When Conservatives See Red but Liberals Feel Blue: Why Labeler-Characteristic Bias Matters for Data Annotation

Nora Webb Williams* Andreu Casas† Kevin Aslett‡ John Wilkerson§

Abstract

Human annotation of data, including text and image materials, is a bedrock of political science research. Yet we often overlook how the identities of our annotators may systematically affect their labels. We call the sensitivity of labels to annotator identity "labeler-characteristic bias" (LCB). We demonstrate the persistence and risks of LCB for downstream analyses in two examples, first with image data from the United States and second with text data from the Netherlands. In both examples we observe significant differences in annotations based on annotator gender and political identity. After laying out a general typology of annotator biases and their relationship to inter-rater reliability, we provide suggestions and solutions for how to handle LCB. The first step to addressing LCB is to recruit a diverse labeler corps and test for LCB. Where LCB is found, solutions are modeling subgroup effects or generating composite labels based on target population demographics.

Word count: 9,761

*University of Illinois at Urbana-Champaign: nww3@illinois.edu
†Vrije Unversiteit Amsterdam: andreu.casas@gmail.com
‡University of Central Florida: kevin.aslett@ucf.edu
§University of Washington: jwilker@uw.edu

1 Introduction

Human annotation (also referred to as human labeling or coding) of data is a bedrock of political science research. We read news articles and annotate for partisan bias (e.g. Peterson, Goel, and Iyengar 2021; Budak, Goel, and Rao 2016). We parse judicial decisions for agreement with past precedents (e.g. Segal and Spaeth 1996). We rate images for the presence of violence (e.g. Steinert-Threlkeld, Chan, and Joo 2022) and for whether politicians look competent (e.g. Todorov et al. 2005). We watch campaign ads and note patriotic symbolism (e.g. Kahn and Kenney 1999). From bills (e.g. Gamm and Kousser 2010) and party platforms (e.g. Dolezal et al. 2016) to social media posts (e.g. King, Pan, and Roberts 2013) and interview transcripts (e.g. Putnam 1971): we could make a very long list of data sources that can be annotated to answer important political science research questions.

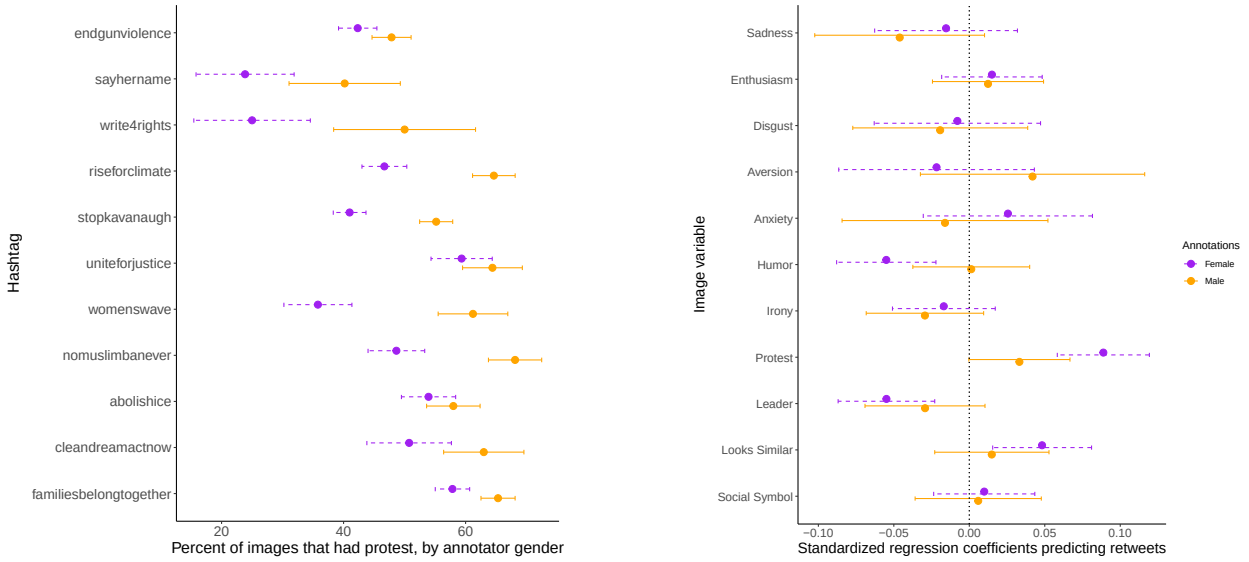
The strength of our conclusions drawn from annotated data depends on our confidence in those annotations. For example, if we wanted to analyze which news sources are more likely to have partisan biases, we need to be confident that we have a correct measure for whether or not articles have a partisan slant. It is not difficult to imagine that who our annotators are might affect the decision to label an article as partisan or not. The task of labeling *for* partisanship might itself be subject *to* partisanship – a liberal might be less likely to identify a liberal slant than a conservative, for example. Indeed, as we demonstrate in this paper, the identity of labelers can have a dramatic impact on their labels. Although this paper does not contain a machine learning application, concerns about label biases are especially prominent in that area (Bolukbasi et al. 2016; Gordon et al. 2021; Hube, Fetahu, and Gadiraju 2019). If we have biased labels from biased labelers, we will have biased models.[1] Most worryingly,

[1] A point made by many scholars, including Hopkins and King (2010) and Benoit, Laver, and Mikhaylov (2009). The correction procedures proposed for these errors, as in Hopkins and King (2010), typically assume that there is one true answer from which annotators have deviated. Our point is that there may be systematic deviations that cannot be corrected in a manner that assumes random error.

we find an impact of labeler identity not just on the types of subjective tasks that we might already suspect would be impacted by labeler identities. Consider Figure 1, which comes from the image analysis study described in detail in subsequent sections. The study contains images shared on Twitter using mobilization hashtags. The left panel shows the average responses from male and female annotators on a question about if an image contains a protest. In other words, it shows the average rate of seeing protest by the gender of the annotator. We see that even on this relatively objective task there are clear differences in responses based on the annotator gender: the male respondents see protest more often, though the magnitude of the difference varies by hashtag. If we were to use these responses in a regression analysis, we might find that our results and conclusions would differ depending on our labeler pool. The right panel of Figure 1 (with a box drawing attention to the main point) demonstrates just this concern. We see that in a regression model predicting retweets based on the images included with tweets, we would conclude that there was a positive significant effect of protest images on retweets if we had an all-female labeling pool. However, if we had an all-male labeling pool, we would conclude that the effect was not significant.

As noted, it is perhaps unsurprising that who labelers are matters to our annotations and therefore to our research conclusions. After all, political scientists are attentive to questions about the validity and reliability of our the labels (see, e.g. Grimmer and Stewart 2013). We want to know that the tasks we set for our annotators result in "good" data and that the annotations reflect stable concepts. The concept of inter-rater (or inter-coder) reliability (herein IRR) is likely familiar to most political scientists as a way to support claims of annotation validity and reliability. However, despite this general awareness of IRR and validity, we find a great deal of variability in whether and which measures of IRR are reported in research published in top political political science journals, based on an original meta-analysis of IRR-reporting. We analyzed 97 articles in-depth that relied on original

Figure 1: Annotator Gender Influences Annotations for "Protest" (panel a) and Regression Results Based on those Annotations (panel b)



(a) (b)

human annotation for content analysis or supervised machine learning (for details on the data collection and analysis, see Appendix A). Only 56% percent of the identified articles (54 out of 97 papers) included any sort of IRR statistic. This low rate of IRR reporting was surprising, especially as the terms "inter-rater reliabilty" and "inter-coder reliability" were used to identify relevant articles. The rate of IRR reporting was relatively consistent over time in our meta-analysis – if anything, the rate of reporting has slightly decreased. Of papers from the 1990s (roughly 9% of the total sample), 66% reported any IRR statistics. Of papers from the 2010s (roughly 28% of the total sample), 63% reported IRR statistics.

Aiming for strong IRR (whether or not we report it in our papers), political scientists design research procedures with attention to *how* labels are generated (especially for crowd-sourced annotation on platforms like Mechanical Turk). We pilot our labeling forms; train coders; drop labels from annotators who speed through the tasks or fail attention checks; drop labels from coders whose annotations consistently do not match our own; use anchoring vignettes, and so on (for recent innovations on ways to improving labeling, see for exam-

ple: Barberá et al. 2021; Struthers, Hare, and R. Bakker 2020; Winter, A. G. Hughes, and Sanders 2020; DeBell 2017; Ying, Montgomery, and Stewart 2022; Benoit, Conway, et al. 2016).

