Sensitivity Analysis for Causal Mediation through Text: an Application to Political Polarization

Graham Tierney and **Alexander Volfovsky**
Department of Statistical Science
Polarization Lab
Duke University
{graham.tierney,alexander.volfovsky}@duke.edu

Abstract

We introduce a procedure to examine a text-as-mediator problem from a novel randomized experiment that studied the effect of conversations on political polarization. In this randomized experiment, Americans from the Democratic and Republican parties were either randomly paired with one-another to have an anonymous conversation about politics or alternatively not assigned to a conversation — change in political polarization over time was measured for all participants. This paper analyzes the text of the conversations to identify potential mediators of depolarization and is faced with a unique challenge, necessitated by the primary research hypothesis, that individuals in the control condition do not have conversations and so lack observed text data. We highlight the importance of using domain knowledge to perform dimension reduction on the text data, and describe a procedure to characterize indirect effects via text when the text is only observed in one arm of the experiment.

1 Introduction

Increasing large and varied text corpora are becoming available to researchers. Especially in the field of computational social science, text data are yielding new insights and opening new areas of study (Grimmer and Stewart, 2013; Gentzkow et al., 2019a; Salganik, 2019). Text are being used to study who sets the political agenda (Barberá et al., 2019), measure partisanship in congressional speeches (Gentzkow et al., 2019b), and legislator attitudes expressed on Twitter (Spell et al., 2020).

An understudied area of this rapidly expanding field is how to perform causal inference with text data, especially when text is the treatment or outcome (Keith et al., 2020). Text data pose unique challenges due to the complex and high-dimensional structure that they impose on the already complex task of causal inference. For instance, concerns that social media might be causing political polarization by increasing ideological segregation (Bakshy et al., 2015; Barberá, 2015) or spreading disinformation (Lazer et al., 2018) must disentangle the effect of exposure to social media text on polarization and the propensity of highly polarized individuals to be active on social media.

The contributions of this paper are two-fold: (1) We introduce a sensitivity analysis for mediation when the information on the mediator is observed for only one group in an experiment. (2) We demonstrate the difficulty of using unsupervised text models for performing a causal analysis.

To illustrate these contributions, we analyze data from a novel randomized experiment studying the dynamics of political polarization (Bail, 2021; Combs et al., 2021). American voters were randomly paired across party lines to have meaningful, political conversations. Of particular interest is the text of the conversations. We want to characterize the kind of conversations that cause depolarization. This is inherently a question of *causal mediation*: we are interested in both the direct effect of simply having a conversation and the indirect effect of the conversation content.

The study we analyze follows a common design where information must be observed asymetrically for the different treatment arms. Specifically, text data are only available for the units who had conversations and are not available for control individuals, making explicit causal mediation impossible. This asymmetry is required by the research question. The original study sought to estimate the effect of *having a conversation with an out-partisan about politics*, which necessitates the control group having no conversation. Contrasting treatment with a control that had a non-political conversation, for example, would only estimate the effect of having a conversation with an out-partisan *about politics*.

This asymmetry is also a common feature of peer encouragement or counseling studies where treated units have a conversation with a peer or

Proceedings of CI+NLP: First Workshop on Causal Inference and NLP, pages 61–73

professional and control units have no interactions (Anderson et al., 2005; Carandang et al., 2020; Malchodi et al., 2003). The natural followup to such studies is to identify what kind of conversations produce the best results. However, the experimental data cannot answer such questions because those mediators are missing for the control group. To accommodate such designs, we develop a procedure for a *mediation sensitivity analysis* designed to benchmark the observed correlation with the causal mediation effects that could be observed if control units had produced text data.

Our work also highlights the difficulty of using unsupervised models of text in the context of causal inference. In our application, we have expert knowledge of likely causal pathways, but, because these models do not leverage this expert knowledge, they are unable to identify semantically meaningful conversation features. Moreover, the features they do identify appear to not play a significant role in the causal story of the experimental data.

The outline of the paper is as follows: First, we introduce our illustration data and describe the causal and text complications. Section 3 describes the causal inference framework, concentrating on effect mediation and on how to incorporate text data into causal inference pipelines, and prior work on these issues. Section 4 demonstrates our finding that partner politeness is correlated with depolarization. Section 5 develops the mediation sensitivity procedure for this result. Section 6 demonstrates the failure of unsupervised text models to capture causal mechanisms. Section 7 concludes.

2 Polarization and Insights from Text

A growing body of research raises concerns that social media are increasing ideological segregation and incivility between political parties in the United States (Bakshy et al., 2015; Barberá, 2015). A particular concern, is that increasing exposure to elites from the other side may produce a backlash effect that increases polarization (Bail et al., 2018). In this paper we revisit data from a study that experimentally tested whether prolonged out-group contact could decrease polarization in the U.S. by having Democrats and Republicans engage on an anonymous chat platform (Bail, 2021; Combs et al., 2021).

In February 2020, approximately 1,500 Republicans and Democrats were recruited by a prominent survey firm and given a survey to measure their political views. The questions covered both issue-based polarization (how close one is to each party on policy views) and affect-based polarization (one's sentiment toward the other party). Individuals randomized to treatment were sent a seemingly-unrelated invitation to download a mobile chat application within 48 hours of completing the survey. Each person who logged into the app was randomly assigned a partner from the other party and prompted to discuss either gun control or immigration. Politics were not mentioned in the recruitment dialog or prior to logging into the app. After the conversations were finished, another seemingly unrelated survey was sent to all study participants that contained the same questions as the first survey. Within-person depolarization was measured by averaging the difference in post- and pre-survey responses to all polarization questions. The original study reveals that individuals assigned to treatment became significantly more depolarized (Combs et al., 2021; Bail, 2021, Appendix).

This paper focuses on analyzing the text of the conversations. We want to identify what conversation features caused the depolarization. However, we cannot rely on the randomization to identify causal effects for two reasons. The first and primary reason is that the control group did not have conversations, so we cannot compare treatment and control text. The original experiment was designed to estimate the effect of *having a conversation about politics* with an out-partisan, and as such, individuals in the control group had no conversations.

