AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Causal Inference with Latent Treatments

Christian Fong University of Michigan
Justin Grimmer Stanford University

Abstract: *Social scientists are interested in the effects of low-dimensional latent treatments within texts, such as the effect of an attack on a candidate in a political advertisement. We provide a framework for causal inference with latent treatments in high-dimensional interventions. Using this framework, we show that the randomization of texts alone is insufficient to identify the causal effects of latent treatments, because other unmeasured treatments in the text could confound the measured treatment's effect. We provide a set of assumptions that is sufficient to identify the effect of latent treatments and a set of strategies to make these assumptions more plausible, including explicitly adjusting for potentially confounding text features and nontraditional experimental designs involving many versions of the text. We apply our framework to a survey experiment and an observational study, demonstrating how our framework makes text-based causal inferences more credible.*

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

Social scientists often seek to estimate the effects of complex, high-dimensional interventions, such as texts, audio, images, and videos. However, when examining these high-dimensional interventions, they are almost always interested in estimating the effect of a *latent treatment* within these interventions, such as the topic of a speech, an expressed opinion, or the tone of a message. By latent treatment, we mean a treatment that cannot be manipulated directly, but can only be manipulated indirectly by manipulating the text. As a result, the treatment is measured in a text by applying a many-to-one function that maps a higher dimensional intervention to a lower dimensional vector.[1]

As an example, suppose that we are interested in assessing the effect of negative advertising in a political campaign on the decision to turnout to vote (Ansolabehere, Iyengar, and Simon 1999; Arceneaux and Nickerson 2010). In our framework, we decompose the effect of high-dimensional interventions, such as advertisements, into *measured* and *unmeasured* latent treatments. The measured latent treatment of interest is whether there is negative information about the opponent included in the advertisement, whereas other features of the advertisement that could affect an individual's participation choice—such as information about where to vote—may not be measured by the analyst.

We provide a framework for identifying and estimating the causal effect of measured latent treatments from high-dimensional interventions in the presence of potentially confounding unmeasured treatments. We show that the inability to directly randomize the latent treatment of interest and the presence of unmeasured treatments implies that the familiar condition for an experiment to identify causal effects—random assignment of individuals to treatment conditions—is insufficient to identify the causal effect of the latent treatment. Even if individuals are randomly assigned to read particular texts, unmeasured treatments might confound the estimate of the latent treatment effect, because the unmeasured treatments could be correlated with the latent treatment of interest. We find that identifying the average treatment effect (ATE) for a latent treatment requires assumptions analogous to those used in observational research to rule out omitted-variable bias: Either

Christian Fong, Assistant Professor, Department of Political Science, University of Michigan, 6640 Haven Hall, 505 S. State Street, Ann Arbor, MI 48109 (cjfong@umich.edu). Justin Grimmer, Professor, Department of Political Science and Senior Fellow, Hoover Institution, Stanford University, 212 Encina Hall West, 616 Jane Stanford Way, Stanford, CA 94305 (jgrimmer@stanford.edu).

The authors thank Naoki Egami, Jonathan Mummolo, Margaret Roberts, Brandon Stewart, Sean Westwood, anonymous reviewers at the *AJPS*, and the editor for their thoughtful comments and suggestions.

[1] Online Appendix B, p. 2, provides a glossary of terms.

American Journal of Political Science, Vol. 00, No. 0, XXXX 2021, Pp. 1–16

DOI: 10.1111/ajps.12649

TABLE 1 Extensive Use of Latent Treatments

Experiment type	Count	Latent	Aliased	Single vignette	Both
Survey experiment	29	100%	66%	97%	62%
Field experiment	13	92	54	77	46
Conjoint experiment	5	100	0	100	0
Lab experiment	4	100	75	100	75

Note: Aliased refers to designs in which the latent treatment of interest may be confounded by another unmeasured latent treatment of interest. Single vignette refers to designs in which there is only one text per treatment condition, which yields a local average treatment effect.

the unmeasured treatments have no effect on the outcome or the unmeasured and measured treatments are independent of the measured treatment. If our identifying assumptions hold, then the difference-in-means estimator is consistent for the ATE.

Reliance on these assumptions is pervasive, though implicit, in prior observational and experimental studies. Regardless of how the measured treatments within a document are coded—whether by hand, unsupervised, or supervised methods, whether known in advance of assignment of units to texts or discovered afterward—the analyst must make an assumption that ensures the estimated effects are attributable to the measured treatments of interest rather than other features of the text. Table 1 compiles 51 articles published in the *American Journal of Political Science* and the *American Political Science Review* that use a text-based treatment from 2015 to 2019. It shows that when scholars estimate the effects of texts, they focus almost exclusively on latent treatments.[2]

We apply our framework to vignette survey experiments and demonstrate that the designs of most existing studies face two potential new threats to inference. First, unmeasured latent treatments that vary across texts might alias the effect of measured latent treatments, conflating the effect of the measured and unmeasured latent treatments. We show that vignette experiments implicitly make a *no aliasing* assumption: The effect of any unmeasured latent treatment that varies with measured latent treatments is zero. Moreover, even if there is no aliasing and researchers are able to design manipulations that affect only the measured latent treatments, we show that vignette survey experiments estimate a local treatment effect: A single text is usually used to deliver each condition, so the estimated effects are conditional on the unmeasured treatments in the vignettes. Extrapolating the survey results to a broader set of texts requires a *no interaction assumption*: The effects of measured treatments do not depend on the unmeasured treatments in the text. The problems of aliased treatments and the use of only a single vignette are pervasive across text-based experiments.[3] Thus, most experiments rely on implicit, stringent, and untested assumptions to identify the effect of measured latent treatments in the presence of unmeasured treatments.

To demonstrate how our framework facilitates more credible inferences, we apply it to an original vignette experiment and an observational study of how actual political rhetoric affects the public. In the vignette experiment we assess the *no aliasing* and *no interaction assumptions* with an unconventional vignette experimental design: providing many vignettes per latent treatment. Specifically, we examine how information about prior U.S. legal commitments affects support for protesters in Hong Kong, an experiment that we run twice to ensure the robustness of our findings. We construct vignettes to limit the chance for aliasing and show how including many vignettes per treatment enables us to adjust for previously unmeasured treatments. Across two replications of the experiment, we find that information about U.S. commitments has a substantial effect on the desire to support the Hong Kong protesters.

Our second empirical example examines the public response to President Donald Trump's rhetoric, using Trump's actual messages posted to Twitter. We apply and extend a procedure developed by Fong and Grimmer (2016) to discover and then estimate the effect of latent treatments in texts. We find limited evidence of a differential partisan response to some rhetoric from Trump (Zaller 1992), but the direction of each feature's effect is the same across partisan groups (Coppock, Ekins, and Kirby 2018). Most surprisingly, we find evidence that Republicans evaluate Trump's tweets lashing out against opponents and the media negatively, which is suggestive evidence that Trump's idiosyncrasies are liabilities rather than assets. We show how the effect of plausible unmeasured treatments can be assessed and demonstrate that

[2]The lone exception is the work of Kalla and Broockman (2017), which estimates the effect of specific campaign messages on vote choice.

[3]Online Appendix A, p. 2, provides information on all 51 articles.

our inferences are robust to one detectable violation of our assumptions.

Our approach to causal inference with latent treatments builds on prior work on causal mediation (Acharya, Blackwell, and Sen 2016; Imai et al. 2011; Pearl 2001), but it has critical differences, which implies that the insights from mediation are not directly applicable to the problem of latent treatments. In standard causal mediation problems, the goal is to understand how an intervention's effect on an outcome goes "through" mediators, similar to our measured latent treatments. Unlike in most applications of mediation, however, the measured and unmeasured latent treatments are deterministic functions of the texts, the functions are determined by the analyst, and all individuals reading the text have the same value of the measured and unmeasured treatments. This contrasts with standard mediation analysis, where the mediators are intermediate outcomes that vary across individuals. This implies that a different set of identification and estimation challenges will be present when estimating the effects of latent treatments.[4]

Our framework also builds on recent and important work on the treatments that subjects infer from vignettes, though we introduce a distinct theoretical issue that implies a different preferred research design. Dafoe, Zhang, and Caughey (2018) (DZC) highlights that treatments may have unintended effects on respondents' background beliefs, which may confound estimates of the treatment of interest. In this article, the issues that we explicate are distinct from, and occur prior to, the respondent-based inferences in DZC. In fact, we show that the issues with text-based treatments are present even if the confounding in DZC is completely absent. As a result, we make a different set of research design recommendations. DZC prefers an embedded natural experiment to make inferences about treatments of interest independent of background beliefs, but this is often impossible with text and that even if it is successfully accomplished, the embedded natural experiment requires a strong assumption eliminating effect heterogeneity. Instead, our framework emphasizes how variation in natural language can enable analysts to design studies to adjust background features beforehand—similar to DZC's covariate control method—but importantly, our text-based approach enables us to measure and adjust for additional confounding features of text after the experiment. We provide a more thorough comparison to DZC in Online Appendix D, pp. 4–5.

