Text as Causal Mediators: Research Design for Causal Estimates of Differential Treatment of Social Groups via Language Aspects

Katherine A. Keith, Douglas Rice, and Brendan O'Connor
University of Massachusetts Amherst
kkeith@cs.umass.edu,drrice@legal.umass.edu,brenocon@cs.umass.edu

Abstract

Using observed language to understand interpersonal interactions is important in high-stakes decision making. We propose a causal research design for observational (non-experimental) data to estimate the natural direct and indirect effects of social group signals (e.g. race or gender) on speakers' responses with separate aspects of language as causal mediators. We illustrate the promises and challenges of this framework via a theoretical case study of the effect of an advocate's gender on interruptions from justices during U.S. Supreme Court oral arguments. We also discuss challenges conceptualizing and operationalizing causal variables such as gender and language that comprise of many components, and we articulate technical open challenges such as temporal dependence between language mediators in conversational settings.

1 Introduction

Interactions between individuals are key components of social structure (Hinde, 1976). While we rarely have access to individuals' internal thoughts during these interactions, we often can observe the language they use. Using observed language to better understand interpersonal interactions is important in high-stakes decision making—for instance, judges' decisions within the United States legal system (Danescu-Niculescu-Mizil et al., 2012) or police interaction with citizens during traffic stops (Voigt et al., 2017). In these settings, analysts may be interested in understanding the behavior of decision makers as individuals or at the subgroup or aggregate level.

Important decision makers sometimes treat some social groups (e.g. women, racial minorities, or ideological communities) differently than others (Gleason, 2020). Yet, quantitative analyses of this problem often do not account for all possible mechanisms that could induce this differential treatment. For instance, one might ask, *During U.S. Supreme Court oral arguments, is a justice interrupting female advocates more because of their gender, because of the content of the advocates' legal arguments, or because of the advocates' language delivery* (Fig. 1B)? Accounting for these language mechanisms could help separate and estimate the remaining "gender bias" of justices.

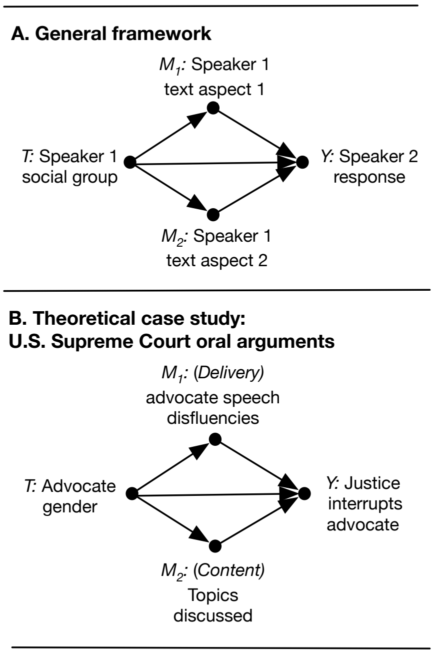
A. General framework
M_1: Speaker 1 text aspect 1
T: Speaker 1 social group
Y: Speaker 2 response
M_2: Speaker 1 text aspect 2
B. Theoretical case study: U.S. Supreme Court oral arguments
M_1: (Delivery) advocate speech disfluencies
T: Advocate gender
Y: Justice interrupts advocate
M_2: (Content) Topics discussed

Figure 1: Causal diagrams in which nodes are random variables and arrows denote causal dependence for **A.** proposed general framework for *differential treatment of social groups via language aspects* and **B.** instantiation of the framework for a theoretical case study of U.S. Supreme Court oral arguments. In both diagrams, T is the treatment variable, Y is the outcome variable, and M are mediator variables. This is a simplified schema; see Fig. 2 for an expanded diagram.

We reformulate the previous question as a general *counterfactual* query (Pearl, 2009; Morgan and Winship, 2015) about two speakers: *How would Speaker 2 respond if the signal they received of Speaker 1's social group flipped from A to B but Speaker 1 still used language typical of social*

Proceedings of CI+NLP: First Workshop on Causal Inference and NLP, pages 21–32
November 10, 2021. ©2021 Association for Computational Linguistics

group A? Here, our question is about the direct causal effect of *treatment*—Speaker 1's signaled social group—on *outcome*—Speaker 2's response—that is not through the causal pathway of the *mediator*—an aspect of language (Fig. 1A).[1]

The fundamental problem with this and any counterfactual question is that we cannot go back in time and observe an individual counterfactual while holding all other conditions the same (Holland, 1986). Furthermore, in many high-stakes, real-world settings (e.g. the U.S. Supreme Court), we cannot run experiments to randomly assign treatment and approximate these counterfactuals. Instead, in these settings, causal estimation must rely on *observational* (non-experimental) data.

In this work, we focus on this observational setting and build from causal mediation methods (Pearl, 2001; Imai et al., 2010; VanderWeele, 2016) to specify a research design of causal estimates of *differential treatment of social groups via language aspects*. Other work has used causal mediation analysis to better understand components of natural language processing (NLP) models (Vig et al., 2020; Finlayson et al., 2021). However, this work is more closely aligned with studies that focus on causal estimation in which text is one or more causal variables (e.g., Veitch et al., 2020; Roberts et al., 2020; Keith et al., 2020; Zhang et al., 2020; Pryzant et al., 2021).

