The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)

Uncovering Latent Biases in Text: Method and Application to Peer Review

Emaad Manzoor*, **Nihar B. Shah**†
Carnegie Mellon University
*emaad@cmu.edu, †nihars@cs.cmu.edu

Abstract

Quantifying systematic disparities in numerical quantities such as employment rates and wages between population subgroups provides compelling evidence for the existence of societal biases. However, biases in the *text* written for members of different subgroups (such as in recommendation letters for male and non-male candidates), though widely reported anecdotally, remain challenging to quantify. In this work, we introduce a novel framework to quantify bias in text caused by the visibility of subgroup membership indicators. We develop a nonparametric estimation and inference procedure to estimate this bias. We then formalize an identification strategy to causally link the estimated bias to the visibility of subgroup membership indicators, provided observations from time periods both before and after an identity-hiding policy change. We identify an application wherein "ground truth" bias can be inferred to evaluate our framework, instead of relying on synthetic or secondary data. Specifically, we apply our framework to quantify biases in the text of peer reviews from a reputed machine-learning conference before and after the conference adopted a double-blind reviewing policy. We show evidence of biases in the review ratings that serves as "ground truth", and show that our proposed framework accurately detects the presence (and absence) of these biases from the review text *without* having access to the review ratings.

Introduction

Societal biases against individuals based on race, gender, and other attributes can lead to disparities in hiring (Bertrand and Mullainathan 2004), wages (Blau and Kahn 2017) and incarceration rates (Alesina and La Ferrara 2014), among other socioeconomic outcomes. Uncovering evidence of such biases informs the creation of policies to eliminate long-standing gaps between different population subgroups.

Many important socioeconomic outcomes are influenced by written feedback, such as academic hiring that is influenced by reference letters, and employee appraisals that rely on managerial performance reviews. Previous studies have reported biases in the *text* of such feedback (Mitchell and Martin 2018; Madera et al. 2019; Correll et al. 2020). Mitchell and Martin (2018) find that student evaluations of female faculty teaching were more likely to focus on communication ability. Madera et al. (2019) find that academic letters of recommendation for women were more likely to contain "doubt-raising" phrases. In the context of managerial feedback provided during employee appraisals, Correll et al. (2020) find that "women were more likely to receive vague feedback that did not offer specific details of what they had done well and what they could do to advance".

Such biases in text can have severe economic consequences. For example, differences in the media framing of natural disasters were found to be associated with large disparities in the amount of allocated foreign aid (Strömberg 2007; Kweifio-Okai 2014). Similarly, online reviews of Asian and Indian restaurants declared them "unauthentic" when they defied the negative stereotypes of uncleanliness, leading to significant losses in revenue (Kay 2019).

Biases in text are more prevalent than those in numerical feedback due to the absence of enforced structure (Mackenzie, Wehner, and Correll 2019), and are more difficult to detect due to their subtle manner of expression (Morstatter et al. 2018). In addition, the inherently unstructured nature of text makes it challenging to quantify the biases it contains. Without the ability to quantify and provide credible evidence of such biases, society remains at risk of widening socioeconomic gaps fueled by the unchecked expression of prejudices in written feedback.

In this work, we propose a framework to quantify biases in text, provided data from time periods both before and after an identity-hiding policy change. Our proposed framework extends the difference-in-differences causal inference methodology introduced by Card and Krueger (1994) (which compares differences in *numerical* quantities over time) to handle differences in *unstructured text* over time. We motivate and evaluate our framework in the setting of scholarly peer review, which is a key mechanism of feedback and quality assurance in scientific research. Specifically, we assemble a dataset of peer reviews from the International Conference of Learning Representations (ICLR), which switched from single-blind to double-blind reviewing in 2018. We test for biases in the peer review text, which is prone to prejudice due to the lack of enforced structure. Importantly, our dataset enables estimating "ground truth" biases using the peer review ratings (a numerical quantity). The goal of our proposed framework is to quantify biases in the peer review text that are consistent with the "ground truth", without having access to the review ratings.

Our main contributions are:

I. We formalize bias as a *causal estimand* — the disparity in the peer review text between the subgroups *caused* by the visibility of author identities — that relies on a weaker assumption than "no unobserved confounders". We propose a nonparametric estimation and inference procedure to quantify this bias. Our procedure makes no assumptions on the data-generating process of the peer review text and requires no feature engineering.

II. We apply our proposed framework to quantify the bias in the text of peer reviews from the International Conference on Learning Representations (ICLR). We detect a statistically significant bias with respect to the authors' *affiliation*, but find no evidence of bias with respect to the authors' *perceived gender*.

III. Our chosen application is motivated by the opportunity to evaluate our proposed framework on "ground truth" biases derived from review ratings. Specifically, we evaluate our proposed framework by comparing the estimated biases in the peer review text with the biases in the peer review ratings estimated using the difference-in-differences methodology (Card and Krueger 1994; Angrist and Pischke 2008). We show that the biases in the peer review text estimated using our proposed framework are consistent with the "ground truth".

In the appendix, we also evaluate an alternative measure of disparity in the text proposed by Gentzkow, Shapiro, and Taddy (2019) and empirically show that our proposed measure of disparity has greater statistical power. Though presented in the context of peer review, our proposed framework can be applied to test for biases in employee feedback (Goldin and Rouse 2000) and new employee hiring (Capowski 1994), among other settings.

Reproducibility: We make our code and data publicly available at http://emaadmanzoor.com/biases-in-text/.

Related Work

Our work complements previous studies that investigated biases in peer review (Goldberg 1968; Ceci and Peters 1982; Swim et al. 1989; Lloyd 1990; Blank 1991; Garfunkel et al. 1994; Snodgrass 2006; Ross et al. 2006; Budden et al. 2008; Webb, O'Hara, and Freckleton 2008; Walker et al. 2015; Okike et al. 2016; Seeber and Bacchelli 2017; Bernard 2018; Tomkins, Zhang, and Heavlin 2017; Stelmakh, Shah, and Singh 2019; Stelmakh et al. 2021; Salimi et al. 2020). In contrast with our work, these studies focused on biases in numerical quantities such as review ratings.

