Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022, pages 83–94
Marseille, 20 June 2022

The *Measuring Hate Speech* Corpus: Leveraging Rasch Measurement Theory for Data Perspectivism

Pratik S. Sachdeva[1], Renata Barreto[1,2], Geoff Bacon[3], Alexander Sahn[4], Claudia von Vacano[1], Chris J. Kennedy[5]
[1]D-Lab, University of California Berkeley
[2]School of Law, University of California, Berkeley
[3]Google
[4]Center for the Study of Democratic Politics, Princeton University
[5]Center for Precision Psychiatry, Harvard Medical School
pratik.sachdeva@berkeley.edu

Abstract

We introduce the *Measuring Hate Speech* corpus, a dataset created to measure hate speech while adjusting for annotators' perspectives. It consists of 50,070 social media comments spanning YouTube, Reddit, and Twitter, labeled by 11,143 annotators recruited from Amazon Mechanical Turk. Each observation includes 10 ordinal labels: sentiment, disrespect, insult, attacking/defending, humiliation, inferior/superior status, dehumanization, violence, genocide, and a 3-valued hate speech benchmark label. The labels are aggregated using faceted Rasch measurement theory (RMT) into a continuous score that measures each comment's location on a hate speech spectrum. The annotation experimental design assigned comments to multiple annotators in order to yield a linked network, allowing annotator disagreement (perspective) to be statistically summarized. Annotators' labeling strictness was estimated during the RMT scaling, projecting their perspective onto a linear measure that was adjusted for the hate speech score. Models that incorporate this annotator perspective parameter as an auxiliary input can generate label- and score-level predictions conditional on annotator perspective. The corpus includes the identity group targets of each comment (8 groups, 42 subgroups) and annotator demographics (6 groups, 40 subgroups), facilitating analyses of interactions between annotator- and comment-level identities, i.e. identity-related annotator perspective.

Keywords: hate speech, item response theory, Rasch measurement theory, measurement, annotator identity

1. Introduction

The application of machine learning on increasingly diverse and difficult tasks has required the curation and annotation of new, large-scale datasets (Bender and Friedman, 2018). These tasks, particularly in natural language processing (NLP), can exhibit low *intersubjectivity*, in which observer variability may be high: annotators may assign different labels to a data sample (Basile et al., 2021b; Basile et al., 2021a). Such disagreement may stem from differences in how annotators interpret the task, their knowledge and understanding of the data sample, or their subjective opinion on the label to assign. Typically, annotator agreement metrics (Krippendorff, 2018) are used to assess the "quality" of *gold labels*, in which a single label is assigned to a data sample based on the input of one or more annotators. At the same time, these tasks are often constructed around binary or ordinal labels which may be limited in their ability to capture complex phenomena.

Data perspectivism (Basile et al., 2021a) argues that annotator disagreement is an informative feature of the data, rather than noise that must be tamped down. Thus, disaggregated datasets, containing the labels provided by all annotators to each sample, are preferable. Data perspectivism, however, requires the development of new methods to facilitate the analysis and training of models on disaggregated datasets.

Measurement theory, a framework in which latent attributes of observed data are estimated, is well suited to the data perspectivist paradigm. In particular, Rasch measurement theory provides a framework to construct a measurement scale to a problem, develop annotation tasks appropriate for that measurement scale, and fit a probabilistic model whose parameters detail important contributions to the scale (Engelhard and Wind, 2017; Hambleton et al., 1991; Rasch, 1968). Specifically, *faceted* Rasch measurement (Linacre, 1994) allows one to capture multiple features ("facets") that influence the generation of a label, including content of the data sample, the annotator's perspective, and the task at hand. The features are measured on a continuous scale, providing more information than binary or ordinal labels generally encountered in NLP corpora. Rasch measurement theory, therefore, motivates not just the development of disaggregated datasets suitable for perspectivist analysis, but those suitable for *measurement*.

In this work, we introduce the *Measuring Hate Speech* (MHS) corpus, a dataset created to measure hatefulness in social media comments. Hate speech detection, particularly in social media comments, has become an increasingly studied and prevalent problem. We chose to study hate speech due to its importance as both a computational social science and human rights problem. Furthermore, hate speech is a complex linguistic phenomena, with no unified definition, which limits