The second reason we cannot rely on the randomization to identify causal effects is that the conversation text is properly thought of as a mediator variable, something causally affected by treatment that also affects the outcome. Randomized treatment does not guarantee that there is no unobserved confounding of the mediator-outcome relationship. The conversation itself is both an outcome and exposure of interest. Use of text in both of those contexts is challenging, and the subject of much of our literature review.

We address these challenges by first, relying on prior work that identifies politeness or civility as a key feature in producing persuasive text, measuring politeness of messages received by individuals in treatment, then linking that measure to the outcome of interest, depolarization. Next, we impute politeness measures for control individuals in a way that

preserves the observed relationship between politeness and depolarization among the treated and test how large the difference in politeness between treatment and control must be to observe significant mediation.

3 Causal Inference and Prior Work

The general goal behind causal inference is to understand "what if?" questions — what happens to an outcome of interest Y if one intervenes to set variable T. Information is frequently collected alongside the intervention and outcome; modern causal inference is often concerned with how this additional information interacts with the causal effects of interest. In this paper we are concerned with one such mechanism — that of causal mediation — and how text acts as part of that mechanism.

3.1 Mediation Analysis

Causal mediation analysis attempts to understand the mechanisms through which exposures affect an outcome (Pearl, 2014; VanderWeele, 2015). A mediator M is a variable that is causally affected by treatment T and also affects the outcome of interest Y. The directed acyclic graph (DAG) of such a process is depicted in Figure 1.

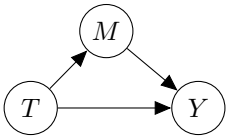

Figure 1: Causal DAG for treatment T, mediator M, and outcome Y.

This can be clearly represented in terms of potential outcomes, let $Y_i(T_i)$ note the potential outcome for unit i if treatment were T_i (Rubin, 1974). The total effect (TE) or average treatment effect is defined as $E[Y_i(1)] - E[Y_i(0)]$ (where the expectation is taken over the finite experimental population or over some global population distribution). This effect can be *identified* from simple observable quantities if the following three assumptions are satisfied: the potential outcome for unit i depends only on T_i (SUTVA), treatment is assigned independent of the potential outcome $Y_i(0), Y_i(1) \perp T_i | X_i$ (conditional independence or unconfoundedness), and for all i, $P(T_i = t) > 0$ for all t (positivity). A randomized experiment guarantees that conditional independence and positivity hold.

For mediation, we consider the decomposition of this total effect into *the natural direct effect* and the *natural indirect effect*, hereafter referred to as simply direct and indirect effects (Pearl, 2001). The direct effect measures the difference in expected outcome when changing treatment but holding the mediator at its natural level were treatment fixed, and the indirect effect measures the difference in expected outcome when holding treatment fixed and allowing the mediator change as it would naturally were treatment changed. These introduce a second level of potential outcome notation, let $M_i(T_i)$ note the potential outcomes for the mediator value for unit i when treatment is set to T_i. Additionally, let $Y_i(T_i, M_i(T_i))$ be the potential outcome Y_i as a function of both treatment and mediator for unit i. Note $Y_i(T_i) = Y_i(T_i, M_i(T_i))$. The total effect can be decomposed into:

$$TE = E[Y(1, M(1))] - E[Y(0, M(1))] + \quad (1)$$
$$E[Y(0, M(1))] - E[Y(0, M(0))] \quad (2)$$

The first line is the direct effect, the second is the indirect effect.[1] The direct effect corresponds to the effect measured along the direct $T \rightarrow Y$ path in the DAG in Figure 1, while the indirect effect is measured along the $T \rightarrow M \rightarrow Y$ path.

Additional assumptions are required to estimate the direct and indirect effects as functions of observed data. For details, see VanderWeele and Vansteelandt (2009); Nguyen et al. (2020). Essentially, the same assumptions regarding T_i and $Y_i(T_i)$ for estimating the TE are required for the mediator-outcome and treatment-mediator relationships with an additional assumption regarding cross-world quantities $M_i(1)$ and $Y_i(0, M_i(1))$. Importantly, randomizing T_i does not guarantee the $M \rightarrow Y$ relation from Figure 1 is unconfounded.

A variety of estimation methods exist for mediation analysis, which depend on the target estimands and modeling assumptions. For a full overview see Nguyen et al. (2021). Imai et al. (2010) provide a commonly used general method and framework for estimating both parametric and non-parametric models for mediation analysis. To facilitate inference in this paper we will consider a set of structural equation models (Eq. 3 and 4, which are discussed in detail in Section 5).

[1] The decomposition can also be performed by adding and subtracting $E[Y(1, M(0))]$, resulting in slightly different expressions of the direct and indirect effects. Under commonly used linear structural equation models that we apply in Section 5.2, these expressions are equivalent (VanderWeele, 2016).

3.2 Causal Inference with Text Data

Text data are extremely rich and can play important and complex roles in a causal pipeline. However, this richness makes satisfying the necessary assumptions that much harder. In large part, the difficulty arises from the fact that we want the actual text to play the role of outcome Y, treatment T, confounder X, or mediator M, but during analysis we must rely on summary measures of the text. These summaries might be parameters in an unsupervised text model, word counts, sentiment measures or other objects, but this requires replacing the existing assumptions with ones that include the summarization procedure.

Much of the prior work on text in causal settings has focused on text as a confounder, e.g. Saha et al. (2019) and Roberts et al. (2020). See Keith et al. (2020) for a full overview of the text-as-confounder literature. In general, these methods attempt to compare treated and control units with *similar* text or features inferred from text. In our work, we are interested in comparing polarization for participants who saw *different* text from their partner.

Methods directly related to our question of interest, linking the conversation features to depolarization, come from viewing text-as-treatment. These procedures generally transforming the high-dimensional text data into some low-dimensional representation, interpretable as treatments, and model outcome Y as a function of those treatments.

Wood-Doughty et al. (2018) studies the use of text classifiers to infer binary treatment when treatment is affected by measurement error or partial missingness. Fong and Grimmer (2016) discover latent binary treatments from text, which are then used to estimate causal effects using a supervised Indian Buffet Process (sIBP). Egami et al. (2018) develop a framework for causal inference with text as either treatment or outcome. They recommend using a train-test split where a low-dimensional representation of the text is learned on the training set, then causal quantities are estimated on the test set. We explore this procedure in Section 6.