Our approach compares single-blind and double-blind reviews for gender and affiliation-based population subgroups. A number of previous studies have quantified biases using the peer-review ratings provided for different subgroups under single-blind and double-blind reviewing policies. Blank (1991) conducts a randomized control trial and finds no significant difference in the review ratings for male and female authors under single and double-blind reviewing. Ross et al. (2006) compare single and double-blind reviewing in different years at a medical conference and find that the association between abstract acceptance and whether the authors were affiliated to institutions in the USA reduces significantly when reviewing is double-blind. Madden and DeWitt (2006) compare single and double-blind reviewing for the SIGMOD conference in different years and find that the mean number of accepted of papers by a "prolific" author remained largely similar before and after the switch to double-blind reviewing. In contrast, Tung (2006) find a significant reduction in the *median* number of accepted papers by a "prolific" author after SIGMOD switched to double-blind reviewing. More recently, Tomkins, Zhang, and Heavlin (2017) conduct a semi-randomized controlled trial with the WSDM 2017 conference and do not find a significant association between a paper's single-blind review rating and whether its authors were women, or whether its authors were affiliated to institutions in the USA. Salimi et al. (2020) compare single and double-blind reviewing in several conferences and find a significant effect of institutional prestige on review ratings when reviewing was single-blind (and none when reviewing was double-blind). Our work complements these studies by focusing on biases in the review *text*.

Our proposed framework is closely related to the bias discovery method proposed by Field and Tsvetkov (2020), who train a machine learning model to identify differences in online comments addressed towards men and women. Their method also relies on the idea that text which is predictive of gender is likely to contain bias. They show that their method can detect gender bias on a labeled dataset from a different domain than the one their model was trained on. However, the method proposed by Field and Tsvetkov (2020) crucially depends on the "no unobserved confounders" assumption. This assumption is restrictive and unlikely to hold in practice. In contrast, our proposed framework relies on a weaker assumption that remains valid under a large class of unobserved confounders (though requiring additional data, from two different time periods). We also overcome a key limitation in (Field and Tsvetkov 2020) by evaluating our proposed framework on "ground truth" derived from the same peer review process used for bias estimation, instead of on secondary datasets or tasks.

Problem Definition

We assume the availability of peer reviews from a conference in two different years, with author identities (their names, emails, and affiliations) visible to reviewers during the peer review process in exactly one of the two years. Our goal is to quantify biases in the *text* of the peer reviews written for papers belonging to two pre-specified subgroups based on a selected identifying attribute of their authors.

Previous studies have reported systematic *disparities* in the text written for different population subgroups. For example, Madera et al. (2019) report that recommendation letters for women are more likely to contain "doubt-raising" language then those for men. However, such disparities by themselves are not sufficient evidence of bias (Rathore and Krumholz 2004). While disparities are observed differences in the review text, biases are observed differences that are *caused* by author identity visibility, and not other factors.

In a *counterfactual* universe where the author identities are hidden, the bias must be zero (since its cause no longer exists) but the observed disparity can be nonzero. For example, if we partition papers into subgroups based on their first author's affiliation country, disparities in the review text could arise due to country-specific preferences for different research topics. When defining bias, our goal is to separate the disparity caused by author identity visibility from the disparity caused by other factors. We now formalize bias as a causal estimand with the potential outcomes framework (Imbens and Rubin 2015), and derive an expression that relates bias to the disparities observed in different time periods.

Consider papers submitted to a conference in the years t_{SB} and t_{DB}, where the conference employed single-blind reviewing in year t_{SB} and double-blind reviewing in year t_{DB}. We partition the papers into two subgroups G_0 and G_1 based on a selected identifying attribute of their authors (such their affiliation or perceived gender). Our goal is to formalize bias as the disparity in the review text for papers in each subgroup caused by the visibility of this identifying attribute to reviewers during the peer review process.

We denote by Δ_t the *observed disparity* in the review text in year $t \in \{t_{\text{SB}}, t_{\text{DB}}\}$. While we propose a careful nonparametric formulation of Δ_t in Section , Δ_t can be viewed as any measure of the difference in the review text in year t between subgroups G_0 and G_1. We denote by Δ_t^{SB} and Δ_t^{DB} the *counterfactual disparities* in the review text in year t that would have been observed had author identities been visible to and hidden from reviewers, respectively. Only one of the quantities Δ_t^{DB} and Δ_t^{SB} is visible in each year. When $t = t_{\text{SB}}$ and reviewing was single-blind, $\Delta_{t_{\text{SB}}}^{\text{DB}}$ is unobserved and quantifies the disparity in year t_{SB} had reviewing been double-blind instead. When $t = t_{\text{DB}}$ and reviewing was double-blind, $\Delta_{t_{\text{DB}}}^{\text{SB}}$ is unobserved and quantifies the disparity in year t_{DB} had reviewing been single-blind instead.

We define the *bias* (our causal estimand) as a difference in counterfactual disparities:

$$\text{bias} = \Delta_{t_{\text{DB}}}^{\text{DB}} - \Delta_{t_{\text{DB}}}^{\text{SB}}. \tag{1}$$

Eq. (1) subtracts the disparity $\Delta_{t_{\text{DB}}}^{\text{SB}}$ (caused by both author identity visibility and other factors) from the disparity $\Delta_{t_{\text{DB}}}^{\text{DB}}$ (caused by other factors only) to isolate the disparity caused by author identity visibility only. Note that the bias could also have been defined as $\Delta_{t_{\text{SB}}}^{\text{DB}} - \Delta_{t_{\text{SB}}}^{\text{SB}}$ (the change in the disparity that would have been observed had reviewing in year t_{SB} been double-blind instead). Either definition is valid and applicable to our framework (after minor algebraic changes).

