bias against female scholars in IR would be strongest if women wrote identical articles to men but were then cited less often. However, men and women tend to write about different topics within IR, use different methods, and have different epistemological commitments. Because these factors might affect citation counts, it is possible the lower citation counts of women reflect bias against certain topics and approaches, rather than against women themselves. Maliniak, Powers and Walter (2013) address this challenge using information from the TRIP Journal Article Database to control for the broad sub-field of each article, the issue areas covered, the general methodology, paradigm,[14] and epistemology. In this application, we attempt to use automated methods rather than hand coding to address this confounding.[15]

[13] D'Amour et al. (2017) provide a reassessment of positivity in the high-dimensional context and show that this assumption can often fail. This is another way in which high-dimensional data are a challenge for current matching approaches.

[14] Scholarship in International Relations is sometimes organized into "paradigms," or schools of thought about which factors are most crucial for explaining international relations. The predominant paradigms are Realism, Liberalism, and Constructivism, though others exist.

[15] The Maliniak, Powers and Walter (2013) hand coding has an additional benefit of allowing us to validate our approach.

With the help of JSTOR's Data For Research Program, Roberts, Stewart and Nielsen (2018) obtain the full text of 3,201 articles in the IR literature since 1980, 333 of which are authored solely by women.[16] In what follows, we use the data collected by Roberts, Stewart and Nielsen (2018) to explore a variety of different methods for conditioning on the text of the articles while examining the influence of gender on citations. We discuss the drawbacks and benefits to each of these methods in what follows.

7.0.1 Regression Adjustments for Text Confounders

The most common way of adjusting for confounding in social science data analysis is including the confounding variable in a regression. Therefore, the simplest extension to adjusting for confounding with a high-dimensional variable like text is including it in a regression, but using a shrinkage method to select the best subset of the regressors for the application at hand. As discussed in Chapter 5, penalized regression estimators such as ridge regression and lasso can select the covariates that are highly predictive of the outcome, shrinking or eliminating covariates that are less important. If these estimators select the relevant words to adjust for with confounding, then we could use them in adjusting for text confounders.

We take this first very simple approach for conditioning on the text of the articles to estimate the impact of gender. We estimate a lasso where we regress citations of each article on the term-document matrix of the articles and all of the other covariates the Maliniak, Powers and Walter (2013) include within their analyses, such as article age, tenure, and journal. We also include an indicator variable for treatment: whether the article was written by only female authors. We do not include a penalty on the treatment coefficient, as this is the coefficient of interest. We use cross-validation to choose the optimal penalty on the remaining coefficients.

[16] They analyze more articles than Maliniak, Powers and Walter (2013) because the TRIP database has coded more articles since 2013. However, they are missing data for a few articles used by Maliniak, Powers and Walter (2013) because they are not in JSTOR's data set.

The advantage of this approach is that the model selects important words and covariates that are highly predictive of the outcome and therefore discards the many words that are not of interest. Figure 7 shows the words that are ultimately included in the model, and thus highly predictive of citations, including words related to funding like "Minerva," words related to methodology such as "cause–effect" and "zscore", and words related to substance such as "surgenc" and "democ." Citations of particular authors also seem to be important in predicting citations, such as "weingast", and "shepsl". The model also suggest particular covariates coded by Maliniak, Powers and Walter (2013) to be important, such as whether the article was published in the American Political Science Review – the primary political science venue – or whether the article has a positivist framework, both of which predict higher citations. The coefficient estimated on female authorship is negative, indicating that adjusting for the words selected by Lasso and the other confounders, female authorship has a negative relationship with citations.

Using a Lasso is relatively straight forward way to select words that could be confounding, but it has some important drawbacks. For example, by only including unigrams in our model, we rule out interactions between words using a lasso because we are using the bag of words approach. It could be that when "minerva" is used with words like "democracy", it produces a lot of citations, but when it's used with "autocracy" it has very few. Interactions between words can be important – language involves words interacting to each other.

To adjust for more substantively meaningful confounding, we could control for the context of the text by controlling for a low-dimensional representation of the text. Say, for example, that we used a topic model to represent each document as a combination of topics. We could then estimate the influence of gender on citations, controlling for topical content. Using topics as control variables might be more realistic than conditioning on unigrams, as male and female authors in political science are likely to use different words because they write about different topics. In this case, we would condition on the context rather than just on the word itself.

Figure 8 shows the regression output of citations regressed on gender conditioning on the topic proportions estimated within each document as well as the other confounders included within Maliniak, Powers and Walter (2013). Many of the topics should be recognizable to political scientists – the authors discuss topics ranging from war to economic development to foreign policy. Topic 11, associated with international organizations have overall higher citations counts than Topic 12, associated with the Soviet Union and the military. The female coefficient is similar to that in the Lasso, indicating that controlling for topic women still receive fewer citations than men.

There are many approaches one could take to controlling for lower-dimension representations of the data and the analyst would want to select the one that they thought was most important in confounding the relationship between treatment and outcome. For example, principal components analysis (PCA) or singular value decomposition (SVD) could be used in place of topics to control for clusters of words that explain variance in the documents. However, PCA, SVD, and vanilla topic modeling approaches do not contain information about the outcome or the treatment in the model. The topics in [topics figure] do not contain the word "minerva" even though this word highly predicts citations. Including information about the outcome or the treatment when selecting covariates may be useful in ensuring that the model is picking up confounders. For example, it could be useful to use the supervised LDA or STM and include a covariate for treatment in order to ensure that the topics found are relevant to treatment status.