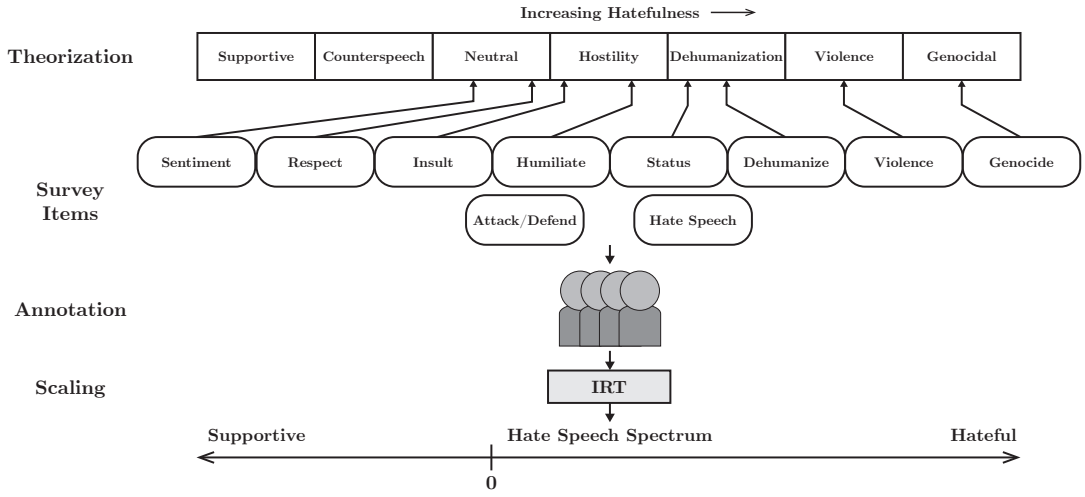

Figure 1: **Measuring hate speech requires the theorization of a construct, development of survey items, annotation, and scaling.** The major steps in developing a measurement scale consist of theorization, developing survey items, annotation, and scaling. **Theorization.** Seven theorized levels of hatefulness, ranging from supportive to genocidal speech. These levels increase in hatefulness from left to right. **Survey Items.** Survey items, or labeling tasks for annotators, consisted of 10 questions that interrogated the data samples at various points along the construct. **Annotators.** Each annotator provided labels on the 10 survey items for some subset of the comments. **Scaling.** Annotator responses are passed as input to an *item response theory* model, resulting in parameter estimates capturing, for example, the hatefulness of each comment, the annotator bias, and the level of hatefulness captured by each survey item. These parameter estimates formulate a hate speech spectrum, centered around 0.

the use of classical gold-label corpora while motivating data perspectivist approaches (Sellars, 2016).

This paper is organized as follows. First, in Section 2, we discuss related work in hate speech detection/measurement and data perspectivism. We introduce the *Measuring Hate Speech* corpus in Section 3, providing a details on the data collection, annotation, and a discussion on Rasch measurement theory. We provide exploratory analyses on the MHS corpus in Section 4. We conclude with a discussion in Section 5.

2. Related Work

Scholars in the emerging field of data perspectivism have identified a number of assumptions about the data generation process, such as the idea that there is only one truth resulting in the creation of gold standard ground-truth datasets and that disagreements among annotators "should be avoided or reduced" (Aroyo and Welty, 2015). Beginning in computational linguistics and spreading in other ML applications, empirical analyses operationalizing data perspectivism theories have found that annotator disagreements are not statistical noise, but rather indicative of ambiguities (Plank et al., 2014) and driven by background and lived experiences (Akhtar et al., 2019). Researchers have found that for annotations of highly subjective tasks, namely offensive language, labelers' different decisions should all be considered correct (Basile et al., 2021b).

The literature in particular acknowledges that disagreement is more likely to occur in tasks such as "detecting affect, aggression, and hate speech" (Davani et al., 2022)–in other words, in tasks modulated by social factors that are "highly polarizing" (Akhtar et al., 2019). In a novel, mixed-methods study, Sang and Stanton (2022) carried out interviews with 170 annotators in a hate speech task to understand where these differences come from. They found that "age and personality differences were connected with the dimensional evaluation of hate speech". To handle these disagreements, researchers have developed methods that incorporate this signal into their models. Akhtar et al. (2019) create a metric of polarization at the individual comment level, which is used to weight samples during training. Other methods have used multi-task or multi-label models to capture annotator differences (Davani et al., 2022). Our work builds on the recognition that annotator disagreements are useful at the data, model, and audit level.

Several existing corpora similarly capture multiple aspects of hate speech beyond a binary label (Waseem and Hovy, 2016; Zampieri et al., 2019; Cercas Curry et al., 2021) and label multiple identity targets (Kennedy et al., 2022; Röttger et al., 2021). However, to our knowledge, the MHS corpus is the only corpus created for hate speech measurement.

3. The *Measuring Hate Speech* Corpus

The *Measuring Hate Speech (MHS)* corpus, created by Kennedy et al. (2020), consists of annotations on social media comments designed to construct a measurement scale for hate speech. In contrast to traditional hate speech corpora, the MHS corpus contains multi-

ple hate-informative labels for each annotator's review of a comment. These labels reflect a theoretical construct of hate speech, which captures degrees of "hatefulness" on a continuous spectrum rather than a yes/no dichotomy (Fig. 1).

We organize this section to first broadly introduce Rasch measurement theory, the motivating theory behind the MHS corpus (Section 3.1), followed by a summary of the data collection and preprocessing (Section 3.2). We then roughly follow the outline shown in Figure 1, discussing key components of the datasets in the context of Rasch measurement theory, including construct theorization, survey items, data annotation, and scaling procedure. While we highlight many of the components of data collection, annotation, and preprocessing, we refer the reader to Kennedy et al. (2020) for additional details.

The MHS dataset is publicly available on HuggingFace[1]. Additionally, the code used to conduct the analyses and create the figures shown in this paper is publicly available on GitHub[2].

3.1. Rasch Measurement Theory

The goal of measurement theory is to measure a latent attribute of a particular unit, such as a social media comment. Measurement frameworks allow one to transform observations–such as examination responses, or in this context, annotations–into variables that reflect an underlying scale. Rasch measurement theory is a measurement framework capable of assessing multiple contributions to the observed labels via the development of a measurement scale, coupled with a multilevel probabilistic model that explicitly captures separate contributions to the ratings in its parameters (Engelhard and Wind, 2017; Hambleton et al., 1991; Rasch, 1968). It simultaneously places the fitted parameters on a common, continuous scale that represents the task at hand.

Critically, Rasch measurement theory requires one to obtain data (in this case, annotations on comments) that fits a proposed model, rather than proposing a model to suit the data. To be clear, one must first develop a theorization for the measurement scale, as well as a labeling instrument (i.e., survey items) that allow one to measure along the theorized scale. Annotations must be obtained *given* this theorization, to which a measurement scale can be obtained (Fig. 1).