To our knowledge, the only paper specifically about text-as-mediator is Veitch et al. (2020).[2] They develop a method for causal text embeddings by fine-tuning a pre-trained BERT network to estimate propensity scores and conditional outcomes. In our experiment, if control units had produced text, this method could decompose the treatment effect into the direct effect and total text-based mediation. However, the method would not produce interpretable measures of which text features are driving the depolarization. At the core of many mediation analyses is a desire to understand the causal pathways through which treatment operates.

3.3 Measuring Politeness in Conversations

The importance of politeness in conversations has been studied in the computational linguistics literature: Danescu-Niculescu-Mizil et al. (2013) identify politeness as a signifier of social power, Niu and Bansal (2018) and Firdaus et al. (2020) focus on generating polite dialogues, Kang and Hovy (2019) show that politeness can be delineated from the context, and Madaan et al. (2020) use style transfer techniques to translate polite and impolite messages while preserving the underlying context.

Politeness and incivility have also been studied in political contexts. Jaidka et al. (2019) find Twitter's increase of the character limit significantly improved the politeness of replies to politicians, Mutz and Reeves (2005) find that incivility in televised political disagreement harms citizens' political trust, and Papacharissi (2004) and Theocharis et al. (2016) highlight the role of civility in online political discourse in both citizen-citizen and citizen-politician interactions. All of the above emphasise the importance of civility and politness as necessary components of democratic deliberation.

4 Preliminary analysis of text data

The work on politeness and incivility lead us to focus on politeness as a potential mediator of depolarization. Individuals exposed to more polite partners may change their prior beliefs about the other side more than those exposed to less polite partners. We use the R package `politeness` to identify linguistic features of politeness with sentence parsing and dictionary methods (Yeomans et al., 2018). This package extends the work of Danescu-Niculescu-Mizil et al. (2013) and Voigt et al. (2017) by combing the indicators of politeness in both works.[3]

[2] Vig et al. (2020) apply causal mediation analysis to large, neural network-based language models to identify gender bias in the network itself. While this paper uses mediation tools, it is not focused on estimating causal effects with text as a mediator.

[3] When using the `convokit` Python library to identify politeness features, our results are unchanged (Chang et al., 2020). We use `politeness` in our main results because it identifies a broader set of features. The original application of these tools was not to political discussions of gun control and

To measure exposure to politeness, we extract all of the politeness features for each message sent by participants. Then, we standardize each feature to be mean zero variance one, and sum the results across all messages received by each user. *We focus on messages received rather than sent because we want to measure how one's conversation partner affects one's own depolarization.* We sum rather than average across politeness features because we want to measure total experienced politeness rather than a per-message measure. We wish to characterize a short and polite conversation as less exposure than a long and polite conversation. We refer to this constructed measure as the *politeness index*. Note that the random assignment of partners makes the no unmeasured confounding assumption plausible.

First, we examine the relationship between the politeness index and depolarization among the treated units who finished their conversations (completed at least 10 exchanges). We regress depolarization on politeness both in the full sample and split by party, controlling for demographic features. The results are shown in Table 1. We observe that participants with more polite partners depolarized significantly more; a 1 standard deviation increase in partner politeness is associated with a 0.069 standard deviation increase in depolarization. This relationship is stronger among Democrats with polite Republican partners.

We also look at how politeness and the relationship between depolarization and politeness differ by conversation topic. While participants assigned to talk about immigration were significantly more polite, there was no significant difference in the relationship between depolarization and politeness by topic. Figure 2 shows scatter-plots of politeness and the depolarization index by topic. Adding conversation topic to the regressions in Table 1 does not change any of results. Nevertheless, because control individuals do not have conversation topics assigned to them, we do not adjust for topic in the remaining analysis in the paper.

Crucially, we cannot conclude that the above are causal relationships based on these regressions. While partners were randomly assigned, partner *politeness* was not. If partner politeness is affected by one's own messages, as is likely the case, then there could be unmeasured confounding whereby certain participant features cause both polite replies by the conversation partner and depolarization. We believe our demographic controls cover many of the potential confounders but must rely on the untestable assumption that there are no other confounders. In the next section we develop a framework for understanding the causal nature of the relationship between conversations, the text and depolarization.

immigration. As such, we only use the context-independent politeness features, e.g. whether a participant expresses gratitude, not the specific projection of those features learned in a different, non-political context.

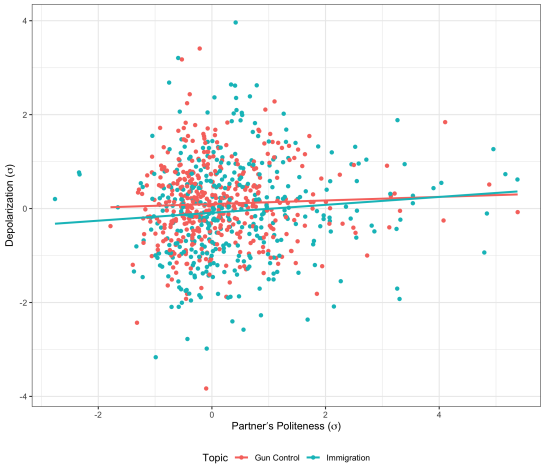

Figure 2: Depolarization and partner politeness. All units are in standard deviations. While those assigned to talk about immigration were 0.23 standard deviations more polite than those assigned to talk about gun control ($p < 0.01$), the relationship between politeness and depolarization was not significantly different across conversation topics (difference in slopes 0.04, $p \approx 0.50$).

5 Mediation Sensitivity

We would like to use politeness as a mediator and estimate the direct and indirect effects (introduced in Section 3) using the structural equation model:

$$Y_i = \tau T_i + \alpha M_i + \gamma X_i + \epsilon_i \quad (3)$$

$$M_i = \beta T_i + \theta X_i + \nu_i \quad (4)$$

where Y is the outcome (depolarization), M is the mediator (politeness), T is a binary treatment indicator, X is measured covariates, and ϵ and ν are error terms. The natural direct effect is τ, the difference in expected outcome while changing treatment and holding M fixed, the natural indirect effect is $\alpha\beta$, the difference in expected outcome while holding treatment fixed and allowing politeness to change as it would if treatment was changed. The total effect, identified by regressing Y on T *without* M, is $\tau + \alpha\beta$.