Our focus is on the research design, and we therefore intentionally do not present empirical results. Instead, we discuss the potential promises and challenges of this causal research design with both general examples and concrete examples from a theoretical case study of U.S. Supreme Court arguments. This aligns with Rubin (2008) who argues "design trumps analysis" in observational studies and emphasizes the importance of conceptualizing a study before any outcome data is analyzed.

Overall, we make the following contributions:

- We propose a new causal research design to estimate the natural indirect and direct effects of social group signal on speakers' responses with separate aspects of language as causal mediators (§3).

- We illustrate the promises and challenges of this framework via a theoretical case study of the effect of an advocate's gender on interruptions by justices during U.S. Supreme Court oral arguments. (§2).

- We discuss challenges researchers might face conceptualizing and operationalizing the causal variables in this research design (§4).

- We directly address critiques of using social groups (e.g. race or gender) as treatment and construct gender and language as *constitutive* variables, building from Sen and Wasow (2016); Hu and Kohler-Hausmann (2020) (§4.2 and §4.4).

- We articulate potential open challenges in this research design including temporal dependence between mediators in conversations, causal dependence between multiple language mediators, and dependence between social group perception and language perception (§5).

2 Theoretical Case Study: Gender Bias in U.S. Supreme Court Interruptions

To motivate our causal research design and illustrate challenges that arise with it, we focus on a specific theoretical case study—the effect of advocate gender on justice interruptions via advocates' language during United States Supreme Court oral arguments (Fig. 1B). The substantive motivation for this theoretical case study is built from previous work examining the role of interruption and gender on the Court. Patton and Smith (2017) found female lawyers are interrupted earlier in oral arguments, allowed to speak for less time, and subjected to longer speeches by justices; Jacobi and Schweers (2017) found female justices are interrupted at disproportionate rates by their male colleagues; and Gleason (2020) found justices are more likely to vote for the female advocate's side when the female advocate uses emotional language.

Counterfactual questions. We present a novel causal approach to understanding gender bias in Supreme Court oral arguments that corresponds to the following counterfactual questions:

1. *(NDE)*: How would a justice's interruptions of an advocate change if the signal of the advocate's gender the justice received flipped from male to female, but the advocate still used language typical of a male advocate?

2. *(NIE)*: How would a justice's interruptions of an advocate change if a male advocate used

[1] See §4.2 for a discussion on when and how social groups (e.g. gender or race) can be used as causal treatments.

(A) Case: Kennedy v. Plan Administrator for DuPont Sav. and Investment Plan (2008-07-636)

Mark Irving Levy: [...] The QDRO provision is an objective checklist that is easy for – for plan administrators to follow.

Antonin Scalia: What if they had agreed to the waiver apart from [...] We'd be in the same suit that you're - - that you say we have to avoid, wouldn't we?

Mark Irving Levy: I don't think so. I mean I think that would be an alienation.

Antonin Scalia: Well, if it's an alienation, but his point is that a waiver is not an alienation.

(B) Case: Lozano v. Montoya Alvarez (2013-12-820)

Ann O'Connell Adams: Well - -

Antonin Scalia: I mean, it seems to me it just makes that article impossible to apply consistently country to country.

Ann O'Connell Adams: - - No, I don't think so. And - - and, the other signatories have - - have almost all, I mean I think the Hong Kong court does say that it doesn't have discretion, but it said in that case nevertheless it would, even if it had discretion, it wouldn't order the children returned. But the other courts of signatory countries that have interpreted Article 12 have all found a discretion, whether it be in Article 12 or in Article 8. And if I - -

Antonin Scalia: Have they exercised it? Have they exercised it, that discretion which they say is there?

Table 1: Selected utterances from the oral arguments of two U.S. Supreme Court cases, A (Oyez, a) and B (Oyez, b), with advocates Mark Irving Levy (male) and Ann O'Connell Adams (female) respectively. Justice Antonin Scalia responds to both advocates. Hedging language is highlighted in blue. Speech disfluencies are highlighted in red. Gray-colored utterances directly proceed the target utterances (non-gray colored) in the oral arguments.

> language typical of a female advocate but the signal of the advocate's gender the justice received remained male?

which we show correspond to the *natural direct effect* (NDE) and *natural indirect effect* (NIE) respectively in §3. In §4, we walk through the theoretical conceptualization and empirical operationalization of advocate gender (treatment), interruption (outcome), and advocate language (mediators).

Intuitive example. We describe intuitive challenges of our causal research design by contrasting Examples A and B in Table 1. Levy—a male advocate—is not interrupted by Justice Antonin Scalia, but Adams—a female advocate—is interrupted (Oyez, a,b). *Why was the female advocate interrupted? Was it because of her gender or because of* what *she said or* how *she said it*? We hypothesize one causal pathway between gender and interruption is through the mediating variable hedging—expressions of deference or politeness.[2] Suppose we operationalize hedging as certain key phrases, e.g. "I don't think so" and "I mean I think." An initial causal design might assign a binary hedging indicator to utterances and then compare average interruption outcomes for male and female advocates conditional on the hedging indicator.

However, advocate utterances matched on this hedging indicator could have a number of latent mediators and confounders. In Table 1, Adams has speech disfluencies ("and - - and" and "have - - have" shown in red) which might cause Scalia to get frustrated and interrupt. The cases are from different areas of the law,[3] and Scalia may interrupt more during cases that are in areas he has more personal interest. The advocate utterance in Ex. B is longer (more tokens) and longer utterances may be more likely to be interrupted. In Ex. B, Scalia interrupts Adams just prior to the target utterance which possibly indicates a more "heated" portion of the oral arguments during which interruptions occur more on average. With these confounding and additional mediator challenges, a simple causal matching approach (e.g. Stuart (2010); Roberts et al. (2020)) is unlikely to work and we advocate for the causal estimation strategy presented in §3.4. We move from this case study to a formalization of our causal research design in §3.