3.2. Data Collection and Preprocessing

We sourced comments from three major platforms–YouTube, Twitter, and Reddit–performing collection between March and August 2019. We only considered comments that were written primarily in English and were between 4 and 600 characters. Additionally, we aimed to source 40% of the corpus from Reddit, 40% from Twitter, and 20% from YouTube. We used fewer comments from YouTube for two reasons: first, scraping from YouTube was comparatively more difficult, resulting in a smaller comment pool, and second, YouTube comments tended to be shorter and simpler, with less complex language.

To build the corpus, we leveraged each platform's public API to download comments posted on each site. On Reddit, we collected all comments from posts on the real-time stream of `/r/all`, which contains all public posts on the site. For Twitter, we collected tweets from Twitter's streaming API, which is a random sample of all tweets on Twitter. For YouTube, we first searched for videos within proximity of the top 300 most populated U.S. cities, which were most likely to contain English comments with U.S.-based authors. We then downloaded all comments and responses on these videos. Once we scraped comments, we applied a simple preprocessing pipeline, removing URLs, phone numbers, and contiguous whitespace and accents.

In order to account for the fact that hate speech is relatively rare, we subsampled the scraped comments to create the final corpus. We used two predictive algorithms (multilayer perceptron and a random forest: see Kennedy et al. (2020) for more details) to bin comments into five groups: (i) irrelevant, (ii) relevant but not hateful, (iii) moderately hateful, (iv) very hateful, and (v) extremely hateful. We stratified sampling from each bin, heavily oversampling bins (ii), (iv), and (v), resulting in the comment set.

3.3. Construct Theorization

Developing a measurement scale for a problem requires the theorization of a *construct* that represents the underlying scale (Wilson, 2004). The construct represents an effort to make an underlying scale for a phenomenon explicit. In the context, this amounts to theorizing levels of comments: what are the character of comments that are increasingly hateful, culminating in the most hateful content? Developing a construct requires a rigorous qualitative evaluation of example hate speech comment.

We theorized a construct as follows. From a manual review of social media comments, we curated a *reference set*, a small corpus of example text for each conceptual level. We selected 10 comments to serve as examples of each of the theoretical levels, totalling 70 comments. In concert with construct development using existing literature, we manually reviewed thousands of comments from our corpus. We also selected reference set comments for each level that yielded a diversity of target groups, text length, and linguistic styles. Iteratively, we selected comments that we felt best exemplified levels of hate speech, and when we found ambiguities, used the comments to refine the definitions of each level.

The theorized levels we constructed are shown at the top of Figure 1. The levels build off a *Neutral* level,

[1] `https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech`

[2] `https://github.com/dlab-projects/hate_measure_data`

or speech not evidently positive or negative, in opposite directions. The levels toward the right on the scale designate hate speech of increasing severity: *Hostility*, *Dehumanization*, *Violence*, and *Genocidal*. We placed speech supporting genocide, the systematic killing of a specific group, as the most severe form of hate speech (ADL, 2016; Stanton, 2013). We constructed the remaining levels as pathways to genocide, with special attention to threats of violence and dehumanization that may justify violence. On the other side of neutral speech are two levels denoting speech positive in nature: *Counterspeech*, or speech that explicitly seeks to counter hateful content, and *Supportive* speech.

3.4. Labels and Data Annotation

With the theorized levels of the construct in place, we then developed a *labeling instrument*. The labeling instrument contained three components: (i) a set of 10 *survey items* that allowed the annotator to interrogate the comment along several distinguishing features of hate speech, (ii) specification of any identity groups targeted by the comments, and (iii) questions about the annotator's demographic information. The data annotation process was approved by the University of California, Berkeley Institutional Review Board. Annotators were allowed to omit any demographic information, and all data samples were anonymized to protect annotator privacy.

We recruited annotators from Amazon Mechanical Turk to complete the labeling instrument. We used each worker's IP address to ensure that we only recruited annotators within the United States. Each annotator was provided 26 comments, of which 6 were reference set comments. Thus, the reference set comments generally received the most annotations. The median time to complete the survey was 49 minutes. We compensated participants $7, yielding a median pay rate of $8.57 per hour, which is above the minimum wage in the United States. We provided annotators the opportunity to provide feedback on the labeling process. A manual review of their feedback revealed high satisfaction with the compensation for the task, and appreciation that the results would contribute to an understanding of social media conversations.

Annotators provided ratings on five-level Likert-style scales for 10 different survey items, capturing the following aspects of hate speech: sentiment, respect, insult, humiliation, dehumanization, violence, genocide, attacking/defending, inferior/superior status, and a binary hate speech classification. These survey items were designed to roughly span the hate speech construct (Fig. 1: survey items). In each case, a higher rating on the item 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).

We examined the quality of each annotator's labels with an *infit mean-squared statistic*, a rater fit diagnostic that is calculated during the Rasch scaling. This statistic ranges from 0 to infinity, with an expected value of 1. Annotators whose infit mean-squared statistic was greater than 1 had more randomness or noise in their responses than expected by an IRT model, with values of 2 or greater seen as degrading the measurement system. Those with a statistic less than 1 had less randomness than expected, suggesting that they may have favored certain response options. We chose to exclude raters whose infit mean-squared statistic fell outside $[0.37, 1.9]$. This range corresponded to the previously mentioned heuristics and excluded a cluster of annotators whose infit mean-squared statistic was too low (see the Appendix of Kennedy et al. (2020) for more details). We additionally removed annotators with extreme severity parameters, completed the task too quickly, or whose IP addresses were either geolocated to outside the United States, linked to known proxy services, or associated with more than 4 annotation tasks. Lastly, we excluded raters who did not tag a sufficient number of targeted identity groups, specifically on samples known to contain a targeted identity group. Application of these criteria left 8,472 annotators, with 39,565 accompanying comments.