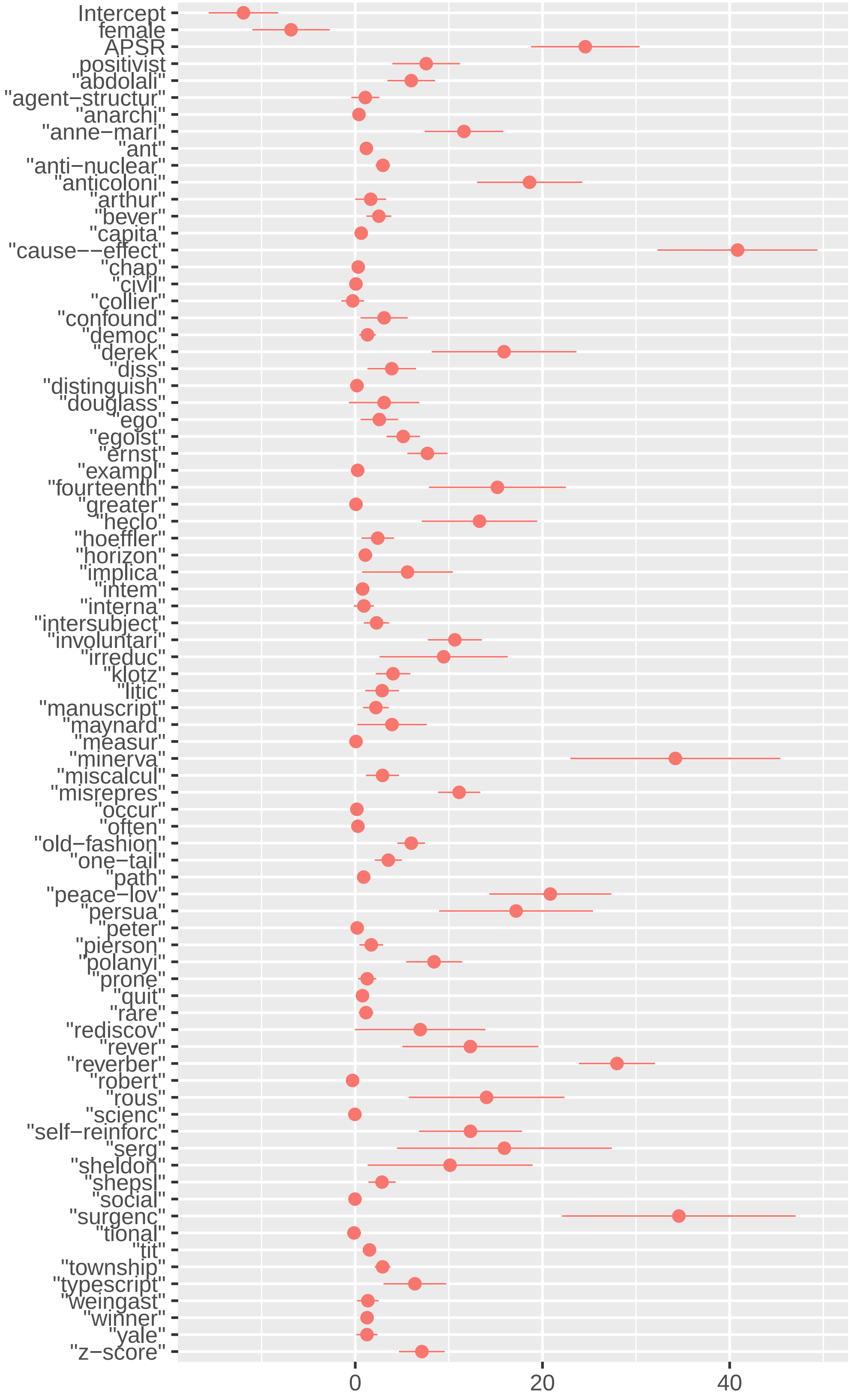


Figure 7: Regression of citations on gender with words and covariates selected by a lasso model.