Yet we are less attentive to the threats to validity and reliability that arise because of *who* our labelers are.[2] Our meta-analysis found that only 25% percent of the 97 papers reported any information about who their labelers were in terms of demographic or identity characteristics (e.g. gender, race, socio-economic status). Most often the only information conveyed was when authors reported that the coding had been done by undergraduate or graduate research assistants, which provides information on the education status of labelers. Roughly 63% of the 24 articles that had any information about demographics of labelers referred to education level in some way (see Table 1). Also relatively common were references to language abilities, particularly for Comparative Politics papers where it was noted that annotators were fluent in or native speakers of relevant languages. Only 2 of the articles reported the gender of the coders, and only 1 reported information about partisanship or party identification of labelers.

Table 1: Rate of Mentioning Specific Characteristics in 24 Articles with Demographic Information on Annotators

Characteristic	Proportion of Papers
Education	0.62
Language	0.17
Race	0.17
Nation	0.12
Gender	0.08
Partisanship	0.04

[2]The question of labeler bias based on demographics has received some attention in the machine learning literature – see for example Y. Chen and Joo (2021), Yang et al. (2022), and Steephen, Mehta, and Bapi (2018). Yet there is an important distinction between our work and these prior pieces. These works start from the perspective that annotators respond differently based on the demographics of the people *in* the pictures. That is, annotators may be less likely to recognize anger in an image of a female face compared to a male face. Out point is that female annotators may respond differently to pictures, not that pictures of females may elicit different responses from (gender-unspecified) annotators.

We call the potential threats to research validity that arise from the differing demographic attributes of annotators "labeler-characteristic bias (LCB)." The potential harm of LCB for inference depends on the nature of the annotation task and, more fundamentally, on the research question at hand. The goals of this paper are fourfold. First, we define and characterize LCB as a function of labeling tasks. Second, we demonstrate the existence of LCB on specific tasks and show that in some cases it is resistant to standard training strategies used to improve IRR. Third, we show the effect that ignoring LCB has on downstream analyses. Finally, we make recommendations for how to proceed with research given what we have shown about LCB.

To be sure, some labeling tasks are not subject to LCB or are only minimally impacted, as we discuss below. Worryingly, however, it is often difficult to know ex ante which tasks will be subject to LCB. Our opening example (see Figure 1) comes from what we thought was a straightforward annotation task: whether or not an image includes a protest. We initially assumed that annotating images for the presence of protests would not be strongly affected by identity. Yet in looking at the data we found that female and male annotators, as well as Republican and Democratic annotators, saw protests in the images at significantly different rates (see Figure 1).

It is not always possible to mitigate LCB using the familiar IRR-boosting tools of more training, better labeling forms, and so on. Particularly for tasks that are highly subjective or highly influenced by identity, no amount of training may make a difference. For example, a more detailed codebook may not change fundamental differences in emotional responses to political advertisements (e.g. "Is this a negative ad?") between conservatives and liberals.

Because the effects of LCB can be difficult to predict ex ante, we argue that political scientists should intentionally recruit representative pools of annotators and collect annotator demographic information. Researchers should also intentionally test for LCB and adjust their methodologies to account for it. In this paper we illustrate the LCB threat using image

and text data. It is relatively easy to test for LCB by asking whether labels correlate with demographic characteristics of labelers. When LCB is detected or anticipated, we recommend two interventions. Researchers can generate composite labels by weighting responses according the demographics of the underlying target research population. For example, if the target population is 30 percent female and 70 percent male, a composite measure for protest would weight a female response by 0.3 and a male response by 0.7 to generate a single protest score. Alternatively, researchers can explicitly model differences in subgroup responses. For example, in a regression predicting whether or not an image is retweeted, a researcher might include separate "protest" variables from Republican and Democratic labelers. This models the association between protest imagery and retweets depending on the audience.

We begin by proposing a typology of annotation tasks with variation on two dimensions: *subjectivity* and *sensitivity to identity characteristics*. Using this typology, we hypothesize about which types of tasks should produce higher and lower IRR in general. We also hypothesize about which types of tasks should produce greater annotation divergences from labelers with different identities – in other words, which tasks should be more or less subject to LCB. The tasks that are not subject to LCB are hypothesized to see improved IRR with interventions such as detailed codebooks and training. We then test this framework and hypotheses with an image-annotation analysis, where the annotators were recruited using an online, crowdsourcing platform and asked to label social media images. We find evidence of LCB across a wide range of labeling tasks as well as evidence of its potential impact for downstream analyses. We then further explore LCB in a text-annotation study where undergraduate coders completed multiple labeling task that represent the different quadrants of our typology of labeling tasks. We assess how overall IRR and discrepancies among labelers of different identities (e.g. ideology and gender) varies for each task in the typology under three conditions: no prior training, limited training, and intensive training. Once again, we

find evidence of LCB in some of the text tasks, as well as evidence that more training does not always alleviate LCB. Finally, we also demonstrate the impact of LCB on downstream analyses for this case of text data. Both examples offer clear diagnoses about how to mitigate unanticipated LCB effects in political science research.

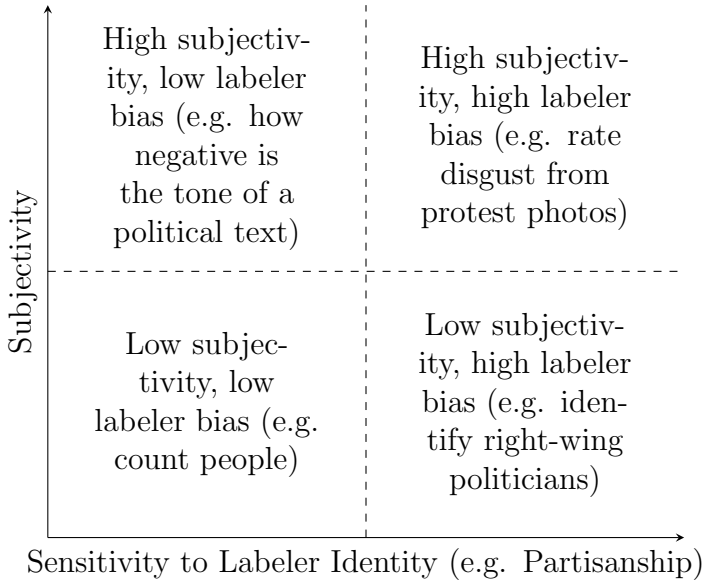
2 A Conceptual Framework for Manual Annotation Tasks

The most familiar annotation tasks involve identifying basic concepts. For text, for example, this might mean noting whether the text of a bill relates to national security. For images, this task often falls under the heading of "object recognition," or noting the presence of objects. For an image labeling task, this might mean answering prompts such as: "Does this picture include a person?" Yet even tasks like these that seem on their face to be completely objective may be subject to interpretation. Does a cartoon drawing of a person, for example, count as a person? Even the most detailed of instructions may still leave room for interpretation.

As the labeling tasks become more subjective, disagreements among labelers become more likely. For example, inspired by Todorov et al. (2005), we might ask an annotator to rate how competent the person in the picture appears. Different coders might judge competence differently. Some annotators might consider age to be a relevant consideration, for example, while others do not. But it might also be the case that their judgements are systematically influenced by their identities. If the person pictured is a well-known politician, judgements of competence might correlate with annotator party affiliation.[3] We know, for example, that partisans report differing emotional reactions to politicians' smiles depending on whether

[3]In the referenced study from Todorov et al. (2005), this issue was accounted for by only showing subject candidates they did not know.

Figure 2: Typology of Human Labeling Tasks on Two Axes: Subjectivity and Sensitivity to Labeler Identity



they recognize the politician as a member of their party (Homan, Schumacher, and B. N. Bakker n.d.).

Figure 2 presents a general typology of potential annotator tasks along two dimensions. The first dimension is the degree of subjectivity of the task. While no task will ever be completely objective, there are relative degrees of subjectivity. For example, tasks at a lower level of subjectivity include object recognition in image analysis. Answering the question "Are there police officers in this picture?" is less subjective than "Does this picture make you feel angry?" Simple counts ("How many people are in the picture?") or text descriptors ("Are there any pronouns in the text?") also have fairly objective answers. Reducing the subjectivity of tasks can be as easy as rewording the task question to add clarity. For example, "Is there a police officer in the picture?" could be made less subjective by adding more detail: "Is there a uniformed police officer in the picture (drawings or other artistic renderings of police do not count, nor do officers in plainclothes)?"

The second dimension in the Figure 2 typology is the degree of sensitivity to labeler

identity. That is, how much might we expect labelers from different demographic groups to perform differently on the task? For political scientists, an example of a relevant labeler characteristic is partisanship. There may be some tasks on some types of data where an annotator's political preference systematically affect their labels – Republicans may be better at identifying pictures of Republican politicians, for example, while Democrats may be better at identifying Democrats. This difference in responses based on identity is an example of labeler characteristic bias (LCB). In different contexts, different aspects of identity may become relevant. For example, when labeling texts for the presence of hate speech, the racial identity of labelers may affect whether or not they label instances of African American Vernacular English as hate speech (Sap et al. 2019). To be clear: the relevant identity characteristics will vary depending on the domain of the research, which encompasses the source of the data, the broad research question, and the specific labeling tasks. If the annotation task is to identify gender issues in campaign speeches, for example, male and female annotators may not agree on which issues are related to gender.

