Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus

Pratik S. Sachdeva*
pratik.sachdeva@berkeley.edu
D-Lab, University of California, Berkeley
Berkeley, CA, USA

Renata Barreto*
School of Law, University of California, Berkeley
Berkeley, CA, USA

Claudia von Vacano
D-Lab, University of California, Berkeley
Berkeley, CA, USA

Chris J. Kennedy
Center for Precision Psychiatry, Harvard Medical School
Boston, MA, USA

ABSTRACT

Content Warning: This paper contains content considered profane, hateful, and offensive.

Annotators, by labeling data samples, play an essential role in the production of machine learning datasets. Their role is increasingly prevalent for more complex tasks such as hate speech or disinformation classification, where labels may be particularly subjective, as evidenced by low inter-annotator agreement statistics. Annotators may exhibit observable differences in their labeling patterns when grouped by their self-reported demographic identities, such as race, gender, etc. We frame these patterns as *annotator identity sensitivities*, referring to an annotator's increased likelihood of assigning a particular label on a data sample, conditional on a self-reported identity group. We purposefully refrain from using the term *annotator bias*, which we argue is problematic terminology in such subjective scenarios. Since annotator identity sensitivities can play a role in the patterns learned by machine learning algorithms, quantifying and characterizing them is of paramount importance for fairness and accountability in machine learning. In this work, we utilize item response theory (IRT), a methodological approach developed for measurement theory, to quantify annotator identity sensitivity. IRT models can be constructed to incorporate diverse factors that influence a label on a specific data sample, such as the data sample itself, the annotator, and the labeling instrument's wording and response options. An IRT model captures the contributions of these facets to the label via a latent-variable probabilistic model, thereby allowing the direct quantification of annotator sensitivity. As a case study, we examine a hate speech corpus containing over 50,000 social media comments from Reddit, YouTube, and Twitter, rated by 10,000 annotators on 10 components of hate speech (e.g., sentiment, respect, violence, dehumanization, etc.). We leverage three different IRT techniques which are complementary in that they quantify sensitivity from different perspectives: separated measurements, annotator-level interactions, and group-level interactions. We use these techniques to assess whether an annotator's racial identity is associated with their ratings on comments that target different racial identities. We find that, after controlling for the estimated hatefulness of social media comments, annotators tended to be more sensitive when rating comments targeting a group they identify with. Specifically, annotators were more likely to rate comments targeting their own racial identity as possessing elements of hate speech. Our results identify a correspondence between annotator identity and the target identity of hate speech comments, and provide a set of tools that can assess annotator identity sensitivity in machine learning datasets at large.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Social media**.

KEYWORDS

annotation, hate speech, annotator sensitivity, item response theory, differential rater functioning

ACM Reference Format:
Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. 2022. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3531146.3533216

1 INTRODUCTION

Algorithmic bias in machine learning algorithms is well-studied across a variety of domains. This bias has been attributed to many sources, including systemic biases during the underlying data generation process, inherited bias from pre-trained models, annotation bias, and others [43]. Annotation bias is of significant interest, particularly as machine learning continues to be applied on increasingly diverse and difficult tasks, which generally require the labeling of new, large-scale datasets [7, 24, 29]. Since the labels of these datasets serve as a "ground truth" for supervised learning, their quality and consistency is of paramount importance to ensure that these models–which may be deployed in real-life settings–properly learn the intended relationships within the problem [50]. However, obtaining labels is a labor-intensive process, and is typically performed by workers via services such as Amazon Mechanical Turk or Figure Eight [63]. The role of these annotators, and the influence

*Co-first author

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533216

of their background and perspective on label generation, has long been underappreciated within the machine learning community [7, 49].

The notion of a "ground truth" label suggests a universally agreed upon label for each data sample. In reality, however, annotators may express disagreement on labels. This is especially true in tasks dealing with complex, nuanced concepts, such as hate speech or disinformation classification. In these cases, a measure of "annotator agreement", such as Kripendorff's alpha, is often used to assess the quality of the dataset [37]. Greater agreement suggests higher quality labels, indicating the data is suitable for predictive modeling. This viewpoint, however, treats the different perspectives offered by the annotators on each task as noise that must be tamped down. An alternative perspective, rather, acknowledges that a ground truth label may not exist [6]. Instead, the annotator disagreement could in part reflect the different viewpoints of the annotators. These viewpoints may stem from differences in how annotators interpret the task, their own opinions on the correct label, or their understanding and perception of the data sample under examination. Even when annotators have a high level of agreement during one labeling process, there is no guarantee that future annotators would interpret the labeling question in exactly the same way or with the same level of agreement. Therefore, a consistent "ground truth" may only be stable over short time periods.

An annotator's disagreement relative to the average annotator rating–when it persists across data samples–could be viewed as a *statistical* bias. In scenarios where there is no clear ground truth label, however, denoting such patterns as bias is problematic. Thus, we abstain from referring to the observations we characterize in this work as "annotator biases" and instead call them *annotator sensitivities*. Annotator sensitivities refer to an annotator's higher likelihood of assigning a particular label relative to the rest of the annotator pool (we expound on this in the hate speech context shortly). Annotator sensitivities become ethically and socially concerning when they reinforce algorithmic oppression against minoritized or marginalized populations (currently and historically) [47]. For example, past work has identified correspondences between an annotator racial or gender identity and their rating patterns–or annotator *identity* sensitivities–particularly when racial or gender identity is prevalent in the task [3, 24, 26, 38, 60]. Since annotator identity sensitivity can play a role in the training of machine learning algorithms, identifying and quantifying it in labeled datasets is critical for better understanding the sources of algorithmic bias and developing mitigation strategies.

The aforementioned approaches can be viewed within the general field of measurement theory, which aims to obtain estimates of latent attributes. Approaches which rely on notions of annotator agreement can be cast within classical test theory (CTT) [48, 64]. However, CTT has been superseded by item response theory (IRT), which offers a more robust modeling formalism [19, 28, 55]. IRT models can be constructed that incorporate the diverse sources influencing a label, such as the sample itself, the annotator, the task at hand, etc. The contributions of these facets to the labels are captured via the parameters of a latent-variable probabilistic model. Thus, IRT offers a novel set of tools to examine annotated datasets in machine learning [19].

In this work, we propose leveraging techniques from IRT to assess annotator sensitivity in labeled datasets. As a case study, we examine annotator sensitivity within hate speech measurement, since it is an important human rights problem studied in computational social science, with well-documented cases of annotator disagreement and algorithmic bias in trained machine learning models [14, 44, 60]. As discussed above, we abstain from using the term "annotator bias", which can be ambiguous, with statistical or social interpretations, and suggests a ground truth that does not exist in this setting. We instead refer to the patterns we aim to quantify as annotator sensitivities. In this context, higher annotator sensitivity would imply that an annotator may be more likely to rate a comment as exhibiting aspects of hate speech relative to the remaining annotator population. We specifically focus on annotator identity sensitivities, in which annotator identity may correspond with differences in sensitivities: e.g., annotators of different racial identities may exhibit different sensitivities when labeling hate speech comments. We acknowledge that the term "sensitivity" brings its own set of norms and assumptions, with the possible framing of "overly sensitive annotators". We see annotator sensitivity, when it serves uplift minoritized communities in the face of algorithm oppression and hate speech, as a *desirable quality* in annotator labeling [4]. On the other hand, not all annotator sensitivity is warranted. For example, we reject annotator sensitivity favorable toward positions that reinforce white supremacy and white fragility, and aim for our methods to support annotator interventions in these cases.

We utilized a previously gathered hate speech corpus containing hate speech annotations to 50,000 social media comments by 10,000 annotators [32]. Importantly, the corpus contains demographic information about the annotators and tracks granular identity groups of each comment's target(s). Thus, we can suitably test the hypothesis that there exists a relationship between an annotator's identity and their sensitivity toward rating speech targeting a particular identity group. Using three different approaches rooted in IRT, we demonstrate that annotators are typically more sensitive to comments that target their own identity groups. For example, an annotator self-identifying as a particular racial identity is more likely to rate a comment targeting that racial identity as exhibiting features of hate speech. We further demonstrate that similar relationships hold for other identity groups, emphasizing how multiple identity groups can correspond to annotator sensitivity. Together, these results highlight the importance of considering annotator identity in downstream machine learning.

This paper is structured as follows. First, we discuss related work within annotation for machine learning, IRT, and IRT for machine learning (Section 2). Next, we briefly discuss the hate speech corpus we utilize, provide an introduction to IRT, and detail the three approaches we use to quantify annotator sensitivity (Section 3). We then present our results utilizing these three approaches (Section 4). We conclude with a discussion on limitations and future work (Section 5).

2 RELATED WORK

As noted by Geiger et al. [23], many machine learning datasets lack information on annotator demographics, annotator independence,

reliability metrics, and compensation details. Some work has investigated the demographic makeup of virtual workers on Amazon Mechanical Turk, finding that they skew toward white and male [17, 56]. Additional work has investigated the relationship between the demographic characteristics of annotators and the labels they generate. In one of the earliest works on the role of annotators' demographics in hate speech classification, Waseem [66] found that CrowdFlower annotators were more likely to label an observation as hate speech relative to annotators who were feminist and anti-racism activists. This paper builds off of prior work which included the predicted gender of the user into the classification model [67]. Garten et al. [22] developed embeddings that represented the demographic information of the speaker and include this information as features in the model, resulting in improved accuracy.

In an examination of annotator-generated text, Geva et al. [24] found that annotators introduced particular biases into classifiers that could not be attributed to the difficulty of the text that they labeled and, furthermore, the authors warned of the dangers of having a few workers account for the majority of the data. Similarly, Al Kuwatly et al. [3] examined the impact of classifiers trained on data labeled by annotators from different demographic groups, namely gender, native language, age group, and education. While they do not find any differences in the male vs. female classifiers, they do find that native language, age groups, and education levels impact the perception of hate speech. Larimore et al. [38] find that the racial identity of the annotator influences their perception of tweets that are about Black people, with the largest difference between white and Black annotators focusing on the latent topic modeling anti-racist politics. Gold et al. [25] show that for political and personal content targeting women, female labelers are more likely than men to rate it as hate speech and also to classify it as more virulent than men. Sap et al. [61] found similar correspondences between annotator identity and ratings on toxic speech, and connected this to an annotator's held beliefs. Our work follows the legacy of critical social scientists that evaluate the values present in machine learning datasets [8, 16, 33, 62]. In this work, we apply a computational social science lens to a dataset that provides annotator information to critically assess annotator identity sensitivity within hate speech measurement.