	All	Dem	Rep
	(1)	(2)	(3)
Politeness	0.069*	0.119*	0.002
	(0.035)	(0.046)	(0.052)
Constant	0.160	0.565	−0.294
	(0.215)	(0.288)	(0.312)
Observations	819	408	411

Table 1: Partner Politeness and Depolarization. Columns show results from regressing depolarization on partner politeness and demographic control variables. Column 1 uses all participants, column 2 only Democrats, and column 3 only Republicans. Participants with more polite partners significantly depolarized, especially Democrats. All regressions control for demographic factors: gender, age, education, race, and geographic region. Stars indicate statistical significance at the 5% level.

While control individuals clearly have a perceived level of politeness for the opposite party, because they did not participate in a conversation, there is no politeness to measure for them. This means β cannot be estimated. What we can do is benchmark the strength of the relationship identified in Table 1 by simulating politeness values for control units that preserve the relationships encoded in $p(Y, M|T = 1, X)$. The remainder of this section describes how the simulation and estimation should be conducted. Section 5.2 applies the procedure to the experimental data. We refer to simulated mediator values for control units as $\widetilde{M}$.

5.1 Procedure Description

The total effect $E[Y_i(1)] - E[Y_i(0)] = \tau + \alpha\beta$ will remain the same regardless of what values are imputed for the control mediator values $\widetilde{M}$ because it is estimated using Y and T alone. Similarly, α is a structural parameter that captures the relationship between politeness and depolarization—if simulated $\widetilde{M}$ changed structural parameter α, the estimated indirect effect, $\alpha\beta$, will not measure indirect effects as they would occur in real data. As such, we study the sensitivity of the decomposition of direct and indirect effects to the size of β, how much does treatment affect the mediator, while keeping $\tau + \alpha\beta$ and α fixed.

A naive approach of simulating $\widetilde{M}$ as random draws from observed M_i among treated units will not reflect information learned about α. With this simulation, Y and M are independent among control units, and α will shrink towards zero in the full sample. In this case, $p(M|T = 0)$ and $p(M|T = 1)$ will match, but the conditional outcome distributions will not, $p(Y|T = 0, M) = p(Y|T = 0)$ whereas among treatment M does provide information about Y.

In principle, one can use any model to learn the joint distribution $p(Y, M|T = 1, X)$, then using observed X and Y among control, impute or simulate $\widetilde{M}$ and continue. Using the outcome is often important to ensure that the mediator-outcome relationship is preserved. However, depending on the estimation method, it is not always necessary.

If one sets $\widetilde{M}$ to the linear projection of M on X learned among treated units ($\widetilde{M} = X_c(X_t^T X_t)^{-1} X_t^T M_t$) and uses OLS to estimate 3 and 4, then (a) the estimated $\widehat{\alpha}$ on the full sample will be exactly equal to the $\widehat{\alpha}$ estimated among treatment only, (b) the estimated $\widehat{\tau}$ will be exactly equal to the total effect, and (c) the estimated $\widehat{\beta}$ will be exactly 0. For proof see Appendix A. The intuition is that OLS estimates $\widehat{\alpha}$ and $\widehat{\tau}$ are learned from the residual variability in Y and M after removing the linear effects of X (Frisch and Waugh, 1933; Lovell, 1963). Because of the imputation, control units have no variability in M after removing the effects of X, so $\widehat{\alpha}$ is learned only from variability in treatment, preserving the structural parameter. $\widehat{\tau}$ and $\widehat{\beta}$ results follow because residual variability in M across T is zero.[4]

Implicit in this construction is the assumption that $p(Y|M)$ is the same for $T = 1$ and $T = 0$. In Equation 3, this assumption is encoded in the parameter α and the fact that there is no treatment-mediator interaction term. Note that α is not a parameter directly related to an intervention; it is the relationship between how polite an out-partisan is to someone and the change in that someone's feelings towards the other party. Simulations that change α among control units, i.e. do not assume that $p(Y|M)$ is the same for $T = 1$ and $T = 0$, would be simulating data that are inconsistent with observed relationships between variables among

[4]When linear projection is not desirable for imputing $\widetilde{M}$, matching methods can be used. Each control unit should be matched with treated unit(s) based on covariates X and Y, and $\widetilde{M}$ imputed from the matched treatment units where M is observed. This will not exactly preserve the relationships. As such, we recommend practitioners look at results such as Table 2 to ensure that the estimated relationship between Y and M is not too different between the full sample and treatment units.

treated units. In our application, $p(Y|M, T = 0)$ is not observable because the mediator cannot be measured for $T = 0$, so a direct assessment of the assumption is impossible. This assumption becomes testable only by changing the underlying research questions. For example, imagine if control participants were matched with an out-partisan to have a conversation about a non-political topic on which they disagreed. The assumption in question would imply that partner politeness affects depolarization the same whether the partner is talking about politics or not.

Once $\widetilde{M}$ has been generated, the remaining task is to test different β values to observe when significant mediation effects are detected. The implementation will depend on the particular mediation model specified. In Equation 4, we can set $\beta = c$ by subtracting constant c from all $\widetilde{M}$ values. Given our specification, c could be subtracted from control or added to treatment with identical results. However, for more complex models, this equivalency may not hold. We recommend changing the imputed values $\widetilde{M}$ rather than observed M among treatment to avoid changing any observed data. The range of values for β should be from 0 to the value such that the direct effect τ is zero. For the structural equation set up here, that is when $\beta = TE/\alpha$ where TE is the total effect, also estimable as the direct effect when $\beta = 0$.

5.2 Procedure Application

We set $\widetilde{M}$ equal to the linear projection described above. We show the empirical validation of the imputation properties in Table 2. Column 1 reports coefficients from regressing Y on M and X for only treated units, column 2 coefficients from regressing Y on T and X for all units, and column 3 the results for regressing Y on T, X, and M with the imputation as described above.

Next, we set β values between 0 and $\widehat{\tau}/\widehat{\alpha}$ from column 3 of Table 2, and compute direct and indirect effects (Tingley et al., 2014). We note that once α and β are fixed, point estimates of mediation effects are immediate. However, uncertainty quantification is not. We can construct $1 - \phi\%$ confidence intervals using a non-parametric bootstrap and identify the smallest β such that the null hypothesis that the indirect effect $\alpha\beta = 0$ is rejected with confidence level ϕ.