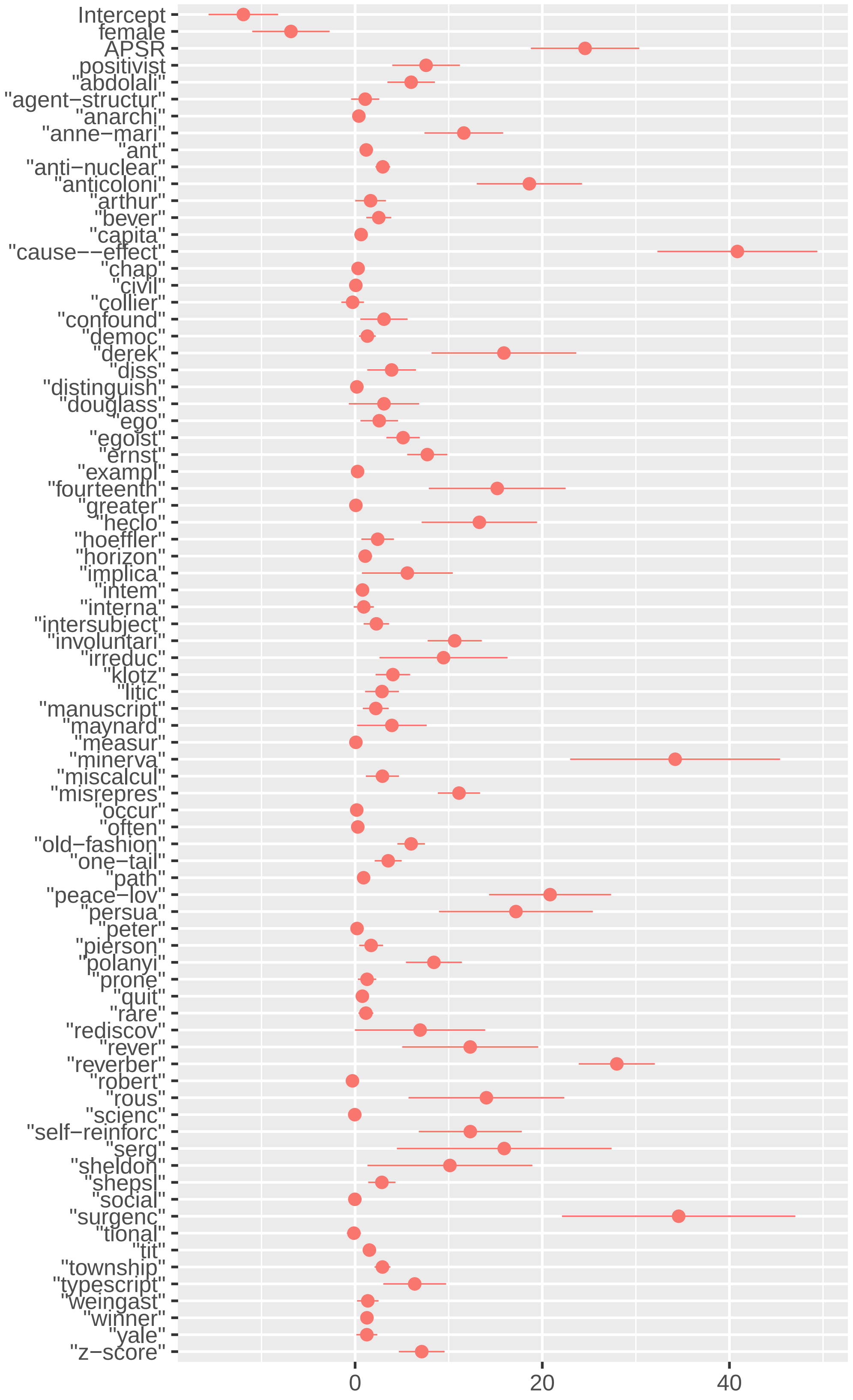


Figure 8: Regression of citations on gender controlling for topic proportions and other confounders.

7.0.2 Matching Adjustments for Text

An alternative to using a regression adjustment is to use text analysis to find matches between treated and control units that write about similar topics or use similar words. Roberts, Stewart and Nielsen (2018) argue that the advantage of matching is that comparable units are more easily verified – matching produces a similarity metric that allows users to find documents that the matching method would consider similar, read them, and evaluate whether these observations would be good counterfactuals for each other. Text models can fail, so as with all text analysis models, we want to be able to validate our results (Grimmer and Stewart, 2013). By reading matched treatment and control an analyst can judge whether they are sufficiently similar that treatment assignment is plausibly "as-if random." This manual validation complements formal balance checks that we describe below by validating some things the formal checks cannot, such as the usefulness of a term-document-matrix representation of the text. The low-dimensional summaries of text that we use for matching could also be used for regression or weighting, but are not as intuitive to evaluate whether the procedure has worked with these other approaches.

There are hundreds of matching approaches, each with strengths and weaknesses. Roberts, Stewart and Nielsen (2018) focus on adaptions for matching leveraging propensity score matching (PSM) and coarsened exact matching (CEM), not because they are optimal, but because they highlight common matching strategies: *modeling treatment assignment*, and *coarsening*. When analysts use PSM to model treatment assignment, they weight some variables in $\boldsymbol{X}$ more than others. When they use CEM to coarsen variables, they replace the variables in $\boldsymbol{X}$ with coarsened versions that treat a range of values for $\boldsymbol{X}$ as equivalent. In the next two sections we consider modeling treatment assignment and coarsening as possible approaches for dealing with a high-dimensional $\boldsymbol{W}$ and describe why one, absent of the other, leads to poor matches.

Modeling Treatment Assignment PSM is a prominent approach to modeling treatment assignments which weights the elements of $\boldsymbol{X}$ by their value for predicting T. PSM summarizes this information in a minimally sufficient statistic for the part of $\boldsymbol{X}$ that is correlated with T, called the *propensity score* (Rosenbaum and Rubin, 1983). The propensity score is the the probability of treatment conditional on $\boldsymbol{X}$, or:

$$\pi_i = p(T_i = 1|X_i) \tag{7.1}$$

with $\hat{\pi}_i$ typically estimated via logistic regression of T on $\boldsymbol{X}$. All variables in $\boldsymbol{X}$ are generally included as predictors "unless there is consensus that it is unrelated to the outcome variables or not a proper covariate" (Rubin, 2006, 269). The estimated probabilities $\hat{\pi}_i$ from this regression, or the linear predictor, are used to match.

Modeling treatment assignment leads us to consider ways to match only on words that predict treatment status. Because the number of words is generally larger than the number of observations ($J \gg N$), Roberts, Stewart and Nielsen (2018) use inverse regression to estimate propensity scores (Cook and Ni, 2005). Faced with the challenge of estimating $p(T|\boldsymbol{W})$, inverse regression posits a parametric model for the inverse problem, $p(\boldsymbol{W}|T)$, which produces a sufficient reduction of the information in $\boldsymbol{W}$ about the conditional distribution $p(T|\boldsymbol{W})$. Penalized regression to estimate $p(T|\boldsymbol{W})$ directly is also possible, but we prefer inverse regression because it text's count data format and is computationally efficient to compute.