IRT has been used in limited capacity within the machine learning literature, with most work highlighting its usage in secondary analyses on trained classifiers [42, 53, 71], or to improve model interpretability [34]. Assessing annotator perspective in the IRT literature is well developed [19, 20, 45, 46, 58, 70], with techniques often applied in educational assessment, such as writing [5, 35] or speaking examinations [35]. These settings are typically characterized by a small number of annotators. To our knowledge, IRT techniques have not previously been leveraged to assess annotator sensitivity in larger-scale machine learning datasets.

3 METHODS

The code used to conduct the analyses and create the figures shown in this paper is publicly available on Github [59].

3.1 Hate speech corpus

We utilized a hate speech corpus created by Kennedy et al. [32] consisting of 50,000 social media comments annotated by 10,000 annotators. We provide brief details on the dataset, and refer the reader to the original paper for further information. In contrast to the majority of prior hate speech corpora, this corpus consists of annotations for several different *survey items* (labels) per comment. These items reflect a theoretical construct of hate speech, which captures degrees of "hatefulness" on a continuous spectrum rather than a yes/no dichotomy. For example, a comment espousing genocide, according to the construct, is inherently more hateful than a comment that expresses negative bias toward an identity group. A binary classification of hate speech would not indicate differences in the degree of hatefulness–only that, according to the annotator, the comment satisfies the criteria of hate speech [14, 15, 30, 51]. Thus, additional annotations included in this dataset provide greater flexibility in interrogating the inherent hatefulness of the comments.

Annotators provided ratings on a five-level Likert-style scales for ten different survey items, capturing the following aspects of hate speech: sentiment, respect, insult, humiliation, dehumanization, violence, genocide, attacking/defending, status, and a ternary hate speech classification. In each case, a higher rating on the Likert scale aligned with "more hatefulness". For example, on survey item "respect", a higher rating implies that the annotator feels the comment expresses a greater degree of disrespect (with disrespect being aligned with more hatefulness). Ratings were collapsed for specific survey items in cases where the granularity of the rating scale did not match the empirical distribution of ratings. See Appendix A for details on the hate speech construct and survey items.

The dataset additionally includes information on the target of each social media comment. Specifically, annotators were asked to provide labels on various identity groups targeted by the comments. These identity groups spanned race, religion, national origin, gender identity, sexual orientation, age, physical disability, and political ideology. Importantly, annotators were allowed to select multiple target identities for each comment. Annotators were also asked to voluntarily provide their demographic information, including their racial identity, age, gender identity, educational level, income, sexual orientation, religious affiliation, and their political ideology. Thus, this dataset captures two key identity variables interacting in the hate speech annotation task: the identity(ies) targeted by the comment and the annotator identities.

3.2 Item response theory provides a framework for measurement

Annotation tasks can be cast within the general field of measurement theory, the goal of which is to measure a latent attribute of a particular unit (here, a social media comment). Classical test theory posits that observed scores via annotation or testing reflect a true underlying score, with additional measurement error [48, 64]. However, it makes no assumptions about the sources of the measurement error, such as the difficulty of the task or annotator influence. Item response theory, meanwhile, provides a framework to assess these contributions to the observed score via a latent-variable probabilistic model that explicitly captures separate contributions to the ratings in its parameters [19, 28, 55]. It simultaneously places the

fitted parameters on a common, continuous scale that represents the task at hand. In this case study, IRT allows for the construction of a scale that reflects different degrees of underlying "hatefulness" [32].

Developing a measurement scale for a problem requires the conceptualization of a construct that represents the underlying scale [68]. Then, an IRT model of this construct is developed using the responses to *survey items* (labels) that interrogate each comment along the construct. For example, Kennedy et al. [32] proposed a hate speech construct that encompasses supportive, counterspeech, neutral, biased, hostile, dehumanizing, violent, and genocidal speech. The ten survey items mentioned in the prior section, then, were specifically designed to test this construct via multiple annotators' responses to the survey items. Thus, the IRT model aims to synthesize a score capturing the hatefulness of comments given the construct, which is measured via annotator responses to the survey items.

In this work, we specifically use the many-facet Rasch model as our IRT model of choice [19, 39, 41, 45]. Our model captures the interaction of three "facets": (i) the comments, each with their own degree of "hatefulness", (ii) the annotators, who may exhibit their own internal thresholds in labeling comments as hate speech, and (iii) the survey items, which provide information on different regions of the "hatefulness" spectrum. The model measures each facet via parameters in a probabilistic model. We fit all IRT models using the Facets software package [40].

The IRT model captures the decision of opting for rating k (say, "strongly agree") versus rating $k-1$ ("agree"). Specifically, let p_{nijk} be the probability that rater j assigns comment n a rating k on survey item i. Similarly define $p_{nij(k-1)}$, but for rating $k-1$. Then, the model can be written as:

$$\log\left[\frac{p_{nijk}}{p_{nij(k-1)}}\right] = \theta_n - \delta_i - \alpha_j - \tau_k. \tag{1}$$

In equation (1), each term is in logit units. By exponentiating both sides, we can view the left hand side as an odds ratio (OR) denoting the likelihood that an annotator "bumps up" their rating:

$$\mathrm{OR}_{nijk} = \exp(\theta_n - \delta_i - \alpha_j - \tau_k). \tag{2}$$

We reiterate that all survey items are aligned in their numerical code ordering. Thus, "bumping up" a rating *always* corresponds to a higher degree of hatefulness. A larger odds ratio implies that the annotator is more likely to rate a comment as possessing some aspect of hate speech. Intuitively, the odds ratio should depend on the facets at hand:

- θ_n, or the **hate speech score** of comment n. Higher values of θ_n indicate a more inherently hateful comment. This corresponds to a larger odds ratio, since a more hateful comment should elicit higher survey item ratings from the annotator.
- δ_i, or the **difficulty** of survey item i. The difficulty sets a scale for the hate speech score. Specific survey items, such as "genocide" inherently capture higher degrees of "hatefulness", and would thus possess a larger difficulty. Thus, it is more "difficult" for a comment to exhibit aspects of genocide in the sense that themes of genocide are more hateful on the underlying hate speech scale.
- α_j, or the **severity** of rater j. Higher values of severity will reduce the odds ratio, indicating that the annotator is less likely to provide higher ratings of hate speech aspects. If we negate the severity, we can instead interpret it as a term capturing "sensitivity." Without estimating such a term, existing annotation approaches assume equal severity across annotators.
- τ_k is the **difficulty** of response k, or an indicator of the rarity of response k relative to $k-1$. This term allows the distances between each response option to vary by item, rather than, for example, "strongly agree" being at the same location on the scale for every item.

This model captures multiple factors that may influence the rating on a sample in a hate speech corpus: the inherent hatefulness of the comment, the sensitivity of the annotator, and the task at hand (e.g., the labeling). The model separates the content of the comment from any annotator-level modulation, allowing the examination of each facet separately.

3.3 Approaches to assess sensitivity in an IRT model

A strength of the IRT model is its ability to capture an annotator's sensitivity via the severity parameter α_j. This sensitivity is calculated in the context of the hate speech scale, directly quantifying an annotator's tendency to rate comments as possessing elements of hate speech that cannot be explained by the comment or task. However, the severity does not indicate how such rating tendencies may vary across particular subgroups. Here, we are interested in the intersection between annotator identity and target identity. Thus, we must use new methods that interrogate the modulation of the severity parameter as a function of identity subgroups.

We use a suite of techniques called *differential rater functioning* (DRF), which seek to determine whether ratings are invariant over annotators, across particular subgroups [19, 46, 58]. Here, we describe three DRF approaches: separated measurements, annotator-level interactions, and group-level interactions. These approaches are complementary in that they assess annotator sensitivity from different perspectives.

3.3.1 Separated Measurements. Separated measurements tests for DRF via examination of multiple IRT models, each corresponding to a different subgroup [20, 70]. Specifically, we partition the corpus $\mathcal{D}$ into separate datasets $\{\mathcal{D}_p\}_{p=1}^{P}$, each corresponding to a particular subgroup p. Then, we fit an IRT model to each sub-corpus $\mathcal{D}_p$. Difficulty parameters are held constant across fits to ensure that all models are fit to a common scale. Differences in annotator behavior can be assessed, then, by examining how the severity parameters vary across the p fits.

As a direct example, suppose we aim to identify whether annotators exhibit preference toward comments targeting one of two identity groups. We obtain two sub-corpora, $\mathcal{D}_1$ and $\mathcal{D}_2$, which correspond to samples targeting identity groups 1 and 2, respectively. We fit an IRT model to each, obtaining severity estimates $\alpha_{j,1}$ and $\alpha_{j,2}$ for annotator j. We quantify the annotator's sensitivity toward one identity group as the *annotator lean*:

$$\Delta\alpha_j = \alpha_{j,1} - \alpha_{j,2} \tag{3}$$

The sign of the annotator lean indicates preference toward one identity group or the other. Specifically, for annotator j, if

$\Delta\alpha_j < 0$: more sensitive to comments targeting group 1

$\Delta\alpha_j > 0$: more sensitive to comments targeting group 2.

Furthermore, the annotator lean $\Delta\alpha_j$ can be interpreted as a change in the odds ratio. Specifically, for annotator j, we can estimate the change in the odds ratio when the target identity changes from group 1 to group 2:

$$\mathrm{OR}_{j,2} = \mathrm{OR}_{j,1} \cdot \exp(\Delta\alpha_j) \tag{4}$$

In this way, the annotator lean quantifies how much likelier an annotator is to rate a comment as exhibiting elements of hate speech when the target is in identity group 2, relative to identity group 1.

3.3.2 Annotator and Target-Identity Interactions. A complementary approach to assessing annotator sensitivity involves extending the IRT model to incorporate interaction terms [5, 18, 35, 69]. Specifically, we can include an interaction term β_{jm} in equation (1) between the target identity (m) and each unique annotator (j):

$$\log\left[\frac{p_{nijkm}}{p_{nij(k-1)m}}\right] = \theta_n - \delta_i - \alpha_j - \tau_k + \beta_{jm} \tag{5}$$

Thus, the interaction term can be viewed as an adjustment to each annotator's severity, conditional on the target identity of the comment. The sign of β_{jm} indicates whether an annotator is more or less sensitive to target identity group m relative to the baseline. This approach is distinct from separated measurements in that we obtain a separate term quantifying annotator sensitivity for *each* target identity group. A t-test can be performed to assess whether the labeling patterns by annotator j on group m warrant a non-zero β_{jm} term.