Figure 3 shows the results of our procedure with 95% confidence intervals. The total effect, as expected, remained constant, and the direct and indirect effects each range from 0 to the total effect. The smallest β where we reject the null hypothesis that the indirect effect is zero is 0.09 standard deviations. *This means that having a conversation with someone from the other political party only needs to move beliefs about politeness by about 0.09 standard deviations for politeness to be a significant mediator*. The imputation sets $\widetilde{M}$ to the best guess as to how a typical member of the other party interacts with each control participant using a linear projection from their demographic variables. If the baseline level of politeness was 0.09 or more standard deviations lower among control units, then we would have observed a statistically significant indirect effect from partner politeness, representing mediation of about 4% of the total effect. If treatment moved politeness by the same amount that switching the topic from gun control to immigration does (0.23 standard deviations), we would still have observed a statistically significant indirect effect, representing mediation of about 10% of the total effect.

	Treatment Only	Total Effect	Imputed Politeness
	(1)	(2)	(3)
Treatment		0.156*	0.156*
		(0.076)	(0.076)
Politeness	0.069*		0.069*
	(0.035)		(0.035)
Constant	0.159	0.119	0.134
	(0.214)	(0.202)	(0.202)
N	819	1,037	1,037

Table 2: Politeness Imputation Results. Results from regressing depolarization on (1) partner politeness among treatment only, (2) treatment indicator among all units, and (3) treatment indicator and partner politeness (imputed for control) among all units. All regressions control for demographic factors: gender, age, education, race, and geographic region. The imputation is computed using the linear projection learned among treatment units. Stars indicate statistical significance at the 5% level.

6 Raw Text Methods

In this section, we adapt two text-as-treatment models that perform dimension reduction on the con-

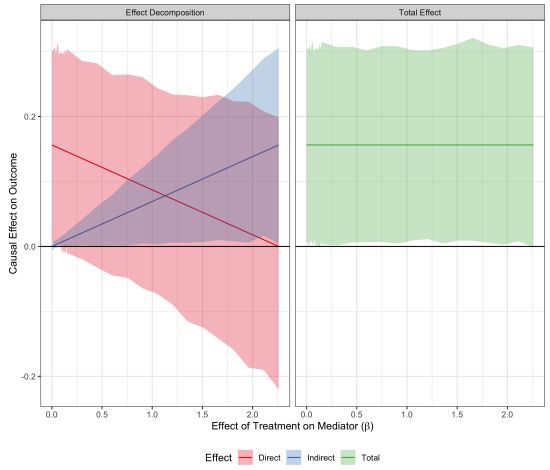

Figure 3: Decomposition of treatment effects by simulated effect on mediator. Figures shows point estimates and 95% confidence intervals for the direct (τ), indirect ($\alpha\beta$), and total ($\tau + \alpha\beta$) effects. Significant indirect effects are first detected at $\beta = 0.09$.

versation data to study if unsupervised methods can identify alternatives or improvements to "politeness." We refer to these methods as "unsupervised" because they take as an input, the raw text of the conversations, rather than using an analyst-specified dimension reduction as in the previous section. *These methods do not perform well and highlight the need for domain knowledge and prior work to select conversation features to explore as potential mediators*. It is possible that the small size of the data plays a role in this poor performance: randomized experiments are expensive and lead to datasets that might be too small for such unsupervised models. However, high-quality experiential data are extremely useful for learning causal effects as treatment-related unconfoundedness assumptions are guaranteed by randomization.

We apply standard LDA (Blei et al., 2003) topic modeling without using any outcome information to learn topics and the supervised Indian Buffet process (sIBP) of Fong and Grimmer (2016), which does use the outcome when learning treatments in the documents. We do not claim these methods are exhaustive, but they are representative of common practice, see e.g. Roberts et al. (2020) and Egami et al. (2018).[5]

For both methods, we combine all messages received and treat those as the documents for analysis. To create a document-word matrix as required by both methods, we pre-process the data by making all words lower case, removing standard stopwords, and dropping words that appear in fewer than 1% documents. The result is 819 documents covering a vocabulary of 3,715 words. For both methods, we need only explore the effect of the output of the text model on depolarization. Because no meaningful relationship is identified, the mediation sensitivity procedure is not needed.

LDA. We estimate the LDA model using 4 through 12 topics without considering the outcome data. Then, we extract posterior estimates of topic prevalence for each set of received messages, and use those in a regression analysis along with the same demographic control variables.

The top panel of Figure 4 shows point and interval estimates for the effect of each topic on depolarization for each model.[6] The only significant results across all models are three topics from the 10 topic model. Under the global null hypothesis that all true topic coefficients are 0, we'd expect about 2 coefficients to be significant due to random chance. The top 10 words for these topics are shown in Table 3. While these topics might suggest mediation is occurring, Topic 7 really only identifies immigration words, one of the two assigned discussion issues, and the other two do not have any clear interpretation. Discussion of immigration is really a measure of compliance, adhering to the researcher-assigned topic, rather than mediation.

Topic 7 ($K = 10$): immigration, people, country, illegal, immigrants, border, legal, wall, agree, law
Topic 8 ($K = 10$): people, agree, feel, good, make, issue, country, issues, change, things
Topic 9 ($K = 10$): people, pay, country, money, work, good, $, jobs, things, president

Table 3: Top Words for Significant Topics. The 10 highest probability words are shown for the three topics with $p < 0.05$. All are from the LDA, 10 topic model.

sIBP. Fong and Grimmer (2016) note that interpreting marginal effects of topic prevalence in the LDA model is difficult. LDA topic prevalence in a document is a point on the simplex, so it must

[5]The LDA model is estimated with default parameter settings using the `topicmodels` package (Grün and Hornik, 2011). The sIBP model has hyperparameters optimized over a search grid using provided functionality in the `texteffect` package (Fong, 2019).

[6]Note that because topic prevalence must sum to one, for a model with K topics, only $K - 1$ coefficients are estimable.