$\boldsymbol{W}$ is a matrix of word counts which we model using the Multinomial Inverse Regression (MNIR) framework developed in Taddy (2013*a*) and introduced in Chapter 4 of this book.

As a review, each document is modeled as:

$$\begin{aligned} \boldsymbol{W}_i &\sim \text{Multinomial}(n_i, \boldsymbol{\pi}_i) \\ \pi_{ij} &= \frac{\exp(\mu_{0j} + \mu_{1j}T_i)}{\sum_{l=1}^{J} exp(\mu_{0l} + \mu_{1l}T_i)} \end{aligned} \tag{7.2}$$

but now instead of Y_i we are using T_i, a ℓ-length vector containing a categorical encoding of the treatment variable (in ℓ categories) for document i, $\boldsymbol{\pi}_i$ is a vector of event probabilities whose elements are $\pi_{i,j}$ and n_i is the number of trials in the multinomial distribution. The coefficients μ are often given a sparsity-inducing regularizing prior, a point which we return to below. Mechanically, MNIR amounts to estimating a multinomial logistic regression with words as outcomes and treatment status as a predictor variable. After estimating the model we can calculate a sufficient reduction score:

$$\pi_i = \mu'(w_i/n_i) \tag{7.3}$$

which implies $T_i \perp\!\!\!\perp w_i, n_i | \pi_i$, as shown in Propositions 3.1 and 3.2 of Taddy (2013*a*) which establish the classical sufficiency properties of the projection. This implies that under the model in Equation 7.2 we can condition on π_i and discard the higher dimensional data w_i.

MNIR results in efficiency gains when estimating propensity scores with high-dimensional $\boldsymbol{W}$. This efficiency comes from making fairly strong assumptions about the generative process of the predictor $\boldsymbol{W}$. In addition to the usual assumptions for propensity scores, we also introduce assumptions about the suitability of the generative model. Under the standard propensity score model, the variance in the MLE of the coefficients for the propensity score model decreases in the number of documents. However with MNIR the variance decreases with the number of total words (Taddy, 2013*b*, See Proposition 1.1). Taddy (2013*a*) provides a complete description of technical properties.

Multinomial inverse regression requires estimating many coefficients because the coefficient matrix μ has one column per word in the vocabulary. We model treatment assignment by estimating the coefficients with a regularizing prior (see details in the Supplemental Information). A sparsity-promoting prior focuses our attention on a subset of words that have substantially different rates of use in the treated group in comparison to the control group. Following estimation of the MNIR model, we can estimate propensity scores using the forward regression, in other words transform π_i to have support between zero and 1. However, the forward regression provides no new information for matching so we skip it and match directly on the sufficient reduction.

The primary strength of this approach is that it simplifies $\boldsymbol{W}$ efficiently and achieves balance. However, the propensity score approach limits our ability to do human balance-checking. Matching on the MNIR-generated propensity score often results in matches that are not obviously similar to human readers because PSM only provides balance in distribution and does not necessarily recover (nearly) exactly matching pairs. Two very different topics can be matched because both have high probability of treatment.. If balance is achieved and identification assumptions hold, we can still obtain unbiased estimates of our causal effects. However, we lose the ability to read matched document pairs and assess similarity using expert judgment which is an important reason to adopt a matching approach to text.

Coarsening Coarsening offers an alternative for reducing the dimensions of a standard covariate matrix $\boldsymbol{X}$. CEM (Iacus, King and Porro, 2011) is an approach to coarsening which applies exact matching to a modified version of $\boldsymbol{X}$ in which variables have been replaced by coarsened summaries. In most applications, these summaries are created by the analyst, who coarsens each variable into substantively meaningful bins and then performs exact matching within strata defined by these bins. For example, an analyst matching survey respondents based on years of education might coarsen the many-valued variable *years of education* into substantively meaningful bins: *no high-school degree*, *high-school degree*, *college degree* and *post-graduate degree.* Units that are identical according to these coarsened categories are in the same stratum, and thus are matched. When no matches are available for a unit it is dropped from the data, changing the estimand from the Average Treatment Effect on the Treated to a treatment effect specific to the treated units which remain. CEM is monotonic imbalance bounding (MIB), meaning that the researcher bounds the differences between treated and control to the extrema of the strata (Iacus, King and Porro, 2011).

CEM is generally only feasible when the set of matching variables is small relative to the number of observations. Consider an application with a single matching variable coarsened into four categories. This results in four strata that should ideally be populated with treated and control units. Adding a second variables of four categories leads to $4 \times 4 = 16$ strata. Now consider a very small text corpus with only 100 unique words. Even if we apply the broadest possible coarsening — replacing the term frequencies with a binary variable indicating whether or not the document contains the word — we get 2^{100} strata. Unless combinations of words are almost perfectly correlated, the number of strata is so large that there will generally be no matches. If a document contains even a single unique word, then no coarsened exact matches are possible.

To develop a coarsening approach for high-dimensional matching problems with the covariate matrix $\boldsymbol{W}$, weextend the logic of CEM. Rather than grouping units with similar values on individual variables into the same strata, we could group the variables themselves. For example, we might assume that when the words "democracy" and"democratic", occur in a set of articles, they all have approximately the same referent and can be grouped together into one concept ("democ") and represented by a single indicator (which now takes on the value of 1 if either word is present and is 0 otherwise). This procedure is called stemming and is already widely used for dimension reduction in text analysis. Crucially, CEM retains the MIB property with stemmed text data because distance between texts remains bounded:

any matched documents must have equal counts of the stem "democ," although counts of the unstemmed words may not be equal.