3.3.3 Group-level Interactions. The previous approaches allow for the assessment of annotator sensitivity on an annotator-by-annotator basis. Group-level behavior can be assessed, meanwhile, via secondary analyses on the differential effects. However, these approaches are limited in their ability to compare multiple identity groups interacting with each other. To perform multiple group-level comparisons, we to extend equation (1) to incorporate a group-level interaction term between all annotator identity groups, and all comments targeting particular identity groups. Specifically, we include two additional facets to the model:

$$\log\left[\frac{p_{nijkmn}}{p_{nij(k-1)mn}}\right] = \theta_n - \delta_i - \alpha_j - \tau_k + \mu_{lm} \tag{6}$$

where l denotes the annotator's identity, and m denotes target's identity. The term μ_{lm} can be viewed as an interaction between two additional facets capturing both annotator and target identities, without modeling the individual facets. In practice, we can view $\exp(\mu_{lm})$ as an adjustment to the odds ratio that depends on both annotator identity and target identity. Specifically, if $\exp(\mu_{lm}) > 1$, annotators of identity group m are more likely to rate comments targeting group n as exhibiting features of hate speech, and vice versa for $\exp(\mu_{lm}) < 1$. Of particular interest is assessing self-interactions: when annotators rate comments targeting their own identity group, i.e. when $l = m$. Lastly, the group-level interaction term can be assessed via a chi-squared test to determine if its inclusion improves the goodness-of-fit.

4 RESULTS

We sought to determine whether we could leverage techniques from item response theory to assess annotator identity sensitivity. To this end, we evaluated a hate speech corpus containing 50,000 social media comments that target different identity groups, labeled by annotators whose identities are also included in the corpus (Section 3.1). We specifically examined whether there existed a relationship between annotator identity and the target identity of a hate speech comment. We initially present results on Black- and white-targeting comments, since these two racial identities were the most represented in the dataset. We include additional experiments on other identity groups in Appendix B. However, our choice of these two racial identity groups should not be interpreted as placing them as equally situated targets of hate speech: indeed, hate speech has historically been levied largely against non-white groups. We first consider a simple empirical analysis on example comments to fully motivate the problem. Then, we show results from three IRT-based approaches discussed in the Methods to assess annotator sensitivity in various contexts.

4.1 Observable differences on hate speech survey items align with annotator identity

To motivate further analyses, we first conducted an empirical analysis examining whether annotators exhibited differences in their rating behaviors as a function of their own identity. We identified two comments that respectively targeted on the basis of race (specifically, a Black person) and gender (specifically, women) (Fig. 1: titles) that were annotated by more than 50 annotators (i.e. "reference set" comments from [32]). For each comment, we identified the largest two relevant annotator identity groups (Fig. 1a: Black and white annotators; Fig. 1b: female and male annotators). We examined how each annotator group, on average, rated the respective comments on several example survey items. For example, in Figure 1a, Black and white annotators exhibited significantly different behavior in rating whether the comment expressed (dis)respect, insult, humiliation, and hate speech. Black annotators generally found the comment more disrespectful, insulting, humiliating, and hateful. Meanwhile, annotators generally agreed on the level of violence expressed by the comment, which we might expect given its content. As for Figure 1b, female annotators found the comment significantly more disrespectful, insulting, and humiliating. However, female and male annotators generally agreed that the comment did not express violence and did not satisfy the criteria of the binary hate speech item.

These examples demonstrate that annotator identity can correspond to different labeling patterns on comments, whether the annotator characterized the comment as hate speech or not. Importantly, the additional survey items–which span a spectrum of hatefulness–revealed these differences, rather than a binary hate speech label. Since there is no clear "ground truth" rating for these comments, characterizing these differences as annotator *bias* is problematic. We instead frame them as differences in annotator identity sensitivity. That is, in Figure 1a, Black annotators tended to exhibit greater sensitivity to the comment in that they are more likely to characterize it as containing features of hate speech.

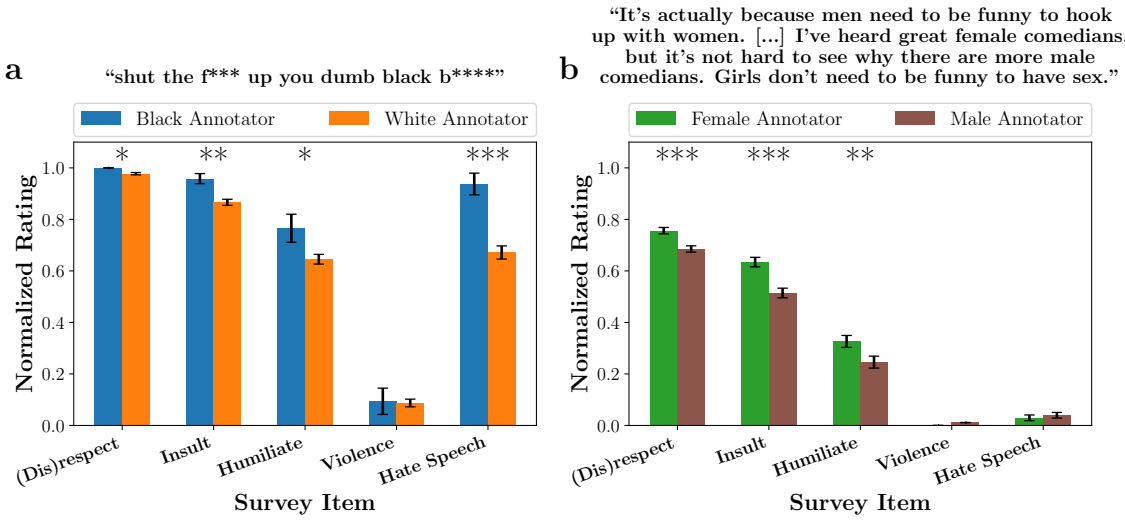

Figure 1: Observable differences on hate speech survey items align with annotator identity. The distribution of ratings on two example comments, for separate annotator groups, across several survey items. Plot titles denote the specific social media comments under consideration. Survey items (a subset of the 10 labeled items) are denoted on the x-axis. The y-axis denotes the average rating normalized to the maximum value, since not every survey item was measured on the same Likert scale. Rating scales are oriented such that a larger rating always corresponds to a higher degree of "hatefulness". a. The distribution of ratings by Black and white annotators on an example comment that targets Black identity. b. The distribution of ratings by female and male annotators on a comment that targets women. Ratings by non-binary annotators are not included due to their relative rarity, limiting statistical comparison. Error bars denote 95% confidence intervals. Asterisks denote a statistically significant difference in ratings between the two annotator groups ($*$: $p < 10^{-1}$, $**$: $p < 10^{-2}$, $***$: $p < 10^{-3}$; Mann-Whitney U Test).

We additionally considered annotator agreement metrics, such as Krippendorff's alpha [36], which have been used ascertain differences in annotator subgroup rating patterns [38]. We found that Black and white annotators exhibited comparable alpha values across survey items, which in turn were similar to the agreement between all annotators (Appendix B: Table S3). Furthermore, both groups of annotators exhibited low to moderate agreement, demonstrating the limitations of using classical test theory to assess systematic differences in annotator behavior.

Conducting these empirical analyses requires each comment receive a sufficiently large number of ratings by unique annotators in each identity group. However, in many hate speech corpora, only a few annotators (e.g., 2-5) typically rate each comment. Furthermore, the trends we observe in Figure 1 may not manifest as clearly in every comment. Additional methods are needed to assess whether annotator sensitivity varies throughout the entire corpus. Thus, we turn to IRT-based approaches, which provide a suitable alternative via the parameters of a probabilistic model.

4.2 Annotator identity often corresponds to their lean toward Black- or white- targeting comments

We assessed whether there existed a correspondence between annotator sensitivity and the target identity of each comment by performing separated measurements, specifically in the context of racial identity (Section 3.3.1). We extracted two sub-corpora: one with Black-targeting comments (19,686 total annotations) and the other with white-targeting comments (7,333 total annotations). We performed separate IRT fits to each dataset, obtaining severity estimates $\alpha_{j,b}$ and $\alpha_{j,w}$ for each annotator j. We calculated the *annotator lean* $\Delta\alpha_j = \alpha_{j,w} - \alpha_{j,b}$ as the difference in their severity estimates for each sub-corpora (see Section 3.3.1). An annotator whose lean is positive, i.e. $\Delta\alpha_j > 0$, is more sensitive toward Black-targeting speech: they are more likely to rate Black-targeting comments as exhibiting hateful features, relative to white-targeting speech. Meanwhile, a negative annotator lean $\Delta\alpha_j < 0$ implies the opposite: the annotator is more sensitive toward white-targeting speech, relative to Black-targeting speech.

We identified 4,276 annotators who annotated both Black- and white-targeting comments and calculated their annotator leans. Then, we examined the distribution of the leans according to various subgroups of the annotators. For example, we compared the distribution of annotator leans for Black annotators to those of white annotators (Fig. 2a: left-most boxplots). We found a wide distribution of annotator leans for both Black and white annotators, indicating that annotators in both groups exhibited sensitivity to both Black- and white-targeting comments. However, at the population level, we observed a significant difference in the median annotator lean (Kruskal test: $p = 1.3 \times 10^{-4}$). Specifically, Black annotators in general exhibited annotator leans in the positive direction (median = 0.35), implying they were typically more sensitive

toward Black-targeting speech. Meanwhile, white annotators generally exhibited annotator leans in the negative direction, though with a smaller effect size (median = −0.03).

The annotator lean can be more easily interpreted as a change in the odds ratio. Specifically, the term $\exp(\Delta\alpha_j)$ denotes the change in the odds ratio, equation (2), when a comment targets a Black person rather than a white person (see Section 3.3.1). If this quantity is greater than 1, an annotator is more likely to rate a comment as exhibiting elements of hate speech if it targets a Black person, relative to one targeting a white person (holding the other facets constant). We calculated each annotator's change in odds ratio, finding that Black annotators were 1.42 times more likely to rate Black-targeting speech as exhibiting hateful content, while white annotators were 0.97 times as likely (Fig. 2b).