3 Causal Mediation Formalization, Identification, and Estimation

Many causal questions involve *mediators*—variables on a causal path between treatment and outcome. For example, what is the effect of gender[4] (treatment) on salary (outcome) with and without considering merit (a mediator)? If one intervenes on treatment, then one would activate both the "direct path" from gender to salary *and* the "indirect path" from gender through merit to salary. Thus, a major focus of causal mediation is specifying conditions under which one can separate estimates of the *direct effect* from the *indirect effect*—the former being the effect of treatment on outcome *not*

[2] Previous work has shown hedging is used more often by women (Lakoff, 1973; Poos and Simpson, 2002), and we hypothesize judges might respond more positively to more authoritative language (less hedging) from advocates.

[3] The Supreme Court Database codes Ex. A as "economic activity" and Ex. B as "civil rights" (Spaeth et al., 2021).

[4] See §4 for discussion of operationalizing difficult causal variables such as gender.

through mediators and the later the effect through mediators.

We use this causal mediation approach to formally define our framework. For each unit of analysis (see §4.1), i, let T_i represent the treatment variable—the social group, e.g. gender of an advocate—and Y_i represent the outcome variable—the second speaker's response, e.g. a judge's interruption or non-interruption of an advocate. For each defined mediator j, let M_i^j represent the mediating variable—an aspect of language, e.g. an advocate's speech disfluencies or the topics of an utterance. Let X_i represent any other confounders between any combination of the other variables.

We use the potential outcomes framework (Rubin, 1974) to define the natual direct and indirect effects.[5] Let $M_i(t)$ represent the (counterfactual) potential value the mediator would take if $T_i = t$. Then $Y_i(t, M_i(t'))$ is a doubly-nested counterfactual that represents the potential outcome that results from both $T_i = t$ and potential value of the mediator variable with $T_i = t'$. With this formal notation, we define the individual *natural direct effect (NDE)* and *natural indirect effect (NIE)*:[6]

$$\text{NDE}_i = Y_i(1, M_i(0)) - Y_i(0, M_i(0)) \quad (1)$$

$$\text{NIE}_i = Y_i(0, M_i(1)) - Y_i(0, M_i(0)) \quad (2)$$

These correspond to the two counterfactual questions from §2 if $T_i = 0$ and $T_i = 1$ represent the gender signal of the advocate being male and female respectively.

3.1 Estimands

We second the advice of Lundberg et al. (2021) and recommend researchers explicitly state their estimand of interest. As we briefly touch on in the introduction, some studies may be interested in the estimand as the *individual*-level natural direct and indirect effects (Equations 1 and 2). For example, a legal scholar may be interested in an individual U.S. Supreme Court case and estimate the individual NIE and NDE for this single case in order to evaluate how "fair" the case was with respect to the gender of an advocate. Machine learning approaches to estimating individual-level causal effects are promising (Shalit et al., 2017) but may not be applicable to all datasets. In contrast, more feasible—and potentially equally substantively valid—estimands may be at the *subgroup* level (e.g. effects of all cases about civil rights or all cases for a particular justice) or aggregate level. Here, the estimands are some kind of aggregation over Equations 1 and 2. Thus, in Section 3.4, we provide estimators for general population-level (*not* individual-level) estimands.

3.2 Interpretation of the NDE as "bias"

Many applications of causal mediation aim to quantify "implicit bias" or "discrimination" via the natural direct effect. However, if all relevant mediators are not accounted for, one cannot interpret the estimand of the natural direct effect as the actual direct causal effect (Van der Laan and Rose, 2011, p.135). Nevertheless, if we separate the total effect into the proportion that is the NDE and the NIE with the mediators to which we have access, our analysis moves *closer* to estimating the true direct effect between treatment and outcome. Thus, in this work we emphasize the value of having interpretable mediators (i.e. language aspects) for which the NIE is a meaningful quantity to analyze in itself.

3.3 Identification

Like any causal inference problem, we first examine the *identification assumptions* necessary to claim an estimate as causal. The key assumption particular to causal mediation is that of *sequential ignorability* (Imai et al., 2010):

1. Potential outcomes and mediators are independent of treatment given confounders

$$\{Y_i(t', m), M_i(t)\} \perp\!\!\!\perp T_i \mid X_i = x \quad (3)$$

2. Potential outcomes are independent of mediators given treatment and confounders

$$Y_i(t', m) \perp\!\!\!\perp M_i(t) \mid \{T_i = t, X_i = x\} \quad (4)$$

for $t, t' \in \{0, 1\}$ and all values of x and m.

Mediator Independence Assumption:[7] For our particular framework, we make an additional assumption that for each language aspect we study,

[5] Pearl (2001) notes *do*-notation cannot represent causal mediation questions, since they concern counterfactual paths, not interventions of variables.

[6] Pearl et al. (2016) defines the NDE and NIE in terms of the non-treatment condition, $T = 0$. Others (e.g. Imai et al. (2010) and Van der Laan and Rose (2011)) give alternate definitions of these quantities in terms of $T = 1$. We follow Pearl et al.'s definitions in the remainder of this work.