Tasks that fall into each quadrant of Figure 2 are subject to different forms of potential labeling bias and therefore to different solutions. In both the lower and upper left quadrants, the labeling tasks are not expected to vary significantly with labeler demographics. This would include, for example, the classic image task of identifying different animals. In these cases, it is possible to claim there is a real ground truth in the material that can be revealed by annotators. Moving from the upper left quadrant to the lower left quadrant (e.g. reducing bias attributed to task subjectivity) can be addressed with research design. That is, on these tasks labels that differ from the ground truth – labels that are incorrect or "biased" – can be attributed to issues with research practices. Wrong answers can be avoided in advance by reducing the ambiguity of the coding task, providing an extensive codebook with examples, or by training annotators, as our annotator studies illustrate. After labeling is done, errors can be corrected by filtering out "bad" labelers who did not read or view the materials the

way the researcher intended.

The tasks on the right hand side of Figure 2 have a strong potential for LCB. In these instances, the identities of the labelers may have an outsized impact on annotations. For example, the upper right quadrant of "high subjectivity, high sensitivity to identity" could include tasks like "How angry does this text make you feel?" If the material being labeled comes from political protests, labelers reported degree of anger may be impacted by whether they support the demands of the protesters. In short, their partisan identity or political ideology may bias their responses in ways that can impact research findings that rely on their labels.

Labeler bias is not a problem per se. We may in fact want to know whether partisanship impacts how people react emotionally to protests or political violence. The point is that researchers need to recognize the potential for labeler bias effects in downstream tasks, including supervised machine learning. If a task has no ground-truth, no amount of training will achieve high IRR. Although it may be difficult to predict which annotation tasks are subjective or prone to labeler bias, our meta analysis of articles provided examples of tasks for each category of the typology. An example of a task that was not subjective and had low identity-dependence (bottom left square in Figure 2) asked annotators to determine whether vote tallies had been altered in an image of vote tabulations (Cantú 2019). An example of an annotating task that was highly subjective, but with low potential for labeler bias (top left square in Figure 2), Baerg and Lowe (2020) asked annotators to judge whether a central bank statement was dovish or hawkish. Rating the importance of legislation (e.g. Boussalis et al. 2021) is likely highly subjective and also highly sensitive to identity (top right square in Figure 2). Judging whether tweets are election-related or make references to violence (Benoit, Munger, and Spirling 2019) does not seem very subjective on its face, but may be influenced by partisanship (or other characteristics), potentially leading to high labeler bias (bottom right square in Figure 2).

Another way of thinking about the information in the typology is in terms of opinion data. When surveyors ask citizens about their voting preferences, they do not expect respondents to provide the same answer. Instead, they are specifically interested in whether and how the demographics of respondents correlate with their responses. The typology above makes the point that some annotation tasks may be more like answering an opinion survey than recording ground truths. When this is the case, simply reporting IRR statistics without testing for systematic labeler bias may be a problem.

As an alternative way of thinking about the differences in annotation biases, Table 2 summarizes sources of annotation biases and the solutions to those biases. Here we organize entries based on the problem, the general source of the bias (annotator or researcher), what IRR indicator might show the presence of that bias, and potential solutions. The non-LCB biases are likely familiar to many researchers – speeding, inattentive annotators is a common problem – as are the solutions. The low IRR indicator is also likely familiar. In the presence of low initial IRR during piloting, researchers can tweak their forms, hire different annotators, etc. LCB can be indicated by stubbornly low IRR (as demonstrated in our text example in this paper). That is, even after trying other solutions that might solve the problem of low IRR, LCB-sourced biases remain. In the case of LCB, we have two proposed solutions in Table 2. If the demographic breakdown of the target population is known, we can create composite, reweighted scores for the variable of interest that match our annotator demographics to the population demographics. If the demographics of the target population are not known, we can take a subgroup approach and separately model the different demographics from our annotator sample. We explore both solutions in the image example below. Before turning to the image example, we briefly describe why gender and partisanship are potential demographic factors associated with LCB.

Table 2: Sources of Biased Labels and Solutions

Problem	**Source of bias**	**Indicator**	**Solution**
Distracted, speeding annotators	Annotators	Low IRR, fast annotation completion time	Attention checks; hire qualified/experienced/motivated labelers; filter responses completed too quickly; have ground-truth examples to filter respondents
Highly complex or unclear concepts for labeling	Researchers	Low IRR, incorrect responses compared to ground-truth	Pilot questions and forms; update materials based on early feedback; provide examples and training; have ground-truth examples to filter respondents
Task requires specific knowledge to complete	Researchers	Low IRR, incorrect responses compared to ground-truth	Hire knowledgeable coders; provide examples and training
Task vulnerable to LCB, underlying population demographic proportion known	Annotators	Stubbornly low IRR despite training and other measures	Collect labeler demographics and test for systematic variation in responses; reweight responses based on population proportions to create composite measures
Task vulnerable to LCB, underlying population demographic proportion unknown	Annotators	Stubbornly low IRR despite training and other measures	Explore the "why" of different labels; explicitly model differences in responses based on demographics

2.1 Partisanship and Gender as Sources of LCB

To our knowledge, no prior research has explicitly examined whether partisanship impacts large scale image and text annotation tasks. Yet the expectation that differences in terms of political identity can be important for perception is supported by a substantial literature on partisan differences. Ahn et al. (2014), for example, find significant differences in disgust responses by partisanship. Similarly, Schaffner and Luks (2018) and Bullock and Lenz (2019) (among many others) detect significantly different responses by partisanship on opinion surveys. Existing research also reports differing gender responses to a variety of treatments (Ksiazkiewicz, Window, and Friesen 2020; Deng et al. 2016), including stronger emotional reactions to treatments among women compared to men (Brown 2014; Deng et al. 2016). But this research also does not specifically examine image or text labeling tasks, as we do.

3 LCB in Image Annotation

Image analysis has long been of interest to political scientists but has generated increasing interest given the importance of social media in politics. Our first illustration addresses the potential issue of LCB for those studying the impact of images shared on social media on social movement mobilization. There is a well established literature on this subject (Casas and Webb Williams 2018; Tufekci and Wilson 2012; Kharroub and Bas 2015). Here we ask whether tweets promoting social movement political action are more likely to be retweeted if their accompanying images contain particular content or evoke particular emotions. We did not originally collect these data to explore LCB. However, because we collected demographic information about our labelers, the data are well suited to exploring whether labeler characteristics affect the labels assigned for both content and emotions. If they do, then we can also ask whether conclusions about the impact of images on retweets differs depending on whose labels we use.

Our images are drawn from tweets associated with left-leaning social movements in the United States. As mentioned above, researchers may not know in advance which labeler characteristics systematically affect their labels. We expected that gender and party identification might matter for our task, but we did not limit the demographic information collected to just these characteristics. Having more information (e.g. education, income, religion) allows for a more comprehensive consideration of potential confounders.

To build a dataset of images associated with social movements on social media, we collected tweets from January 2018 to mid-2019 by tracking the Twitter accounts of a wide range of US-based public affairs organizations (a full description of the data collection is available in online Appendix B). We then automatically collected tweets from any Twitter account that used any of the hashtags promoted by these organizations. Our focus here is restricted to a limited number of hashtags that we interpreted as mobilization attempts. To count as potentially mobilizing, a hashtag needed to be used in tweets that asked readers to engage in specific offline or online political action (see Appendix C for details).

The eleven hashtags we selected for the purposes of this study are all left-leaning and cover a range of issues. #familiesbelongtogether, #cleandreamactnow, #abolishice, and #nomuslimbanever focus on immigration. #Womenswave addressed women's rights while #uniteforjustice and #stopkavanaugh opposed Brett Kavanaugh's confirmation to the U.S. Supreme Court. #riseforclimate supported action on climate change. #Write4rights encourages people to write letters of support for political prisoners around the world. #Sayhername memorializes black women killed by police; and #endgunviolence advocates for gun control regulations. The full corpus includes about 650,000 deduplicated images. To lower labeling costs for the current analysis, we used an unsupervised visual clustering method (Peng 2020) to construct a stratified sample of about 7,500 images. This ensured that we sampled images across a wide array of topics and account popularity.