We examined the two aforementioned quantities–the annotator lean and the change in the odds ratio–for various annotator identity groups. Specifically, we asked whether there existed a correspondence between annotator lean and a host of non-racial annotator identities, including political ideology, religion, gender, sexual orientation, income, and education level (see Section 3.1). We found annotator lean significantly differed as a function of political ideology ($p = 1.3 \times 10^{-15}$), religion ($p = 2.4 \times 10^{-4}$), gender ($p = 4.7 \times 10^{-3}$), and sexual orientation ($p = 5.9 \times 10^{-3}$), but not across income or education levels (Fig. 2a). Identity subgroups exhibiting sensitivity toward Black-targeting speech included liberals (relative to conservatives), non-religious annotators (relative to Christians), women (relative to men), and queer annotators (relative to straight annotators) (Fig. 2b).

The multitude of identity groups corresponding to annotator lean raises the question of whether they could be explained by correlations with Black identity. That is, if Black annotators are more likely to be liberal, or non-religious, etc., within the scope of the dataset, then the observations in Figure 2 could be a byproduct of these intersectional correlations. We examined the correlations between Black identity and other identity subgroups, including annotators identifying as liberal, non-religious, women, and queer, finding only weak correlations (Spearman correlations below 0.1: Appendix B, Fig. S3). Together, these results demonstrate that multiple annotator identities correspond to sensitivity toward Black-targeting comments relative to white-targeting comments. In particular, Black and white annotators are each more lenient toward comments targeting their own identity group, with the largest effect sizes.

We conducted separated measurements analyses on various identity groups, including gender identity and political ideology (Appendix B: Fig. S1, Fig. S2). In each case, we found that various annotator identities corresponded toward increased sensitivity to different target identities, demonstrating that the above observations hold beyond racial identity.

4.3 Annotator-level interaction terms corroborate annotator leans, while identifying annotators with high sensitivity

We next sought to explicitly test whether annotators exhibited sensitivity toward Black- and white-targeting comments by incorporating annotator-level interaction terms in the IRT model. These interaction terms can be viewed as an adjustment to each annotator's severity that is dependent on the target identity of the comment (see Section 3.3.2). The sign of the term indicates whether the annotator is more or less sensitive to comments targeting a particular identity group, and the magnitude of the term quantifies the degree of their sensitivity. Furthermore, a t-test can be performed to assess whether the identity group differences are statistically significant, allowing a more conclusive evaluation of whether an annotator may be exhibiting a particular lean toward an identity group.

We again assessed whether annotators exhibited sensitivity toward Black- or white- targeting comments, now by examining annotator-level interaction terms. We extracted a sub-corpus containing 27,019 annotations on Black- and white-targeting comments. Then, we fit equation (5) to the data, obtaining single severity estimates α_j for each annotator j, along with interaction terms that indicate target identity dependent adjustments. We specifically focused on the 4,276 annotators that identified as either Black or white, and rated both Black- and white-targeting comments.

For simplicity, we examined the sign of the interaction term to determine whether the annotators were more or less sensitive to Black- and white-targeting comments. Specifically, we calculated the percentages of Black annotators who exhibited more sensitivity, less sensitivity, or no lean toward Black-targeting comments (Fig. 2a, blue bars). We calculated similar percentages of white annotators on Black-targeting comments (Fig. 2b). We found that a sizeable fraction of annotators exhibited no lean on Black targeting comments (Fig. 3a: middle columns). However, more Black annotators than white annotators had greater sensitivity in rating Black-targeting comments (Fig. 3a: left columns). We performed a similar comparison on white-targeting speech, finding the opposite: more white annotators than Black annotators exhibited greater sensitivity toward white-targeting comments (Fig. 3a: left columns). Together, these results demonstrate than Black and white annotators tended to be more sensitive toward comments targeting their own identity group, corroborating the findings from the separated measurements.

Each of the interaction terms can be evaluated with a t-test to determine whether their value differed significantly from zero. We found that a sizeable fraction of annotators exhibited no particular sensitivity toward Black- or white-targeting speech, suggesting that their p-values are close to 1. We examined the distribution of p-values across annotators for both interaction terms (Fig. 3c). A majority of annotators exhibited large p-values, suggesting that their interaction terms were not significantly different from zero.

In order to determine whether the above observations may be a byproduct of multiple comparisons, we examined the annotators whose p-values implied a significant lean ($p < 0.10$; Fig. 3c: red dashed line). Within this smaller pool, we similarly examined the fraction of Black and white annotators who exhibited more or less sensitivity toward Black-targeting speech (Fig. 2d). We found that white annotators within this pool were considerably less sensitive to Black-targeting speech. Furthermore, when examining white-targeting speech, we found that Black annotators were considerably less sensitive than white annotators (Fig. 3e). Together, these findings demonstrate that annotator-level interaction terms corroborate

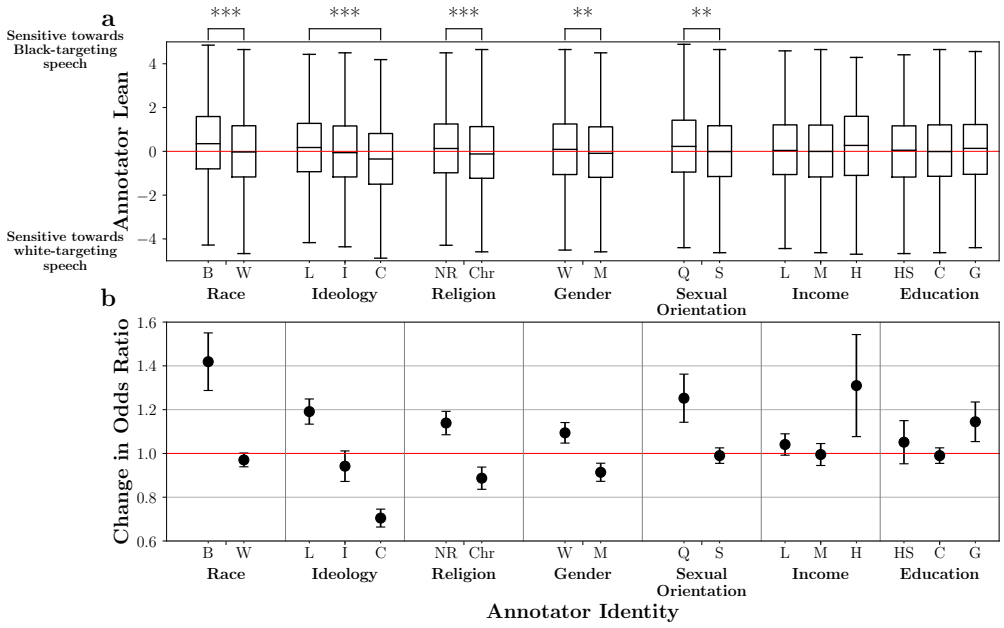

Figure 2: Annotator identity often corresponds to their lean toward Black- or white- targeting comments. The x-axis captures several different identity groups, including (i) Race: Black (B) vs. white (W) annotators; (ii) Ideology: Liberal (L) vs. Independent (I) vs. Conservative (C) annotators; (iii) Religion: Atheist and non-religious annotators (NR) vs. Christians (C); (iv) Gender: Women (W) vs. Men (M); (v) Sexual Orientation: Queer (Q) vs. Straight (S) annotators; (vi) Income: Low-income (L), Middle-income (M), and High-income (H); and (vii) Education: annotators with High School (HS), College (C), or Graduate (G) education. Identity groups are sorted by increasing p-value. a. The distribution of annotator leans, equal to the difference in severities obtained from sub-corpora of Black-targeting comments and white-targeting comments. A positive annotator lean implies that the annotator is more likely to rate Black-targeting speech as exhibiting elements of hate speech. Significance markers denote Kruskal test of medians ($***$: $p < 10^{-3}$; $**$: $p < 10^{-2}$). b. The change in odds ratio of rating a comment that targets Black-identity vs. white-identity. Higher values indicate a higher likelihood of rating Black targeting speech as hateful. Error bars denote 95% confidence intervals on the median odds ratio.

annotator leans, and identify a small subset of annotators with high sensitivity.

When we assessed annotator sensitivity with the annotator lean, we used a difference of severity estimates $\Delta\alpha_j$ to quantify an annotator's sensitivity toward Black- or white-targeting comments. We devised a similar metric using the interaction terms. Specifically, we define the *interaction lean* as the difference in interaction terms for Black- and white-targeting speech, i.e. $\Delta\beta_j = \beta_{jb} - \beta_{jw}$. Similar to the annotator lean, the interaction lean quantifies the change in the odds ratio when annotator j rates Black-targeting speech, compared to white-targeting speech. We should expect the annotator lean and interaction lean to be closely related to each other. We compared the two terms across annotators for Black- and white-targeting speech, finding that they are strongly correlated (Pearson correlation: $r = 0.77$; Appendix B: Fig. S5), demonstrating that distinct IRT approaches are consistent in their evaluation of annotator sensitivity.

4.4 Group-level interaction terms reveal sensitivity toward an annotator's own racial identity

Thus far, we have examined whether annotators exhibited differences in their labeling patterns for comments targeting members of only two racial identities. However, hate speech targets a variety of other racial identities including, but not limited to, Latinx, Asian, Middle Eastern, Native American, and Pacific Islander identities, as well as multiple sub-groups in other identity groups (e.g., gender identity, sexual orientation, etc.). Thus, it is necessary to develop methods capable of detecting differences in sensitivity across multiple identity groups. However, performing annotator-level examinations of these additional identity groups can prove difficult due to low sample sizes. Therefore, to examine differences in rater sensitivity across multiple annotator groups, we turn to a group-level analysis.