Proposed Framework

Our goal is to estimate the bias defined in Eq. (1) given the peer reviews from a conference in years t_{SB} and t_{DB}, with author identities visible to reviewers during the peer review process in year t_{SB} and hidden in year t_{DB}. In this section, we first provide an identification proof to link the causal estimand in Eq. (1) (that contains unobservable counterfactual quantities) with an empirical estimand (that contains observable quantities only). We then propose a nonparametric estimation and inference procedure to estimate the bias from the available peer review data.

Identification

The bias as defined in Eq. (1) contains the unobservable counterfactual disparity $\Delta_{t_{\text{DB}}}^{\text{SB}}$ (the disparity in year t_{DB} had reviewing been single-blind), and cannot be estimated without further assumptions; this is the fundamental problem of causal inference (Holland 1986). The process of linking a causal estimand defined in terms of unobserved counterfactual quantities with an empirical estimand defined in terms of observed quantities is called *identification*, and relies on one or more *identification assumptions*. We make the following identification assumption:

Assumption 1. *The disparity in $t = t_{DB}$ had author identities been visible is equal to the disparity in $t = t_{SB}$ when author identities were indeed visible: $\Delta_{t_{DB}}^{SB} = \Delta_{t_{SB}}^{SB}$.*

Assumption 1 implies that the change in disparity from year t_{SB} to t_{DB} was caused only by the author identities being hidden in year t_{DB}, and not other factors. Assumption 1 remains valid in the presence unobserved confounders that affect the review text and (i) that do not vary from t_{SB} to t_{DB}, or (ii) that vary from t_{SB} to t_{DB} but affect the review text for both subgroups identically (such as a more critical reviewer pool in year t_{DB}). Hence, it is less restrictive than the "no unobserved confounders" assumption in prior work (Field and Tsvetkov 2020). We further discuss the validity of Assumption 1 for our setting in the appendix.

Let Δ_t be the *observed* disparity in year $t \in \{t_{\text{SB}}, t_{\text{DB}}\}$. Given Assumption 1, we link the bias (that contains an unobservable counterfactual disparity) with an empirical estimand (that contains observable disparities only) with the following *identification proof*:

$$\text{bias} \overset{(i)}{=} \Delta_{t_{\text{DB}}}^{\text{DB}} - \Delta_{t_{\text{DB}}}^{\text{SB}} \overset{(ii)}{=} \Delta_{t_{\text{DB}}}^{\text{DB}} - \Delta_{t_{\text{SB}}}^{\text{SB}} \overset{(iii)}{=} \Delta_{t_{\text{DB}}} - \Delta_{t_{\text{SB}}} \tag{2}$$

where the equation (i) follows from the definition in Eq. (1), equation (ii) follows from Assumption 1, and equation (iii) follows from the fact that reviewing was indeed double-blind in year t_{DB} ($\Delta_{t_{\text{DB}}}^{\text{DB}} = \Delta_{t_{\text{DB}}}$) and single-blind in year t_{SB} ($\Delta_{t_{\text{SB}}}^{\text{SB}} = \Delta_{t_{\text{SB}}}$).

Estimation and Inference

Having defined the bias in the review text in terms of observable disparities in Eq. (2), we now focus on estimating this bias from the available peer review data in years t_{SB} and t_{DB}. Formalizing the disparities $\Delta_{t_{\text{SB}}}$ and $\Delta_{t_{\text{DB}}}$ in the *text* in a manner that is both substantively meaningful and that permits estimation and inference is non-trivial. In this section, we formalize the disparities in the text and propose a nonparametric procedure to estimate them from the peer reviews in years t_{SB} and t_{DB}.

Intuitively, the disparity Δ_t in the review text in year t is a measure of how the text of the reviews written for G_0 differ from those written for G_1. A simple approach to quantify the disparity is to select a "feature" of the review text (such as its "politeness"), annotate the review text based on this feature (either manually or via natural language processing methods) and then compare the value of this feature in the reviews for papers in each of the two subgroups G_0 and G_1.

However, the disparities and bias quantified in this manner are sensitive to feature selection and annotation.

In contrast, we propose measuring the Δ_t nonparametrically, without any feature selection and annotation. We rely on the intuition that if the text of the reviews written for G_0 differs systematically from the text of those written for G_1, a binary machine-learning classifier should be able to distinguish between the reviews written for each subgroup using the *review text*. Hence, we could use any measure of the performance (such as the accuracy, precision, or recall) of such a classifier as a measure of the disparity Δ_t.

However, disparities in the review text may also be caused by differences in the research topics pursued by each subgroup. To "control for" subgroup differences in research topics, we rely on the following intuition: subgroup differences in research topics should be reflected in the text of their paper abstracts. Hence, we quantify subgroup differences in research topics by the ability of a binary machine-learning classifier to distinguish between the papers belonging to each subgroup using their *abstract text*.

We now define the disparity in the review text based on the intuition discussed previously. Let $f(\cdot)$ be a binary classifier mapping a paper's review text to its subgroup and $g(\cdot)$ be a binary classifier mapping a paper's abstract text to its subgroup. Let perf$(f;t)$ and perf$(g;t)$ be the chosen measures of classification performance of $f(\cdot)$ and $g(\cdot)$ respectively, such as their area under the ROC curve (AUC), accuracy or precision. We measure the disparity in the review text as the ratio of the performances of the two classifiers on the peer review data in year t:

$$\Delta_t = \text{perf}(f;t)/\text{perf}(g;t). \quad (3)$$

Normalizing perf$(f;t)$ by perf$(g;t)$ as in Eq. (3) "controls for" subgroup differences in research topics: if perf$(f;t)$ is high due to subgroup differences in research topics, perf$(g;t)$ will also be high. While any binary classifier may be used for $f(\cdot)$ and $g(\cdot)$, poor classifiers are more likely to underestimate the bias (due to equally poor classification performance in both t_{SB} and t_{DB}).