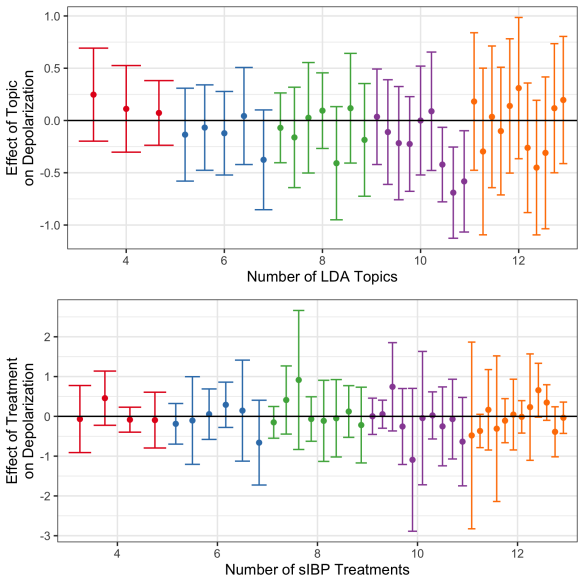

Figure 4: Point and 95% confidence intervals are shown for the effects of topics and treatments on depolarization. No effects are significant at the 5% level. Each regression controls for the following variables: conversation topic, political party, age, gender, race, and education.

sum to one. An increase in one topic must be offset with a decrease in another. Thus, they propose defining a K dimensional binary vector for each document that captures whether each of the topics is present in that document. Per their implementation, the data are split into a 50/50 training and test set. Topics are learned jointly with the document text and the outcome variable in the training set. Then, topics are inferred from the text in the test set, and effects corresponding to each topic are learned using only the test set. We augment their procedure by also including control variables in the test-set estimation of treatment effects.

The bottom panel of Figure 4 shows the results. None of the effects are statistically significant. In part, this is because the procedure requires splitting the data, with effects being estimated using only half of the data that the other models are learned on. Manual review of the top words from the topics shows some have semantic coherence, successfully splitting gun control and immigration issues, but their lack of meaningful correlation with the outcome limits their usefulness in understanding how the conversations cause depolarization.

7 Conclusion

This paper contributes to the methodological literature on text in causal inference by considering text as a mediator, i.e. considering text simultaneously as both a treatment and outcome, and the substantive literature on political polarization and exposure to members of opposing parties.

We develop a sensitivity procedure for mediation analysis when the mediator is only observed for one arm of the experiment. Our procedure allows researchers to assess the strength of correlations between the outcome and mediators by determining how much treatment would have to affect the mediator to observe a significant indirect effect. When applied to text data, domain knowledge is especially important to both guide selection of potential mediators and assess practical significance of the required treatment effect on the mediator.

Beyond the experiment analyzed here, this procedure is useful in marketing applications when one wants to assess the alignment of digital advertisements with surrounding web-page context in reference to control units shown no ads (Zanjani et al., 2011). Observational studies sometimes collect data on treated and control units from different sources with different sets of covariates, as in observational analysis of the Lallonde data (Dehejia and Wahba, 1999). This procedure can also be used in a power analysis to guide sample size selection for future studies with similar treatments that will collect mediator data for all participants.

References

Alex K Anderson, Grace Damio, Sara Young, Donna J Chapman, and Rafael Pérez-Escamilla. 2005. A randomized trial assessing the efficacy of peer counseling on exclusive breastfeeding in a predominantly latina low-income community. *Archives of Pediatrics & Adolescent Medicine*, 159(9):836–841.

Chris Bail. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing*. Princeton University Press.

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, 115(37):9216–9221.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239):1130–1132.

Pablo Barberá. 2015. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, 23(1):76–91.

Pablo Barberá, Andreu Casas, Jonathan Nagler, Patrick J. Egan, Richard Bonneau, John T. Jost, and Joshua A. Tucker. 2019. Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review*, 113(4):883–901.

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

Rogie Royce Carandang, Akira Shibanuma, Junko Kiriya, Karen Rose Vardeleon, Edward Asis, Hiroshi Murayama, and Masamine Jimba. 2020. Effectiveness of peer counseling, social engagement, and combination interventions in improving depressive symptoms of community-dwelling filipino senior citizens. *PloS one*, 15(4):e0230770.

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 57–60, 1st virtual meeting. Association for Computational Linguistics.

Aidan Combs, Graham Tierney, Brian Guay, Friedolin Merhout, Christopher Bail, D. Sunshine Hillygus, and Alexander Volfovsky. 2021. Anonymous cross-party conversations can decrease political polarization: A field experiment on a mobile chat platform. *Unpublished Manuscript*.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 250–259, Sofia, Bulgaria. Association for Computational Linguistics.

Rajeev H Dehejia and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

Naoki Egami, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2018. How to make causal inferences using texts. *arXiv*, pages 1–47.

Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Incorporating politeness across languages in customer care responses: Towards building a multi-lingual empathetic dialogue agent. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4172–4182, Marseille, France. European Language Resources Association.

Christian Fong. 2019. *texteffect: Discovering Latent Treatments in Text Corpora and Estimating Their Causal Effects*. R package version 0.3.

Christian Fong 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)*, pages 1600–1609, Berlin, Germany. Association for Computational Linguistics.

Ragnar Frisch and Frederick V. Waugh. 1933. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. 2019a. Text as data. *Journal of Economic Literature*, 57(3):535–74.

Matthew Gentzkow, Jesse M. Shapiro, and Matt Taddy. 2019b. Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Justin Grimmer 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.

Bettina Grün and Kurt Hornik. 2011. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30.

Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A General Approach to Causal Mediation Analysis. *Psychological Methods*, 15(4):309–334.

Kokil Jaidka, Alvin Zhou, and Yphtach Lelkes. 2019. Brevity is the soul of twitter: The constraint affordance and political discussion. *Journal of Communication*, 69(4):345–372.

Dongyeop Kang and Eduard H. Hovy. 2019. xslue: A benchmark and analysis platform for cross-style language understanding and evaluation. *CoRR*, abs/1911.03663.

Katherine Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344, Online. Association for Computational Linguistics.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A.

Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Michael C. Lovell. 1963. Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304):993–1010.

Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabas Poczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W Black, and Shrimai Prabhumoye. 2020. Politeness transfer: A tag and generate approach. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1869–1881, Online. Association for Computational Linguistics.

Carolyn S Malchodi, Cheryl Oncken, Ellen A Dornelas, Laura Caramanica, Elizabeth Gregonis, and Stephen L Curry. 2003. The effects of peer counseling on smoking cessation and reduction. *Obstetrics & Gynecology*, 101(3):504–510.