We modified the base IRT model by incorporating an interaction term μ_{lm} that captures both annotator racial identity and the target racial identity simultaneously (see Section 3.3.3). This interaction term captures behavior at the *group level*, in contrast to the term from the prior section, which modeled at the *annotator level*. It can

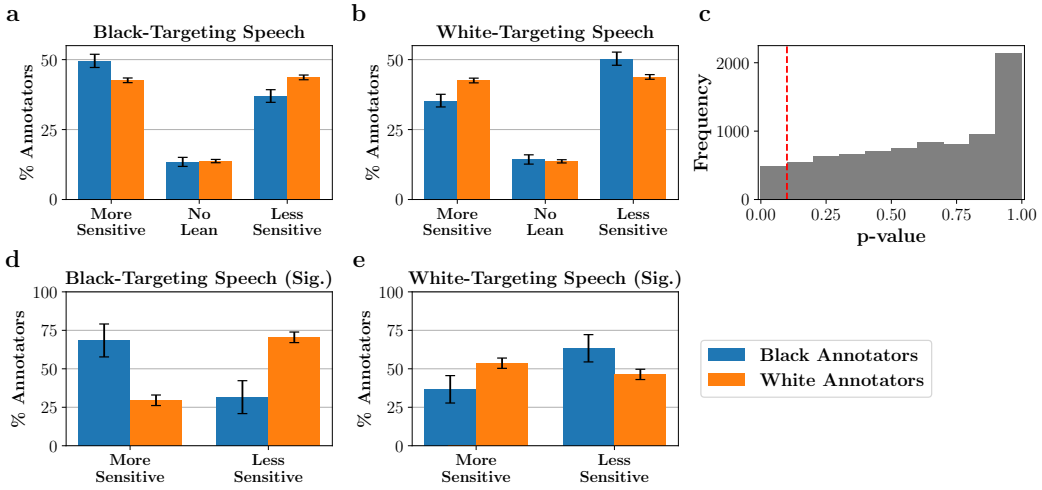

Figure 3: Annotator-level interaction terms corroborate annotator leans, while identifying annotators with high sensitivity. a-b. The fraction of Black annotators (blue bars) and white annotators (orange bars) who were more sensitive, less sensitive, or exhibited no lean toward Black-targeting speech (a) and white-targeting speech (b). Sensitivity was assessed via the sign of each annotator's interaction term. c. The distribution of p-values across all interaction terms. Smaller p-values denote that the annotator's interaction term was significant different than zero (i.e., provides evidence of sensitivity). Red dashed line denotes $p = 0.1$. d-e. Same as a-b, but for annotators whose interaction terms were significant ($p < 0.1$).

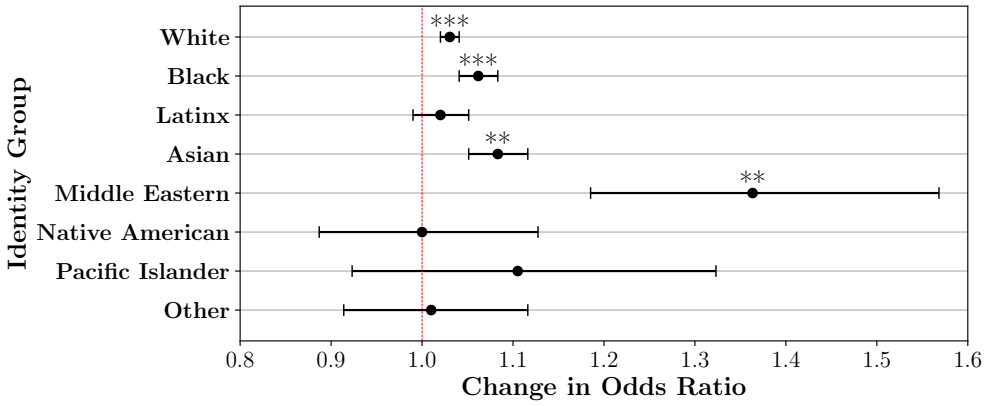

Figure 4: Group-level interaction terms reveal annotators' sensitivity in rating comments targeting their own racial identity. The change in odds ratio induced on a label when an annotator rates a comment targeting their own racial identity. An odds ratio greater than 1 indicates that annotators of an identity group are more likely to rate comments targeting their racial identity as exhibiting elements of hate speech, compared to a different racial identity. Error bars denote 95% confidence intervals. Asterisks denote significance ($***$: $p < 10^{-3}$; $**$: $p < 10^{-2}$). Red dashed line demarcates an odds ratio of one, with no change in probability of rating a comment (which served as the null hypothesis). Identity groups are listed in descending order of sample size.

be interpreted as an adjustment to the odds ratio when an annotator of identity group l rates a comment targeting identity group m (in contrast to the single annotator j). It quantifies group-level trends for a variety of identity groups, simultaneously.

We obtained 55,603 samples within the hate speech corpus that targeted on that basis of any racial identity, including white, Black, Latinx, Asian, Middle Eastern, Native American, Pacific Islander, or some Other identity group. Then, we fit an IRT model to this dataset that incorporated a group-level interaction term. The racial identity groups we used for the annotators matched those of the target identities. Thus, the interaction term was composed of $8 \times 8 = 64$ parameter estimates, each of which can be evaluated with a t-test. We focused in particular on the case where $l = m$, or when an annotator rated comments targeting their own racial identity. We calculated the change in the odds ratio suggested by these terms, i.e., $\{\exp(\mu_{lm})\}_{l=m}$, to quantify group-level sensitivities.

We found that, in most cases, annotator groups exhibited an odds ratio greater than 1 (Fig. 4). This implies that annotators generally demonstrate greater sensitivity toward comments that target their own racial identity. Several self-interaction terms were statistically significant (with the null hypothesis denoting an odds ratio equal to 1), including that of white annotators ($p = 1 \times 10^{-4}$), Black annotators ($p = 7 \times 10^{-4}$), Asian annotators ($p = 2.2 \times 10^{-2}$) and Middle Eastern annotators ($p = 2.1 \times 10^{-2}$). The remaining self-interaction terms were not statistically significant, which may be due in part to the limited number of samples available to these identity groups. Latinx annotators represented an outlier among the annotator groups in that there were an abundance of samples, but their rating behavior was not statistically significant, suggesting that Latinx annotators may exhibit diverse sensitivities in annotating comments targeting their own racial identity. Lastly, the inclusion of the group-level interaction term in the model was found to be highly significant (chi-squared goodness of fit: $p = 7.6 \times 10^{-8}$).

We highlighted the self-interaction terms in Figure 4. However, we can examine the entire set of interaction term coefficients to assess whether, at the group level, specific identity groups tend to rate comments targeting other racial identities differently. While we found some significant relationships–e.g., Black and Latinx annotators are slightly less sensitive to white-targeting comments (Appendix B: Fig. S4)–most interactions were not significant, with small effect sizes. This may be due in part to the smaller sample sizes for cross-interaction terms.

5 DISCUSSION

Identifying and quantifying the sources of algorithmic bias is necessary for mitigating their impacts in machine learning algorithms. Understanding an increasingly prevalent source of algorithmic bias–the "bias" inherited by the annotators' labels–is a critical part of this process, particularly in tasks with greater subjectivity, such as hate speech and disinformation classification. In this work, we we recast annotator bias as *annotator identity sensitivity*, and aimed to quantify it in an existing hate speech corpus. Specifically, we leveraged multiple techniques from item response theory to assess whether there existed a relationship between an annotator's identity and their labels on social media comments targeting various identity groups. We found that annotators tended to be more sensitive when rating comments targeting groups that they identify with: that is, annotators were more likely to rate comments targeting a group they identify with as possessing elements of hate speech. Our results demonstrate that annotator perspective, shaped partially by their identity groups and lived experience, can influence dataset annotation, thereby having important implications for the development of downstream machine learning algorithms.

Our approaches relied on a preprocessing step in which we determined sub-corpora targeting various identity groups. As with the survey items, specification of the targeted identity group relied on labels provided by annotators. In general, annotators expressed higher agreement on this task relative to the hate speech items. However, annotators still expressed moderate disagreement (Krippendorff's alphas ranging from $0.60 - 0.75$: Table S4). Just as we found annotator sensitivity on the hate speech items varied with target identity, so may specification of targeted identity groups. Thus, future work should more closely examine the disagreement expressed by annotators on specifying targeted identity groups, and the degree to which it may correspond to annotator identity.

We discovered correspondences between the annotator identity and the target identity of the comment. However, the identity of the comment's author also plays an important role in shaping their rating of the comment. In particular, members of particular identity groups may use specific vernacular or terminology that may be misunderstood by annotators not within the identity group [60]. One example is reclaimed language, in which slurs or degrading phrases are used by the members of the targeted group in a positive or colloquial manner [10]. Annotators not identifying with the target identity group may, for example, misinterpret usage of reclaimed language as hate speech. Generally, determining the identity of the comment author is difficult since social media comments contain limited or no information about the author. Some past work has probabilistically inferred the racial identity of comment authors [9, 13], though there are both demonstrated biases [31] and ethical risks with race imputation [54]. More recent work has aimed to include content authors within the annotation process to determine race [52].

We evaluated annotator identity sensitivity in constrained settings, where we only considered a single racial identity at a time: e.g., Black or white annotators rating Black- or white-targeting speech. However, annotators are not defined solely by their racial identity [12]. Multiple identities–race, gender, sexuality, religion, political affiliation, and others–can intersect to influence how an annotator assigns labels in a corpus, with notable impacts on trained classifiers [11, 21, 33]. At the same time, social media comments can target multiple identity groups (e.g., Fig. 1a's comment targets a Black woman) and multiple targets within an identity group (e.g., a comment can target multiple racial identities). If a comment targets multiple identity groups, an annotator's rating may synthesize the treatment of all groups, or correspond to some subset. Bringing an intersectional lens to the analyses discussed in this work–for both the annotator identities and target identities–is necessary to more accurately characterize annotator sensitivity. This entails both reexamining the differential rater functioning quantities across intersecting identities as well as improving surveys to allow annotators to better characterize their reasoning for assigning ratings when multiple identities are at play. Measurement evaluation methods designed for complex, intersectional subgroups, such as Rasch trees [65] and latent class analysis [57], should be explored in future work.

Our approach relied heavily on an abundance of annotator information available in the hate speech corpus. However, including annotator information is not commonly done in social computing datasets [23]. In order to facilitate future analyses, datasets should include additional information beyond the data samples and labels, such as annotator demographics, compensation rates, recruitment materials and language, selection procedure on the crowdsourcing platform, training procedures, and quality assurance practices. These changes should be accompanied by continued improvements in the labeling instruments: the survey items and accompanying responses. Improved survey design (precise definitions within the

survey items, clearer constructs, etc.) could serve to improve annotator ratings by bringing clarity to how annotators should rate each data sample.