[4]For example, a sequential ignorability assumption is not necessary because there is no variation across individuals in the levels of measured treatments. Moreover, it is impossible to define mediation quantities such as the average natural direct effect or direct effects (Acharya, Blackwell, and Sen 2016; Imai et al. 2011), because it is impossible to define the relevant cross-world effects when, holding the text constant, we cannot have some observations where the treatment is present and others where it is absent. Our framework is perhaps best thought of as a framework with deterministic mediators of the text, wherein we make assumptions about the measured mediators to avoid confounding from the unmeasured mediators. See Online Appendix C, p. 2, for more details.

Confounding by Unmeasured Treatments

As Table 1 shows, when political scientists use text as a treatment, they tend to focus on latent treatments. In this section, we show that estimation of the effects of latent treatments, even in cases wherein the texts are randomly assigned to respondents, requires another set of strong assumptions about how the measured treatments interact with unmeasured treatments, and that this assumption is similar to assuming that there is no omitted-variable bias in observational research.

Defining the Estimand with Text as Treatment

We first introduce our notation to define our estimand of interest, and then we enumerate the assumptions for identification and provide a more thorough justification and motivation for our framework. Suppose a researcher is interested in understanding how users respond to a collection of texts $\mathcal{X}$. Define the potential outcomes of the texts $Y : \mathcal{X} \rightarrow \mathbb{R}$, with $Y_i(\boldsymbol{X}_i)$ representing the response respondent i gives upon reading text $\boldsymbol{X}_i$. Although an important application of text-based experiments defines causal effects directly using the text-based counterfactuals (Offer-Westort, Coppock, and Green 2019), Table 1 shows that political scientists are interested in estimating the effect of latent treatments in the texts. Let $g : \mathcal{X} \rightarrow \{0, 1\}$ be an analyst-defined codebook function that measures the presence or absence of the latent treatment in any given document based on its text *and only its text.*[5] Using this function, we say that if $g(\boldsymbol{X}_i) = 1$, then the

[5]This means that our framework is robust to individuals failing to perceive that a treatment is present or interpreting g differently than the analyst. If either a failure to perceive or a different interpretation occurs, it would alter the interpretation of the effect but would not change whether the analyst codes a particular latent treatment as present. Our framework deals with the construction of the text, rather than how an individual's beliefs about the world govern their reaction to a particular treatment (Dafoe, Zhang, and Caughey 2018).

latent treatment is present in text $\boldsymbol{X}_i$, and if $g(\boldsymbol{X}_i) = 0$, then the latent treatment is not present in $\boldsymbol{X}_i$. We define $Z_i \equiv g(\boldsymbol{X}_i)$ as the latent treatment in the text assigned to respondent i.

Of course, there are often more relevant features in a text than a latent treatment. For example, campaign advertisements vary in not only whether they are negative, but also in whether they focus on a candidate's policy positions or character, whether they are light-hearted or ominous, whether they are overt or subtle, and many other potential unmeasured treatments. We label these other components of the text as unmeasured treatments to distinguish them from the explicitly measured treatment of interest. We assume that there exists some set of unmeasured latent treatments described by function $h : \mathcal{X} \rightarrow \mathcal{B}$, which, together with the latent treatments of interest from g, capture all features of a document that affect responses. Unlike g, we suppose the analyst does not know h. Define $\boldsymbol{B}_i \equiv h(\boldsymbol{X}_i)$. Because the combination of measured and unmeasured treatments captures all the relevant features of the text, we write the outcome as $Y_i(\boldsymbol{X}_i) = Y_i(g(\boldsymbol{X}_i), h(\boldsymbol{X}_i)) = Y_i(Z_i, \boldsymbol{B}_i)$. The second equality simply notes that Z_i and $\boldsymbol{B}_i$ are shorthands for $g(\boldsymbol{X}_i)$ and $h(\boldsymbol{X}_i)$, respectively. We then marginalize over the possible values of the unmeasured treatments to define our estimand, the ATE:

$$ATE = \sum_{b \in \mathcal{B}} (\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = \boldsymbol{b})] - \mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = \boldsymbol{b})])\, Pr(\boldsymbol{B}_i = \boldsymbol{b}), \quad (1)$$

where $Pr(\boldsymbol{B}_i = \boldsymbol{b})$ is determined by the structure of the texts and the assignment of texts to individuals $Pr(\boldsymbol{X}_i)$. Defining the estimand this way avoids issues of overlap and undefined conditionals that plague recent studies in computer science (Pryzant et al. 2020).

If we could directly randomize the latent treatment of interest, then identification and consistent estimation of the ATE is straightforward even if we never directly control for the unmeasured treatments. We could sample texts from the population, randomize whether the treatment is present or absent, and then randomly assign texts to respondents. Unfortunately, it is impossible to manipulate the treatment without also manipulating the text. When analysts do manipulate the texts, they risk inadvertently manipulating an unmeasured treatment, potentially confounding the effect of the measured treatment of interest.

Given our inability to directly manipulate the latent treatment of interest, we now formally derive the conditions under which the ATE as defined in Equation (1) can be estimated from observed data.

Formal Framework for the Causal Effects of Latent Treatments

Consider a finite population of documents, $\mathcal{X}$, and suppose there are N respondents, $i = 1, 2, \ldots, N$, each of whom observes a document, $\boldsymbol{X}_i \in \mathcal{X}$. Let $\boldsymbol{X} = \{\boldsymbol{X}_i\}_{i=1}^N$ be the collection of documents observed by all respondents.

Suppose that each individual has a potential outcome function that represents how the individual would respond if each subject in the experiment were assigned a particular document, $Y_i : \mathcal{X}^N \rightarrow \mathbb{R}$. The first assumption is that an individual's response depends on only the document they are assigned.

Assumption 1. *For all individuals i and any $\boldsymbol{X}$, $\boldsymbol{X}'$ such that $\boldsymbol{X}_i = \boldsymbol{X}'_i$, $Y_i(\boldsymbol{X}) = Y_i(\boldsymbol{X}')$.*

This allows the potential outcomes function to compactly be written as a function of the document assigned to i rather than of all documents, $Y_i : \mathcal{X} \rightarrow \mathbb{R}$.

Further, assume that the texts are either randomly assigned in an experimental setting or the document assignment is independent of the potential outcomes.[6]

Assumption 2. *For all individuals i, $Y_i(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{X}_i$ and $Pr(\boldsymbol{X}_i = \boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$.*

Assumption 2 is about who reads a text, not about the contents of a particular text. The assumption is violated if individuals select the particular text they read. For example, if strong partisans are both more responsive to negative advertisements and more likely to encounter positive rather than negative advertisements due to their media consumption habits, then Assumption 2 would not be credible in an observational study of the effect of exposure to negative campaign advertisements on participation.

The Codebook Function, Unmeasured Treatments, and a Text's Effect on an Outcome. Following Fong and Grimmer (2016), we suppose the researcher has defined a codebook function that labels the measured latent treatments that are present or absent in a document based on the text *and only the text*. In machine learning applications, g is usually explicitly defined through the machine learning algorithm, but even if the researcher is reading the documents and hand-labeling them, the hand-labeling process implies an implicit g. Formally, let $g : \mathcal{X} \rightarrow \mathcal{Z}$ map texts to binary feature vectors. We will suppose that $\mathcal{Z} \equiv \{0, 1\}$ throughout this section, but we

[6]Our analysis generalizes straightforwardly if instead there exists a set of observed covariates, $\boldsymbol{C}_i$, such that $Y_i(\boldsymbol{x}) \perp\!\!\!\perp \boldsymbol{X}_i | \boldsymbol{C}_i$ and $\Pr(\boldsymbol{X}_i = \boldsymbol{x} | \boldsymbol{C}_i) > 0$ for all $\boldsymbol{x} \in \mathcal{X}$. We focus on the case of unconditional independence in the text to reduce notational clutter.

show in Online Appendix G, p. 9, that our analysis generalizes to settings in which $\mathcal{Z} \equiv \{0, 1\}^K$, where we use as an estimand the multitreatment analog of the ATE, the average marginal component effect (AMCE).[7] A particular z represents the latent treatment present in a document.