Diana C. Mutz and Byron Reeves. 2005. The new videomalaise: Effects of televised incivility on political trust. *American Political Science Review*, 99(1):1–15.

Trang Quynh Nguyen, Elizabeth B. Sarker, Ian Schmid, Noah Greifer, Elizabeth L. Ogburn, Ina M. Koning, and Elizabeth A. Stuart. 2021. Clarifying causal mediation analysis: From simple to more robust strategies for estimation of marginal natural (in)direct effects.

Trang Quynh Nguyen, Ian Schmid, Elizabeth L. Ogburn, and Elizabeth A. Stuart. 2020. Clarifying causal mediation analysis for the applied researcher: Effect identification via three assumptions and five potential outcomes.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Zizi Papacharissi. 2004. Democracy online: Civility, politeness, and the democratic potential of online political discussion groups. *New Media and Society*, 6(2):259–283.

Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 411–420.

Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459.

Margaret E Roberts, Brandon M Stewart, and Richard A Nielsen. 2020. Adjusting for confounding with text matching. *American Journal of Political Science*, 64(4):887–903.

Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Amrita Saha, Ghulam Ahmed Ansari, Abhishek Laddha, Karthik Sankaranarayanan, and Soumen Chakrabarti. 2019. Complex program induction for querying knowledge bases in the absence of gold programs. *Transactions of the Association for Computational Linguistics*, 7:185–200.

Matthew J Salganik. 2019. *Bit by bit: Social research in the digital age*. Princeton University Press.

Gregory Spell, Brian Guay, Sunshine Hillygus, and Lawrence Carin. 2020. An Embedding Model for Estimating Legislative Preferences from the Frequency and Sentiment of Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 627–641, Online. Association for Computational Linguistics.

Yannis Theocharis, Pablo Barberá, Zoltán Fazekas, Sebastian Adrian Popa, and Olivier Parnet. 2016. A bad workman blames his tweets: the consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of communication*, 66(6):1007–1031.

Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. mediation: R package for causal mediation analysis. *Journal of Statistical Software*, 59(5):1–38.

Tyler VanderWeele. 2015. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

Tyler J. VanderWeele. 2016. Mediation Analysis: A Practitioner's Guide. *Annual Review of Public Health*, 37:17–32.

Tyler J VanderWeele and Stijn Vansteelandt. 2009. Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface*, 2(4):457–468.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 919–928. PMLR.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Rob Voigt, Nicholas P. Camp, Vinodkumar Prabhakaran, William L. Hamilton, Rebecca C. Hetey, Camilla M. Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L. Eberhardt. 2017. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4586–4598, Brussels, Belgium. Association for Computational Linguistics.

Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness Package: Detecting Politeness in Natural Language. *The R Journal*, 10(2):489–502.

Shabnam H. A. Zanjani, William D. Diamond, and Kwong Chan. 2011. Does ad-context congruity help surfers and information seekers remember ads in cluttered e-magazines? *Journal of Advertising*, 40(4):67–84.

A Proof of Linear Projection Imputation Properties

Let $\mathbf{Y}_t$, $\mathbf{Y}_c$ be the outcome among treatment and control respectively. Let $\mathbf{X}_t$ and $\mathbf{X}_c$ be $n_t \times p$ and $n_c \times p$ matrices of covariates for treatment and control, including an intercept term. Let $\mathbf{M}_t$ be observed mediator values for treatment. $\mathbf{M}_c$ is not observed. Let $\mathbf{T}$ be the vector of binary treatment indicators. The above variables without a subscript refer to column-stacked (treatment on top of control) versions, e.g. $\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_t \\ \mathbf{Y}_c\end{pmatrix}$.

If $\mathbf{M}_c$ is imputed using the linear regression learned among treatment, $\mathbf{M}_c = \mathbf{X}_c(\mathbf{X}_t^T\mathbf{X}_t)^{-1}\mathbf{X}_t^T\mathbf{M}_t = \mathbf{X}_c\widehat{\beta}_M$, then the OLS regression of $\mathbf{Y}$ on $\mathbf{T}$, $\mathbf{X}$ and $\mathbf{M}$, will (a) estimate coefficient $\widehat{\alpha}$ on the mediator that is exactly equal to the coefficient when regressing $\mathbf{Y}_t$ on $\mathbf{M}_t$ and $\mathbf{X}_t$ and (b) estimate coefficient $\widehat{\tau}$ on the treatment indicator that is exactly equal to the coefficient when regressing $\mathbf{Y}$ on $\mathbf{X}$ without $\mathbf{M}$. And (c), the OLS regression of $\mathbf{M}$ on $\mathbf{X}$ and $\mathbf{T}$ will estimate a coefficient on $\mathbf{T}$ of 0.

Proof. This proof makes extensive use of the Frisch–Waugh–Lovell theorem (FWL) (Frisch and Waugh, 1933; Lovell, 1963). The theorem states that the multivariate OLS regression coefficient for any predictor is exactly equal to the coefficient from a univariate regression of the residualised outcome on the residualised predictor, where the residuals are computed from regressions of the outcome and predictor on all other predictors in the multivariate regression. Specifically, the OLS coefficient β_1 estimated from regression equation $Y = X_1\beta_1 + X_2\beta_2 + e$ is exactly equal to the OLS coefficient estimated from $A_{X_2}Y = A_{X_2}X_1 + A_{X_2}e$ where A_{X_2} is the annihilator matrix for X_2, $A_{X_2} = I - P_{X_2}$, $P_{X_2} = X_2(X_2^TX_2)^{-1}X_2^T$. P_{X_2} is the projection matrix onto the column space of X_2, A_{X_2} is the projection matrix onto the complement of the column space of X_2.

$\mathbf{Y}$ can be expressed as:

$$\mathbf{Y} = \mathbf{X}\widehat{\beta}_Y + \mathbf{T}\widehat{\tau} + \mathbf{M}\widehat{\alpha} + \mathbf{e},$$

where $\mathbf{e}$ is linearly independent from all other predictors, e.g. $\mathbf{e}^T\mathbf{M} = \mathbf{0}$. OLS on the full sample will estimate coefficients of $\widehat{\beta}_Y$, $\widehat{\tau}$, and $\widehat{\alpha}$ with residuals $\mathbf{e}$.