We used the Qualtrics panel service to recruit self-identified Republicans and self-identified

Table 3: Image Label Variables

Image Label	Measure Type	Reaction or Content?
Leader/celebrity	Binary	Content
Protest	Binary	Content
Social symbols	Binary	Content
Someone who looks like me	Agree-disagree, 5 point scale	Content
Humor	Binary	Reaction
Irony	Binary	Reaction
Emotions (hope, enthusiasm, pride, anger, resentment, bitterness, hate, worry, scared, afraid, disgust, sadness)	0-10 point scale	Reaction

Democrats (2,140 total respondents). All of the annotators were over 18, English speakers, and based in the United States. Prior to labeling, respondents answered a set of demographic and media use questions and had to pass an attention check. Each annotator then answered questions about 8 images associated with a hashtag (see Table 3).[4] The questions asked respondents about image content that might predict political mobilization, such as whether the image included any celebrities or leaders, a protest, social symbols (e.g. a flag), or someone who "looks like me." We were also interested in reactions to images, including whether the image was humorous or ironic and 10 evoked emotions: hope, enthusiasm, pride, anger, resentment, bitterness, hate, worry, scared, afraid, disgust, and sadness (Marcus, Neuman, and MacKuen 2000; Casas and Webb Williams 2018). The full survey of questions is available upon request from the authors. At least one Republican and one Democrat annotated each image.[5]

Because our responses were crowdsourced, we have many labelers but few data points for a given labeler, and little-to-no overlap between pairs of coders – as such we do not have traditional IRR statistics to report. However, we can test how average labeling responses differ between demographic groups. We can also test whether findings about the impact of

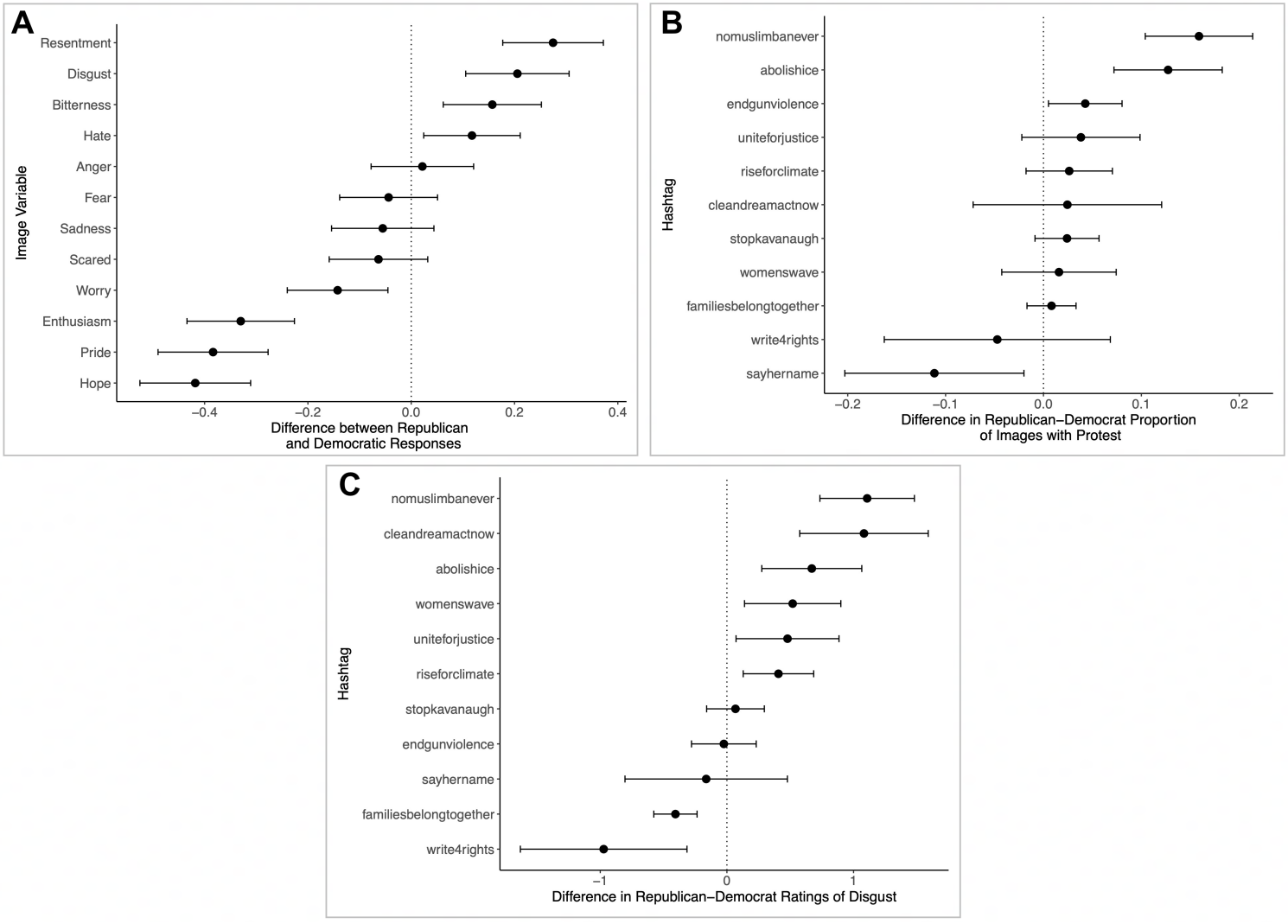
[4]Annotators could only complete the survey once per hashtag, but they could take the survey for multiple hashtags.

[5]In the rare cases where an image was labeled by more than one Republican or Democrat, we averaged scores for respondents of the same party.

different types of images on retweets varies depending on whose labels are used.

Figure 3.A shows the difference in the average of twelve emotional reactions to all of the images by partisan affiliation, with 95% confidence intervals around the differences in means. Points to the right of dashed line indicate stronger reactions from Republicans while points to the left indicate stronger reactions from Democrats. As we might expect for images drawn from tweets using hashtags shared by left-leaning organizations, there are significant partisan differences for several emotional responses to the same images. Democrats were more likely to respond that images elicited enthusiasm, pride, hope, and worry. Republicans were more likely to respond that images elicited disgust, resentment, bitterness, and hate. We see no significant differences in sadness, scared, fear, or anger. The largest difference is about 0.4 on an 11-point scale, which while not extremely large (it represents about 10% of a standard deviation for the emotions tasks) is still notable.

Figure 3: **Differences in Image Annotation by Partisanship:** The average differences between Republicans and Democrats in labeling images from different hashtags, with 95% confidence intervals around the differences in means. Panel A displays the average partisan difference of twelve emotional reactions to all of the image from all hashtags. Panel B displays differences between Republicans and Democrats in the proportion of images where labelers saw a protest, by hashtag. Panel C highlights differences in disgust between Republicans and Democrats by hashtag. Points to the right of the dashed line indicate stronger emotional reactions or more frequent identification of protests by Republicans.



More unexpected than the differences in reactions to the images are differences in content. For example, Figure 3.B shows the differences in the rates of images where Republican and Democratic labelers saw protest. The figure breaks down the differences by hashtag

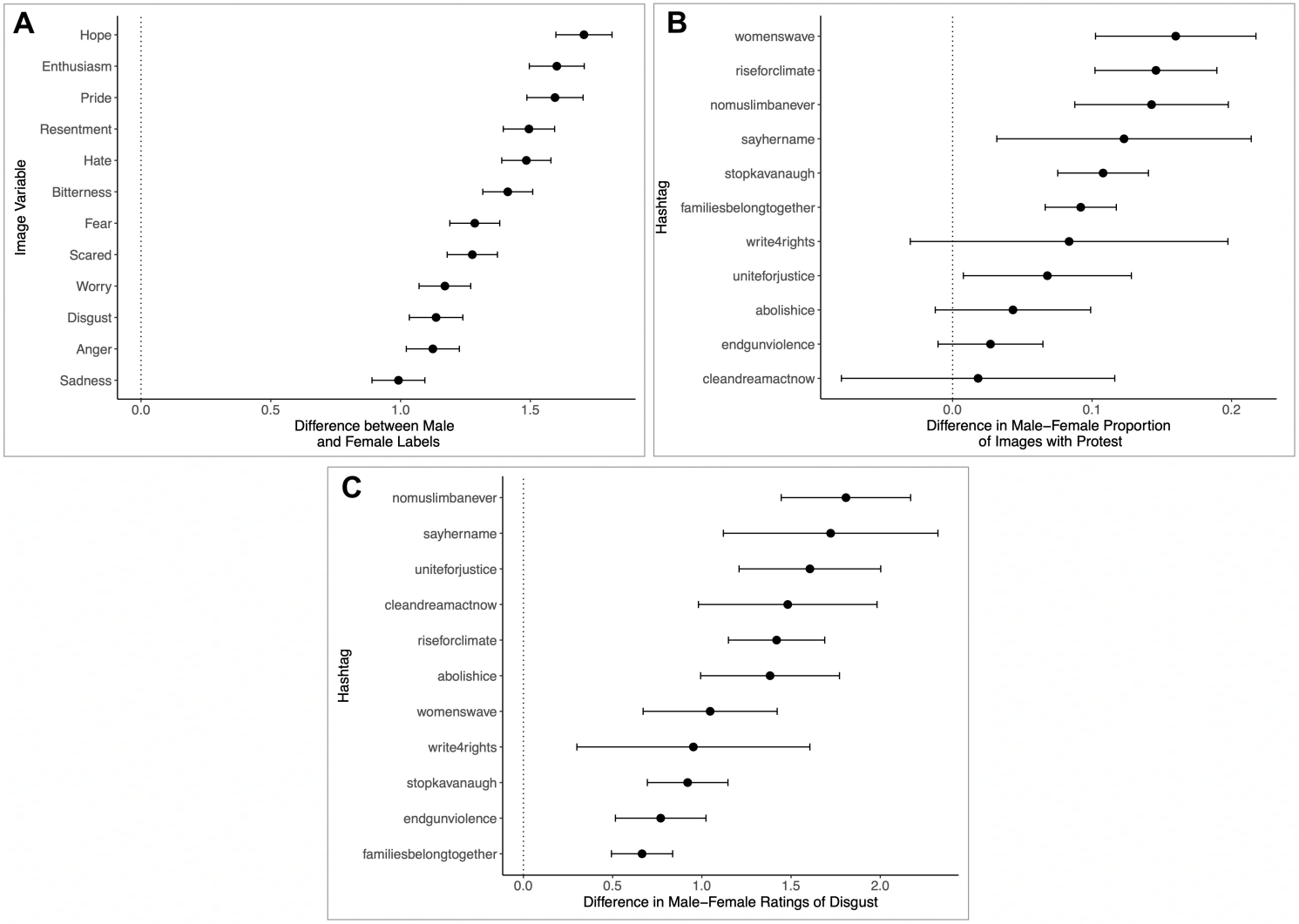
to show that the differences in seeing protest are not uniform across social movement content. Democrats saw protests more often in images associated with the #sayhername hashtag. Republicans saw protests more often in images associated with the #nomuslimbanever, #abolishice, and #endgunviolence hashtags.