In Section , we report results with multinomial Naive Bayes classifiers for $f(\cdot)$ and $g(\cdot)$ and the AUC as our chosen measure of classification performance. We estimate the value of perf$(f;t)$ and perf$(g;t)$ using k-fold cross-validation. To eliminate any dependence on the choice of cross-validation folds, we repeat the bias estimation procedure many times with the data belonging to each fold randomized uniformly in each iteration. We use the empirical distribution of bias estimates from these iterations to construct confidence intervals on the estimated bias.

A final issue we address is that perf$(f;t)$ and perf$(g;t)$ can differ in t_{SB} and t_{DB} due to differences in the sample size (number of reviews) or due to differences in the proportion of papers belonging to each subgroup in t_{SB} and t_{DB}. Hence, when estimating $\Delta_{t_{\text{DB}}}$ we downsample the available peer review data in year t_{DB} such that (i) the number of peer reviews or abstracts is equal to that in t_{SB}, (ii) the proportion of abstracts or peer reviews written for papers in subgroup G_0 is equal to that in t_{SB}, and (iii) the proportion of abstracts or peer reviews written for papers in subgroup G_1 is equal to that in t_{SB}. As with the cross-validation folds, the downsampling is randomized uniformly in each iteration.

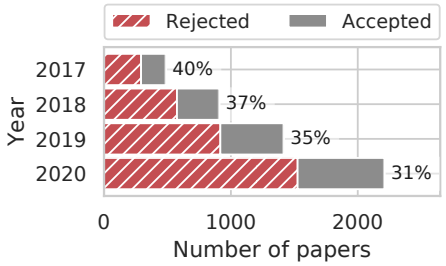


Figure 1: Papers submitted, accepted and rejected from ICLR 2017 through 2020: The proportion of papers accepted in each year is reported to the right of each bar.

	Subgroup Definition	
Year	Affiliation-based	Gender-based
2017	24.1%	25.5%
2018	24.1%	28.5%
2019	26.0%	—
2020	30.5%	—

Table 1: Proportion of submitted papers belonging to subgroup G_1 in each year for different subgroup definitions.

Data

We assemble a dataset of 16,880 peer reviews from the OpenReview platform for all the 5,638 papers submitted to the International Conference on Learning Representations (ICLR) from 2017 to 2020. Each paper receives 3 peer reviews on average (with a standard deviation of 0.3). Each peer review contains textual comments and a numerical rating, from 1 to 10 in ICLR 2017–2019 and in $\{1, 3, 6, 8\}$ in ICLR 2020. Fig. 1 reports the number of papers submitted and the proportion of papers accepted in each year.

We investigate the existence of biases in the peer review text with respect to two types of author attributes: (i) the country of their affiliation, and (ii) their perceived gender. Affiliation and gender biases have been a recurring theme in prior work on improving fairness in peer review (Blank 1991; Ross et al. 2006; Tomkins, Zhang, and Heavlin 2017; Salimi et al. 2020). We focus on testing for affiliation and gender bias in the review text to complement prior findings, though biases with respect to other attributes may also exist.

ICLR was co-founded by researchers affiliated with institutions in the USA and Canada. In addition, the general, program and area chairs of ICLR were almost exclusively affiliated with institutions in the USA and Canada since its inception in 2013. Motivated by the possibility of "in-group bias" (Taylor and Doria 1981), we test for reviewer biases caused by visible author identities in favor of (or against) papers having at least one author with an affiliation in the USA or Canada. We partition the submitted papers into two subgroups, G_0 and G_1, based on the countries of the affilia-

tions of their authors. We allocate all papers having at least one author affiliated to a university, organization, or company in the USA or Canada to G_0, and all other papers to G_1. Author affiliations are extracted from the submitted paper PDFs in ICLR 2017, and from the authors' registered emails on the OpenReview platform in ICLR 2018, 2019 and 2020. The goal of this affiliation-based partitioning is to quantify the extent to which reviewers are biased by visible author affiliations.

In addition to affiliation-based subgroups, we consider subgroups based on the authors' gender *as perceived by the reviewer* (and not self-reported). Since we do not observe how reviewers infer gender from authors' names, we approximate the perceived gender of each author using the following protocol. We first use historical self-reported gender records from the U.S. Social Security Administration to compute the probability of an author's first name being reported as male.[1] If this probability is greater than 90%, we annotate the author's perceived gender as *male*. If this probability is less than 10%, we annotate the author's perceived gender as *non-male*. If this probability is between 10% and 90%, an external human annotator manually infers the gender of the author (male or non-male) using visible information on their homepage and Google Scholar profile (found with a Google search). We expect our annotation protocol to approximate the authors' gender as perceived by reviewers[2].

We allocate all papers having at least one author perceived to be non-male to G_1, and those with all authors perceived to be male to G_0. The goal of this gender-based partitioning is to quantify the extent to which reviewers are biased in favor of (or against) papers having at least one author perceived to be non-male. Table 1 reports the proportion of papers submitted to ICLR in each year that belong to subgroup G_1 for affiliation-based and gender-based subgroup definitions. Since our external human annotations of gender only span ICLR 2017 and 2018, our analyses of bias with gender-based subgroups excludes data from ICLR 2019 and 2020.

A key policy change during this period is ICLR's switch to double-blind reviewing from 2018 onwards. We exploit this policy change to estimate the bias while eliminating the impact of a large class of unobserved confounders, as discussed in Section . We also exploit this policy change in Section to construct a "ground truth" measure of bias in the peer review ratings using the difference-in-differences methodology. We then evaluate whether the biases estimated by our proposed framework are consistent with the presence and absence of the "ground truth" bias in each year.

Since ICLR permits non-anonymized submissions to arXiv and other preprint servers while the paper is under review, it is likely that some author identities were visible even during the double-blind reviewing process in ICLR 2018, 2019 and 2020. Hence, we expect our bias estimates to be conservative (attenuated towards zero).

[1]We do this using the `gender` package available at https://github.com/ropensci/gender.