However, traditional stemming does not normally reduce dimensionality enough to facilitate matching with high-dimensional text data.[17] Instead, Roberts, Stewart and Nielsen (2018) use *Topically Coarsened Exact Matching* (TCEM) in which they estimate a topic model and then apply coarsened exact matching to the estimated topics. Under the topic model, each word in the corpus has a latent topic assignment and two words with the same topic assignment are stochastically equivalent.[18] TCEM results in matched documents that have comparable amounts of stochastically equivalent words, although the specific observed words may differ. In our running example, we could use TCEM to find documents that have the same amount of a topic that might likely be censored, while not conditioning directly on precisely which words were used.

TCEM maintains the monotonic imbalance bounding property in the topic space, a fact which we interpret in two ways. First, it may be the case that treatment assignment is, in fact, based on the topics of texts, rather than on the exact wording. If so, the MIB property on the topics is exactly what we need. Alternatively, we may believe that treatment assignment is based on features that are not perfectly summarized by latent topics. Even so, matching on the density estimate of the topic proportions is a way of reducing variance at the risk of introducing a small amount of bias. From this perspective, TCEM is appropriate if two observations with a common density estimate are stochastically equivalent and deviations between them are essentially random noise (an objective that the topic model is trying to achieve).

Topic models require that the analyst choose the number of topics. This choice can be

[17] Stemming is also underdeveloped for languages like Chinese and Arabic(see Lucas et al., 2015).

[18] Under the data generating process the observed word is drawn from the same multinomial distribution. Thus the difference in the observed word is a product of stochastic noise rather than a systematic difference in the latent variable.

fraught for cases where semantic interpretation of topics is the primary concern (Grimmer and Stewart, 2013) because automated methods like choosing the best fitting model may not result in semantically coherent topics (Chang et al., 2009; Roberts et al., 2014). For matching, we require that topics estimate the joint density well, but do not need them to be semantically coherent. This means we can select the number of topics using model fit statistics such as the highest held-out likelihood. The number of topics should be sufficient for matched documents to be good counterfactuals for each other, as determined by the demands of the research design. In general, more topics will result in closer matches. Redundant topics will decrease efficiency but will not cause bias, so the risks of choosing too few topics are much greater than choosing too many.

TCEM ensures that matched texts will be topically similar which facilitates balance-checking via manual comparison of matched documents. However, TCEM is weak precisely where the MNIR approach is strong: the topics are estimated without information about treatment status so TCEM can fail to detect sets of words that predict treatment assignment. This happens because topic models will capture the subject matter of the document rather than, say, the sentiment, even though sentiment may be a strong predictor of treatment assignment.

The advantage of matching is that comparable units are more easily verified – matching produces a similarity metric that allows users to find documents that the matching method would consider similar, read them, and evaluate whether these observations would be good counterfactuals for each other. If, based on their substantive knowledge, the researcher has reason to believe that the documents are not comparable, then they could adjust the similarity metric so that it produces better matches.

Topical Inverse Regression Matching To address both the drawbacks of TCEM and MNIR, Roberts, Stewart and Nielsen (2018) propose a solution to the problem of high-dimensional text matching that they call topical inverse regression matching (TIRM). To capture the benefits of modeling treatment assignment and coarsening, they use the Structural Topic Model (STM) to jointly estimate topics (for coarsening) and document-level propensity scores (for modeling treatment assignment), and then match on both. To do this, they show that using the treatment as a *topical content covariate* in an STM model effectively combines the MNIR and TCEM frameworks developed above into a single *joint estimation*.

In the MNIR framework, the analog to the propensity score is created by estimating a projection from the model that was a sufficient reduction of the information in the word counts about the probability of treatment. This is also possible in the STM model. Following Taddy (2013*a*), Roberts, Stewart and Nielsen (2018) derive a sufficient reduction of the information contained in the word counts about treatment, but because we include topics, the projection now represents the information about the treatment *not* carried in the topics.