In the interest of space, we will not discuss differences for all of the possible annotations and hashtags. However we will mention the results for an emotion that has been a particular interest of political scientists in recent years: disgust (Kam and Estes 2016; Aarøe, Petersen, and Arceneaux 2017; Ksiazkiewicz, Window, and Friesen 2020; Ahn et al. 2014). Figure 3.C indicates that Republicans reported higher rates of disgust than Democrats when viewing images associated with the hashtags #nomuslimbanever, #cleandreamactnow, #abolishice, #womenswave, #uniteforjustice and #riseforclimate. Democrats reported higher rates of disgust when viewing images associated with the hashtags #familiesbelongtogether and #write4rights. There are no significant partisan differences for remaining three hashtags. While we can only speculate as to why these particular hashtags elicited such varied reactions, it is interesting to note that there is not one type of movement that had uniformly more (or less) disgust. Three of the four hashtags relating to immigration (#nomuslimbanever, #cleandreamactnow, and #abolishice) had higher rates of disgust from Republicans. But the fourth, #familiesbelongtogether, had higher rates of disgust from Democrats. Exploring the "why" of these differences is an important area of future research.

Whereas the partisan effects above vary by hashtag and image content, the differences in labels assigned by men and women are clear and consistent. Across the board, the men report stronger emotional reactions (Figures 4.A and 4.C). Men were also more likely to see protests in images (Figure 4.B). Given these consistent differences, it would be relatively straightforward to standardize responses between men and women to produce one "true" response, for example, by systematically assigning less weight to emotional scores from men or more weight to the scores of women. However, we should also not lose sight of the

substantive finding and the question it raises for future study: why do these emotion labeling differences exist between men and women?

Figure 4: **Differences in Image Annotation by Gender:** The average differences between men and women in labeling images from different hashtags, with 95% confidence intervals around the differences in means. Panel A displays the average gender difference of twelve emotional reactions to all of the image from all hashtags. Panel B displays differences between men and women in the proportion of images where labelers saw a protest, by hashtag. Panel C highlights differences in disgust between men and women by hashtag. Points to the right of the dashed line indicate stronger emotional reactions or more frequent identification of protests by men.



Ultimately, we are interested in whether image content and reactions lead to more or less

mobilization (as indicated by retweets). Would our conclusions differ depending on whose labels we used? Put another way, what if we were not attentive to our annotators' demographics? Here we compare linear regression results where the dependent variable is the logged number of retweets a message received at least two weeks after it was first posted.[6] We only consider tweets with labeled images. Each regression includes the same full range of potentially-relevant image label variables (evoked emotions, presence of protest, etc. (see Table 3). However, following prior research on the mobilizing role of emotions (Marcus, Neuman, and MacKuen 2000; Casas and Webb Williams 2018) we collapse emotional responses onto three main dimensions: *Enthusiasm* (hopeful, enthusiastic, proud); *Aversion* (angry, resentful, bitter, hateful); and *Anxiety* (worried, scared, and afraid). Our control variables include the number of account followers; the time of day of the tweet; the day of the week of the tweet; and the type of tweet (original, retweet, or quote tweet), as well as fixed effects for hashtag. Of interest is what happens when we vary the labels used for the image-feature variables in Table 3. Our five regressions consider pooled labels (all labels), Democrats' labels only, Republicans' labels only, men's labels only and women's labels only. Does the variation in labelers lead to different conclusions about the mobilizing power of image content as measured by retweets?

These are toy models intended to demonstrate that a scholar using this common approach to quantitative modeling would potentially come to different conclusions based on whose image labels were collected. We are not trying to draw definitive conclusions about associations between the images and mobilization. As a check that our toy models with all possible variables are not exaggerating coefficient sensitivity, we include in Appendix E two alternative model specifications. The first alternative includes the control variables listed above and only image variables deemed to be about "content": showing protest; a leader;

[6]Because many tweets were deleted between initial collection and the two-week check for retweets, the number of observations for this analysis drops to just over 3,600.

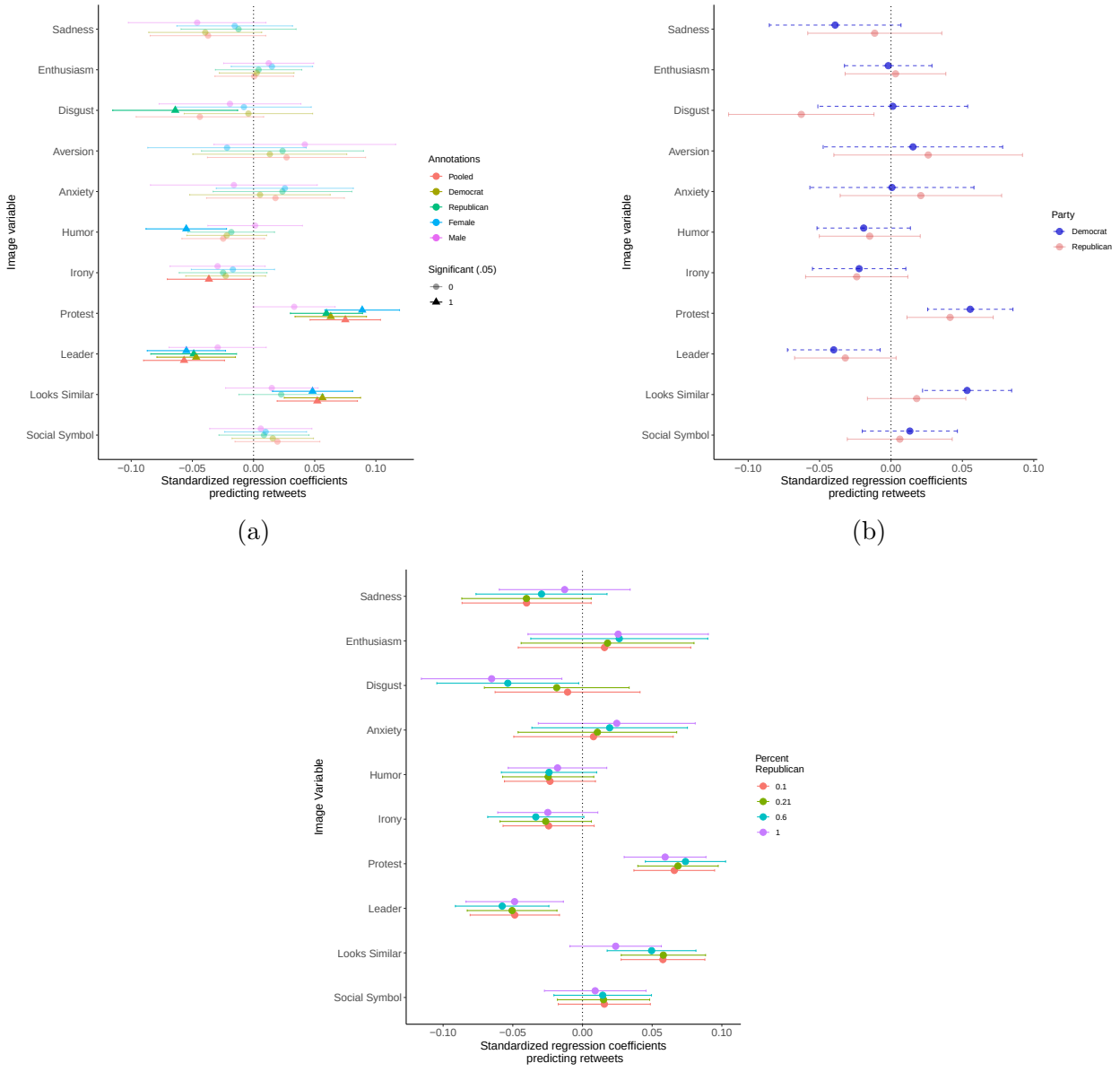
someone who looks like the labeler; or a social symbol. The second includes the control variables and image variables deemed to be about "reactions": the emotions variables; humor; and irony. As with the fully-specified model in Figure 5.A, we see coefficients changing based on whose labels are included in these alternative specifications. The alternative specifications also allay concerns about whether the content and reaction variables should be included in the same model. As the content arguably is what drives the reactions (e.g. seeing a protest could be what makes the respondent feel scared), it may not be appropriate to include both sets of variables in a single model.