Aside from the latent treatments measured through the codebook function, there may be unmeasured features of the document that affect the response. For example, in a study of the effects of negativity on whether a campaign advertisement increases or decreases turnout, we may fail to measure whether the issue focuses on policy positions.

Accordingly, we assume that there is some set of unmeasured latent treatments that, together with the measured latent treatment, fully characterize the features individuals consider when responding to the document. We are not yet assuming anything substantive about the nature of those unmeasured treatments, just that there exists some true set of unmeasured treatments that actually account for the relationship between the text and the outcome.

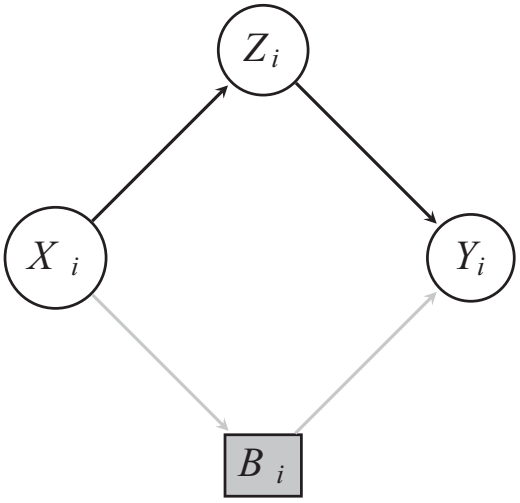
Let $h : \mathcal{X} \rightarrow \mathcal{B}$, where $\mathcal{B}$ is a finite set. Unlike g, the analyst does not know h. We assume that if two documents have the same measured and unmeasured treatments, then respondents respond to them in the same way, on average. Additionally, we assume that any vector of unmeasured latent treatments that occurs in the population occurs in documents with and without the latent treatment.

Assumption 3. *There exists some function* $h : \mathcal{X} \rightarrow \mathcal{B}$ *such that if* $g(\boldsymbol{x}) = g(\boldsymbol{x}')$ *and* $h(\boldsymbol{x}) = h(\boldsymbol{x}')$ *for* $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, *then* $\mathbb{E}[Y_i(\boldsymbol{X}_i = \boldsymbol{x})] = \mathbb{E}[Y_i(\boldsymbol{X}_i = \boldsymbol{x}')]$. *Additionally,* $0 < Pr(Z_i = 1|\boldsymbol{B}_i = \boldsymbol{b}) < 1$ *for all* $\boldsymbol{b} \in \mathcal{B}$.

The first part of Assumption 3 can also be understood as an independence assumption. Given the measured and unmeasured treatments within a text, the text and the outcome are independent. For expositional clarity, we abuse the notation and make a slightly stronger assumption so that we can write $Y_i(\boldsymbol{X}_i) = Y_i(g(\boldsymbol{X}_i), h(\boldsymbol{X}_i)) = Y_i(Z_i, \boldsymbol{B}_i)$. We use $Y_i(Z_i = z, \boldsymbol{B}_i = \boldsymbol{b})$ when defining theoretical quantities for its superior clarity; when we discuss estimators, we use $\mathbb{E}[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = z, h(\boldsymbol{X}_i) = \boldsymbol{b}]$ to emphasize how documents are used. Nevertheless, it is important to note that Assumption 3 implies that these are just two different notations for the same quantity.

[7]For latent treatments, we define the AMCE of the kth component as $\sum_{\boldsymbol{z}_{-k} \in \mathcal{Z}_{-k}} \sum_{\boldsymbol{b} \in \mathcal{B}} \{\mathbb{E}[Y_i(Z_{i,k} = 1, \boldsymbol{Z}_{i,-k} = \boldsymbol{z}_{-k}, \boldsymbol{B}_i = \boldsymbol{b})] - \mathbb{E}[Y_i(Z_{i,k} = 0, \boldsymbol{Z}_{i,-k} = \boldsymbol{z}_{-k}, \boldsymbol{B}_i = \boldsymbol{b})] \times Pr(\boldsymbol{B}_i = \boldsymbol{b}) \times m(\boldsymbol{z}_{-k})$, where $-k$ is an index that indicates all but the kth components of $\boldsymbol{Z}_i$ and $\mathcal{Z}$. $m(\boldsymbol{z}_{-k})$ is an analyst-specified density on measured treatments besides the latent treatment of interest.

FIGURE 1 Directed Acyclic Graph for Causal Text Diagram



Note: The text, $\boldsymbol{X}_i$, causes both the latent treatment of interest, Z_i, and the unmeasured latent treatments, $\boldsymbol{B}_i$. These latent treatments, in turn, cause the outcome, Y_i.

The second part of Assumption 3 is the common support assumption frequently used in causal inference, including propensity score analysis. It ensures that, for any text, there exists some text in the population that allows the analyst to answer the question, "What would the individual's response have been if they had seen a text with with the same unmeasured latent treatments, but with a different value of the measured latent treatment?" Without Assumption 3, it is difficult to define a causal estimand for latent treatments.[8]

Figure 1 provides a directed acyclic graph consistent with this assumption. In terms of this graph, our goal is to estimate the effect of the text, $\boldsymbol{X}_i$, that goes through the measured latent treatments, Z_i. The gray box around $\boldsymbol{B}_i$ indicates that these treatments are not explicitly observed. Figure 1 shows that Assumption 3 implies that the text only affects responses through the measured treatment, Z_i, and the unmeasured treatments, $\boldsymbol{B}_i$.[9]

[8]If there is an unmeasured treatment that affects the outcome and can occur only when the measured treatment is present (or can occur only when it is absent), the analyst has two choices. They may either explicitly measure and adjust for the treatment or they may incorporate it as part of their definition of the treatment of interest and marginalize over it.

[9]We focus on latent treatments as a consequence of the text, rather than latent treatments based on authors' intentions when writing the text (Pryzant et al. 2020). We focus on latent treatments based

If $\boldsymbol{B}_i$ were observed and there were sufficient data, we could adjust for it to identify the effect of $\boldsymbol{X}_i$ through Z_i. However, because $\boldsymbol{B}_i$ is not observed, we will need an additional assumption to ensure that the unmeasured treatments do not confound the estimated effects of the measured treatment.

To estimate the ATE as described in Equation (1), we make one of two assumptions—either $\boldsymbol{B}_i$ and Z_i are independent or $\boldsymbol{B}_i$ does not affect the outcome:

Assumption 4. *At least one of the following is true:*

- *The measured and unmeasured latent treatments are independent:* $Pr(Z_i = z, \boldsymbol{B}_i = \boldsymbol{b}) = Pr(Z_i = z)Pr(\boldsymbol{B}_i = \boldsymbol{b})$.
- *The unmeasured treatments are unrelated to the outcome:* $\mathbb{E}[Y_i(Z_i = z, \boldsymbol{B}_i = \boldsymbol{b})] = \mathbb{E}[Y_i(Z_i = z, \boldsymbol{B}_i = \boldsymbol{b}')]$ *for all* $z \in \{0, 1\}$ *and all* $\boldsymbol{b}, \boldsymbol{b}' \in \mathcal{B}$.

The conditions in Assumption 4 are analogous to the conditions required to avoid omitted-variable bias in observational research. In linear regression, omitted-variable bias arises if there is an omitted variable that is correlated with the independent variables in the regression and has an effect on the outcome. In our setting, the first condition is that the measured treatments are independent of the unmeasured treatments. If Z_i and $\boldsymbol{B}_i$ are independent under the allocation of texts $Pr(\boldsymbol{X}_i)$, the distribution of $\boldsymbol{B}_i$ is identical between the treatment and control groups, so differences in the outcome between the treatment and control groups cannot be attributed to differences in the distribution of $\boldsymbol{B}_i$. The second condition is that the unmeasured treatments have no effect on the outcome, on average. If $\boldsymbol{B}_i$ is irrelevant to the outcome in expectation, it does not matter if the treated and control groups have completely different distributions of $\boldsymbol{B}_i$; regardless of the distributions of $\boldsymbol{B}_i$ in treatment versus control, Equation (1) will always give the same answer.