[7] This is similar to the assumptions Pryzant et al. (2021) make for linguistic properties of text as treatment.

the mediators are independent conditional on the treatment and confounders

$$\forall j, j' : \; M_i^j(t) \perp\!\!\!\perp M_i^{j'}(t) \mid \{T_i = t, X_i = x\} \quad (5)$$

With this assumption, we can estimate the NIE and NDE of each mediator successively, ignoring the existence of other mediators. (Imai et al., 2010; Tingley et al., 2014). We discuss the validity of this assumption in §5.

These assumptions correspond to the causal relationships of a graph similar to Fig. 1, with the addition of confounder X as a parent of all T, M^j, and Y (to be more precise may require a richer formalism; e.g. Richardson and Robins (2013)).

3.4 Estimation

Given the satisfaction of sequential ignorability, mediator independence, and other standard causal identification assumptions,[8] we propose using the following estimators of population-level natural direct and indirect effects for each mediator j (Imai et al., 2010; Pearl et al., 2016):

$$\text{SA-NDE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\Big(\hat{f}^j(Y|M_i^j = m, T_i = 1, X_i = x) - \hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x)\Big)\hat{g}^j(m|T_i = 0, X_i = x) \quad (6)$$

$$\text{SA-NIE}^j = \frac{1}{N}\sum_{i=1}^{N}\sum_{x\in\mathcal{X}}\sum_{m\in\mathcal{M}^j}\hat{f}^j(Y|M_i^j = m, T_i = 0, X_i = x) \quad (7)$$
$$\Big(\hat{g}^j(m|T_i = 1, X_i = x) - \hat{g}^j(m|T_i = 0, X_i = x)\Big)$$

Each is a **S**ample **A**verage estimate from N data points, relying on models trained to predict mediator and outcome given confounders and treatment: $\hat{g}^j$ infers mediator j's probability distribution, while $\hat{f}^j$ infers the expected outcome conditional on mediator j. The estimators marginalize over confounders and mediators from their respective domains ($x \in \mathcal{X}$, $m \in \mathcal{M}^j$), which for our discrete variables is feasible with explicit sums (see Imai et al. for the continuous case).

Model fitting. When fitting models $\hat{f}$ and $\hat{g}$, we recommend using a cross-sample or cross-validation approach in which one part of the sample is used for training/estimation (S_{train}) and the other is used for testing/inference (S_{test}) in order to avoid overfitting (Chernozhukov et al., 2017; Egami et al., 2018). With text, one must also fit a model for the mediators conditional on text, $h(m|\text{text})$ using S_{train}. In some cases, such as measuring advocate speech disfluencies, h may be a simple deterministic function. However, when using NLP and other probabilistic models (e.g topic models or embeddings), h could be a difficult function to fit and have a certain amount of measurement error. A major open question is whether to jointly fit h and g at training time as advocated by previous work (Veitch et al., 2020; Roberts et al., 2020) or if h and g should be treated as separate modules. At inference time, we do not use the inference text from S_{test} since Eqns. 6 and 7 only rely on the mediators through estimates from $\hat{g}$.

4 Conceptualization and Operationalization of Causal Variables

For any causal research design—and particularly those in the social sciences—there are often challenges *conceptualizing* the theoretical causal variables of interest. Even after these theoretical concepts are made concrete, there are often multiple ways to *operationalize* these concepts. We discuss conceptual and operational issues for our both our general research design and our theoretical case study. In particular, we recommend researchers formalize variables such as gender and language as *constitutive* variables made of multiple components (Fig. 2) as per Hu and Kohler-Hausmann (2020), or Sen and Wasow (2016)'s "bundle of sticks."

4.1 Unit of analysis

As with most causal research designs, one starts by conceptualizing the *unit of analysis*—the smallest unit about which one wants to make counterfactual inquiries. In our framework, the *unit of analysis* is a certain amount of language (L) between speakers of two categories: the first category of speakers, P_1, are those belonging to a group of interest (e.g. advocates) for which treatment values (e.g. female and male) will be assigned; and the second, P_2, is the set of decision-makers responding to the first speakers (e.g. judges).

Operationalizations. There are several possible operationalizations of L: pairs of single utterances—whenever a person from P_1 speaks and a person from P_2 responds; a thread of several

[8] Overlap, SUTVA etc.; see Morgan and Winship (2015).

utterances between persons from P_1 and P_2 within a conversation; or the entire conversation between persons from P_1 and P_2. In §5, we note that selecting the unit of language could have implications for modeling temporal dependence between mediators.

4.2 Treatment

At the most basic level, *treatment*, T, in our research design is *the social group* of persons in P_1 (Fig. 1). However, inspired by the *causal consistency* arguments from Hernán (2016),[9] we examine several competing versions of treatment for our theoretical case study of U.S. Supreme Court oral arguments and explain the reasons we eventually choose version #5 (in bold):

1. Do judges interrupt at different rates based on an advocate's *gender*?
2. Based on an advocate's *biological sex assigned at birth*?
3. An advocate's *perceived gender*?
4. An advocate's *gender signal*?
5. **An advocate's *gender signal* as defined by (hypothetical) manipulations of the advocate's clothes, hair, name, and voice pitch?**
6. An advocate's *gender signal* by (hypothetical) manipulations of their entire physical appearance, facial features, name, and voice pitch?
7. An advocate's *gender signal* by setting their physical appearance, facial features, name, and voice pitch to specific values (e.g. all facial features set to that of the same 40-year-old, white female and clothes set to a black blazer and pants).