Figure 5.A reports the standardized regression coefficients for the image-features in the toy models with all controls and all image variables (full regression tables are available in Appendix D). Coefficients that are statistically significant at the 0.05 level are represented as darker triangles.[7] Importantly, the signs of some coefficients flip depending on whose labels are used (e.g. for Aversion), as does significance (that is, some coefficients that are statistically significant with one set of labelers lose their significance with a different set, e.g. Humor). If we were less attentive to who is doing the labeling, we might come to very different conclusions about the mobilizing effects of image features. For example, the positive and significant association between protest images and retweets disappears if we rely solely on male annotators. These different significance findings are very unlikely to be a function of sample size, as in all these models the number of observations is roughly the same.

In general, the coefficients for the variables where we observed larger differences between labelers are the most sensitive to whose labels are used in downstream analyses. We see differences for the disgust coefficient, for example. The regression coefficients signs for anxiety and aversion flip depending on whose labels are used, though none of the coefficients are statistically significant. For disgust, there is a statistically significant and negative effect for

[7]Standardized regression coefficients and confidence intervals generated using the betaDelta package in R Pesigan, Sun, and Cheung (2023).

Figure 5: **Regression Results Vary by Annotator Demographics**: These figures present standardized regression coefficients from linear regressions predicting the logged number of retweets a tweet received when different compositions of annotators are used to label images. Panel A presents the standardized regression coefficients from five regressions, each with a different set of annotators (all annotators, Democrats only, Republicans only, females only, males only). The image variables are on the y-axis. Statistically significant coefficients are represented by solid triangles. Panel B presents standardized regression coefficients from a model that included Democratic labels (blue) and Republican labels (red) as separate variables, with 95% confidence intervals. Panel C displays the standardized regression coefficients from the same model as Panel A, showing only a small subset of variables. Here each variable is a composite score combining the Republican and Democratic labels, with varying weights for the responses.



(a) (b)

(c)

Republican labelers, but the effect is not significant for any other annotator subgroup.

Figure 5.B suggests a modeling strategy solution in the absence of data on the true demographic variation in the target population (see Table 2). We can report separate coefficients for the reactions of Republicans versus Democrats or for women and men.[8] That is, we can model the variable effects on subgroups within the labelers. In Figure 5.B increased levels of Republican disgust are associated with fewer retweets(full regression table available in Appendix D). Seeing someone who "looks like you" in the picture has a positive association with retweets for Democratic labelers but not for Republicans. Protest, interestingly, has a significant positive coefficient for both Republican and Democratic labelers.

If we do know the underlying demographic breakdown in the target population, we could consider an alternative solution to the issue of LCB. For example, the target population in this toy analysis is Twitter users – we want to know, based on our sample, how Twitter users respond to social movement images. Because we have demographic information associated with the labeling responses, we can reweight our annotations to create a composite label score that approximates the population of interest. Of course, the challenge here is knowing what the true population demographics are. The actual proportion of US Twitter users who are Republican or Democrats is a moving target. For illustrative purposes, we use a 2019 estimate from Pew Research that put Republicans at 21% of Twitter users (Wojcik and A. Hughes 2019). We weight our image labels based on that proportion to create a composite score for each image variable, assuming that that the remaining Twitter population is all Democrats.[9] After generating the composite score (e.g. *composite enthusiasm* = .21**Republican enthusiasm score* + .79**Democratic enthusiasm score*), we rerun the main model

[8] An alternative approach to the subgroup analysis strategy would be to use a model to correct for measurement error in the labels (X. Chen, Hong, and Nekipelov 2011). However, these models often assume that there is a "true" measurement, with bias representing deviation from that truth. In our example, it may be the case that there simply is no one "truth."

[9] This simplifying assumption for demonstrative purposes is clearly wrong – many Twitter users identify as independents but we do not have annotation data for Independents and thus cannot take that into account for our composite score.

with only the single, composite measure for each of the variables of interest. As a sensitivity analysis, we reweight the responses thrice more to generate alternative composite measures, with the Republican proportion set to 0.1, 0.6 and 1.0. In Figure 5.D we show the results of the same regression analysis as before that now uses the composite measures. Depending on how we weight the responses to generate the composite score, we see different regression coefficients in terms of statistical significance for disgust and "looks similar." In general, though, the coefficients from the reweighted variables are fairly stable. As this solution to LCB demonstrates, it is possible to build single, composite scores that account for differences in labelers by weighting responses to match the underling population demographics. However, as we have shown, this solution is sensitive to what population proportions are used. If researchers are not confident that they know the true population demographics, generating a composite measure is more risky than modeling the subgroup effects separately, as in the first solution.

4 LCB in Text Annotation

The image-annotation study, where the data were not initially collected to explore LCB, led to a research design that more systematically assesses LCB. In particular, building on the framework outlined in Figure 2, we wanted to explore: (a) whether higher IRR is more easily achieved for objective and non-identity dependent tasks (versus subjective and/or identity-dependent tasks); (b) whether we would observe systematic annotation differences for labelers of different identities (e.g. ideology and gender) – and for which annotation tasks; and (c) whether these group differences could be mitigated with further training. In addition, we were interested in a different setting - in this case, text rather than images, undergraduate research assistants rather than crowd workers; party/elite communications rather than social movements, training interventions versus no interventions, and the Netherlands instead of

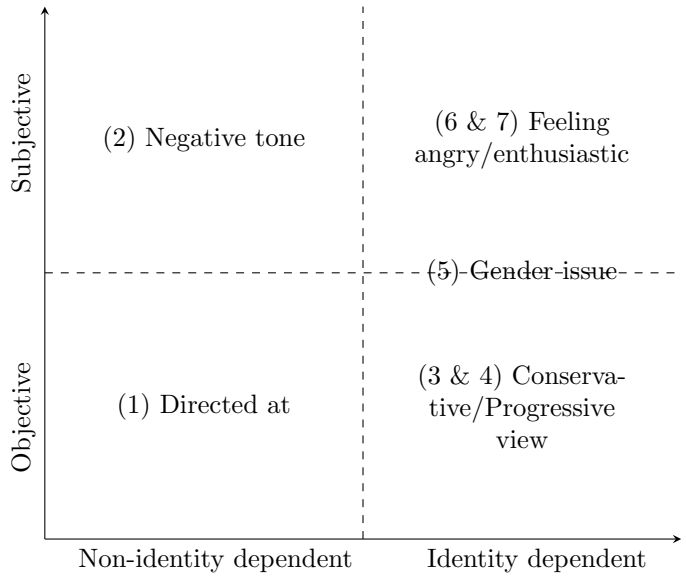
the U.S.

Table 4: Seven Text-Annotation Tasks

Task	**Description**
(1) Directed at	Is this message directed at another person, party, group, company, or organization?
(2) Negative tone	Does the message use a negative tone or criticizes a person, party, group, or organization?
(3) Conservative view	Does the message reflect or contain a conservative (i.e. right-leaning) view?
(4) Progressive view	Does the message reflect or contain a progressive (i.e. left-leaning) view?
(5) Gender issue	Does the message discuss a gender issue?
(6) Feeling angry	Do you feel some anger when reading this message?
(7) Feeling enthusiastic	Do you feel some enthusiasm when reading this message?