3.5. Item Response Theory

The labels for each survey item constitute a set of ordinal responses aimed to interrogate each comment for their placement on the hate speech construct. The goal of item response theory (IRT) is to utilize these ordinal responses to devise the continuous scale corresponding to the construct. This is done via a multilevel probabilistic model that maps the labels onto latent parameters which set the scale. There are a variety of possible IRT models one can use depending on the use case.

We detail a *faceted partial credit model*, as it is the most appropriate IRT model for the MHS corpus. This 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 for rater j assigns comment n a rating k on survey item i. Similarly define $p_{nij(k-1)}$, but for rating $k-1$. The model defines an *odds ratio* as a function of several parameters to be learned from the data:

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

We reiterate that survey items are aligned in their numerical code ordering. Thus, "increasing" a rating (going from $k-1$ to k) *always* corresponds to a higher degree of hatefulness. A larger odds ratio implies that the annotator is more likely to rate a particular comment as possessing some aspect of hate speech. Intuitively, the odds ratio should depend on the following *facets*:

- θ_n, or the **hate speech score** of comment n. Higher values of θ_n indicate a more inherently hateful comment.

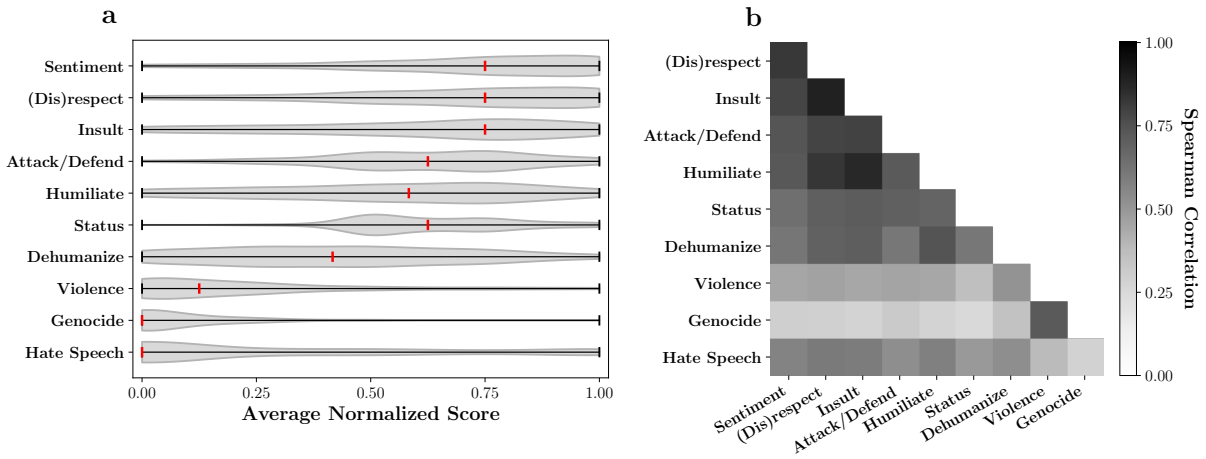

Figure 2: **Survey items allow annotators to evaluate comments at different degrees of hatefulness. a.** The distribution of the survey item ratings, across comments, averaged across annotators for each comment. Each score is normalized to the maximum rating allowed on the Likert scale (4 for all items except "hate speech", which had a maximum of 2). A higher normalized score implies a greater degree of hatefulness. Red lines denote median across comments. **b.** The distribution of Spearman correlations, calculated across comments, between the average ratings of each pair of survey items.

- δ_i, or the **difficulty** of survey item i. The difficulty sets the scale on the hatefulness spectrum. We should expect survey items that probe the higher end of the hatefulness spectrum, such as genocide, to have higher difficulties. In this sense, it is more "difficult" for a comment to exhibit aspects of genocide due to it corresponding to a higher level on the construct.

- α_j, or the **severity** of rater j. We can interpret this quantify as directly quantifying annotator perspective. Specifically, annotators with higher severity are less likely to label comments as possessing features of hate speech: their threshold for "hatefulness" tends to be higher.

- τ_k is also referred to as the **difficulty** of response k. In contrast to the difficulty of the survey item, τ_k is an indicator of the rarity of the ordinal 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.

The faceted partial credit model separates the content of the comment from any modulation stemming from the annotators or survey items, allowing the examination of each facet separately. The distribution of the parameters forms a *hate speech spectrum* (Fig. 1: bottom). The strength of this approach is that comments, survey items, and annotators simultaneously lie on a common scale, allowing one to interpret the model parameters in the context of the construct.

4. Exploratory Analysis of MHS Corpus

We provide exploratory analyses on the annotations and features available in the MHS corpus. Specifically, we show analyses of survey item annotations, target identity annotations, and annotator demographics. Overall, we aim to quantify the intersubjectivity of each set of features, while suggesting potential future analyses on the data. We refer the reader to Kennedy et al. (2020) for an IRT analysis of the data.

4.1. Survey items capture the spectrum of hatefulness

The ten survey items labeled by annotators were designed to align the measurement scale to the theorization proposed in Figure 1. The item responses are chosen such that a higher "value" always aligns with more hatefulness. Survey item responses on different Likert scales, then, can be compared by dividing annotator responses by the maximum possible response, resulting in a *normalized score*. A comment can be summarized in aggregated fashion by taking the mean of normalized scores across annotators, resulting in an *average normalized score*.