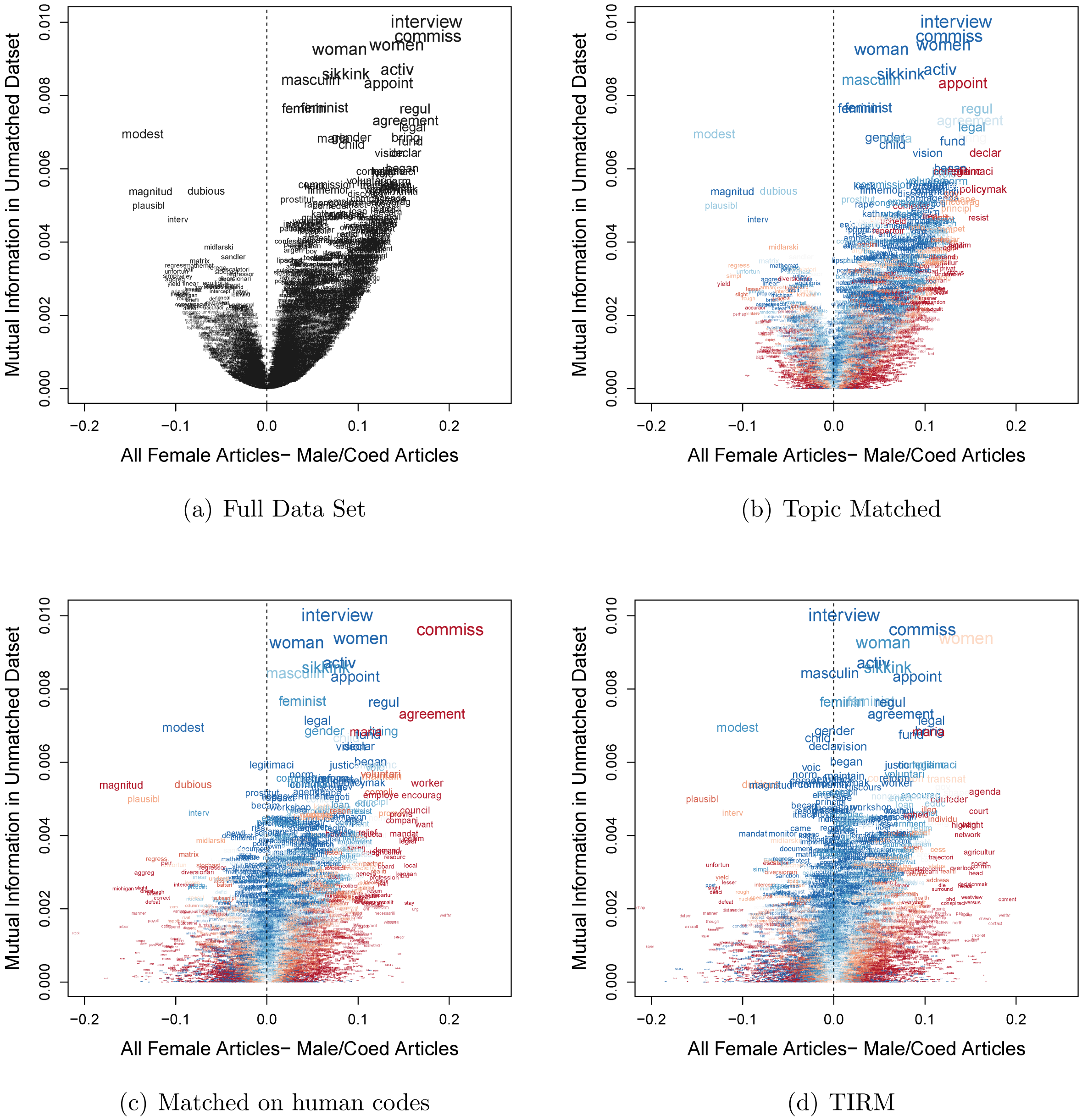
The other estimate they extract from the STM model is the topical content of the documents. A slight complication arises because when topic-treatment intactions are present: the same topic can have different estimated distributions of words under treatment and control. They ensure that topics are comparable irrespective of treatment/control differences by adding a final estimation step to the STM in which they re-estimate the topic proportions of all control as though they were treated. This choice is consistent with an estimand that is a (local) average treatment effect on the *treated*. They now match on both the STM projection and the estimated topic proportions from the STM, which ensures that matches are both topically similar and have similar within-topic probabilities of treatment.

Roberts, Stewart and Nielsen (2018) apply TIRM to the Maliniak, Powers and Walter (2013), specifying 15 topics in the STM portion of the algorithm to recover broad topics. For comparison, they also create matched samples using MNIR, TCEM, and exact matching based on the human-coded data.

They test whether TIRM adequately reduced these differences in the matched sample by identifying the words that have high *mutual information* with author gender in the raw data set as well as the matched data from TIRM, TCEM, and MNIR. The Figure 9 shows the relationship between the difference in word occurrence by gender and the mutual information of each word in each dataset. If perfect balance on all words were possible, we would hope to see every word lined up vertically on the $x = 0$ line (and shaded blue accordingly). However, since not all words can be balanced, balance on words with high mutual information is most important. TIRM — shown in the bottom-right panel — outperforms the others in balancing the high mutual information words. Many high mutual information words such as "interview" that were previously imbalanced are now lined up down the center. TIRM makes the imbalance substantially worse on words with low mutual information, but this is unimportant because balancing these words does not reduce bias. This analysis highlights the benefits of the treatment model because it can identify and address the most imbalanced words – exact matching on human coding and TCEM do not perform as well as TIRM.

They also compare balance on the 15 estimated topics and find that TCEM performs the best on this metric, followed closely by TIRM, as shown in Figure 10. Exact matching on human coding does reasonably well, with the exception of a single topic — foreign policy — where it makes balance worse. MNIR *increases* imbalance for most topics. They also see how well TIRM performs in comparison to the human coded categories and find that TIRM performs well on most categories, particularly the most imbalanced ones.

They then re-estimate the models of Maliniak, Powers and Walter (2013) using the TIRM matched sample and find that gender differences in citations are even more pronounced than those originally reported. They find an average 16 fewer citations to articles written by women, when compared to as-identical-as-possible articles written by men or mixed-gender groups.



(a) Full Data Set

(b) Topic Matched

(c) Matched on human codes

(d) TIRM

Figure 9: Relationship between mutual information and difference in word occurrence (all female - male/coed) a) Full Data Set b) Topic Matched c) Matched on human codes d) TIRM. In panels b, c, and d, words for which matching decreased the absolute value of the difference in word appearance are in blue and words for which matching increased the absolute value of the difference in word appearance are in red.