The methods we used in this work raise the possibility of initiating annotator interventions, which have been utilized in past work on hate speech labeling [60]. For example, in a operationalized setting where annotators continually generate ratings, situations where annotators exhibit significantly different labeling patterns across identity groups may warrant further qualitative examination to better understand why they occur. The second method we discussed–annotator interaction terms–is most amenable to finding such annotators. Thus, if an annotator is deemed to exhibit a significant lean toward an identity group, their labels can be studied to inform an intervention, if necessary. These interventions are a critical component of improving annotation in an equitable manner, particularly to avoid the scenario of bringing on annotators of a minoritized group to label comments targeting their own group, because of their increased sensitivity. Our takeaway from these results is that the majority group (e.g., in the the main results, white annotators) should undergo interventions as needed in order to achieve the sensitivities exhibited by the minoritized groups, since the majority group lacks the lived experience and perspective that the minoritized group brings to the labeling task. Such interventions could provide a more resilient and sustainable approach to improving content moderation.

Machine learning algorithms trained on labeled corpora are susceptible to inheriting identity-related sensitivities exhibited by the annotators. In this context, traditional machine learning algorithms trained on the survey item labels could learn the annotator identity sensitivities demonstrated in this work. Future work should assess the extent to which traditional machine learning algorithms inherit the annotator identity sensitivities that labeled this hate speech corpus. Approaches to quantify the sensitivities include using various fairness metrics previously used in the hate speech literature [44, 60], as well as training models on various sub-corpora of annotator identities [3] (an approach that is similar to separated measurements). If it is necessary to alleviate the impact of different annotator sensitivities, future work could utilize approaches developing novel model architectures that are more closely integrated with item response theory. For example, Kennedy et al. [32] utilize knowledge of the severity parameter in prediction, allowing the classifier to appropriately weight labels provided by specific annotators. They additionally train models to predict hate speech score (rather than, or in addition to, the survey items), which is disentangled from annotator influence. Future work could extend these approaches to incorporate knowledge of the annotator identity sensitivity, via the differential rater functioning metrics described in this work. In cases where a hate speech score is not available, reweighting schemes could be utilized to more heavily weight annotations by less-represented annotator identity groups.

While we focused on analyzing a hate speech corpus in this work, the techniques we invoked can be applied more generally. Specifically, quantifying annotator sensitivity is relevant for other constructs, such as toxicity (closely related to hate speech), disinformation, sentiment analysis, and others [1, 2, 27]. The usage of our approach in these contexts requires the development of new constructs, survey instruments to measure the constructs, and annotation of the survey items on data samples. Taking these steps, while intensive, will result in richer, more informative datasets.

ACKNOWLEDGMENTS

We thank colleagues from the D-Lab for their feedback and discussions.

Funding/Support: The authors declare no additional sources of funding.

REFERENCES

[1] Alim Al Ayub Ahmed, Ayman Aljabouh, Praveen Kumar Donepudi, and Myung Suh Choi. 2021. Detecting Fake News using Machine Learning: A Systematic Literature Review. *arXiv preprint arXiv:2102.04458* (2021).

[2] Qurat Tul Ain, Mubashir Ali, Amna Riaz, Amna Noureen, Muhammad Kamran, Babar Hayat, and A Rehman. 2017. Sentiment analysis using deep learning techniques: a review. *Int J Adv Comput Sci Appl* 8, 6 (2017), 424.

[3] Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*. 184–190.

[4] AHMER ARIF and OS KEYES. 2022. Vulnerability, Trust and AI. (2022).

[5] Aslıhan Erman Aslanoğlu and ŞATA Mehmet. 2021. Examining the Differential Rater Functioning in the Process of Assessing Writing Skills of Middle School 7th Grade Students. *Participatory Educational Research* 8, 4 (2021), 239–252.

[6] Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. Toward a Perspectivist Turn in Ground Truthing for Predictive Computing. *arXiv preprint arXiv:2109.04270* (2021).

[7] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[8] Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2021. The values encoded in machine learning research. *arXiv preprint arXiv:2106.15590* (2021).

[9] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[10] Robin Brontsema. 2004. A queer revolution: Reconceptualizing the debate over linguistic reclamation. *Colorado Research in Linguistics* (2004).

[11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[12] Kimberle Crenshaw. 1990. Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stan. L. Rev.* 43 (1990), 1241.

[13] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516* (2019).

[14] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.

[15] Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*. 86–95.

[16] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, and Hilary Nicole. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society* 8, 2 (2021), 20539517211035955.

[17] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and Dynamics of Mechanical Turk Workers *(WSDM '18)*. Association for Computing Machinery, New York, NY, USA, 135–143. https://doi.org/10.1145/3159652.3159661

[18] Thomas Eckes. 2005. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly: An International Journal* 2, 3 (2005), 197–221.

[19] George Engelhard and Stefanie A Wind. 2017. *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

[20] George Engelhard Jr, Stefanie A Wind, Jennifer L Kobrin, and Michael Chajewski. 2013. Differential Item and Person Functioning in Large-Scale Writing Assessments within the Context of the SAT®. Research Report 2013-6. *College Board* (2013).

[21] Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*. 94–104.

[22] Justin Garten, Brendan Kennedy, Joe Hoover, Kenji Sagae, and Morteza Dehghani. 2019. Incorporating demographic embeddings into language understanding.

Cognitive science 43, 1 (2019), e12701.

[23] R Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 325–336.

[24] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898* (2019).

[25] Michael Wojatzki Tobias Horsmann Darina Gold and Torsten Zesch. 2018. Do women perceive hate differently: Examining the relationship between hate speech, gender, and agreement judgments. (2018).

[26] Nitesh Goyal and et al. 2022. Is Your Toxicity My Toxicity? Exploring the Impact of Rater Identity on Toxicity Annotation. In *The 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing*. ACM.

[27] Isuru Gunasekara and Isar Nejadgholi. 2018. A review of standard text classification practices for multi-label toxicity identification of online content. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*. 21–25.

[28] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory*. Vol. 2. Sage.

[29] Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass* 15, 8 (2021), e12432.

[30] Alvi Md Ishmam and Sadia Sharmin. 2019. Hateful speech detection in public facebook pages for the bengali language. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, 555–560.

[31] Anna Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of studying and processing dialects in social media. In *Proceedings of the Workshop on Noisy User-generated Text*. Association for Computational Linguistics, Beijing, China, 9–18. https://doi.org/10.18653/v1/W15-4302

[32] Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277* (2020).

[33] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. *arXiv preprint arXiv:2005.05921* (2020).

[34] Adrienne Kline, Theresa Kline, Zahra Shakeri Hossein Abad, and Joon Lee. 2020. Using Item Response Theory for Explainable Machine Learning in Predicting Mortality in the Intensive Care Unit: Case-Based Approach. *Journal of Medical Internet Research* 22, 9 (2020), e20268.

[35] Kimi Kondo-Brown. 2002. A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing* 19, 1 (2002), 3–31.

[36] Klaus Krippendorff. 1980. Validity in content analysis. (1980).

[37] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[38] Savannah Larimore, Ian Kennedy, Breon Haskett, and Alina Arseniev-Koehler. 2021. Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*. 81–90.

[39] John M Linacre. 1994. *Many-Facet Rasch Measurement*. MESA press.

[40] John M. Linacre. 2015. Facets computer program for many-facet Rasch measurement.

[41] John M Linacre and Benjamin D Wright. 2002. Construction of measures from many-facet data. *Journal of Applied Measurement* (2002).

[42] Fernando Martínez-Plumed, Ricardo BC Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence* 271 (2019), 18–42.

[43] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[44] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.

[45] Carol M Myford and Edward W Wolfe. 2003. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement* 4, 4 (2003), 386–422.

[46] Carol M Myford and Edward W Wolfe. 2004. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of applied measurement* 5, 2 (2004), 189–227.

[47] Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

[48] Melvin R Novick. 1966. The axioms and principal results of classical test theory. *Journal of mathematical psychology* 3, 1 (1966), 1–18.

[49] Ellie Pavlick, Matt Post, Ann Irvine, Dmitry Kachaev, and Chris Callison-Burch. 2014. The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics* 2 (2014), 79–92.

[50] Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong?. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 507–511.

[51] Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2019. Annotating hate speech: Three schemes at comparison. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, Vol. 2481. CEUR-WS, 1–8.

[52] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1534–1545.

[53] Ricardo BC Prudêncio, José Hernández-Orallo, and Adolfo Martınez-Usó. 2015. Analysis of instance hardness in machine learning using item response theory. In *Second International Workshop on Learning over Multiple Contexts in ECML*.

[54] Megan Randall, Alena Stern, and Yipeng Su. 2021. Five Ethical Risks to Consider before Filling Missing Race and Ethnicity Data. (2021).

[55] George Rasch. 1968. A mathematical theory of objectivity and its consequences for model construction. In *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.

[56] Joel Ross, Andrew Zaldivar, Lilly Irani, and Bill Tomlinson. 2009. Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep* (2009), 49.

[57] Jürgen Rost. 1990. Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement* 14, 3 (1990), 271–282.

[58] Frank E Saal, Ronald G Downey, and Mary A Lahey. 1980. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin* 88, 2 (1980), 413.

[59] P. S. Sachdeva, R. Barreto, C. von Vacano, and C. J. Kennedy. 2022. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. https://github.com/dlab-projects/annotator_sensitivity_irt.

[60] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*. 1668–1678.

[61] Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. *arXiv preprint arXiv:2111.07997* (2021).

[62] Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do datasets have politics? Disciplinary values in computer vision dataset development. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–37.

[63] Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*. 254–263.

[64] Steven E Stemler and Adam Naples. 2021. Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research, and Evaluation* 26, 1 (2021), 11.

[65] Carolin Strobl, Julia Kopf, and Achim Zeileis. 2015. Rasch trees: A new method for detecting differential item functioning in the Rasch model. *Psychometrika* 80, 2 (2015), 289–316.

[66] Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*. 138–142.

[67] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.

[68] Mark Wilson. 2004. *Constructing measures: An item response modeling approach*. Routledge.

[69] Edward W Wolfe, Bradley C Moulder, and Carol M Myford. 1999. Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. (1999).

[70] Benjamin D Wright and Geofferey N Masters. 1982. *Rating scale analysis*. MESA press.