$\mathbf{M}_t$ can be expressed as:

$$\mathbf{M}_t = \mathbf{X}_t\widehat{\beta}_M + \mathbf{r}_t$$

where $\mathbf{r}_t^T\mathbf{X}_t = 0$. Thus, by the assumed construction of $\mathbf{M}_c = \mathbf{X}_c\widehat{\beta}_M$, $\mathbf{M}$ can be expressed as:

$$\mathbf{M} = \mathbf{X}\widehat{\beta}_M + \begin{pmatrix}\mathbf{r}_t \\ \mathbf{0}\end{pmatrix}$$

Proof of (a). Now, consider the regression among just treated units. By FWL, we can consider just the regression of $\mathbf{A}_{X_t}\mathbf{Y}_t$ on $\mathbf{A}_{X_t}\mathbf{M}_t$. Call the estimated coefficient $\widetilde{\alpha}$. We show that $\widetilde{\alpha} = \widehat{\alpha}$. Note that among treatment, $T_i = 1$, so $\mathbf{T}_t$ is in the column space of $\mathbf{X}_t$ because it contains an intercept term. Thus, $\mathbf{A}_{X_t}T_t = 0$.

$$\begin{aligned}
\widetilde{\alpha} &= \left(\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{A}_{X_t}\mathbf{M}_t\right)^{-1}\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{A}_{X_t}\mathbf{Y}_t \\
&= \left(\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{M}_t\right)^{-1}\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{Y}_t \\
&= \left(\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{M}_t\right)^{-1}(\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{M}_t\widehat{\alpha} + \\
&\quad \mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{e}_t) \\
&= \left(\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{M}_t\right)^{-1}\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{M}_t\widehat{\alpha} \\
&= \widehat{\alpha}
\end{aligned}$$

The second to last line is because $\mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{e}_t = 0$. By construction, $0 = \mathbf{M}^T\mathbf{e} = \left(\mathbf{X}\widehat{\beta}_M + \begin{pmatrix}\mathbf{r}_t \\ \mathbf{0}\end{pmatrix}\right)^T\mathbf{e} = \mathbf{r}_t^T\mathbf{e}_t = \mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{e}_t$. Thus, $0 = \mathbf{M}_t^T\mathbf{A}_{X_t}\mathbf{e}_t$.

Proof of (b). Here we consider the regression of $\mathbf{A}_X\mathbf{Y}$ on $\mathbf{A}_X\mathbf{T}$, which estimates the multivariate regression coefficient from regressing $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{T}$ without $\mathbf{M}$. We show that $\widetilde{\tau}$ estimated from this regression is exactly equal to $\widehat{\tau}$.

$$
\begin{aligned}
\widetilde{\tau} &= (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}\mathbf{T}^T\mathbf{A}_X\mathbf{Y} \\
&= (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}\mathbf{T}^T(\mathbf{A}_X\mathbf{M}\widehat{\alpha} + \mathbf{A}_X\mathbf{T}\widehat{\tau} + \mathbf{e}) \\
&= \widehat{\tau} + (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}\mathbf{T}^T\mathbf{A}_X\mathbf{M}\widehat{\alpha}, \\
&\quad \text{because } \mathbf{T}^T\mathbf{e} = 0 \\
&= \widehat{\tau}
\end{aligned}
$$

Again, the above uses $\mathbf{T}^T\mathbf{A}_X\mathbf{M} = 0$. $\mathbf{T}^T\mathbf{A}_X\mathbf{M} = \mathbf{T}^T\begin{pmatrix}\mathbf{r}_t \\ \mathbf{0}\end{pmatrix} = \mathbf{1}^T\mathbf{r}_t = 0$, with the final equality because $\mathbf{r}_t$ are the residuals from the regression of $\mathbf{M}_t$ on $\mathbf{X}_t$, so must sum to zero.

Proof of (c). Consider the regression of $\mathbf{M}$ on $\mathbf{T}$ and $\mathbf{X}$. Using FWL, the coefficient on $\mathbf{T}$ will be a the regression of $\mathbf{A}_X\mathbf{M}$ on $\mathbf{A}_X T$. From the above decomposition of $\mathbf{M}$, $\mathbf{A}_X\mathbf{M} = \begin{pmatrix}\mathbf{r}_t \\ \mathbf{0}\end{pmatrix}$.

$$
\begin{aligned}
\widehat{\tau}_M &= (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}\mathbf{T}^T\mathbf{A}_X\mathbf{M} \\
&= (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}\mathbf{T}^T\begin{pmatrix}\mathbf{r}_t \\ \mathbf{0}\end{pmatrix} \\
&= (\mathbf{T}^T\mathbf{A}_X\mathbf{T})^{-1}(\mathbf{1}^T\mathbf{r}_t) \\
&= 0
\end{aligned}
$$

$\mathbf{1}^T\mathbf{r}_t = 0$ because $\mathbf{X}_t$ contains an intercept term and $\mathbf{r}_t^T\mathbf{X}_t = \mathbf{0}$.

B Ethical Considerations

This study makes extensive use of data collected from the experiment described in detail in Combs et al. (2021), Bail (2021) and Section 2. This study was conducted with approval from an Institutional Review Board. The target population is registered voters in the United States that have a smartphone and lean towards either the Democratic or Republican political party. Participants were given informed consent dialogues regarding compensation, expected time to complete the tasks, and that their data would be used for research purposes before each survey and when invited to download the mobile application. After completing the final post-survey, participants were debriefed on the study purpose. Participants were told that at any time, they may request the study authors destroy any data collected from them.

Participants were compensated $12.50 for the pre- and post-surveys and $17.50 for using the conversation application. These values were set to be approximately 2-3 times the minimum wage in the United States given estimates of survey and conversation completion time. While participants were told they would only be compensated if they completed a conversation with a partner, all app-users were given full compensation so that no one would be denied compensation due to a partner who failed to respond promptly.

Participants shared personally identifying information, such as names and social media profiles, as well as sensitive personal information regarding their experiences with gun violence and immigration. For these reasons, the conversation data and identifying demographic data cannot be shared per the IRB protocol. We do share the code for all analyses and residualized versions of the outcome, mediator, and treatment variables that permit exact replication of the regression results in Section 5 without the identifying demographics.