We recruited a pool of 23 undergraduate students from a Dutch university. In a pre-survey, they provided information about the two individual-level characteristics: their ideology (15 progressive, 8 conservative students) and gender (13 female, 10 males). Then they participated in 3 coding sessions of 3-hours each. In each session, all 23 coders annotated the same set of 150 social media messages (either a Twitter, Facebook, or Instagram post) sent by Dutch politicians during the 2021 electoral campaign. These messages were selected from a larger collection of all messages from all Dutch politicians sent on these three platforms during the campaign. The politicians of interest in the annotation set belonged to one of two progressive parties: GroenLinks (GL, N = 37 messages) and Labour Party (PVDA, N = 38); or to one of two conservative parties: People's Party (VVD, N = 38) and the Party of Freedom (PVV, N = 37). We first selected a random sample of messages sent by politicians from the four parties. We then manually selected 150 messages to ensure enough positive cases for each of the annotation tasks described below. We masked all names, hashtags, and handles referencing a politician, party, or organization to prevent participants from relying on clear partisan cues when performing the annotations. For each session, the 150 messages were randomly sorted (and so annotated) in a different order.

Figure 6: **Text Labeling Tasks on the Subjective/LCB Axes**



As illustrated in Figure 6, we identified a set of annotation tasks that varied in their levels of objectivity and identity-dependence (Table 4). First, the more objective and non-identity dependent task ("Directed at") simply asked annotators to indicate whether the social media message was directed at someone. Second, a more subjective and non-identify dependent task ("Negative tone") asked participants to indicate whether the message had a negative tone. Third, a more objective and potentially more identity-dependent task asked them to indicate whether a message contained a "Conservative view" and/or a "Progressive view" (non-mutually exclusive). Fourth, a somewhat more subjective and identity-dependent task asked them whether a message discussed a "Gender issue." The assumption here is that progressives and conservatives, or women and men, may have different perceptions of what constitutes a progressive/conservative statement, or issues that are gender-related. Finally, we assigned two highly subjective and identify dependent tasks - whether they felt "angry" and/or "enthusiastic" (non-mutually exclusive) when reading a progressive message.

To assess whether additional training might improve IRR across these different types of tasks, we provided more instruction and training in each of the three sessions. At the begin-

ning of the first "Basic" coding session, participants were only given the questions/prompts described in Table 4. Limited instructions are common in projects that rely on crowdsourcing services, such as Mechanical Turk. In the second "Intermediate" session, the annotators were provided with (and asked to read) a codebook (available from the authors by request) with detailed instructions about how they were to complete each task. In the last "Advanced" session, the administrators spent 45 minutes discussing how to complete the tasks using 15 example messages, and answering outstanding questions about the codebook (none of the messages to be coded were discussed). These last two sessions emulate situations where researchers work closely with their own team of research assistants. The question is whether additional training of this type has any impact on IRR for the different tasks.

We expected IRR to be higher between respondents of the same (versus different) ideology, and of the same gender; particularly for the more subjective and (ideology-)/(gender-)identity dependent tasks. We also expected IRR, both overall as well as between pairs of coders of different identities, to improve in each round of training, particularly for the more objective and less identity-dependent tasks. A blinded pre-registration for this study is available at the following **link**. We will mostly focus on annotation differences by ideology because they are the most interesting findings in this particular study (results by gender are available in Appendix G). Importantly, not all of our expectations were supported, which underscores the value of actually testing for annotator differences rather than assuming that they do or do not exist.

We used Cohen's Kappa to measure IRR_{ijsz} for each unique pair ij of coders, annotation session s, and annotation task z. Figure 7 sorts the tasks by their level of objectivity and identity-dependence, so that on the left we have the task we expected to be most objective and less identity-dependent. In Figure 7.A, with two notable exceptions ("Directed at" and "Gender issue"), we observe IRR to be higher on average for more objective ("Conservative view" 0.39 in the first session; "Progressive view" 0.4) and less identity-dependent tasks

("Negative tone" 0.57), than for subjective and identity-dependent tasks ("Feel angry" 0.35, "Feel enthusastic" 0.3). Contrary to our expectations, IRR for what we assumed was the most objective and least identity-dependent task ("Directed at"), was exceptionally low: 0.15. In follow up conversations, some coders said that they struggled in deciding whether a message was directed at someone if a person or organization (or their social media handle) was simply mentioned. Also contrary to our expectations, the highest IRR (0.75) is observed for the "Gender issue" task. In retrospect, Dutch politicians' gender related messages were typically very explicit, perhaps not leaving much room for subjective interpretation of whether issues discussed in messages were gender-related or not.

Figure 7: **Results from the Text-Annotation Exercise:** Panel A presents the average Cohen's Kappa measure between unique pairs of coders at each training level and all of the labeling tasks. Panel B presents the overall improvement in Cohen's Kappa measure between the basic and advanced training sessions for all of the labeling tasks. Panel C presents the difference in the average Cohen's Kappa measure for pairs of coders with the same ideology and pairs of coders with different ideologies at each training level and all of the labeling tasks.

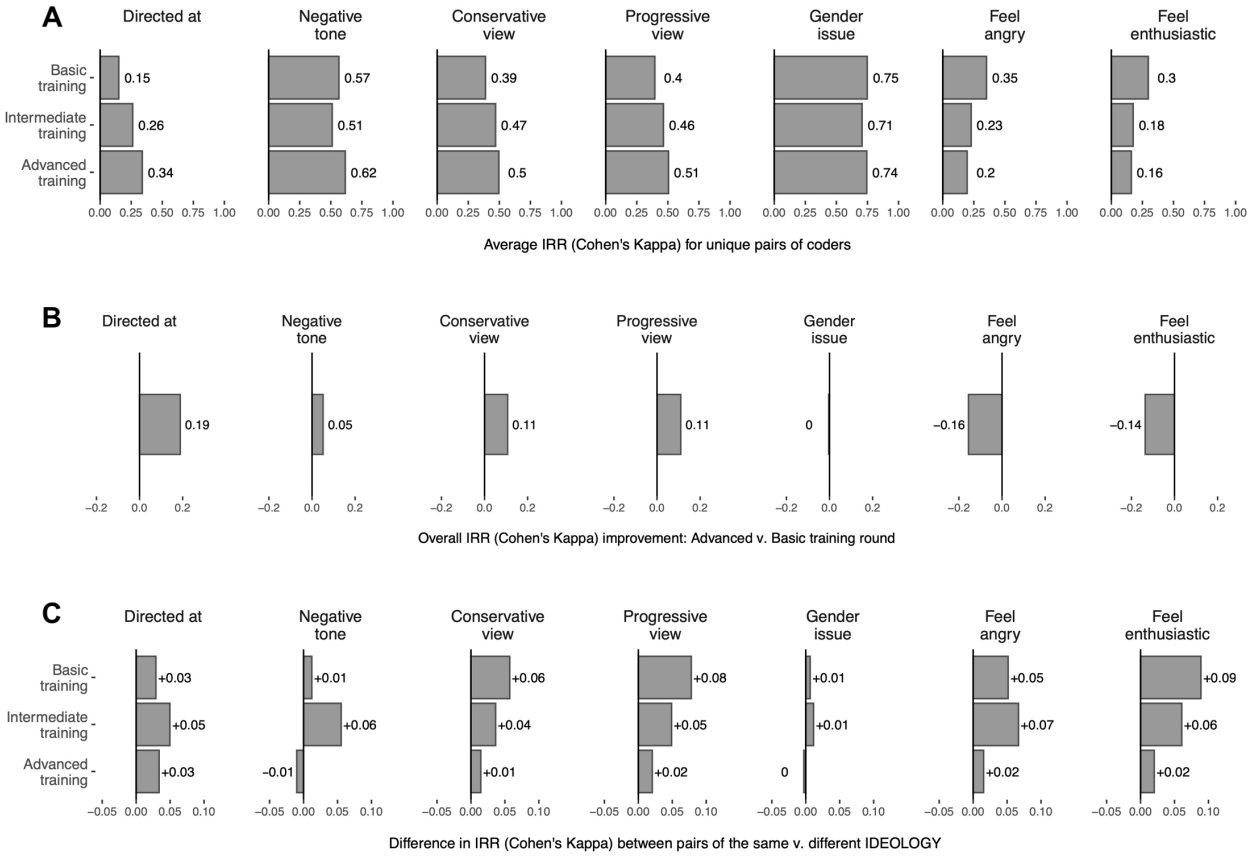


Figure 7.B displays changes in IRR for the different tasks between the "Advanced" and "Basic" training rounds (the difference between the last and first bars in Figure 7.A). As expected, IRR improvement is substantially larger for the most objective and non-identity-dependent task ("Directed at", +0.19). Also as expected, the more subjective and identity-dependent the task, the smaller the improvement in IRR from additional training: more subjective ("Negative tone", +0.05); more identity-dependent ("Conservative view" +0.11,