In critique of treatment version #1, most social groups (e.g. gender or race) reflect highly contextual social constructs (Sen and Wasow, 2016; Kohler-Hausmann, 2018; Hanna et al., 2020). For gender in particular, researchers have shown social, institutional, and cultural forces shape gender and gender perceptions (Deaux, 1985; West and Zimmerman, 1987), and thus viewing gender as a binary "treatment" in which individuals can be randomly assigned is methodologically flawed. In critique of version #2, *biological sex assigned at birth* is a characteristic that is not manipulable by researchers and the "at birth" timing of treatment assignment means all other variables about the individual are post-treatment. Thus, researchers have warned against estimating the causal effects of these kinds of "immutable characteristics" (Berk et al., 2005; Holland, 2008).

Greiner and Rubin (2011) propose overcoming the issues in versions #1 and #2 by shifting the unit of analysis to the *perceived gender* of the decision-maker (#3) and defining treatment assignment as the moment the decision-maker first perceives the social group of the other individual. Hu and Kohler-Hausmann (2020) critique this *perceived gender* variable and emphasize that we, as researchers, cannot actually change the internal, psychological state of decision-makers, but rather we can change the *signal* about race or gender those decision-makers receive (#4). However, as Sen and Wasow (2016) discuss, defining treatment as the *gender signal* (#4) is dismissive of the many components that make up a social construct like gender. Instead, Sen and Wasow recommend articulating the specific variables one would potentially manipulate. For *gender* in our case study, this could mean hypothetical manipulations of an advocate's dress, name, and voice pitch (#5).

Shifting from versions #5 to #6 and #7, we define treatment in terms of more specific manipulations. However, we also enter the realm of Hernán's argument that precisely defining the treatment never ends, and some aspects of #6 and #7 are impossible to manipulate in real-world settings such as the U.S. Supreme Court. What does it mean to manipulate an advocate's "entire physical appearance?"[10] When we define treatment very specifically—e.g. using the same 40-year old white woman as the treatment for "female advocate" (#7)—are we estimating a causal effect of gender *in general*? Thus, we back-off from versions #6 and #7, and advocate using #5 as our definition of treatment.

Constitutive causal diagrams. With these con-

[9] *Consistency* is the condition that for observed outcome Y and treatment T, the potential outcome equals the observed outcome, $Y(t) = Y$ for each individual with $T = t$. Hernán (2016) presents eight versions of treatment for the causal question "Does water kill?" to illustrate the deceptiveness of this apparently simple consistency condition. Hernán points out that "declaring a version of treatment sufficiently well-defined is a matter of agreement among experts based on the available substantive knowledge" and is inherently (and frustratingly) subjective.

[10] Would justices have to interact with advocates through a computer-mediated system in which one could customize avatars of the advocates? We note, using computer-mediated avatars to signal social group identity has been used effectively in other causal studies, e.g. Munger (2017).

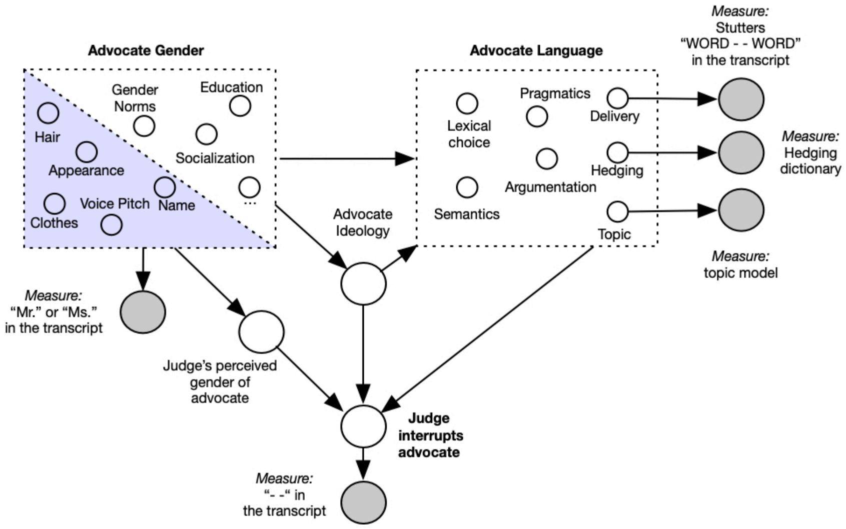

Figure 2: *Constitutive* causal diagram for gendered interruption in U.S. Supreme Court oral arguments. Latent theoretical concepts are unshaded circles and observed operationalizations (measurements) of concepts are shaded circles. We provide alternative operationalizations in the text. The causal variables *gender* and *language* are represented as dashed lines around their constituent parts, building from the arguments of Sen and Wasow (2016); Hu and Kohler-Hausmann (2020). The shaded portion of *gender* consists of the gender variables that one could potentially manipulate in a hypothetical intervention.

siderations, drawing a causal diagram in which a *gender* is represented as a single node seems flawed. Instead, building from Sen and Wasow (2016) and Hu and Kohler-Hausmann (2020), we represent treatment (the social group) as cloud of components (a *constitutive* variable), some of which are latent, some observable, and some manipulable. In Fig. 2, we shade the "outward" components of *gender*—hair, appearance, clothes, voice pitch, and name—that are our hypothetical manipulations and would influence the latent variable of a judge's perceived gender of the advocate. Other "background" components of gender—gender norms, education, and socialization—are the components that could causally influence language.