To better understand the the behavior of the survey item responses along the theorized construct, we examined the distribution of averaged normalized scores across comments in the corpus (Fig. 2a). We found that, generally, the average normalized scores decreased on survey items that probed for increasingly hateful content (Fig. 2a: top to bottom). This implies that, within the MHS corpus, fewer comments tend to exhibit the most hateful content (e.g., violence and genocide), which we may expect as a reasonable prior on the distribution of hateful content on social media.

Since the survey items probe points along the theorized hatefulness spectrum, we should expect item responses closer to each other to correlate more strongly. Thus, we computed the Spearman correlations between averaged normalized scores for each pair of survey items, across comments (Fig. 2b). We found that nearby survey items exhibit strong correlations with each other (Fig. 2b: diagonal). Importantly, pairs of survey item further away on the hatefulness spectrum have markedly lower correlations with each other. For example, "violence" and "genocide" are weakly correlated with the remaining survey items, but exhibit strong correlations with each other. Furthermore, the hate speech survey item showed moderate correlations with all other survey items, indicating that each survey item is capturing some component of hate speech (Fig. 2b: bottom row). Together, these results demonstrate that the chosen survey items adequately probe the theorized hatefulness spectrum.

4.2. Annotators exhibit low agreement on survey item responses

In traditional corpora, labels are aggregated across annotators to assign a "gold label" to each sample (Basile et al., 2021a; Ide and Pustejovsky, 2017). In order to assess the reliability of the gold label, annotator agreement metrics such as Cohen's kappa or Krippendorff's alpha are generally computed (Krippendorff, 2018; Waseem and Hovy, 2016). However, in NLP datasets, these metrics are often low, indicating that annotators do not tend to strongly agree on the label for each data sample (Poletto et al., 2021) This holds particularly true for hate speech corpora: hate speech can be difficult to define, may require intimate knowledge of in-group language or slurs, and generally exhibits low intersubjectivity (Sellars, 2016). In the MHS corpus, we might expect that annotator agreement to be low, given that annotators likely have different interpretations of the survey instrument (e.g., "sentiment" may be interpreted differently by annotators) and they may exhibit subjectivity in assigning different Likert scale ratings (i.e., annotators have different internalized thresholds for each response).

We evaluated the annotator agreement on the responses to each survey item using Krippendorff's alpha. We found that annotators generally exhibited weak agreement on all survey items, with Krippendroff's alphas of less than 0.5 (Fig. 3: light grey bars). Some survey items–such as "attack/defend" and "status"–exhibited markedly lower agreement, indicating that these items are prone to more subjective responses. Meanwhile, the hate speech survey item received a higher Krippendorff's alpha than the remaining survey items. This implies that, while annotators may agree more often on whether a comment is hate speech, they agree less often on the *components* of that hate speech. Thus, the additional survey items allow finer examination on how the levels of the construct contribute to an annotator's perception of hate speech.

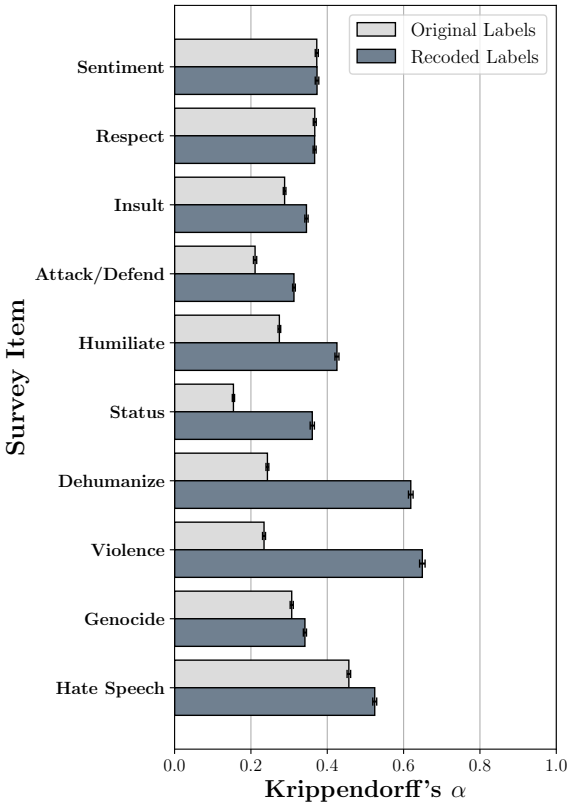

Figure 3: **Annotators exhibit low agreement on survey items, demonstrating the subjectivity of the task.** Annotator agreement on each survey item, as quantified by Krippendorff's α. Error bars denote 95% confidence intervals. Light grey bars denote agreement calculated on the original labels, while dark grey bars denoted agreement calculated on recoded labels, which were coarsened fom the original labels. The cardinalities of each label (before/after) recoding were as follows: sentiment (5/5), respect (5/5), insult (5/4), attack/defend (5/4), humiliate (5/3), status (5/2), dehumanize (5/2), violence (5/2), genocide (5/2).