[71] Yan Xu, Quanlong Li, Huirong Dong, and Yuanlong Chen. 2020. A Learning Ability Evaluation Method Based On Item Response Theory and Machine Learning Method. In *Proceedings of the 2020 The 3rd International Conference on Big Data and Education*. 55–60.

A HATE SPEECH CONSTRUCT

We utilized a construct of hate speech developed by Kennedy et al. [32]. This construct reflects the authors' conceptualization of a hate speech spectrum, informed by qualitative assessment of a reference set of hate speech comments. We reproduce the construct by Kennedy et al. for convenience in Table S1. The construct contains seven levels, spanning the hate speech–counterspeech spectrum. Higher levels on the construct correspond to more hatefulness. Kennedy et al. conceptualized the highest level as condoning genocide toward a target identity. "Bias" constitutes the lower end of the construct (Level 1), which the authors conceptualized as having the lowest degree of hatefulness. The authors then denoted a Level 0 as "Neutral Speech". Comments exhibiting positive or supportive sentiment were also conceptualized along the same hate speech construct, but on the opposite end of the spectrum (in this case, negative levels).

In order to determine where on the hate speech spectrum a comment lies, survey items must be used to interrogate the different levels of the construct. Kennedy et al. devised a set of nine survey questions that aimed to assess the degree of hatefulness expressed by a comment. These survey items are reproduced in Table S2. The authors supplemented these items with an additional question that directly asked for a hate speech classification. Annotator responses to these survey items across a wide range of social media comments could then be passed into an item response theory model to directly quantify the hate speech spectrum implied by Table S1.

Level	Summary	Brief description	Example from social media
5	Genocide	Support for or intention of systematically killing all or a large number of a protected identity group	Every g**k in #LosAngeles should be deported or killed.
4	Violence	Threat or support of physical force or emotional abuse intended to hurt or kill members of a protected identity group	#whitepower...I'll hang you n****r
3	Dehumanization	Depriving a protected group of human-like qualities, such as comparison to an animal, insect, or disease	You're all a bunch of human cockroaches is what you are... You ch**ks...
2	Hostility	Unfriendliness or opposition to a protected identity group, such as through slurs, profanity, or insults	Learn the fucking language you fucking useless immigrant.
1	Bias	Inclination or preference against a protected identity group, including prejudice	@[NAME] They are arabs. Do you need any other explanations?
0	Neutral	Descriptive or other non-harmful references to identity groups	Go get a job at Dick's Sporting Goods and try to work at being a better person
-1	Counterspeech	Response to hate speech that seeks to undermine its impact and standing	No, the chances of a muslim shooting you in America is almost nil. There are over 50K gun deaths every year christian USA...
-2	Supportive	Respectful, prideful, or other solidarity-based messaging about a protected identity group(s)	I'm bi. And a good listener if you need a friend

Table S1: Construct for hate speech–counterspeech spectrum. Increasing levels correspond to a higher degree of hatefulness.

Item name	Prompt	Response options
Sentiment	How would you describe the sentiment of this comment?	Strongly negative, somewhat negative, neutral, somewhat positive, strongly positive
Respect	Is this comment respectful towards the group(s) you previously identified?	Strongly disrespectful, disrespectful, neutral, respectful, strongly respectful
Insult	Do you agree or disagree with the following statement: This comment is insulting towards the group(s) you previously identified.	Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree
Humiliate	Do you agree or disagree with the following statement: This comment is humiliating towards the group(s) you previously identified.	Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree
Status	This comment states that the group(s) you previously identified is:	Strongly inferior, inferior, neither superior nor inferior, superior, strongly superior
Dehumanize	Do you agree or disagree with the following statement: This comment dehumanizes the group(s) you previously identified (e.g., by comparing them to an animal).	Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree
Violence	Do you agree or disagree with the following statement: This comment calls for using violence against the group(s) you previously identified.	Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree
Genocide	Do you agree or disagree with the following statement: This comment calls for the deliberate killing of a large group of people from the group(s) you previously identified.	Strongly disagree, disagree, neither disagree nor agree, agree, strongly agree
Attack-Defend	Is the comment attacking or defending the group(s) you previously identified?	Strongly defending, defending, neither defending nor attacking, attacking, strongly attacking
Hate speech	Does this comment contain hate speech, defined as "bias-motivated, hostile and malicious language targeted at a person/group because of their actual or perceived innate characteristics, especially when the group is unnecessarily labeled?	Yes, no, unclear

Table S2: Survey items used to interrogate hate speech construct

B EXTENDED RESULTS

In this section, we include additional figures supporting the main text.

Annotator Group	Sentiment	Respect	Insult	Status	Dehumanization	Physical Violence	Genocide	Attack Defend	Hate Speech
Black Annotators	0.403	0.421	0.364	0.406	0.372	0.711	0.694	0.374	0.646
White Annotators	0.438	0.498	0.391	0.309	0.366	0.662	0.642	0.381	0.686
All Annotators	0.381	0.378	0.355	0.436	0.371	0.629	0.656	0.348	0.537

Table S3: Krippendorff's alpha calculated on each survey item (columns) for different subsets of annotators (rows): Black annotators, white annotators, and all annotators. These values were calculated on comments that target either Black or white identity.

Target Identity Group	Krippendorff's Alpha
Age	0.341
Disability	0.744
Gender	0.712
National Origin	0.570
Race	0.672
Religion	0.797
Sexual Orientation	0.718

Table S4: Krippendorff's alpha calculated on the annotator labels identifying the targets of each comment. Agreement values are calculated across all comments, indicating whether the degree to which annotators indicated similar binary assignment of target identity groups to each comment.

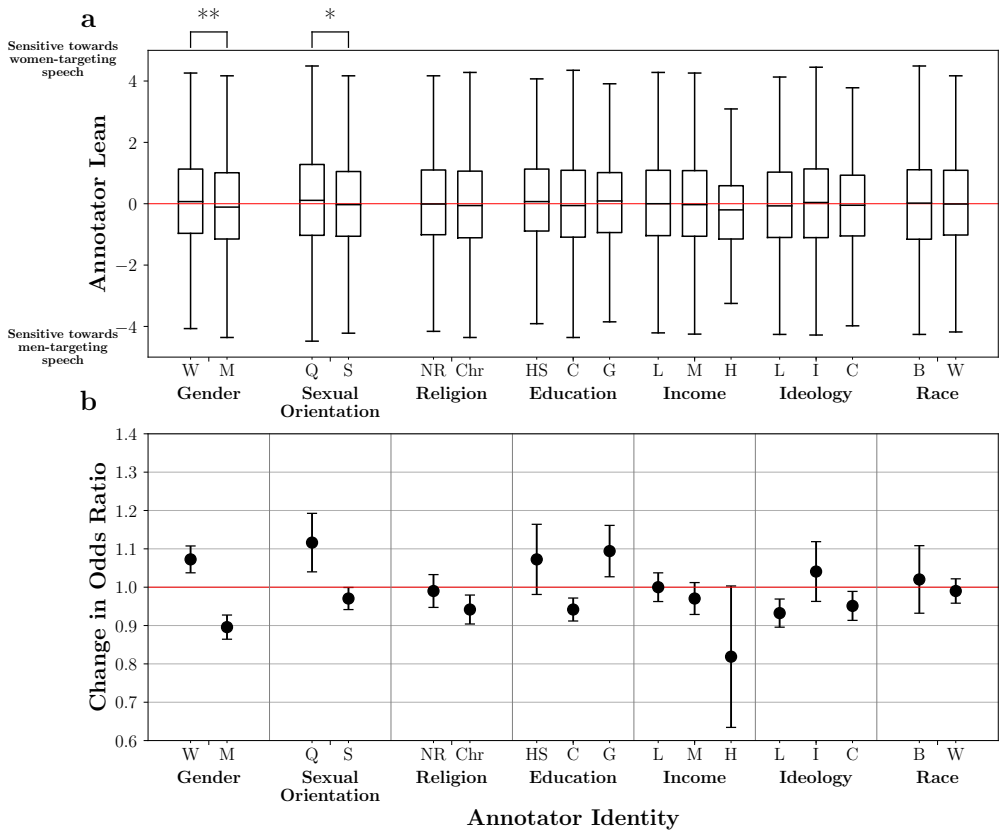

Figure S1: Annotator identity can corresponds to their lean toward Women- or Men- targeting comments. The x-axis captures several different identity groups, including (i) Gender: Men (M) vs. Women; (ii) Sexual Orientation: Straight (S) vs. Queer (Q) annotators; (iii) Religion: Christians (C) vs. Atheist and non-religious annotators (NR); (iv) Education: annotators with High School (HS), College (C), or Graduate (G) education (v) Income: Low-income (L), Middle-income (M), and High-income (H); and (vi) Ideology: Liberal (L) vs. Independent (I) vs. Conservative (C) annotators; and (vii) Race: Black (B) vs. white (W) annotators. Identity groups are sorted by increasing p-value. a. The distribution of annotator leans, equal to the difference in severities obtained from sub-corpora of Women-targeting comments and Men-targeting comments. A positive annotator lean implies that the annotator is more likely to rate Women-targeting speech as exhibiting elements of hate speech. Significance markers denote Kruskal test of medians ($***$: $p < 10^{-3}$; $**$: $p < 10^{-2}$). b. The change in odds ratio of rating a comment that targets female identity vs. male identity. Higher values indicate a higher likelihood of rating Women-targeting speech as hateful. Error bars denote 95% confidence intervals on the median odds ratio.