Case study operationalizations. Even after selecting version #5 as our conceptualization of treatment, there are still multiple operationalizations for our theoretical case study:

Treatment operationalization 1: Previous work operationalizes gender in Supreme Court oral arguments by using norm that the Chief Justice introduces an advocate as "Ms." and "Mr." before their first speaking turn (Patton and Smith, 2017; Gleason, 2020). The advantage of this operationalization is that it is simple, clean, and consistent, and occurs directly before an advocate's first utterance.[11]

Treatment operationalization 2: Alternatively, one could focus on even more specific components of gender for (hypothetical) manipulations. For instance, Chen et al. (2016) and Chen et al. (2019) measure voice pitch when studying gender on the U.S. Supreme Court. While being more cumbersome to measure, this operationalizes gender as a real-valued (instead of binary) variable and thus potentially measures more subtle gender biases.

4.3 Outcome

In our general framework, we define the *outcome*, Y, as *the response of the second speaker* (Fig. 1A), and we intentionally leave this variable vague and domain-specific. However, if making the leap from *differential treatment* to claiming *discrimination* or *bias*, conceptualizing a causal outcome requires normative commitments and a moral theory of what is harmful (Kohler-Hausmann, 2018; Blodgett et al., 2020). In our case study, we conceptualize the outcome variable as a judge interrupting an advocate. This outcome is of substantive interest because, in general, interruptions can indicate and reinforce status in conversation (Mendelberg et al., 2014), and, specifically to the U.S. Supreme Court,

[11] The treatment assignment timing is potentially important for the rest of the causal diagram. If we can define *gender signal* and thus latent *perceived gender* as happening right before an advocate first speaks, and it is not adapted or updated by the judge over the course of the oral arguments, then we can eliminate the causal arrow between variables "language" and "perceived gender."

justice's behavior in oral arguments has been connected to case outcomes.

Outcome operationalization 1: Previous work uses the transcription norm of a double-dash ("- -") at the end of a advocate utterance when a justice interrupts in the next utterance (Patton and Smith, 2017). However, the validity of this operationalization relies on consistent transcription standards.

Outcome operationalization 2: An alternative operationalization could classify interruptions into positive (agreeing with the first speaker's comment), negative (disagreeing, raising an objection, or completely changing the topic), or neutral categories (Stromer-Galley, 2007; Mendelberg et al., 2014). While estimating the effects of only negative interruptions could further refine the causal question—*Do justices negatively interrupt female advocates more?*—this operationalization could also introduce measurement error since it could prove difficult difficult to design an accurate NLP classifier for this task.

4.4 Language Mediators

Our framework explicit focuses on *language as a mediator* in differential treatment of social groups. Yet, language consists of multiple levels of linguistic structure (Bender, 2013; Bender and Lascarides, 2019), so as with social groups (§4.2), it is a variable that is non-modular and we believe it should be represented as constituent parts (Fig. 2).

Mediator Operationalizations: We focus on three potential language aspects for our Supreme Court case study: (A) *hedging*—expressions of deference or politeness—with an operationalization as lexical matches from a single-word hedging dictionary (e.g. Prokofieva and Hirschberg (2014)); (B) *speech disfluencies*—repetitions of syllables, words, or phrases—which we operationalize as the transcript noting a repeated unigram with a double dash, "*word - - word*"; and (C) semantic *topics* operationalized as a topic model (Blei et al., 2003) applied to utterances.

Recommendations. We discuss the choice of these particular language aspects, M^j, for our case study as well as general recommendations for researchers operationalizing language as a mediator.

- Is M^j *interpretable*? Is there a *hypothetical manipulation*[12] of M^j? In contrast to prior work that treats language as a black-box in causal mediation estimates (Veitch et al., 2020), we advocate for using interpretable aspects of language. If language mediators are interpretable, then the NIE is both meaningful (see §3.2) and potentially more fine-grained (we can estimate an NIE for each aspect of language that we are studying instead of a black-box approach that lumps all text into one effect). Furthermore, since identification is essential to claiming an estimate is causal and identification can only be verified qualitatively and through domain expertise, interpretable text mediators will be much easier to evaluate.

- Is there *substantive theory* for causal pathways $T \rightarrow M^j$ and from $M^j \rightarrow Y$? Without such theory, studying certain aspects of language is not meaningful. For example, see §2 for our theoretical reasoning about the causal dependence between gender, hedging, and interruption.

- To what extent does one expect *measurement error* of M^j when using automatic NLP tools? Our operationalizations of hedging lexicons and speech disfluencies are deterministic; however, topic model inferences are probabilistic and sensitive to changes in hyperparameters and preprocessing decisions (Schofield et al., 2017; Denny and Spirling, 2018). These kinds of measurement errors are still open questions although there is recent work that examines measurement error when text is treatment (Wood-Doughty et al., 2018).

- Is M^j *causally independent* from other measured language aspects, $M^{j'}$? If not, our proposed estimator from §3.4 is invalid. Thus, one must scrutinise which aspects of language are separable and thus able to be included in the causal analysis—e.g. we could include content (topics) versus delivery (speech disfluencies) since one could hypothetically modify one without affecting the other. We discuss this assumption further in §5.