Estimator and Identification. Our proposed estimator is the difference in means between respondents who received texts with the treatment and without the treatment:

$$\widehat{ATE} = \mathbb{E}[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 1] - \mathbb{E}[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 0]. \quad (2)$$

Assumptions 1–4 connect Equations (1) and (2). Equation (2) can be estimated from observed data under minimal assumptions, but it does not necessarily provide a credible estimate of the effect of the latent treatment because it does not address possible confounding from the unmeasured treatments. Equation (1) provides a sensible definition of the causal effect of the latent treatment, but it depends on the unmeasured treatments, which, by definition, the analyst does not know and has not measured. If Assumptions 1–4 are satisfied, the difference in means between respondents who received texts with the treatment and without the treatment is equal to Equation (1), and the theoretically satisfying definition of the ATE from Equation (1) can be consistently estimated using the difference-in-means estimator in Equation (2).

Proposition 1. *If Assumptions 1–4 hold, Equation (1) is identified by Equation (2) and can be consistently estimated by the difference in means between individuals who received texts with the treatment and without the treatment,* $ATE = E[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 1] - E[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 0]$.

Proof. See Online Appendix F, p. 8. □

A Simple Example: Campaign Advertising

To demonstrate how the conditions in Assumption 4 enable the identification of the measured latent treatment's causal effect, we examine three stylized examples in this. The first two show that when Assumption 4 is satisfied—either because the measured and unmeasured treatments are independent (Example 1) or the unmeasured treatments have no effect on the outcome (Example 2)—the causal effect of the measured treatment is consistently estimated with a difference in means. Example 3 shows that when these conditions are not satisfied, a difference in means estimator will be inconsistent, because the unmeasured treatments will confound the measured treatments. Across the three examples, we suppose that we are interested in assessing the effect of a negative campaign advertisement on turnout in an election. Further, for the sake of this example, we also suppose that a single unmeasured latent treatment in the advertisements satisfies Assumption 3—whether the advertisement focuses on policy positions or not. We will also suppose that the advertisements have been randomly assigned to individuals.

Example 1 (Independent Measured and Unmeasured Latent Treatments). *We first suppose that the measured and unmeasured treatments are independent* $Pr(Z_i, \boldsymbol{B}_i) = Pr(Z_i)Pr(\boldsymbol{B}_i)$*, specifically assuming that it has the joint distribution in Table 2(a), which is induced from the assignment of texts to individuals'* $Pr(\boldsymbol{X}_i)$*. We suppose that the effect of the advertisements on turnout*

on text-based content because authors' intentions cannot have an effect on the outcome directly, because they only affect the content of the text.

TABLE 2 Example Joint Distributions of Measured and Unmeasured Treatments

(a) Independent Measured and Unmeasured Treatments

	$B_i = 0$	$B_i = 1$
$Z_i = 0$	0.16	0.24
$Z_i = 1$	0.24	0.36

(b) Dependent Measured and Unmeasured Treatments

	$B_i = 0$	$B_i = 1$
$Z_i = 0$	0.20	0.40
$Z_i = 1$	0.30	0.10

Note: This synthetic example considers scenarios in which the measured and unmeasured treatments are independent and in which the measured and unmeasured latent treatments depend on one another.

rates follows Table 3(a). Using Table 3(a), we calculate the effect of negative advertising on turnout rates in advertisements that focus on policy positions, $\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 1)] - \mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = 1)] = 0.35 - 0.40 = -0.05$, *and the effect of negative advertising on turnout rate in advertisements that do not focus on policy positions,* $\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 0)] - \mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = 0)] = 0.39 - 0.5 = -0.11$. *This implies that the true ATE of a negative advertisement in this example is an 8.6 percentage point decrease in the turnout rate, or* $ATE = -0.05 \times 0.40 - 0.11 \times 0.6 = -0.086$. *The difference in means estimator can consistently estimate this quantity. Through the calculations, we find that* $E[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 1] - E[Y_i(\boldsymbol{X}_i)|g(\boldsymbol{X}_i) = 0] = 0.40 \times 0.35 + 0.6 \times 0.39 - (0.4 \times 0.40 + 0.6 \times 0.50) = -0.086$. *The difference in means is consistent because the independence of the measured and unmeasured treatments ensures that the unmeasured treatments do not confound the measured treatment of interest.*

Example 2 (Unmeasured Treatments Do Not Affect the Outcome). *We now suppose that the other condition in Assumption 4 holds: Unmeasured and measured treatments are dependent, as given in Table 2(b), but the unmeasured treatment—a focus on policy positions—does not affect the turnout rate, as given in Table 3(b). In this instance the true effect of negative advertising is a 5 percentage point decrease in turnout rate,* $ATE = -0.05$. *The difference in means estimator provides* $0.75 \times 0.35 + 0.25 \times 0.35 - (\frac{1}{3} \times 0.40 + \frac{2}{3} \times 0.40) = -0.05$. *If the unmeasured treatment does not covary with the outcome, it cannot confound our estimate of the ATE and the difference in means estimator will consistently estimate the ATE.*

Example 3 (Dependent Measured and Unmeasured Treatments and Unmeasured Treatments Affect the Outcome). *If the conditions of Assumption 4 do not hold, we cannot guarantee that a difference in means will consistently estimate the true ATE. Suppose that the unmeasured and measured treatments are dependent, as given in Table 2(b), but the unmeasured treatment also affects the turnout rate, as given in Table 3(a). This implies that negative advertisements decrease the turnout rate by 8 percentage points,* $ATE = -0.08$. *However, the difference in means estimator is inconsistent, with* $\widehat{ATE} = 0.75 \times 0.35 + 0.25 \times 0.39 - (\frac{1}{3} \times 0.4 + \frac{2}{3} \times 0.5) = -0.107$. *The inconsistency occurs because the unmeasured treatment confounds our estimate of the measured treatment, resulting in bias in the estimate, even though the texts are randomly assigned.*

Theoretical Importance of Proposition 1

Building on the intuition from these three examples, we now explain three implications of Proposition 1 for estimating the effect of latent treatments in texts.

Random Assignment of Texts Is Not Sufficient. Proposition 1 demonstrates a crucial point: To estimate the causal effect of a latent treatment, we have to make strong assumptions about how it relates to the unmeasured treatments in the text. The random assignment of texts to individuals addresses confounding by individual-level covariates, but it does not address the possibility that the presence of measured latent treatments in texts might be correlated with unmeasured latent treatments that also affect the outcome. Because it is impossible to alter the latent treatment without also altering the text and, in so doing, possibly altering unmeasured latent treatments, addressing this source of confounding requires the kinds of assumptions researchers have traditionally used to address confounding in observational studies.

Measured and Unmeasured Treatments to Define Intelligible Quantities of Interest. Our second contribution is to highlight the need for Assumption 3. Credibly estimating the effect for a measured latent treatment requires the analyst to assume that there exist some unmeasured treatments that, together with the measured latent treatments, fully characterize the expected response to documents, averaged across individuals. If analysts are unwilling to make this assumption, they would likely need Assumption 4 to be true for *all* possible unmeasured

TABLE 3 Example Potential Outcomes for Measured and Unmeasured Treatments

(a) Unmeasured Treatments, B, Affect Participation

	$\boldsymbol{B_i = 0}$	$\boldsymbol{B_i = 1}$
$Z_i = 0$	$\mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = 0)] = 0.40$	$\mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = 1)] = 0.50$
$Z_i = 1$	$\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 0)] = 0.35$	$\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 1)] = 0.39$

(b) Unmeasured Treatments, B, Do Not Affect Participation

	$\boldsymbol{B_i = 0}$	$\boldsymbol{B_i = 1}$
$Z_i = 0$	$\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 0)] = 0.40$	$\mathbb{E}[Y_i(Z_i = 0, \boldsymbol{B}_i = 1)] = 0.40$
$Z_i = 1$	$\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 0)] = 0.35$	$\mathbb{E}[Y_i(Z_i = 1, \boldsymbol{B}_i = 1)] = 0.35$

Note: This synthetic example considers two additional scenarios in which the unmeasured treatment affects the outcome and in which the unmeasured treatment does not affect the outcome.

treatments.[10] Otherwise, one of those latent treatments could confound their estimate. However, Assumption 4 cannot hold for all possible unmeasured treatments. To see why, consider a simple counterexample wherein the unmeasured treatment is identical to the measured treatment, except that 20% of the treated units are moved to the control group and 20% of the control units are moved to the treatment group. This unmeasured treatment will be correlated with the treatment and appears to affect the outcome; therefore, it will violate Assumption 4. Similarly, recent computer science studies condition on the text explicitly (Pryzant et al. 2020). However, this leads to ill-defined quantities, because it is impossible for the same text $\boldsymbol{X}_i$ to be used to adjust for differences when the latent treatment is present and when it is absent.[11]

Improving the Estimation of Latent Treatments. Proposition 1 provides helpful guidance for improving estimates of latent treatments in texts. Our results suggest that researchers should focus on conceptually distinct unmeasured treatments that are correlated with both the latent treatments and the outcome. As we show in the "Adjusting for Unmeasured Treatments" section, even if h is unknown and unknowable, simply knowing what types of unmeasured treatments are a threat to identification provides practical guidance to researchers.