[2]We note that the name-to-gender annotations derived from the aforementioned protocol are likely to be U.S.-centric.

Evaluation

We now evaluate the ability of our proposed framework to detect biases in the *text* of peer reviews. Evaluating the validity of causal estimates is challenging in general due to the lack of "ground truth" to compare with, which can only be obtained using randomized control trials that are often expensive and time-consuming. Hence, as in prior work (Field and Tsvetkov 2020), evaluation is typically carried out using secondary tasks or semi-synthetic datasets such as the IBM Causal Inference Benchmark (Shimoni et al. 2018).

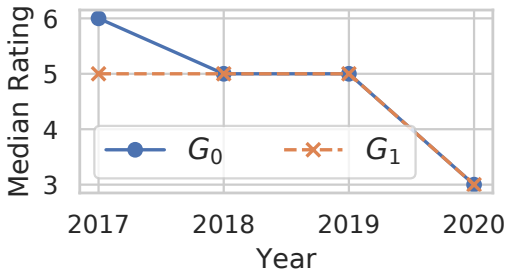
Instead of relying on secondary tasks or semi-synthetic datasets that may not represent peer reviewer behavior in the real world, we construct "ground truth" bias estimates based on the ratings provided by peer reviewers. Specifically, we apply the difference-in-differences methodology to quantify the presence or absence of biases in the review ratings using peer reviews from each consecutive pair of years between ICLR 2017 and 2020. We then apply our proposed framework to estimate biases in the review text.

We evaluate whether (i) the estimated bias in the review text is statistically significant when the estimated bias in the review ratings is statistically significant, and (ii) the estimated bias in the review text is statistically insignificant when the estimated bias in the review ratings is statistically insignificant. The underlying intuition is that the rating of a review must also be reflected in its text (with language expressing praise or criticism, for example). Hence, an accurate textual bias estimation framework must be able to detect biases (when present) using the review text without having access to the review ratings. Our evaluation procedure thus serves as a falsification test of our proposed method.

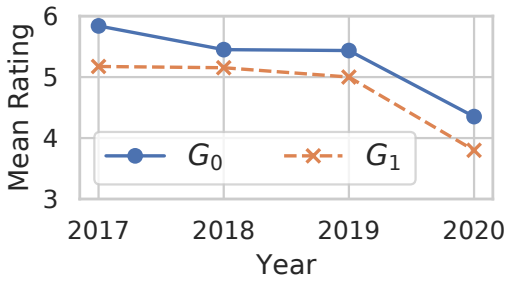
We begin in Section with a detailed discussion on estimating the "ground-truth" affiliation bias using the difference-in-differences methodology and review ratings. We then estimate and evaluate the affiliation bias in the review text in Section . In Section , we estimate and evaluate the bias due to perceived gender. In the appendix, we further discuss the validity of the identification assumptions used in our evaluation. As a comparative baseline, we also evaluate an alternate measure of disparity in the text proposed by (Gentzkow, Shapiro, and Taddy 2019) and empirically show that our proposed measure of disparity has greater statistical power.

Constructing "Ground Truth" Affiliation Biases From Peer Review Ratings

We first provide descriptive evidence in Fig. 2 of reviewer bias in favor of papers having at least one author with an affiliation in the USA or Canada (subgroup G_0) in ICLR 2017, when reviewing was single-blind. Fig. 2a shows that the median review rating for papers in G_0 was 1 unit higher than that for papers in G_1 in ICLR 2017. This median ratings disparity disappears after the switch to double-blind reviewing in ICLR 2018. Similarly, Fig. 2b shows that the mean review rating for papers in G_0 was 0.665 units higher than that for papers in G_1 in ICLR 2017, and 0.297 units higher in ICLR 2018. Thus, switching to double-blind reviewing coincided with a reduction in mean ratings disparity of 0.369 units.



(a) Median review rating for each subgroup



(b) Mean review rating for each subgroup

Figure 2: Ratings disparities in ICLR 2017 through 2020: The ratings disparity is quantified by the difference in (a) median, and (b) mean ratings for papers in G_0 and G_1.

We now estimate the ratings bias using the *difference-in-differences* causal inference methodology (Card and Krueger 1994; Angrist and Pischke 2008). The difference-in-differences methodology can be viewed as an analogue of our proposed framework to quantify biases in numerical quantities, instead of biases in unstructured text. This framework has been previously used to quantify gender biases in peer review ratings and hiring decisions (Blank 1991; Goldin and Rouse 2000), among several other settings.

The difference-in-differences methodology, like our proposed framework, requires peer reviews in two years t_{SB} (with single-blind reviewing) and t_{DB} (with double-blind reviewing). Let r_{ij} be the rating that reviewer i gave to paper j, let $T_j \in \{t_{\text{SB}}, t_{\text{DB}}\}$ be the year in which paper j was submitted, and let $S_j \in \{G_0, G_1\}$ be subgroup that paper j belongs to. The difference-in-differences methodology defines the ratings disparity Δ_t^{rating} in each year $t \in \{t_{\text{SB}}, t_{\text{DB}}\}$ as:

$$\begin{aligned}\Delta_t^{\text{rating}} &= \mathbb{E}[r_{ij} | S_j = G_0, T_j = t] \\ &\quad - \mathbb{E}[r_{ij} | S_j = G_1, T_j = t] \end{aligned} \tag{4}$$

where the expectations are over all papers j and their respective reviewers i. The ratings bias γ is defined as the difference in the ratings disparities between the year with single-blind reviewing and the year with double-blind reviewing:

$$\gamma = \Delta_{t_{\text{DB}}}^{\text{rating}} - \Delta_{t_{\text{SB}}}^{\text{rating}} \tag{5}$$

Interpreting γ as a ratings *bias* – the difference in mean subgroup ratings *caused* by visible author identities – requires the *parallel trends* identification assumption. This assumption states that, had the conference never switched to double-blind reviewing, the change in expected rating for subgroup G_1 from t_{SB} to t_{DB} would have been equal that for subgroup G_0 from t_{SB} to t_{DB}. It is a special case of Assumption 1 when the disparity in each year is defined as a difference in mean subgroup ratings, as in Eq. (4). The parallel trends assumption is less restrictive than the "no unobserved confounders" assumption. We discuss the validity of this assumption in our setting in the appendix.