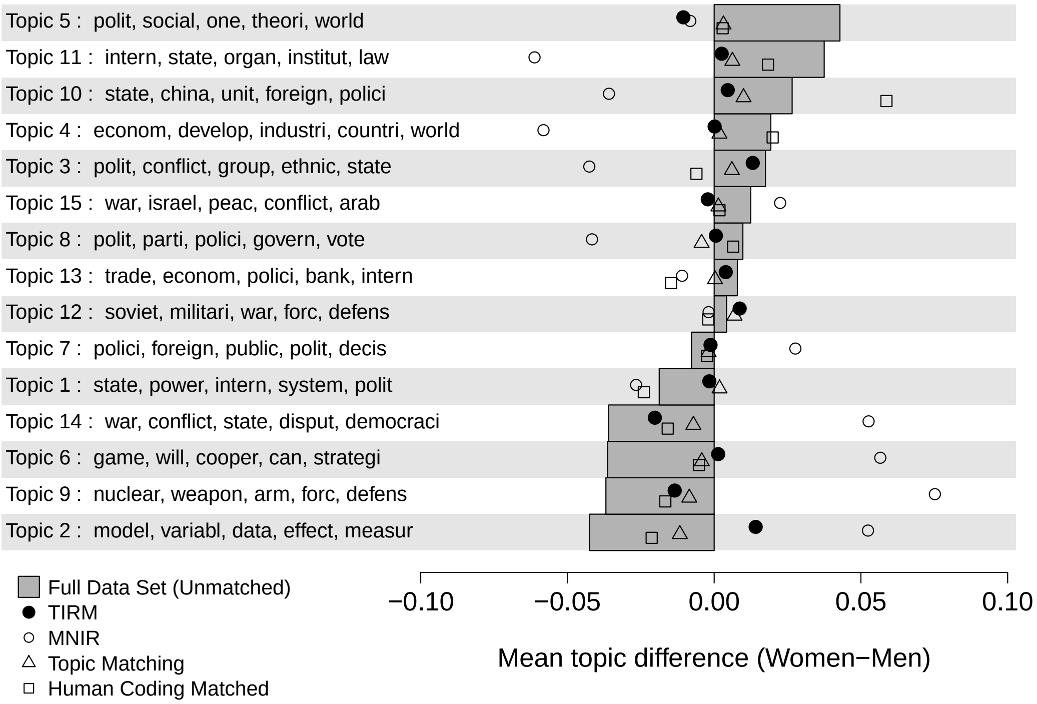


Figure 10: Matching Comparison for Topics

8 Conclusion

Text measures so many social phenomena, but is currently used rarely in analyses to make causal inferences. In the chapter, we developed a framework for thinking about how to use text in causal inferences by spelling out the potential pitfalls researchers may run into when using text as a treatment, outcome, or control variable. We included many examples of using text in casual inference throughout the chapter to demonstrate how vast the potential is for using text to study causation in social science.

Of course, causal identification is difficult, and using text in causal analyses does not solve any of the core problems of causal inference. On the contrary, using text for causal inference may add more, not fewer, complications. We urge readers to take advantage of the large literature in this area to improve their inferences. We suggest, in addition, that analysts take advantage of train/test splits to simulate replication within each experiment, and, as often as possible, repeat their experiments in new populations and time periods.

References

Bartels, Larry M. 2002. "The impact of candidate traits in American presidential elections." *Leaders personalities and the outcomes of democratic elections* pp. 44–69.

Benoit, Kenneth and Michael Laver. 2003. "Estimating Irish party policy positions using computer wordscoring: the 2002 election - a research note." *Irish Political Studies* 18(1):97–107.

Biernacki, Richard. 2012. *Reinventing evidence in social inquiry: Decoding facts and variables.* Springer.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research* 3:993–1022.

Bonica, Adam and Maya Sen. Forthcoming. "The Politics of Selecting the Bench from the Bar: The Legal Profession and Partisan Incentives to Politicize the Judiciary." *Journal of Law and Economics* .

Butler, Daniel M. 2014. *Representing the advantaged: How politicians reinforce inequality.* Cambridge University Press.

Campbell, Rosie and Philip Cowley. 2014. "What Voters Want: Reactions to Candidate Characteristics in a Survey Experiment." *Political Studies* 62(4):745–765.

Canon, David T. 1990. *Actors, Athletes, and Astronauts: Political Amateurs in the United States Congress.* University of Chicago Press.

Carlsmith, Kevin M, John M Darley and Paul H Robinson. 2002. "Why do we punish? Deterrence and just deserts as motives for punishment." *Journal of personality and social psychology* 83(2):284.

Carnes, Nicholas. 2012. "Does the Numerical Underrepresentation of the Working Class in Congress Matter?" *Legislative Studies Quarterly* 37(1):5–34.

Catalinac, Amy. 2016. "From Pork to Policy: The Rise of Programmatic Campaigning in Japanese Elections." *The Journal of Politics* 78(1):1–18.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems.* pp. 288–296.

Cohen, Mark A, Roland T Rust and Sara Steen. 2002. "Measuring public perceptions of appropriate prison sentences: Report to National Institute of Justice." *NCJ Report* (199365).

Cohen, Mark A, Roland T Rust and Sara Steen. 2004. "Measuring perceptions of appropriate prison sentences in the United States, 2000. ICPSR version. Nashville, TN: Vanderbilt University [producer], 2000." *Ann Arbor, MI: Inter-university Consortium for Political and Social Research.[distributor]* .

Cook, R Dennis and Liqiang Ni. 2005. "Sufficient dimension reduction via inverse regression." *Journal of the American Statistical Association* 100(470).