We found that a 5-level item may not be necessary for survey items aligned on the higher end of the hatefulness spectrum. For example, the item responses for "sentiment" (Appendix A) may naturally be suited to increased label granularity due to its lower intersubjectivity. However, a concept such as "genocide" may align more neatly to a lower level Likert item (or simply a binary item), since "genocide" may exhibit higher annotator intersubjectivity. Thus, we considered a recoding scheme in which annotator responses were mapped onto a lower level Likert items. We chose the recodings in order to improve the IRT modeling statistics (Kennedy et al. (2020)). Specifically, we retained the sentiment and respect survey items as is, but recoded insult ($5 \rightarrow 4$ levels), attack/defend ($5 \rightarrow 4$ levels), humiliate ($5 \rightarrow 3$ levels), status ($5 \rightarrow 2$ levels), de-

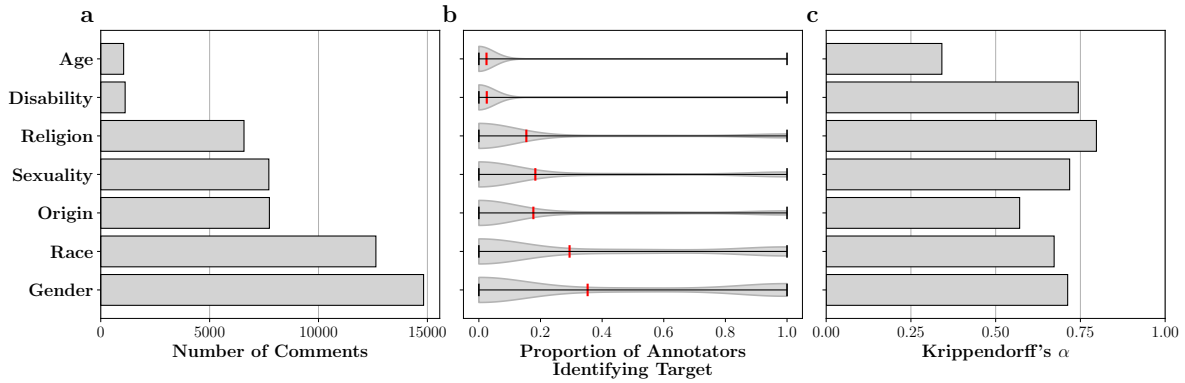

Figure 4: **Annotators recorded identity targets of comments, expressing stronger agreement than the survey items.** For each comment, annotators recorded a binary response specifying whether a particular identity group was targeted by the comment. **a.** The number of comments targeting each identity group, according to a 0.5 annotator agreement threshold. That is, if 50% or more annotators indicated a comment targeted a specific identity group, that comment was assigned a positive label for that group for purposes of the subplot. **b.** The distribution of "annotator agreements" across comments, for each identity group. Annotator agreements were calculated as proportions by averaging annotators' binary responses to whether an identity group was targeted. Red lines denote the mean proportion. **c.** Annotator agreement on each target identity group, quantified by Krippendorff's α.

humanize ($5 \rightarrow 2$ levels), violence ($5 \rightarrow 2$ levels), genocide ($5 \rightarrow 2$ levels), and hate speech ($3 \rightarrow 2$ levels). We found that, under the recoding, Krippendorff's alpha increased for each survey item (Fig. 3: dark gray bars). In particular, we found large increases for the "status", "dehumanize", and "violence" survey items. Thus, recoding schemes can reduce observer variability when the survey item tend to exhibit lower degrees of intersubjectivity.

The low annotator agreement observed in the MHS corpus further motivates the usage of methods better suited to handle disaggregated data. Specifically, item response theory models such as the faceted partial credit model discussed in Section 3.5 are particularly relevant, as they explicitly model the multiple components that may contribute to the results seen in Figure 3.

4.3. Annotation of identity group targets

Hate speech differs from other kinds of toxic or offensive speech in that it specifically targets an identity group(s) (Sellars, 2016; Poletto et al., 2021). Thus, identification of the targeted identity groups is a vital component of a hate speech corpus. Past studies have specified various characterizations of "targeting", such as explicit and implicit rhetoric (Kennedy et al., 2022) or individual and group targeting (Zampieri et al., 2019). While the notion of targeting can be captured by additional labels or possibly a measurement scale, we restricted labeling to the binary identification of pre-specified identity group and sub-groups targeted by a comment, as has been done in previous corpora (Röttger et al., 2021; Kennedy et al., 2022).

Annotators were asked *"Is the [comment] directed at or about any individuals or groups based on..."*, with the option to select among the following eight identity groups: race/ethnicity, religion, national origin or citizenship status, gender, sexual orientation, age, disability status, political identity; along with the option to select "none of the above" (options listed in order presented on the survey). Annotators were further asked to specify identity sub-groups targeted by the comment (see Appendix B). Annotators could select more than one option among these identity groups and sub-groups. Thus, the target identity annotations can be viewed as a multi-label binary variable indicating whether each identity sub-group was targeted or not.

Specification of target identities is a task that exhibits higher rater intersubjectivity than hate speech measurement, because comments often make clear which identity group is targeted. However, hateful content can subtly indicate its target, sometimes using specific vernacular, dog whistles, or vague language that may not be understood or difficult to notice by some annotators (Sellars, 2016). Thus, annotators still express disagreement on identity group targets.

We first examined the number of comments targeting each identity group. As a cursory analysis, we used majority voting across annotators to assign each comment a single binary label specifying whether it targeted any of the 8 identity groups. We found that most comments targeted based on gender and race (Fig. 4a), with the least number of comments targeting age and disability. This distribution likely reflects the true distribution of comments on social media. It also is likely influenced by the sampling procedure, as it is easier to identify hateful comments targeting groups that have larger available hate lexica, such as for race and gender.

We then computed the proportion of annotators labeling each identity group as the target of a comment. If this proportion is 1, all annotators agree that the comment targets the identity group. If the proportion is

0, all annotators agree that the comment does not target the identity group. Values between 0 and 1 indicate some measure of disagreement on the target. We examined the distribution of these proportions across comments for each target identity group (Fig. 2b). We found that, across identity groups, the density of proportions exhibited modes at 0 and 1, indicating that annotators generally agreed on the targets of comments. However, some density spanned between 0 and 1, indicating a sizeable amount of disagreement.