4.5 Non-language Mediators

Returning to §3.2, there is often a tendency to interpret the NDE as something like "pure" *gender bias*—What is the effect of gender on interruption when all other possible causal pathways are

[12] To be precise, the *controlled direct effect* is the estimand in which the mediator is manipulated, $do(M)$ (Pearl, 2001). In contrast, the *natural* direct and indirect effects are counterfactuals on paths. However, we still find thinking through potential manipulations is helpful in refining the conceptualization of a language aspect.

stripped away? Conceptualizing and operationalizing language aspects as mediators (§4.4) moves the NDE towards the desired "gender bias." However, there may be other mediator pathways that explain these effects. For example, in our case-study, two additional mediators of interest are advocate ideology (e.g. liberal or conservative) and the level of "eliteness" of the advocate's law firm. A major validity issue is the *causal independence* of these mediators from the language mediators. For instance, ideology could influence certain aspects of language (topic), and "eliteness" of the advocate's law firm could be a proxy for level of training which could influence the advocate's delivery.

5 Challenges and Threats to Validity

We discuss additional challenges and threats to validity for our research design that should be addressed before implementing the design and claiming the estimates from the design are causal.

Temporal dependence of utterances. So far, we have assumed the "units of analysis" of text are independent (§4.1). However, previous utterances in a conversation often influence the target utterances. For our case study, if Judge A interrupted Advocate B often in $t' < t$, interruption at t is more likely (the two speakers are possibly in a "heated" part of the conversation) and Advocate B's speech disfluencies at t are also more likely (the advocate could be mentally fatigued). Potential avenues forward include changing the unit of analysis to the entire conversational thread between the two target speakers or building extensions to the multiple mediator literature, i.e. Imai and Yamamoto (2013); VanderWeele and Vansteelandt (2014); Daniel et al. (2015); VanderWeele (2016).

Dependence between multiple language mediators. Our framework assumes one can computationally separate aspects of language.[13] However, some sociolinguists argue aspects of language such as "style" cannot be separated from "content" because style originates in the content of people's lives and different ways of speaking signal socially meaningful differences in content (Eckert, 2008; Blodgett, 2021). If our mediator independence assumption (Eqn. 5) is violated, then we would have to turn to alternate estimation strategies to deal with this dependence.

Dependence between social group perception and linguistic perception. Separating the direct and indirect causal paths in our framework relies on there being a *decision-maker's latent perception of social group* variable on the direct path between treatment and outcome and that this variable is independent from a *decision-maker's latent perception of language* variable on the indirect path from treatment through mediators to outcome. However, "indexical inversion" considers "how language ideologies associated with social categories produce the perception of linguistic signs" (Inoue, 2006; Rosa and Flores, 2017). Suppose Judge A perceives Advocate B as female, then Judge A might perceive Advocate B's language as more feminine even if it is linguistically identical to language used by male advocates. Furthermore, latent gender perception and latent language perception might interact in affecting the outcome through mechanisms such as rewarding "conforming to gender norms"—an advocate who is perceived as a man might get penalized for using feminine language whereas an advocate perceived as a woman might get rewarded, e.g. Gleason (2020).

6 Conclusion

In this work, we specify a causal research design for *differential treatment of social groups with language as a mediator*. We believe this research design is important for studying the direct and indirect causal effects in high-stakes decision making such as gender bias in the United States Supreme Court. Separating the indirect effect of treatment on outcome through interpretable language aspects allows us to estimate counterfactual queries about differential treatment when speakers use and do not use the same language. Despite open theoretical and technical challenges, we remain optimistic that researchers can build upon this framework and continue to improve our understanding of decision makers' differential treatment of social groups.

Acknowledgments

The authors thank Abe Handler, Su Lin Blodgett, Sam Witty, David Jensen, and anonymous reviewers from the First Workshop on Causal Inference & NLP for helpful comments. KK gratefully acknowledges funding from a Bloomberg Data Science Fellowship.

[13] This assumption is made in other NLP applications such as style transfer or machine translation (Prabhumoye et al., 2018; Li et al., 2018; Hovy et al., 2020).

References

Emily M. Bender. 2013. Linguistic fundamentals for natural language processing: 100 essentials from morphology and syntax. *Synthesis lectures on human language technologies*, 6(3):1–184.

Emily M. Bender and Alex Lascarides. 2019. Linguistic fundamentals for natural language processing ii: 100 essentials from semantics and pragmatics. *Synthesis Lectures on Human Language Technologies*, 12(3):1–268.

Richard Berk, Azusa Li, and Laura J. Hickman. 2005. Statistical difficulties in determining the role of race in capital cases: A re-analysis of data from the state of maryland. *Journal of Quantitative Criminology*, 21(4):365–390.

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

Su Lin Blodgett. 2021. Sociolinguistically driven approaches for just natural language processing. *PhD Thesis*.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Daniel Chen, Yosh Halberstam, and Alan CL Yu. 2016. Perceived masculinity predicts us supreme court outcomes. *PloS one*, 11(10):e0164324.

Daniel L Chen, Yosh Halberstam, Manoj Kumar, and Alan Yu. 2019. Attorney voice and the us supreme court. *Law as Data, Santa Fe Institute Press, ed. M. Livermore and D. Rockmore*.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney Newey. 2017. Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *WWW*.

Rhian M. Daniel, Bianca L. De Stavola, SN Cousens, and Stijn Vansteelandt. 2015. Causal mediation analysis with multiple mediators. *Biometrics*, 71(1):1–14.