The ratings bias γ in Eq. (5) is typically estimated using a "two-way fixed-effects" regression (Imai and Kim 2020) on peer reviews from the years t_{SB} and t_{DB}:

$$\begin{aligned} r_{ij} = \rho + \alpha \mathbb{I}[T_j = t_{\text{DB}}] + \beta \mathbb{I}[S_j = G_0] \\ + \gamma \mathbb{I}[T_j = t_{\text{DB}}] \times \mathbb{I}[S_j = G_0] + \epsilon_{ij} \end{aligned} \tag{6}$$

where the coefficients ρ, α, β and γ are estimated using ordinary least squares (OLS). The error term ϵ_{ij} is assumed to be Gaussian with zero-mean, which enables deriving asymptotic confidence intervals and p-values for the estimates.

We estimate the ratings bias using the two-way fixed-effects regression model in Eq. (6) on peer reviews from ICLR 2017 (t_{SB}) and 2018 (t_{DB}). Recall that Fig. 2b reports a change in the mean ratings disparity from $\Delta_{2017}^{\text{rating}} = 0.665$ to $\Delta_{2018}^{\text{rating}} = 0.297$, for a total of $\Delta_{2018}^{\text{rating}} - \Delta_{2017}^{\text{rating}} = -0.369$ units after switching to double-blind reviewing. The estimated bias in the first row of Table 2 (left) mirrors this. In addition, the confidence intervals and p-value indicate that the estimated bias is statistically significant ($p = 0.024$).

Recall that reviewing in ICLR was double-blind in the years 2018, 2019 and 2020. Hence, as "placebo tests", we also estimate γ using the two-way fixed-effects regression model in Eq. (6) using peer reviews in the year pair ($t_{\text{SB}} = 2018, t_{\text{DB}} = 2019$), and the year pair ($t_{\text{SB}} = 2019, t_{\text{DB}} = 2020$). The estimates are reported in the second and third rows of Table 2 (left). The ratings bias estimated using either of these year pairs is statistically insignificant. This is consistent with the fact that reviewing was double-blind during both years in the pair, and lends support to the validity of the parallel trends assumption.

Table 2 (left) thus comprises the "ground truth" presence and absence of bias for each year pair, which we expect our proposed framework to uncover from the review text without having access to the review ratings.

Estimating and Evaluating Affiliation Bias in the Review Text

We now estimate the bias in the review text using equations (1) and (3). We use multinomial Naive Bayes classifiers with add-one smoothing for $f(\cdot)$ and $g(\cdot)$ on frequencies of unigrams and bigrams in the review and abstract text respectively. We use the area under the ROC curve (AUC) for both perf$(f; t)$ and perf$(g; t)$, estimated using 10-fold cross-validation. We downsample the reviews and abstracts in year t_{DB} to equalize the sample sizes and subgroup proportions in t_{SB} and t_{DB}, as described in Section . We repeat the bias estimation procedure 1,000 times with downsampling and

	"Ground truth" bias in the review ratings			Estimated bias in the review text		
Years $t_{\text{SB}}, t_{\text{DB}}$	Bias	p-value	95% CI	Bias	p-value	95% CI
2017, 2018	**-0.369** (0.164)	**0.024**	(-0.690, -0.047)	**-0.166** (0.055)	**0.002**	(-0.270, -0.063)
Placebo Tests						
2018, 2019	0.138 (0.112)	0.219	(-0.082, 0.358)	-0.070 (0.068)	0.308	(-0.195, 0.072)
2019, 2020	0.118 (0.099)	0.236	(-0.077, 0.313)	0.012 (0.043)	0.781	(-0.082, 0.083)

Table 2: "Ground truth" and estimated bias with respect to affiliation: "Ground truth" difference-in-difference estimates of the bias in the review ratings (left) and bias in the review text estimated by our proposed framework (right). Standard errors reported in brackets. Estimates in each row are computed using ICLR peer reviews in consecutive years t_{SB} and t_{DB}. Estimates in bold are statistically significant at the 5% level.

the cross-validation folds randomized uniformly in each iteration. We use the empirical distribution of bias estimates from these iterations to construct confidence intervals on the estimated bias. We compute the p-value from the confidence intervals using the analytical method proposed by Altman and Bland (2011).

The estimated biases in the review text for the year pair $(t_{\text{SB}} = 2017, t_{\text{DB}} = 2018)$ and the placebo year pairs $(t_{\text{SB}} = 2018, t_{\text{DB}} = 2019)$ and $(t_{\text{SB}} = 2019, t_{\text{DB}} = 2020)$ are reported in Table 2 (right). The first row of Table 2 (right) reports a statistically significant ($p = 0.002$) estimated bias corresponding to a *reduction* of 0.166 units (and hence, a negative estimate) in the classification performance ratio (see Eq. 3) from ICLR 2017 to 2018. The second and third rows of Table 2 (right) report statistically insignificant bias estimates using peer reviews in the double-blind year pairs $(t_{\text{SB}} = 2018, t_{\text{DB}} = 2019)$ and $(t_{\text{SB}} = 2019, t_{\text{DB}} = 2020)$. The biases in the review text in each year pair estimated using our proposed framework are consistent with the presence and absence of "ground truth" ratings bias in each year pair reported in Table 2 (left). This validates the effectiveness of our proposed framework.

Estimating and Evaluating Bias With Respect To The Authors' Perceived Gender

Different types of biases may be expressed in the text against population subgroups defined in different ways (such as by affiliation, race or gender). Our proposed framework does not rely on linguistic feature-engineering targeted at any specific type of bias. Hence, we evaluate the ability of our proposed framework to test for biases with subgroups defined based on the authors' gender *as perceived by the reviewer* (and not their self-reported gender). We detailed our gender-based subgroup definition and our gender annotation protocol earlier in Section .