Lastly, we computed Krippendorff's alpha in order to quantify annotator agreement for each target identity group. We found that Krippendorff's alpha was greater than 0.60 for every identity group except for age, with religion and disability exhibiting the highest agreement. On the whole, these values are larger than those of the hate speech survey items, indicating that identifying targets of hate speech likely exhibits higher intersubjectivity than the hate speech survey items. Thus, these labels are more amenable for weak perspectivist direct prediction tasks, such as a model that aims to predict the target of the identity group.

4.4. Annotator demographics

A critical aspect of data perspectivism relies on the relationship between an annotator's perspective and the labels they assign to text on a task. Specifically, the various groups that an annotator may identify with can shape their perspective, thereby influencing their interpretation of subjective labeling tasks. Annotator demographics, therefore, are a necessary consideration in taking on a data perspectivist lens on NLP datasets.

The *MHS* dataset contains demographic information about the annotators for several identity groups. Annotators were asked to voluntarily specify their racial identity, gender identity, sexual orientation, religious affiliation, educational level, income, age, and political affiliation. The specific sub-groups annotators were asked to identify within these broad identity groups are specified in Appendix B. Within the race, gender, sexual orientation, and religion identity groups, annotators could select more than one sub-group.

We examined the distribution of annotator identities by calculating the proportion of annotators identifying as each sub-group (Fig. 5). We found that, while many racial identities were represented among the annotators, the vast majority identified as White (over 80% of the entire annotator pool). Among these annotators, roughly 90% identified solely as White (i.e., did not identify as multiracial). With respect to gender, most annotators identified as women (56%), followed by men (43%), with less than 1% of annotators identifying as non-binary. Additionally, nearly all (99%) annotators identified as cisgender. With respect to sexual orientation, most annotators identified as straight (85%). An array of religious affiliations were represented, with a plurality of annotators identifying as Christian (42%) followed by atheist annotators (21%).

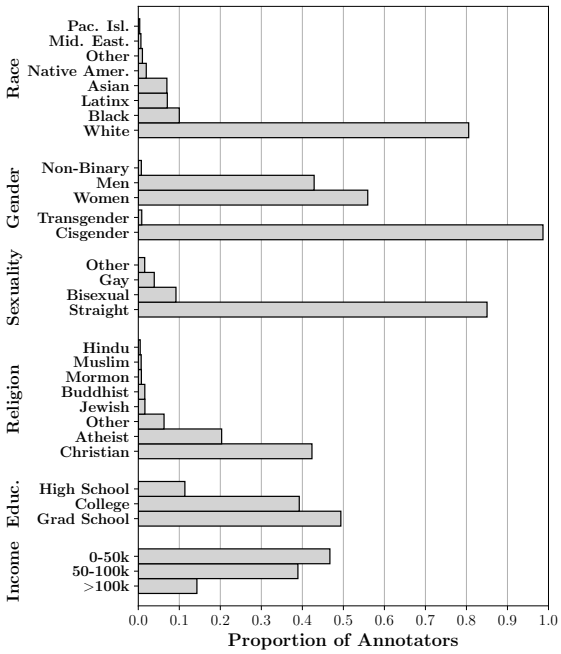

Figure 5: **Annotator demographic specifications span multiple identity group and sub-groups.** The proportion of annotators that identified as specific identity sub-groups. The sub-groups fall in larger identity groups, including race, gender, sexuality, religion, education, and income (y-axis labels). For race, sexuality, and religion identity groups, annotators could select multiple identity sub-groups. The gender identity group consisted of two separate questions, asking for gender identity (man, woman, non-binary) and annotator identification as transgender. Some sub-groups are coarsened from finer sub-group options (e.g., "High School" education constitutes annotators identifying their educational background as "Some High School" or "Completed High School"). For race, gender, sexuality, and religion, identity sub-groups are sorted in order of increasing proportion.

Nearly half of annotators had some level of graduate school education, including a master's degree, professional degree, or doctorate degree. Lastly, the majority of annotators stated that their income was less than $100,000 per year.

5. Discussion

We presented the *Measuring Hate Speech* corpus, a dataset created following Rasch measurement theory to measure hate speech. The 10 component labels, identity target labels, and annotator demographics available in the dataset can support a wide range of subsequent analyses that incorporate knowledge of annotator perspective in studies on hate speech.

The MHS dataset follows data perspectivism by providing the means to capture an annotator's perspective via the severity parameters. Usage of these parameters allow one to sidestep the need to consider annota-

tor agreement, as an annotator's own strictness is explicitly captured in an IRT model. They are also useful in secondary analyses examining whether an annotator's labeling patterns exhibit identity-level interactions. For example, several studies have documented the relationship between an annotator's identity and the labels they assign to comments in hate speech classification tasks (Sap et al., 2021; Geva et al., 2019; Larimore et al., 2021). Item response theory offers avenues to perform similar analyses. For example, Sachdeva et al. (2022) used these techniques in the MHS corpus, finding that annotators were more likely to rate speech targeting groups they identify with as possessing elements of hate speech. Therefore, datasets structured with a measurement scale in mind can be flexibly analyzed to quantify annotator perspective. The ability to conduct such analyses is becoming increasingly important as perspectivist datasets are used in training downstream machine learning algorithms.