[10]In Online Appendix F, p. 8, we show that there is a set of other knife edge conditions that could potentially satisfy this assumption as well. These knife edge conditions also cannot hold for all unmeasured treatments.

[11]In Pryzant et al. (2020), the goal is to estimate the ATE of a text feature T. In this setting, ATE $= \mathbb{E}_X(E[Y|g(\boldsymbol{X}_i) = 1, \boldsymbol{X}_i] - E[Y|g(\boldsymbol{X}_i) = 0, \boldsymbol{X}_i])$. However, either $g(\boldsymbol{X}_i) = 1$ or $g(\boldsymbol{X}_i) = 0$, which implies it is impossible to adjust directly for the text, $\boldsymbol{X}_i$.

Adjusting for Unmeasured Treatments

Although Assumption 4's requirements may be restrictive, intuition from observational research about how to mitigate the effect of omitted variables can be applied to text. We show in Online Appendix G, p. 9, that researchers can overcome violations of Assumption 4 by explicitly measuring and controlling for previously unmeasured treatments, just as researchers can overcome omitted-variable bias by measuring and controlling for previously omitted variables. The familiar advice to use subject area expertise and theoretical considerations to identify potential confounders, measure them, and adjust for them applies here with equal force.

Text-based research, however, has an advantage over researchers grappling with other kinds of omitted-variable bias: Every possible confounder is contained within the text. In other words, if a researcher can conceive of an unmeasured latent treatment, then the researcher can measure and adjust for it. In the "Applications" section, we provide two applications in which we illustrate the process of identifying latent confounders and adjusting for them.

The possibility of measuring and adjusting for confounders poses a new problem: Which types of unmeasured latent treatments should the analyst prioritize, and which types can be safely ignored? We offer two pieces of advice on this matter: Focus on prevalent, consequential features and avoid posttreatment variables.

First, we should prioritize unmeasured latent treatments that are common in the text and plausibly related to the outcome. As we show in Online Appendix F, p. 8, the difference between Equation (2) and Equation (1) becomes smaller as an unmeasured treatment becomes less prevalent and its correlation with the outcome decreases. Intuitively, unmeasured treatments that are rare or have

low correlation with the outcome do not explain enough variation to generate a misleading inference.

Second, avoid posttreatment features. For example, consider the tone of the advertisement. Even if it is possible to somehow deliver a negative advertisement with a positive tone or a positive advertisement with a negative tone, the tone is a consequence of the type of advertisement. Therefore, the tone is not a confounder; rather, it is a part of the latent treatment that makes up negative advertisements.

Vignette Experiments

Vignette experiments are the most common way texts are used as treatments in the social sciences. In a vignette experiment, the researcher constructs a small number of hypothetical situations (the eponymous "vignettes") and delivers one of them to each respondent. The treatments in vignette experiments are "latent" in the sense that they can only be manipulated by manipulating the text. The current best practice recommends designing two texts that are identical in all respects except for the presence or absence of the treatment of interest. Then, researchers attribute differences in the responses to the effect of the treatment.

Our framework from the section "Confounding by Unmeasured Treatments" suggests two threats to validity in vignette experiments. First, because the latent treatment of interest can only be manipulated by manipulating the text, the researcher may inadvertently change the value of some unmeasured treatment while attempting to change the treatment. Second, because the effect of the latent treatment may depend on the values of unmeasured treatments, vignettes estimate only a local ATE. The basic limitation with this design is that even as the number of individuals in the experiment grows, the number of distinct texts remains constant, which makes it impossible to marginalize over unmeasured latent treatments, unlike the setup presented in the section "Confounding by Unmeasured Treatments."

The issues we identify here have long been implicitly appreciated in the design of vignette experiments (Sniderman 2018; Sniderman and Grob 1996). However, our statistical framework offers a fresh perspective that suggests an unconventional solution: providing many vignettes per treatment. This is similar to the advice of running numerous separate vignette experiments (Sniderman 2018), but it does not require several distinct experiments and provides a natural framework for marginalizing over the results.

Aliased Treatments

When constructing vignettes, there is the risk that researchers will change more than just the measured treatment of interest. For example, Valenzuela and Michelson (2016) construct a vignette to assess how appeals to ethnic or community identity affect turnout decisions. To assess the effects of the ethnic/community appeal to turnout to vote, Valenzuela and Michelson (2016) employ an appeal that encourages respondents to recycle. The text with the ethnic appeal—labeled as the treatment in their experiment—differs from the control text—the text without the labeled treatment—in many ways. Only the treatment text includes an explicit encouragement to vote, reminds respondents of the month when the election will be held, and concerns a club good. The control text concerns a public good, is relevant to the environment, and so on. We cannot disentangle the effect of making an ethnic appeal from these other components of the text, even if we explicitly measure them. Similar issues persist in many other survey experiments (e.g., Clayton, O'Brien, and Piscopo 2018; Grimmer, Messing, and Westwood 2012).

The general problem is that the treatment of interest may be *aliased* by other measured or unmeasured latent treatments, which raises concerns about internal validity. Whenever the latent treatment changes, some other measured or unmeasured latent treatments might change simultaneously, rendering it impossible to attribute changes in the response to one or the other. One could redefine the latent treatment of interest to include these other features, but this greatly complicates the connection between the experiment and social science theories. In our framework, the vignette risks violating Assumption 4 because there are unmeasured latent treatments that are correlated with the latent treatment of interest.

Interaction between Measured and Unmeasured Latent Treatments

Vignette experiments only provide a local treatment effect. Extrapolating from this local treatment effect to the population of relevant texts requires the strong assumption that the many possible ways of delivering the latent treatment do not interact with the latent treatment, causing its effect on the outcome to change. For example, if we were to design a vignette experiment for campaign advertisements in which the treatment document was negative and the control document was not, the observed effect would be conditioned on whether the advertisement was policy focused, the complexity of the

language used, the length of the advertisement, the specificity of the claims, and whether the claims are backed by specific examples. If the effect of the negative treatment interacts with any of them, the effect estimated in the vignette experiment will be a local ATE and will not generalize to the desired population of advertisements—even if these unmeasured treatments are constant across the treatment and control documents. In short, if the measured treatments interact with the unmeasured treatments at all, then the estimated effect is not externally valid.

Improving Survey Experiments: Many Vignettes per Treatment

The framework from the "Confounding by Unmeasured Treatments" section suggests a surprising research design to address the issues of aliasing and local treatment effects: constructing many vignettes per latent treatment. To see why many vignettes are useful for addressing these issues, recall that our proof of Proposition 1 marginalizes over unmeasured latent treatments. In a standard vignette design—whether the design focuses on only a single treatment or incorporates many treatments as in a conjoint experiment—it is impossible to perform this marginalization over unmeasured treatments because there is only one vignette per treatment condition. However, with many vignettes per treatment condition, this marginalization is feasible. As a result, if the researcher later discovers an unmeasured latent treatment that is plausibly correlated with both the latent treatment of interest and the outcome, the researcher can then measure it and adjust for it. The study on protests in Hong Kong below illustrates the value of this research design and gives practical guidance for implementing it, such as on how to construct the vignettes, avoid aliasing, and make Assumption 4 credible.

Applications

We illustrate how to apply our recommendations with two applications that have distinct ways of constructing the texts and selecting the measured treatments. These two applications highlight a key advantage of using many texts per treatment condition: It allows the analyst to measure previously unmeasured latent treatments ex post, test whether they lead to a violation of Assumption 4, and adjust for them if necessary.