Kay Deaux. 1985. Sex and gender. *Annual review of psychology*, 36(1):49–81.

Matthew J. Denny and Arthur Spirling. 2018. Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, 26(2):168–189.

Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics*, 12(4):453–476.

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

Matthew Finlayson, Aaron Mueller, Stuart Shieber, Sebastian Gehrmann, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *ACL-IJCNLP*.

Shane A. Gleason. 2020. Beyond mere presence: Gender norms in oral arguments at the us supreme court. *Political Research Quarterly*, 73(3):596–608.

D. James Greiner and Donald B. Rubin. 2011. Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, 93(3):775–785.

Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512.

Miguel A. Hernán. 2016. Does water kill? a call for less casual causal inferences. *Annals of epidemiology*, 26(10):674–680.

R. A. Hinde. 1976. Interactions, relationships and social structure. *Man*, 11(1):1–17.

Paul W. Holland. 1986. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Paul W. Holland. 2008. Causation and race. *White logic, white methods: Racism and methodology*, pages 93–109.

Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. 2020. "You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1686–1690.

Lily Hu and Issa Kohler-Hausmann. 2020. What's sex got to do with machine learning? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 513–513.

Kosuke Imai, Luke Keele, and Dustin Tingley. 2010. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309.

Kosuke Imai and Teppei Yamamoto. 2013. Identification and sensitivity analysis for multiple causal mechanisms: Revisiting evidence from framing experiments. *Political Analysis*, 21(2):141–171.

Miyako Inoue. 2006. *Vicarious language*. University of California Press.

Tonja Jacobi and Dylan Schweers. 2017. Justice, interrupted: The effect of gender, ideology, and seniority at supreme court oral arguments. *Va. L. Rev.*, 103:1379.

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.

Issa Kohler-Hausmann. 2018. Eddie Murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Nw. UL Rev.*, 113:1163.

Robin Lakoff. 1973. Language and woman's place. *Language in society*, 2(1):45–79.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Ian Lundberg, Rebecca Johnson, and Brandon M Stewart. 2021. What is your estimand? defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565.

Tali Mendelberg, Christopher F. Karpowitz, and J. Baxter Oliphant. 2014. Gender inequality in deliberation: Unpacking the black box of interaction. *Perspectives on Politics*, 12(1):18–44.

Stephen L. Morgan and Christopher Winship. 2015. *Counterfactuals and causal inference*. Cambridge University Press.

Kevin Munger. 2017. Tweetment effects on the tweeted: Experimentally reducing racist harassment. *Political Behavior*, 39(3):629–649.

Oyez. a. Kennedy v. Plan Administrator for DuPont Sav. and Investment Plan. Accessed 15 Jul. 2021.

Oyez. b. Lozano v. Montoya Alvarez. Accessed 15 Jul. 2021.

Dana Patton and Joseph L. Smith. 2017. Lawyer, interrupted: Gender bias in oral arguments at the US Supreme Court. *Journal of Law and Courts*, 5(2):337–361.

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

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl, Madelyn Glymour, and Nicholas P. Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Deanna Poos and Rita Simpson. 2002. Cross-disciplinary comparisons of hedging. *Using corpora to explore linguistic variation*, pages 3–23.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876.

Anna Prokofieva and Julia Hirschberg. 2014. Hedging and speaker commitment. In *5th Intl. Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*.

Reid Pryzant, Dallas Card, Dan Jurafsky, Victor Veitch, and Dhanya Sridhar. 2021. Causal effects of linguistic properties. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online. Association for Computational Linguistics.

Thomas S. Richardson and James M. Robins. 2013. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128*.

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.

Jonathan Rosa and Nelson Flores. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, 46(5):621–647.

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

Donald B Rubin. 2008. For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

Alexandra Schofield, Måns Magnusson, and David Mimno. 2017. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436.

Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science*, 19:499–522.

Uri Shalit, Fredrik D Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR.

Harold J. Spaeth, Lee Epstein, Andrew D. Martin, Jeffrey A. Segal, Theodore J. Ruger, and Sara C. Benesh. 2021. Supreme Court Database, Version 2021 Release 01. *Database at http://scdb.wustl.edu/*.

Jennifer Stromer-Galley. 2007. Measuring deliberation's content: A coding scheme. *Journal of public deliberation*, 3(1).

Elizabeth A. Stuart. 2010. Matching methods for causal inference: A review and a look forward. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 25(1):1.

Dustin Tingley, Teppei Yamamoto, Kentaro Hirose, Luke Keele, and Kosuke Imai. 2014. Mediation: R package for causal mediation analysis.

Mark J. Van der Laan and Sherri Rose. 2011. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Tyler VanderWeele and Stijn Vansteelandt. 2014. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115.

Tyler J. VanderWeele. 2016. Mediation analysis: a practitioner's guide. *Annual review of public health*, 37:17–32.

Victor Veitch, Dhanya Sridhar, and David Blei. 2020. Adapting text embeddings for causal inference. In *Conference on Uncertainty in Artificial Intelligence*, pages 919–928. PMLR.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *NeurIPS*.

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.

Candace West and Don H. Zimmerman. 1987. Doing gender, gender and society. *Vol. 1, No. 2.(Jun*, page 125.

Zach Wood-Doughty, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586. NIH Public Access.

Justine Zhang, Sendhil Mullainathan, and Cristian Danescu-Niculescu-Mizil. 2020. Quantifying the causal effects of conversational tendencies. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–24.