The outputs of the IRT model can be used for the development of machine learning algorithms that measure hate speech. For example, Kennedy et al. (2020) developed neural networks to predict the continuous hate speech score for each comment. These networks can be extended to incorporate annotator severity as an additional input. This modification can improve performance, as models can be trained on a fully disaggregated datasets in an annotator-aware fashion. Furthermore, fully trained networks can produce output scores dependent on a desired perspective, with the annotator severity input indicating the network's leniency or strictness in measuring the speech.

Future hate speech datasets, and others, can improve on the construct and labeling instrument of the MHS corpus. For example, the theorized construct can undergo further qualitative review and cognitive interviewing, resulting in more precise measurement. Ordinal responses to survey items exhibiting higher intersubjectivity can be adjusted, preventing the need for recording schemes. Annotator demographic questions can be improved to allow more granular responses (e.g., allowing more options for gender identity). Additional rounds of annotation can include more emphasis on annotator explanations for their choices. This would further facilitate data perspectivist analysis of the corpus, and allow qualitative reviews to inform future iterations of the construct operationalization.

Lastly, the construction of a measurement scale for hate speech motivates usage of Rasch measurement theory in other settings relevant for machine learning. For example, tasks which are prone to lower intersubjectivity, such as assessing toxicity, disinformation, and sentiment would be served well by the development of measurement scales in the Rasch paradigm.

6. Acknowledgements

We thank our colleagues in the D-Lab for useful feedback and discussions.

7. Bibliographical References

ADL. (2016). Empowering young people in the aftermath of hate (in english and en español).

Akhtar, S., Basile, V., and Patti, V. (2019). A new measure of polarization in the annotation of hate speech. In *International Conference of the Italian Association for Artificial Intelligence*, pages 588–603. Springer.

Aroyo, L. and Welty, C. (2015). Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Basile, V., Cabitza, F., Campagner, A., and Fell, M. (2021a). Toward a perspectivist turn in ground truthing for predictive computing. *arXiv preprint arXiv:2109.04270*.

Basile, V., Fell, M., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M., Uma, A., et al. (2021b). We need to consider disagreement in evaluation. In *1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Bender, E. M. and Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Cercas Curry, A., Abercrombie, G., and Rieser, V. (2021). ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Davani, A. M., Díaz, M., and Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Engelhard, G. and Wind, S. A. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Geva, M., Goldberg, Y., and Berant, J. (2019). Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166, Hong Kong, China, November. Association for Computational Linguistics.

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory*, volume 2. Sage.

Ide, N. and Pustejovsky, J. (2017). *Handbook of linguistic annotation*, volume 1. Springer.

Kennedy, C. J., Bacon, G., Sahn, A., and von Vacano, C. (2020). Constructing interval variables

via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Kennedy, B., Atari, M., Davani, A. M., Yeh, L., Omrani, A., Kim, Y., Coombs, K., Havaldar, S., Portillo-Wightman, G., Gonzalez, E., et al. (2022). Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, pages 1–30.

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology*. Sage publications.

Larimore, S., Kennedy, I., Haskett, B., and Arseniev-Koehler, A. (2021). Reconsidering annotator disagreement about racist language: Noise or signal? In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 81–90.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. MESA press.

Plank, B., Hovy, D., and Søgaard, A. (2014). Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511.

Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., and Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Rasch, G. (1968). A mathematical theory of objectivity and its consequences for model construction. In *Report from European Meeting on Statistics, Econometrics and Management Sciences, Amsterdam*.

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., and Pierrehumbert, J. (2021). HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online, August. Association for Computational Linguistics.

Sachdeva, P. S., Barreto, R., von Vacano, C., and Kennedy, C. J. (2022). Assessing annotator sensitivity via item response theory: A case study in a hate speech corpus. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM.

Sang, Y. and Stanton, J. (2022). The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In Malte Smits, editor, *Information for a Better World: Shaping the Global Future*, pages 425–444, Cham. Springer International Publishing.

Sap, M., Swayamdipta, S., Vianna, L., Zhou, X., Choi, Y., and Smith, N. A. (2021). Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Sellars, A. (2016). Defining hate speech. *Berkman Klein Center Research Publication*, 2016(20):16–48.

Stanton, G. (2013). The ten stages of genocide. *Genocide Watch*.

Waseem, Z. and Hovy, D. (2016). Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wilson, M. (2004). *Constructing measures: An item response modeling approach*. Routledge.

Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., and Kumar, R. (2019). Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota, June. Association for Computational Linguistics.

A. Hate Speech Survey Items

We show the hate speech survey items in Table 1.

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 1: Survey items and ordinal responses given to annotators for each comment.

B. Annotator Identity Sub-Groups

We show the annotator identity sub-groups in Table 2.

Identity Group	Identity Subgroups
Race or ethnicity	Black or African American, Latino or non-white Hispanic, Asian, Middle Eastern, Native American or Alaska Native, Pacific Islander, Non-hispanic white
Religion	Jews, Christians, Buddhists, Hindus, Mormons, Atheists, Muslims
National origin or citizenship status	A specific country, immigrant, migrant worker, undocumented person
Gender identity	Women, men, non-binary or third gender, transgender women, transgender men, transgender (unspecified)
Sexual orientation	Bisexual, gay, lesbian, heterosexual
Age	Children (0 - 12 years old), adolescents / teenagers (13 - 17), young adults / adults (18 - 39), middle-aged (40 - 64), seniors (65 or older)
Disability status	People with physical disabilities (e.g., use of wheelchair), people with cognitive disorders (e.g., autism) or learning disabilities (e.g., Down syndrome), people with mental health problems (e.g., depression, addiction), visually impaired people, hearing impaired people, no specific disability

Table 2: Identity group and corresponding subgroups annotators were asked to identify as targets of comments.