Hong Kong Protests

Our first experiment assesses how information about U.S. commitments to Hong Kong affects the public's preference for the U.S. government to support Hong Kong protesters. To make this assessment, we explicitly construct many vignettes per treatment so that the assumptions of the 'Confounding by Unmeasured Treatments" section are credible. In 2019, a new Chinese law that enabled extradition from Hong Kong to Mainland China triggered protests in Hong Kong. We ran an experiment in December 2019 and then replicated the experiment in October 2020.[12] Our experiment constructs texts composed of several statements designed to elicit support for the protesters. Based on congressional floor speeches, we identified several kinds of information often presented alongside our latent treatment of interest (we describe this process in greater detail in Online Appendix H, p. 11). Our constructed statements can include descriptions of the *commitment* the United States made to Hong Kong through the Hong Kong Policy Act of 1992, *bravery* the protesters displayed in risking physical harm, China's *mistreatment* of its own citizens outside of Hong Kong, protesters waving American *flags*, the security *threat* China poses to the United States, Hong Kong's political system and *economy*, and how China's actions are in *violation* of its treaty with the United Kingdom.

The latent treatment of interest is *commitment*. For observations that include this treatment, a description of America's commitment appears alongside one or two of the other latent treatments. In other words, we will say that this treatment is present if any of the vignettes that convey America's commitment to Hong Kong is present in the text. Control observations consist of two or three of the other latent treatments and so will be equal to zero if none of the texts convey America's commitments. Varying the number of latent treatments presented alongside the commitment treatment wherever it appears ensures that the presence of the commitment vignette is not aliased by the absence of another vignette.

To allow us to marginalize over potentially unmeasured latent treatments, we use 100 different versions of the commitment vignette. These versions vary (1) whether the law is described merely as a bill or the Hong Kong Policy Act; (2) the language and verb tense used to describe the commitment; (3) the presentation of the

[12]There is one difference between these experiments. All texts with three arguments included the latent treatment of interest in the original experiment. In the replication, the number of texts and the presence of the latent treatment of interest are independently randomized.

timing of the bill's passage—that is, whether it was in 1992, 27 years ago, during George H.W. Bush's administration, or some time ago; and (4) whether the commitment protects Hong Kong's freedom, autonomy, right to govern itself, or some combination thereof.

Similarly, we use many different versions of the bravery, mistreatment, flags, threat, economy, and violation treatments. Overall, there are 555,660 distinct potential texts that could be assigned.

We randomly construct texts from the constituent arguments, randomly assign texts to individuals, and ensure all individuals see only one text over the course of the experiment, which guarantees that Assumptions 1 and 2 hold. We ask respondents how much they agree with the statement that the United States should help Hong Kong protesters on a scale from 0 to 100. We design our experiment with a predetermined codebook function, g. We call the vignettes each respondent sees $\boldsymbol{X}_i$ and we apply a codebook function $g : \mathcal{X} \rightarrow \{0, 1\}^7$. The respondent receives the commitment treatment if their text includes a commitment vignette. Because the treatments are all independently randomized, we estimate the AMCE—a generalization of the ATE—for each treatment using a linear regression.

Table 4 provides the AMCE for the latent treatments. By marginalizing over the many versions of the treatment we present, we find a statistically significant effect for the commitment treatment in both the original and replication experiments. The AMCEs for the other features are all smaller, and we fail to reject the null that they are zero in both experiments.

This design helps to minimize the risk of aliasing. The information presented alongside the commitment treatment varies by design, and we deliberately vary the text of the commitment argument itself. This allows us to explicitly marginalize over the other measured treatments. For any unmeasured treatments, this preserves our ability to also measure and marginalize over them ex post.

For example, someone could object that many of the texts reference dates in the late 1980s and early 1990s—dates referencing the timings of the Hong Kong Policy Act, transfer of Hong Kong from Britain to China, and Tiananmen Square massacre. We call this unmeasured treatment "dates." There are several plausible mechanisms by which dates could affect the outcome. It might prime memories of a time of American hegemony and thereby encourage aggression, make the relevant commitments and antagonisms seem longstanding and therefore more important, or make them seem antiquated. Our design allows us to measure this hitherto unmeasured latent treatment ex post, test whether it is correlated with the outcome and the treatment of interest, and, if necessary, measure and adjust for it.

TABLE 4 Hong Kong Experiment Treatments

	December 2019	October 2020
Intercept	64.23	69.03
	(3.14)	(1.07)
Commitment	5.23	2.68
	(1.74)	(1.23)
Bravery	−0.72	1.85
	(1.82)	(1.38)
Mistreatment	0.97	0.14
	(1.77)	(1.39)
Flags	0.04	−2.12
	(1.81)	(1.41)
Threat	−2.50	−2.07
	(1.86)	(1.36)
Economy	−0.44	−0.94
	(1.84)	(1.35)
Violation	−0.98	0.75
	(1.81)	(1.38)
N	1,983	2,072

Note: Results come from a linear model, in which the outcome is the degree to which the respondent agrees the U.S. government should help Hong Kong. The left column refers to the original experiment and the right column refers to a replication experiment.

Regressing commitment on dates, we find a statistically significant coefficient in both the original and replication experiments (0.24 with a standard error of 0.02 and 0.22 with a standard error of 0.02, respectively). Regressing the outcome on dates, we find a statistically significant coefficient in the original experiment but not the replication (2.29 with a standard error of 1.24 and 1.32 with a standard error of 1.23, respectively). These relationships suggest that Assumption 4 may be violated. Thus, the safest course is to explicitly measure the dates treatment and include it in the regression of the response on all of the latent treatments to adjust for it. The results of this regression are substantively identical, yielding estimated effects of 4.74 (s.e. 1.83) in the original experiment and 2.54 (s.e. 1.16) in the replication experiment.

The key advantage of this design is that it allows the analyst to address threats to inference without running

a new experiment. If we had instead ran a vignette experiment that compared a speech with the commitment treatment to an otherwise identical speech without it, we could not be certain whether commitment was aliased by, among other things, information about the timing of American policy toward Hong Kong and the invocation of high concepts such as freedom. If information about timing were altogether absent and somebody protested that people would feel differently if they knew that the events in question happened more than 20 years ago or that the presence of information about timing created a priming effect that confounded the treatment effect, the analyst would have no recourse but to rerun the experiment. Because we allow these unmeasured treatments to vary, we can marginalize over them. If an unmeasured treatment seems to threaten the inference, the analyst can always measure it explicitly and adjust for it.

Reaction to President Trump's Messages

As a second example, we perform an analysis that provides us with less control over the content of the messages and uses machine learning to discover the measured latent treatments. We analyze how features of President Trump's tweets affect citizens' evaluations of those messages. A large literature seeks to understand when and how the president is able to affect public opinion (Edwards 2006). The bulk of it merely examines how the public responds to the act of giving a speech rather than to the content of the speech generally (Canes-Wrone 2010; Cohen 1995; Edwards 2006). We use our framework to explore how survey respondents evaluate the content of President Trump's speeches.

We draw data from YouGov's TweetIndex data from February 4, 2017, to October 31, 2017, which regularly presents citizens with recent tweets from President Trump and asks them to rate the tweet as "Great," "Good," "OK," "Bad," or "Terrible." YouGov then aggregates these responses by party identification (Independent, Democrat, and Republican) and rescales them from -200 to 200. We extend and apply the procedure from Fong and Grimmer (2016) to identify features in President Trump's tweets that explain these various party groups' evaluations and then estimate the effect of these discovered features. Each tweet produces three observations: one for the Independents, one for the Democrats, and one for the Republicans.

The goal is to make substantively interesting and statistically rigorous causal claims about why some of Trump's tweets receive more favorable reception than others. We first present the results of an analysis based on a plausible design. We then carefully consider what would have to be true of the experiment and discovered treatments for that analysis to credibly identify causal effects.

Discovering Latent Treatments and Estimating Their Effect in President Trump's Tweets. Because there are many potential treatments, we extend a machine learning procedure to discover treatments that are both common in the corpus and plausibly related to the outcome. Fong and Grimmer (2016) and Egami et al. (2018) show that using the same data to discover treatments and estimate effects leads treatment effects to be undefined. They show that this problem can be avoided by dividing the data into a training set for discovering treatments and a test set for estimating their effects.