Costa, Mia. 2017. "How Responsive are Political Elites? A Meta-Analysis of Experiments on Public Officials." *Journal of Experimental Political Science* 4(3):241–254.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei and Jasjeet Sekhon. 2017. "Overlap in Observational Studies with High-Dimensional Covariates." *arXiv preprint arXiv:1711.02582* .
URL: *https://arxiv.org/pdf/1711.02582.pdf*

DiMaggio, Paul, Manish Nag and David Blei. 2013. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding." *Poetics* 41(6):570–606.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts and Brandon M Stewart. 2018. "How to make causal inferences using texts." *arXiv preprint arXiv:1802.02163* .

Eisenstein, Jacob, Amr Ahmed and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th International Conference on International Conference on Machine Learning.* Omnipress pp. 1041–1048.

Fong, Christian. 2017. *texteffect: Discovering Latent Treatments in Text Corpora and Estimating Their Causal Effects*. R package version 0.1.

Fong, Christian J and Justin Grimmer. 2018*a*. "Exploratory and Confirmatory Causal Inference for High-Dimensional Interventions." *Stanford University Mimeo* .

Fong, Christian and Justin Grimmer. 2016*a*. Discovery of Treatments from Text Corpora. In *ACL (1)*.

Fong, Christian and Justin Grimmer. 2016*b*. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1 pp. 1600–1609.

Fong, Christian and Justin Grimmer. 2018*b*. "Exploratory and Confirmatory Causal Inference for High Dimensional Interventions.".

Franco, Annie, Justin Grimmer and Chloe Lim. 2018. "The Limited Effect of Presidential Appeals." *Stanford University Mimeo* .

Gentzkow, Matthew and Jesse M Shapiro. 2010. "What drives media slant? Evidence from US daily newspapers." *Econometrica* 78(1):35–71.

Gerber, Alan S and Donald P Green. 2012. *Field experiments: Design, analysis, and interpretation*. WW Norton.

Gill, Michael and Andrew B Hall. 2015. "How Judicial Identity Changes The Text Of Legal Rulings.".

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* .

Grimmer, Justin. 2013. *Representational Style in Congress: What Legislators Say and Why It Matters*. Cambridge University Press.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political analysis* 21(3):267–297.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2012. "How words and money cultivate a personal vote: The effect of legislator credit claiming on constituent credit allocation." *American Political Science Review* 106(4):703–719.

Hainmueller, Jens and Daniel J Hopkins. 2014. "Public Attitudes Toward Immigration." *Annual Review of Political Science* 17:225–249.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2013*a*. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hainmueller, Jens, Daniel J Hopkins and Teppei Yamamoto. 2013*b*. "Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments." *Political Analysis* 22(1):1–30.

Hopkins, Daniel J and Gary King. 2010. "A method of automated nonparametric content analysis for social science." *American Journal of Political Science* 54(1):229–247.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter Van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* 21(1):1–20.

Iacus, Stefano M, Gary King and Giuseppe Porro. 2011. "Multivariate matching methods that are monotonic imbalance bounding." *Journal of the American Statistical Association* 106(493):345–361.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Imbens, Guido W and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. Sage.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Leeper, Thomas J. 2017. *MTurkR: Access to Amazon Mechanical Turk Requester API via R*. R package version 0.8.0.

Lowe, Will. 2008. "Understanding Wordscores." *Political Analysis* 16(4):356–371.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer and Dustin Tingley. 2015. "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis* 23(3):254–277.

Maliniak, Daniel, Ryan Powers and Barbara F Walter. 2013. "The gender citation gap in international relations." *International Organization* 67(04):889–922.

Mimno, David and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*. AUAI Press pp. 411–418.

Neuendorf, Kimberly A. 2016. *The content analysis guidebook*. Sage.

Popkin, Samuel. 1994. *The Reasoning Voter: Communication and Persuasion in Presidential Campaigns*. Chicago: University of Chicago Press.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin and Dragomir R Radev. 2010. "How to analyze political attention with minimal assumptions and costs." *American Journal of Political Science* 54(1):209–228.

Quinn, K.M., B.L. Monroe, M. Colaresi, M.H. Crespin and D.R. Radev. 2006. "How To Analyze Political Attention With Minimal Assumptions And Costs." Annual Meeting of the Society for Political Methodology.

Roberts, Margaret E., Brandon M. Stewart and Dustin Tingley. 2017. *stm: R Package for Structural Topic Models*. R package version 1.2.3.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58(4):1064–1082.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* pp. 1–49.

Roberts, Margaret E, Brandon M Stewart and Richard Nielsen. 2018. "Adjusting for Confounding with Text Matching.".

Rosenbaum, Paul R and Donald B Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1):41–55.

Rubin, Donald B. 2006. *Matched sampling for causal effects*. New York: Cambridge University Press.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman et al. 2013. "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS one* 8(9):e73791.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant." *Psychological science* 22(11):1359–1366.

Slapin, Jonathan and Sven-Oliver Proksch. 2008. "A Scaling Model for Estimating Time-Series Party Positions from Texts." *American Journal of Political Science* 52(3):705–722.

Spirling, Arthur. 2012. "US treaty making with American Indians: Institutional change and relative power, 1784–1911." *American Journal of Political Science* 56(1):84–97.

Taddy, Matt. 2013*a*. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association* 108(503):755–770.

Taddy, Matt. 2013*b*. "Rejoinder: Efficiency and Structure in MNIR." *Journal of the American Statistical Association* 108(503):772–774.

van der Laan, Mark J, Alan E Hubbard and Sara Kherad Pajouh. 2013. "Statistical inference for data adaptive target parameters.".