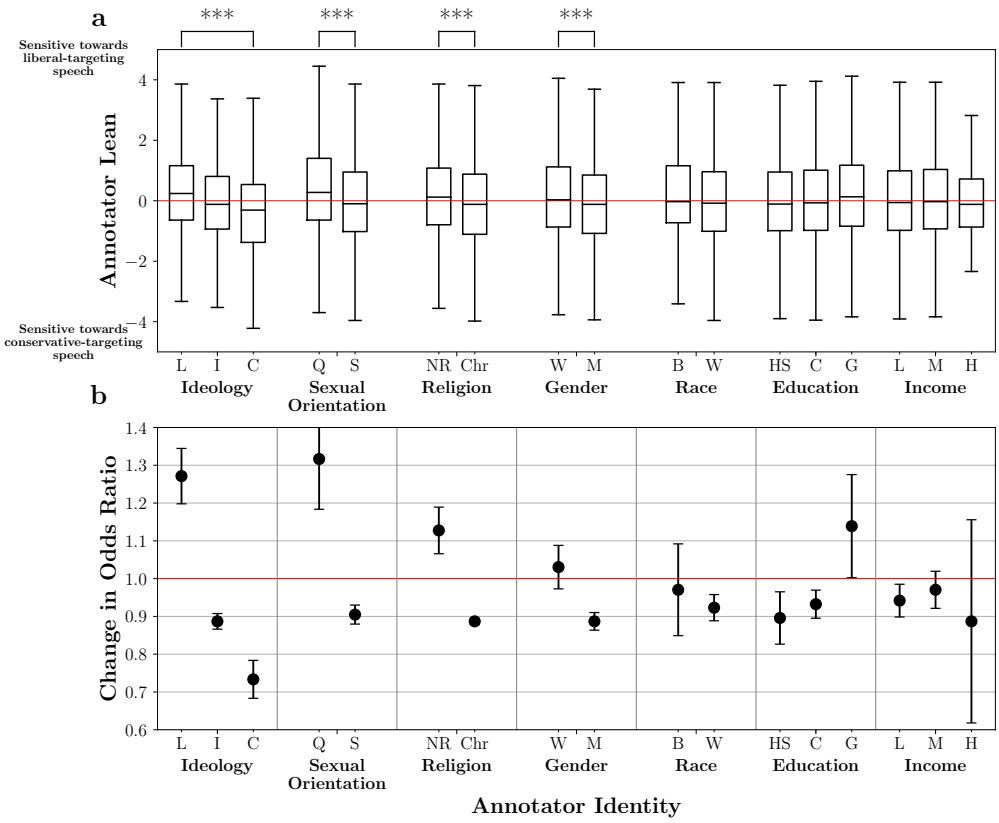

Figure S2: Annotator identity can corresponds to their lean toward liberal- or conservative- targeting comments. The x-axis captures several different identity groups, including (i) Ideology: Liberal (L) vs. Independent (I) vs. Conservative (C) annotators; (ii) Sexual Orientation: Straight (S) vs. Queer (Q) annotators; (iii) Religion: Christians (C) vs. Atheist and non-religious annotators (NR); (iv) Gender: Men (M) vs. Women; (v) Race: Black (B) vs. white (W) annotators; (vi) Education: annotators with High School (HS), College (C), or Graduate (G) education; and (vii) Income: Low-income (L), Middle-income (M), and High-income (H). Identity groups are sorted by increasing p-value. a. The distribution of annotator leans, equal to the difference in severities obtained from sub-corpora of liberal-targeting comments and conservative-targeting comments. A positive annotator lean implies that the annotator is more likely to rate liberal-targeting speech as exhibiting elements of hate speech. Significance markers denote Kruskal test of medians ($***$: $p < 10^{-3}$; $**$: $p < 10^{-2}$). b. The change in odds ratio of rating a comment that targets liberal identity vs. conservative identity. Higher values indicate a higher likelihood of rating liberal-targeting speech as hateful. Error bars denote 95% confidence intervals on the median odds ratio.

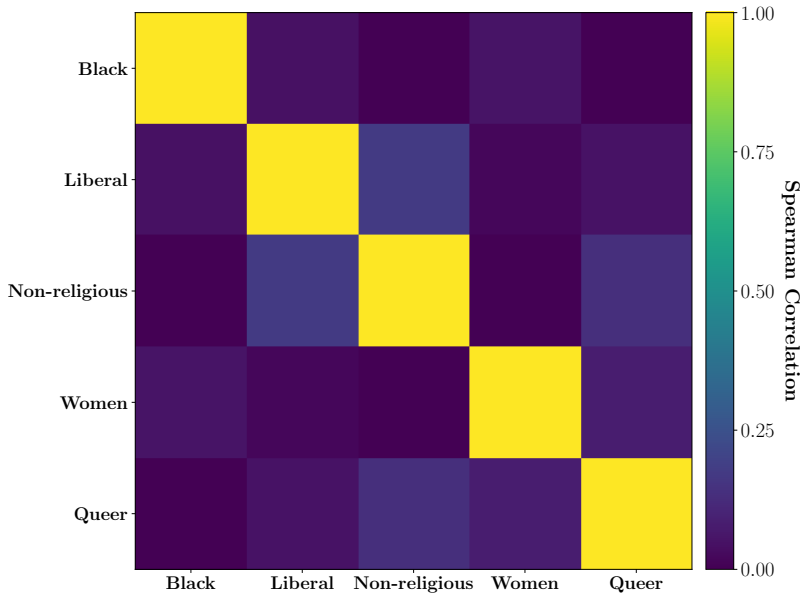

Figure S3: Correlations amongst annotator identity groups. The Spearman correlation between several different annotator identity groups, taken across annotators in the hate speech corpus.

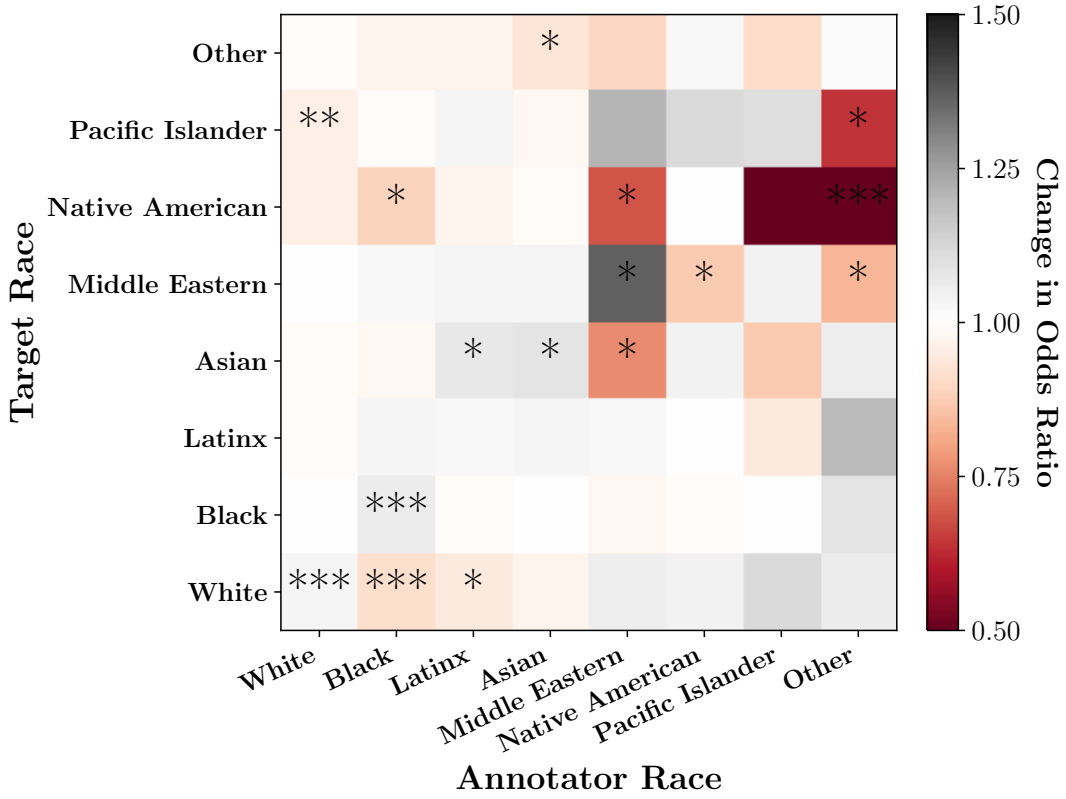

Figure S4: Group-level interaction terms for racial identity. Each shaded square denotes a single parameter estimate within the 8×8 set of interaction terms. The x-axis denotes the annotator race, while the y-axis denotes the target race. Squares are color-coded according to the change in the odds ratio. Higher values (black) indicate a greater sensitivity exhibited by annotators of a racial identity rating comments targeting another racial identity. An odds ratio lower than one (red) implies a reduction in sensitivity for that particular pairing. Significance markers denote p-values ($*$: $p < 10^{-1}$; $**$: $p < 10^{-2}$; $***$: $p < 10^{-3}$.)

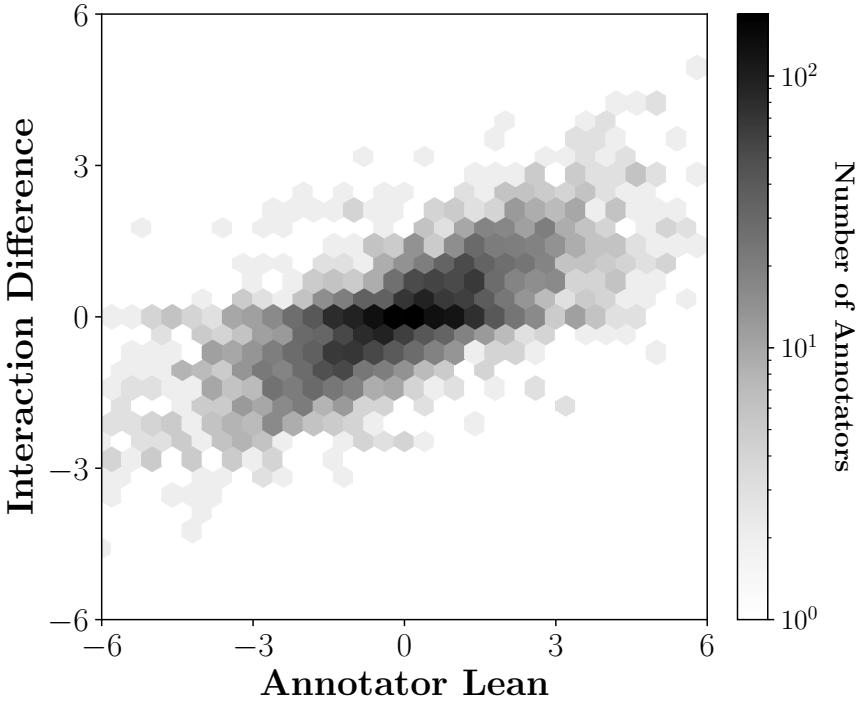

Figure S5: Distinct IRT approaches result in similar quantification of annotator sensitivity. The relationship between annotator lean (calculated from separated measurements) and interaction difference (calculated from annotator interaction terms). Hexbins are shaded according to the number of annotators, on a log-scale.