In this application, we allocate two-thirds of the data to the training set and one-third to the test set. We cluster assignment to the training and test sets at the tweet level, such that if, for example, the Independents' response for a given tweet is in the training sample, then the Republicans' and Democrats' responses to that same tweet are also in the training sample.

Following Fong and Grimmer (2016), we fit a supervised Indian Buffet Process (sIBP) in the training set for different starting points and parameter configurations, and then we manually select the most substantively interesting treatments discovered.[13] We then use the mapping from text to latent treatments discovered in the training set to infer the values of the latent treatment in the test set, and estimate effects using the test data. To accommodate this particular application, we extend the sIBP to discover treatments and estimate effects where we expect an individual's response to vary substantially based on some measured characteristic. This is a reasonable expectation when comparing how Democrats, Republicans, and Independents react to Trump's rhetoric.

Although we might be concerned that we have to impose additional assumptions to justify using a machine learning procedure, it is merely a statistical tool to determine the codebook function. Instead, our primary concern is that there is some latent confounder that is not included in our codebook function whose effects might confound the estimate of the latent treatment of interest's effect.

[13]We tune the parameters via a grid search with five treatments, $\alpha \in \{2, 3, 4\}$, and $\sigma_n^2 \in \{0.50, 0.75, 1\}$. Five different initializations are run for each parameter configuration, and, consistent with the recommendations by Egami et al. (2018), we select the run with the most interesting treatment (in this case, the first run with $\alpha = 3$ and $\sigma_n^2 = 0.50$). These treatments are inferred in the test set, where $Z_{i,j} = 1$ if $\eta_{i,j} > 0.50$ with $\eta_{i,j}$ as defined in Fong and Grimmer (2016).

TABLE 5 Words Most Strongly Associated with Treatments

Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
fake	cuts	obamacare	flotus	prime
news	strange	senators	behalf	minister
media	tax	repeal	anthem	korea
cnn	luther	healthcare	melania	north
election	stock	replace	nfl	stock
story	market	republican	flag	market
nbc	alabama	vote	prayers	china
stories	reform	republicans	bless	executive
hillary	record	senate	ready	prayers
clinton	high	north	players	order

Note: Latent treatments were obtained from a supervised Indian Buffet Process. The listed words are the most characteristic of the latent treatment.

Table 5 shows the words most associated with each of the five discovered treatments. Treatment 1 corresponds to tweets that attack the media or Hillary Clinton (note that the data come entirely from after the 2016 election). Treatment 2 includes a mixture of Trump's economic achievements along with his efforts on behalf of Luther Strange in the special election for the Alabama Senate seat. Treatment 3 focuses on Trump's efforts to repeal and replace the Affordable Care Act. Treatment 4 includes Trump's commentary on the national anthem protests and miscellaneous symbolic topics. Treatment 5 combines Trump's foreign policy (especially in North Korea) with advertising excellent stock market performance. Online Appendix I, p. 12, provides three examples of test set tweets for each treatment.

Because we again have many measured treatments of interest, we focus on the AMCE via linear regression. Hainmueller, Hopkins, and Yamamoto (2013) show that linear regression estimates the AMCE under conditionally independent randomization. Although we cannot guarantee conditionally independent randomization with this design, we show in the Online Appendix I, p. 13, that the correlations between treatments are low, the inclusion of first-order interactions does not improve model fit, and the key results are robust to the inclusion of first-order interactions.

The results of Table 6 reveal surprising patterns in how the public responds to Trump's tweets. One important pattern is that an issue that was supposedly very divisive, the National Football League (NFL) national anthem controversy, was comparatively popular among all groups. Given the extremely low-baseline evaluation for Democrats, it is inappropriate to say that they liked these tweets—it is more accurate to say that they disliked them less. Even so, it is surprising that they are relatively uncontroversial. Importantly, Table 6 shows that divisive language is generally unpopular. Trump's tweets assailing Clinton and the media are less popular with Republicans than his average tweet. This contradicts suppositions that his messaging on "fake news" strongly appeals to his base.

This suggests an intriguing hypothesis. Many have conjectured that Republicans are drawn to Trump's unique positions, including his adversarial relationship with the mass media. Given the relative weakness of Treatment 1 compared with Treatments 2 and 4, our results suggest that Republicans are actually drawn to Trump as a conventional Republican, and his idiosyncrasies are liabilities.

Plausibility of Assumptions. The substantive results of the preceding analysis rest on causal claims: For example, respondents of all partisan affiliations responded less favorably to tweets about fake news because those tweets were about fake news and would have responded differently if they had been about something else. As we have shown, causal identification of the effects of these latent treatments requires four assumptions.

Assumption 1 (stable unit treatment value assumption) requires that a subject's response depend on only the tweet that subject is shown, and not on tweets the other subjects are shown. This assumption would be violated if subjects gave the tweet they were shown a more positive evaluation when Trump had recently made a series of tweets that they liked. This is a strong assumption; many subjects likely see Trump's tweets outside of the context of the study. Those tweets likely affect their evaluation of Trump, and their opinion of Trump likely influences how they evaluate his tweets. The problem would be exacerbated if YouGov used the exact same subject pool to measure the evaluations of each tweet. Further, we envision that the broad reaction to Trump's tweets—including media coverage and

TABLE 6 Regression of Tweet Favorability on Latent Features

	Democrats		Independents		Republicans	
	1	2	3	4	5	6
Intercept	−76.57	−88.50	2.32	−6.33	98.52	91.73
	(1.67)	(2.16)	(1.32)	(1.71)	(1.06)	(1.38)
Treatment 1	−45.22	−33.83	−31.75	−27.27	−18.91	−15.28
	(5.18)	(5.02)	(4.03)	(3.94)	(3.27)	(3.20)
Treatment 2	10.59	8.18	10.89	9.21	16.45	15.09
	(8.91)	(8.55)	(7.02)	(6.78)	(5.63)	(5.45)
Treatment 3	−35.93	−34.42	−25.69	−24.73	−10.27	−9.41
	(4.78)	(4.58)	(3.74)	(3.61)	(3.02)	(2.92)
Treatment 4	18.14	18.98	22.66	23.41	16.89	17.37
	(8.37)	(8.02)	(6.48)	(6.25)	(5.29)	(5.11)
Treatment 5	34.11	33.32	18.84	18.32	7.84	7.39
	(7.40)	(7.09)	(5.82)	(5.62)	(4.68)	(4.52)
Pos. Sent		23.74		17.23		13.50
		(2.89)		(2.29)		(1.84)
N	752		752		752	

Note: Results come from a linear regression of how respondents feel toward a given tweet on the latent treatments. The even-numbered columns add positive sentiment, to test whether the results are robust to possible confounding by this previously unmeasured treatment.

other online conversations—are part of the treatment of Trump's messages.

Assumption 2 (random assignment) requires that the tweets be randomly assigned to subjects, conditional on observed covariates. In this case, the assumption would be violated if YouGov systematically presented tweets to subjects based on how it expected them to respond to those tweets. This is not an explicit part of the sampling design, and so we have good reason to believe that individuals are not assigned tweets based on their likely response.

Assumption 3 (existence of a set of unmeasured treatments that, along with the measured treatments, fully characterize the response) can be cast as a regularity assumption. It cannot be tested empirically, but the idea that there exists some set of latent treatments that fully characterize the expected response to documents seems at least plausible.

Adjusting for Previously Unmeasured Treatments to Address Assumption 4. With regard to Assumption 4, we might be concerned that some tweets in the corpus express positive sentiment whereas others express negative sentiment. Perhaps Treatments 1 and 3 perform poorly because they tend to express negative sentiment, whereas the others perform well because they tend to express positive sentiment. To test this hypothesis, we use the sentiment dictionary developed by Nielsen (2011) and as implemented by Silge and Robinson (2016) to categorize tweets as having either positive or nonpositive sentiment.

Table 7, which reports the results of a regression of positive sentiment on the latent treatments, shows texts from Treatment 1 tend to have less positive sentiment. This negative correlation between Treatment 1 and positive sentiment raises concerns that Assumption 4 is violated.