Since our manual gender annotations only span ICLR 2017 and 2018, we report the estimated bias in the review ratings and text using ICLR 2017 and 2018 in Table 3. The "ground truth" bias in the review ratings (estimated using the difference-in-differences methodology as in Section) is statistically insignificant. The bias in the review text estimated using our proposed framework is also statistically insignificant, and hence, consistent with the "ground truth".

In summary, given our data and choice of gender-based subgroups, we cannot reject the null hypotheses of there being no bias in the review ratings and text against papers with at least one author perceived to be non-male. It is, however, important to note that failing to reject the null hypothesis does not confirm the absence of gender bias.

Source	Bias	p-value	95% CI
Ratings	-0.073 (0.160)	0.647	(-0.386, 0.240)
Text	-0.468 (0.335)	0.163	(-0.862, 0.198)

Table 3: "Ground truth" and estimated bias with respect to perceived gender: Estimated bias in the review text and review ratings with respect to the authors' perceived gender using ICLR peer reviews in the years 2017 and 2018. Standard errors reported in brackets.

Conclusion

Our work addresses an important yet relatively overlooked medium through which biases can harm society: that of text-based communication. We propose a framework to nonparametrically estimate biases expressed in text, which is robust to a larger class of unobserved confounders than prior work. We evaluate our approach in the setting of scholarly peer-review, wherein the "ground truth" bias can be inferred, and show that our proposed framework detects bias in the peer review text that is consistent with the "ground truth". Our framework can be used by policymakers to formulate more effective bias-mitigation policies that improve the equitability of hiring, promotion and other socioeconomic processes.

More generally, our work extends the difference-in-differences methodology to accommodate unstructured text as the "outcome". It operates on text observed in two time periods associated with two population subgroups before and after a (potentially identity-hiding) policy change, such as switching to age-blind recruitment (Capowski 1994), blind performance reviews (Goldin and Rouse 2000) or blind grading (Hanna and Linden 2012). Our proposed framework quantifies the *causal* effect of the policy change on the difference in the text associated with each population subgroup. As such, our work also contributes to the nascent literature on causal inference from text (Roberts, Stewart, and Nielsen 2020; Sridhar and Getoor 2019; Egami et al. 2018; Keith, Jensen, and O'Connor 2020) with "text as the outcome".

Acknowledgements

We thank Morgan Schaming for annotation assistance. This work was supported by NSF CAREER Award 1942124.

Ethics Statement

As a policymaking tool, our proposed framework could potentially be used incorrectly, to allege bias where there is none, or the lack of bias when it exists. Our evaluation study in Section is a detailed example of how the estimates and confidence intervals from our proposed framework are to be interpreted, both when they are statistically significant and insignificant. We expect that, with this example, users of our proposed framework are motivated to employ similar care when interpreting the bias estimates in their setting.

Our proposed framework could also be used incorrectly if the validity of the underlying identification assumption (Assumption 1) is not evaluated with care. If an unobserved confounder exists that violates Assumption 1, our estimates will quantify the change in disparity from t_{SB} to t_{DB} caused by a combination of hiding author identities *and* the unobserved confounder, which cannot be interpreted as bias. While this assumption is empirically untestable (since it involves unobserved counterfactual quantities), placebo tests can be used to empirically support its validity. However, note that while the failure of a placebo test implies that the identification assumption does not hold, the success of a placebo test does *not* confirm that the identification assumption holds.

In our peer review setting, a potential confounder is an increase in research funding *only* for the institutions comprising G_1 from ICLR 2017 to 2018, and not for those comprising G_0. In Fig. 2, note that the subgroup disparity in the mean and median review ratings decreases from ICLR 2017 to 2018 (the mean and median review ratings for the two subgroups become more similar in 2018). However, also note that the reduction in disparity is due to a *decrease* in the mean and median ratings for subgroup G_0 in 2018. Had research funding for the institutions comprising G_1 increased, we would have expected an *increase* in the mean or median ratings for subgroup G_1 in 2018. Hence, the temporal trends in ratings in Fig. 2 contradict the hypothesis of confounding due to an increase in research funding for the institutions comprising G_1. In the appendix, we detail this argument further and show how a combination of substantive reasoning, empirical tests and external evidence must be used to assess the validity of the identification assumption.

We also foresee ethical concerns with releasing our gender annotations publicly. We designed our gender annotation protocol to approximate how reviewers perceive the gender of an author from their name, without knowledge of the author's self-reported gender. As such, our annotations may contradict the true gender of an author and cause them unintended psychological harm. To prevent this, we do not plan to make our gender annotations public. However, we have described our gender annotation protocol in sufficient detail in Section for interested parties to replicate if desired.

References

Alesina, A.; and La Ferrara, E. 2014. A test of racial bias in capital sentencing. *American Economic Review* .

Altman, D. G.; and Bland, J. M. 2011. How to obtain the P value from a confidence interval. *BMJ: British Medical Journal* .

Angrist, J. D.; and Pischke, J.-S. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bernard, C. 2018. Gender Bias in Publishing: Double-Blind Reviewing as a Solution? *Eneuro* .

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* .

Blank, R. M. 1991. The effects of double-blind versus single-blind reviewing: Experimental evidence from the American Economic Review. *American Economic Review* .

Blau, F. D.; and Kahn, L. M. 2017. The gender wage gap: Extent, trends, and explanations. *Journal of Economic Literature* .

Budden, A. E.; Tregenza, T.; Aarssen, L. W.; Koricheva, J.; Leimu, R.; and Lortie, C. J. 2008. Double-blind review favours increased representation of female authors. *Trends in Ecology and Evolution* .

Capowski, G. 1994. Ageism: The new diversity issue. *Management Review* .