Following section "Adjusting for Unmeasured Treatments," we explicitly adjust for sentiment as a previously unmeasured treatment in columns 2, 4, and 6 of Table 6. Although positive sentiment indeed seems to affect the outcome and controlling for it attenuates the effect of Treatment 1 (especially for Democrats), our substantive results are unchanged. This is only a single example of an unmeasured treatment, but it illustrates

TABLE 7 Regression of the Sentiment Score on Treatments

	Coefficient	Standard error
Intercept	0.50	(0.02)
Treatment 1	−0.27	(0.06)
Treatment 2	0.10	(0.11)
Treatment 3	−0.06	(0.06)
Treatment 4	−0.04	(0.10)
Treatment 5	0.03	(0.09)
N	752	

Note: Results come from a linear regression of a tweet's sentiment score on its latent treatments. Sentiment is negatively correlated with Treatment 1.

how our framework allows the analyst to determine whether any given unmeasured treatment threatens identification. If the unmeasured treatment varies in the texts, as will typically be the case when the texts are generated by a real-world process as in this study, the analyst can then measure and adjust for the unmeasured treatment. This exercise could be repeated for any unmeasured treatment that may confound our estimate of a measured treatment of interest.

Conclusion

Randomly assigning texts to individuals is not sufficient to identify the effects of latent treatments, because the latent treatment is not randomly assigned to the text. If there are unmeasured latent treatments correlated with the latent treatment and affect the outcome, then estimates of the ATE will be confounded by those unmeasured treatments. Accordingly, researchers must make an assumption analogous to the no omitted variables assumption commonly used in observational studies: The unmeasured treatments are either uncorrelated with the measured latent treatment of interest or do not affect the outcome.

Researchers who would estimate latent treatment effects in texts have one crucial advantage over observational studies: All latent treatments, measured and unmeasured, are contained within the text. By combining substantive knowledge with a close reading of the texts, the analyst can identify potential unmeasured treatments that might confound the estimates, measure them, test whether they are in fact confounding the estimate, and adjust for them if necessary. Doing so requires a research design that affords access to many different texts. In contrast to traditional vignette experiments, which provide valid pairwise estimates of the effect of one text relative to another but risk aliasing and estimate local ATEs, the many text designs allow estimation of the ATE and permit the researcher to measure and adjust for potential confounders, even after the experiment has already been run. We highlight the utility of this approach with two applications, one in which we design the texts to make the identification assumptions as plausible as possible and another in which we study the effects of naturally occurring texts.

Although we focus on text applications, our analysis and recommendations also apply to images, video, and audio data. Latent treatments also come from more exotic data sources—ideologies from roll-call vote matrices (Poole and Rosenthal 2000), personalities from personality tests (Hurtz and Donovan 2000), and aptitudes from test scores (Hoxby and Terry 1999). Although the strategies for making our assumptions more plausible were designed with texts in mind and would be difficult to apply in these other settings, the analytic framework in the section "Confounding by Unmeasured Treatments" applies in these situations as well and provides a springboard for more narrowly tailored methods. Moreover, our approach to identify the effect of measured latent treatments in the presence of unmeasured latent treatments is similar to the approach of Sävje, Aronow, and Hudgens (2017), who focus on defining estimands and identification when there is interference across units. These authors define an estimand similar to Equation (1), but their estimand marginalizes over different treatment assignment vectors, capturing different potential consequences from interference across units rather than unmeasured treatments in a higher dimensional treatment. Both approaches marginalize over high-dimensional interventions and suggest a way to make progress on previously intractable causal inference problems: If we can make defensible simplifying assumptions about how higher dimensional interventions affect responses, then we can define estimands of interest and devise strategies for estimation that would otherwise be impossible to define.

References

Acharya, Avidit, Matthew Blackwell, and Maya Sen. 2016. "Explaining Causal Findings without Bias: Detecting and Assessing Direct Effects." *American Political Science Review* 110(3): 512–29.

Ansolabehere, Stephen D., Shanto Iyengar, and Adam Simon. 1999. "Replicating Experiments Using Aggregate and

Survey Data: The Case of Negative Advertising and Turnout." *American Political Science Review* 93(4): 901–09.

Arceneaux, Kevin, and David W. Nickerson. 2010. "Comparing Negative and Positive Campaign Messages: Evidence from Two Field Experiments." *American Politics Research* 38(1): 54–83.

Canes-Wrone, Brandice. 2010. *Who Leads Whom?: Presidents, Policy, and the Public*. Chicago, IL: University of Chicago Press.

Clayton, Amanda, Diana Z. O'Brien, and Jennifer M. Piscopo. 2018. "All Male Panels? Representation and Democratic Legitimacy." *American Journal of Political Science* 63(1): 113–29.

Cohen, Jeffrey E. 1995. "Presidential Rhetoric and the Public Agenda." *American Journal of Political Science* 39(1): 87–107.

Coppock, Alexander, Emily Ekins, and David Kirby. 2018. "The Long-Lasting Effects of Newspaper Op-Eds on Public Opinion." *Quarterly Journal of Political Science* 13(1): 59–87.

Dafoe, Allan, Baobao Zhang, and Devin Caughey. 2018. "Information Equivalence in Survey Experiments." *Political Analysis* 26(4): 399–416.

Edwards, George C. 2006. *On Deaf Ears: The Limits of the Bully Pulpit*. New Haven, CT: Yale University Press.

Egami, Naoki, Christian Fong, Justin Grimmer, Margaret E. Roberts, and Brandon Stewart. 2018. "How to Make Causal Inferences Using Texts." Princeton University Mimeo. https://scholar.princeton.edu/sites/default/files/bstewart/files/ais.pdf

Fong, Christian, and Justin Grimmer. 2016. "Discovery of Treatments from Text Corpora." In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1600–1609.

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): 703–19.

Hainmueller, Jens, Daniel J. Hopkins, and Teppei Yamamoto. 2013. "Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices Via Stated Preference Experiments." *Political Analysis* 22(1): 1–30.

Hoxby, Caroline M., and Bridget Terry. 1999. "Explaining Rising Income and Wage Inequality Among the College Educated." National Bureau of Economic Research. https://www.nber.org/papers/w6873.pdf.

Hurtz, Gregory M., and John J. Donovan. 2000. "Personality and Job Performance: The Big Five Revisited." *Journal of Applied Psychology* 85(6): 869.

Imai, Kosuke, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. "Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies." *American Political Science Review* 105(4): 765–89.

Kalla, Joshua L., and David E. Broockman. 2017. "The Minimal Persuasive Effects of Campaign Contact in General Elections: Evidence from 49 Field Experiments." *American Political Science Review* 112(1): 1–19.

Nielsen, F. Å. 2011. "AFINN." http://www2.imm.dtu.dk/pubdb/p.php?6010

Offer-Westort, Molly, Alexander Coppock, and Donald P. Green. 2019. "Adaptive Experimental Design: Prospects and Applications in Political Science." https://onlinelibrary.wiley.com/doi/10.1111/ajps.12597.

Pearl, Judea. 2001. "Direct and Indirect Effects". In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc. pp. 411–20.

Poole, Keith T., and Howard Rosenthal. 2000. *Congress: A Political-Economic History of Roll Call Voting*. Oxford: Oxford University Press.

Pryzant, Reid, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2020. "Causal Effects of Linguistic Properties." arXiv preprint arXiv:2010.12919 .

Sävje, Fredrik, Peter M. Aronow, and Michael G. Hudgens. 2017. "Average Treatment Effects in the Presence of Unknown Interference." arXiv preprint arXiv:1711.06399 .

Silge, Julia, and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *The Journal of Open Source Software* 1(3): 37.

Sniderman, Paul M. 2018. "Some Advances in the Design of Survey Experiments." *Annual Review of Political Science* 21: 259–75.

Sniderman, Paul M., and Douglas B. Grob. 1996. "Innovations in Experimental Design in Attitude Surveys." *Annual Review of Sociology* 22(1): 377–99.

Valenzuela, Ali A., and Melissa R. Michelson. 2016. "Turnout, Status, and Identity: Mobilizing Latinos to Vote with Group Appeals." *American Political Science Review* 110(4): 615–30.

Zaller, John R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: The Extensive Use of Latent Treatments, Extended Table
Appendix B: Table of Terminology
Appendix C: Relationship to Causal Mediation
Appendix D: Relationship to Dafoe, Zhang, and Caughey (2018)
Appendix E: Alternative Representation of DAG
Appendix F: Proof of Proposition 1
Appendix G: Generalization to the AMCE
Appendix H: Hong Kong Message Experiment
Appendix I: Trump Message Experiment