Card, D.; and Krueger, A. B. 1994. Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania. *The American Economic Review* .

Ceci, S. J.; and Peters, D. P. 1982. Peer review: A study of reliability. *Change: The Magazine of Higher Learning* .

Correll, S.; Weisshaar, K.; Wynn, A.; and Wehner, J. 2020. Inside the Black Box of Organizational Life: The Gendered Language ofP Performance Assessment. *American Sociological Review* .

Egami, N.; Fong, C. J.; Grimmer, J.; Roberts, M. E.; and Stewart, B. M. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163* .

Field, A.; and Tsvetkov, Y. 2020. Unsupervised Discovery of Implicit Gender Bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Garfunkel, J. M.; Ulshen, M. H.; Hamrick, H. J.; and Lawson, E. E. 1994. Effect of institutional prestige on reviewers' recommendations and editorial decisions. *JAMA* .

Gentzkow, M.; Shapiro, J. M.; and Taddy, M. 2019. Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica* .

Goldberg, P. 1968. Are women prejudiced against women? *Trans-action* .

Goldin, C.; and Rouse, C. 2000. Orchestrating impartiality: The impact of" blind" auditions on female musicians. *American Economic Review* .

Hanna, R. N.; and Linden, L. L. 2012. Discrimination in grading. *American Economic Journal: Economic Policy* .

Holland, P. W. 1986. Statistics and causal inference. *Journal of the American Statistical Association* .

Imai, K.; and Kim, I. S. 2020. On the use of two-way fixed effects regression models for causal inference with panel data. *Political Analysis* .

Imbens, G. W.; and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Kay, S. 2019. Yelp Reviewers' Authenticity Fetish Is White Supremacy in Action. *Eater New York* URL https://ny.eater.com/2019/1/18/18183973/authenticity-yelp-reviews-white-supremacy-trap.

Keith, K.; Jensen, D.; and O'Connor, B. 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 (ACL)*.

Kweifio-Okai, C. 2014. Media distortion and western bias – why do some disasters attract more cash? *The Guardian* URL https://www.theguardian.com/global-development/2014/dec/02/students-speak-media-distortion-western-bias-disasters.

Lloyd, M. E. 1990. Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis* .

Mackenzie, L.; Wehner, J.; and Correll, S. J. 2019. Why Most Performance Evaluations Are Biased, and How to Fix Them. *Harvard Business Review* URL https://hbr.org/2019/01/why-most-performance-evaluations-are-biased-and-how-to-fix-them.

Madden, S.; and DeWitt, D. 2006. Impact of double-blind reviewing on SIGMOD publication rates. *ACM SIGMOD Record* .

Madera, J. M.; Hebl, M. R.; Dial, H.; Martin, R.; and Valian, V. 2019. Raising doubt in letters of recommendation for academia: gender differences and their impact. *Journal of Business and Psychology* .

Mitchell, K. M.; and Martin, J. 2018. Gender bias in student evaluations. *PS: Political Science & Politics* .

Morstatter, F.; Wu, L.; Yavanoglu, U.; Corman, S. R.; and Liu, H. 2018. Identifying framing bias in online news. *ACM Transactions on Social Computing* .

Okike, K.; Hug, K. T.; Kocher, M. S.; and Leopold, S. S. 2016. Single-blind vs double-blind peer review in the setting of author prestige. *JAMA* .

Rathore, S. S.; and Krumholz, H. M. 2004. Differences, Disparities, and Biases: Clarifying Racial Variations in Health Care Use. *Annals of Internal Medicine* .

Roberts, M. E.; Stewart, B. M.; and Nielsen, R. A. 2020. Adjusting for confounding with text matching. *American Journal of Political Science* .

Ross, J. S.; Gross, C. P.; Desai, M. M.; Hong, Y.; Grant, A. O.; Daniels, S. R.; Hachinski, V. C.; Gibbons, R. J.; Gardner, T. J.; and Krumholz, H. M. 2006. Effect of blinded peer review on abstract acceptance. *JAMA* .

Salimi, B.; Parikh, H.; Kayali, M.; Getoor, L.; Roy, S.; and Suciu, D. 2020. Causal relational learning. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*.

Seeber, M.; and Bacchelli, A. 2017. Does single blind peer review hinder newcomers? *Scientometrics* .

Shimoni, Y.; Yanover, C.; Karavani, E.; and Goldschmnidt, Y. 2018. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046* .

Snodgrass, R. 2006. Single-versus double-blind reviewing: an analysis of the literature. *ACM SIGMOD Record* .

Sridhar, D.; and Getoor, L. 2019. Estimating causal effects of tone in online debates. In *IJCAI*.

Stelmakh, I.; Shah, N.; and Singh, A. 2019. On Testing for Biases in Peer Review. *Advances in Neural Information Processing Systems (NeurIPS)* .

Stelmakh, I.; Shah, N.; Singh, A.; and Daumé III, H. 2021. Prior and Prejudice: The Novice Reviewers' Bias against Resubmissions in Conference Peer Review. In *CSCW*.

Strömberg, D. 2007. Natural disasters, economic development, and humanitarian aid. *Journal of Economic Perspectives* .

Swim, J.; Borgida, E.; Maruyama, G.; and Myers, D. G. 1989. Joan McKay versus John McKay: Do gender stereotypes bias evaluations? *Psychological Bulletin* .

Taylor, D. M.; and Doria, J. R. 1981. Self-serving and group-serving bias in attribution. *The Journal of Social Psychology* .

Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Science* .

Tung, A. K. 2006. Impact of double blind reviewing on SIGMOD publication: a more detail analysis. *ACM SIGMOD Record* .

Walker, R.; Barros, B.; Conejo, R.; Neumann, K.; and Telefont, M. 2015. Bias in peer review: a case study. *F1000Research* .

Webb, T. J.; O'Hara, B.; and Freckleton, R. P. 2008. Does double-blind review benefit female authors? *Trends in Ecology & Evolution* .