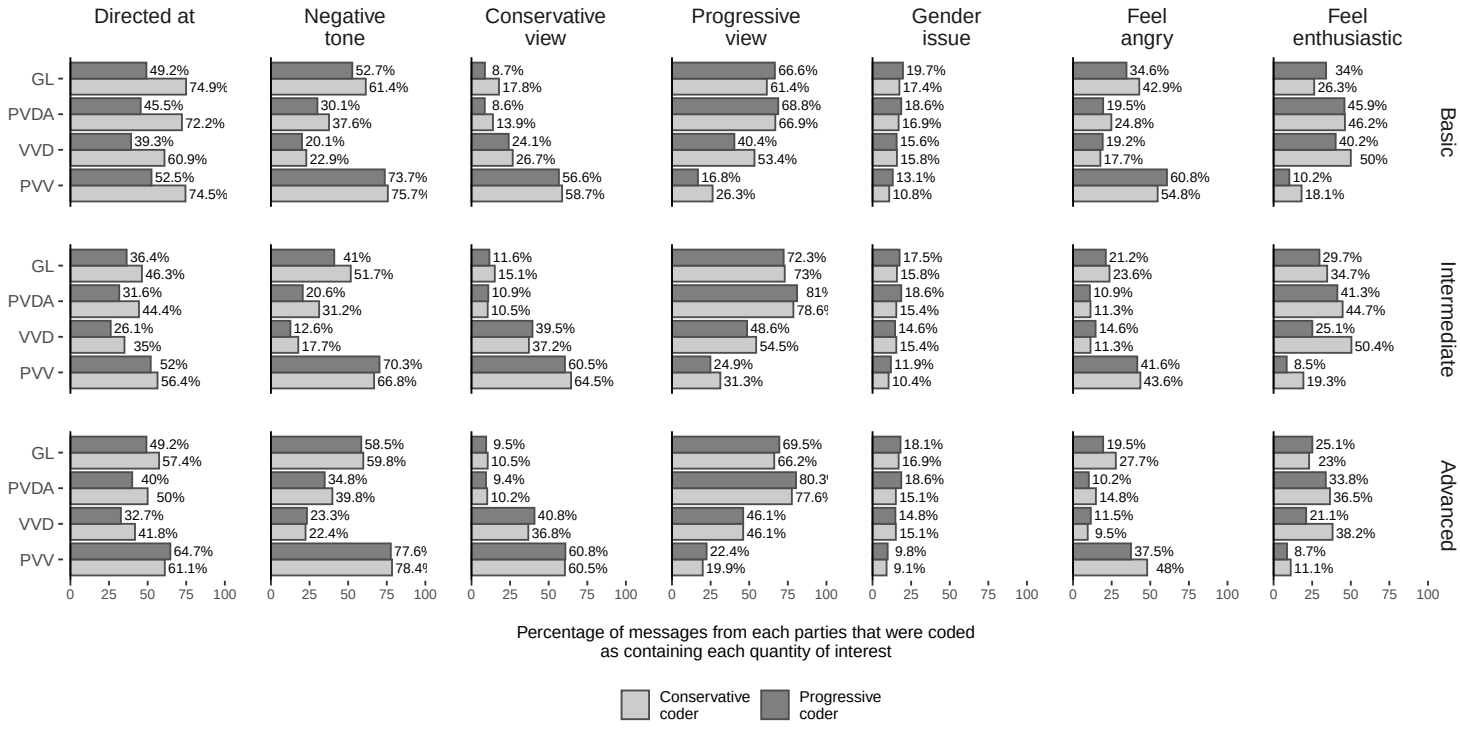
"Progressive view" + 0.11); and more subjective and identity-dependent ("Feel angry" -0.16, "Feel enthusiastic" -0.14) tasks.

Finally, we compared IRR_{ijsz} for pairs of coders of the same versus different ideologies. In Figure 7.C higher values indicate relatively higher IRR among coders of the same ideology compared to coders of different ideologies. As expected, in the first "Basic" round of coding, this IRR difference is greater for (ideological-)identity-dependent tasks ("Conservative view" +0.06, "Progressive view" +0.06, "Feel angry" +0.05, "Feel enthusastic" +0.09), compared to the most objective and non-identity-dependent task ("Directed at" +0.03). Figure 7.B, indicated that additional training improved IRR for most annotation tasks (with the exception of the two most subjective and identity-dependent tasks). Figure 7.C indicates that training also reduced average IRR difference between coders of the same and different ideologies. In the final "Advanced" round of coding, the time spent training coders helped to mitigate LCB and lower IRR even in more identity-dependent tasks ("Conservative view" (+0.01) and "Progressive view" (+0.02)). In Appendix G we show that these descriptive findings hold in a regression framework where we use linear regressions to predict IRR_{ij} as a function of whether a given pair ij is of the same gender and ideology, the training session s, and the annotation task z (where we include random intercepts for each unique pair ij to account for the nested structure of the data).

Mirroring what we did for the images study, in Figure 8 we briefly discuss some effects the LCB shown in 7 can have on downstream analyses. We posit a general research question of: "How does messaging differ between Dutch parties leading up to an election?" Our downstream task is to report the descriptive results of differences in messaging, as measured by our seven annotation tasks. As mentioned above, in each round the participants annotated an equal number of messages from two progressive (GL and PVDA), and two conservative (VVD and PVV) parties. In Figure 8 the parties are sorted by ideology, with the most left-leaning one at the top. Keep in mind that the final 150 messages were not sampled at

random, and so that no meaningful substantive conclusions can be drawn from this analysis. The only purposes is to use this toy dataset to discuss potential effects of LCB on downstream analyses more generally.

Figure 8: **Downstream Analysis of the Annotated Messages, Based on whose Annotations are used:** The percent of messages from each political party that are coded as containing each quantity of interest by progressive and conservative labelers.



First, we observe, particularly in the first "Basic" round of coding, many meaningful differences between the ratings of conservative and progressive participants. For example, conservative respondents annotated more messages from progressive parties as having a negative tone (e.g. 61.4% of GL's messages, versus 52.7% annotated by progressive), as containing more conservative views (e.g. 17.8% for GL, versus 8.7%) and fewer progressive views (e.g. 61.4% v. 66.6%), and as talking less often about gender issues (e.g. 19.7% for GL, v. 17.4%). In the "Basic" round we also see that conservative (versus progressive) participants more often rate the messages as being directed at someone (e.g. 74.9% v.

49.2%), which we did not expect. Finally, we observe these ideological labeling differences mostly washing away in the "Advanced" coding round, after the annotators had received two rounds of training. In regards to the previous examples, conservatives (versus progressive) participants indicated that GL used a negative tone in 59.8% (versus 58.5%) of messages, portrayed conservative and progressive views in 10.5% (versus 9.5%) and 66.2% (versus 69.5%) of their messages, and to mention a gender issue in 16.9% (versus 18.1%). Contrary to our initial expectations, as they received further training, participants harmonized their criteria even in these more ideology-identity-dependent tasks. However, in line with our expectations, we still observe meaningful ideological differences in the annotations of the "Advanced" coding round when it comes to the most subjective and identity-dependent tasks (feeling angry/enthusiastic): for example, conservative participants still feel much more enthusiastic about the posts from the VVD (38.2% versus 21.1%), and much more angry when reading the messages from GL (27.7% versus 19.5%).

The text illuminates two points about human annotation in general and LCB in particular. First, it emphasizes the difficulty of knowing ex ante which coding tasks will be susceptible to LCB (e.g. we see surprisingly high LCB for the "Directed at" task). Our proposed typology helps researchers think through the potential sources of bias in human annotation tasks, but the text example demonstrates that researchers may not correctly understand which quadrant their task fits into. Here our suggestion for the problem is to recruit a diverse pools of annotators to see if there is indeed LCB present. Alternatively, researchers should specify very clearly what their population of interest is and ensure that labeler demographics reflect that population.

A second point of note is that coder training can make a difference in improving IRR, even if it cannot entirely remove LCB. Often researchers rely on crowdsourcing services and labelers with minimal training, as we did in the image example. In this case, researchers often disregard "bad" responses to improve IRR. However, this practice may simply remove from

the crowdsourced pool of annotators those with different sociodemographic backgrounds. This can lead to a homogeneous pool of coders with a high IRR but annotations that are systematically biased. The results in the text example suggest intentionally building teams of research assistants you can work with closely, and gradually train, rather than getting quick labels via a crowdsourcing service.

Where LCB persists even with good training, our two general solutions still apply. A researcher could create composite measures by weighting labeler responses to match population demographics. Or we can separately model the annotations from different subgroups of coders.

5 Discussion and Conclusion

Human annotation has long been an important research tool in political science. Political scientists are well aware of and attentive to concerns about validity and reliability of coding results. These concerns are increasingly important as the field turns to machine learning and "big data" tools. An algorithm trained on biased data will reproduce and often exacerbate that bias (see, e.g. Bolukbasi et al. 2016).

Recent studies in machine learning point to individual characteristics of labelers, such as their identities and personal views, as potentially having a strong impact on data annotations (Gordon et al. 2021; Hube, Fetahu, and Gadiraju 2019). Our fundamental point is that political scientists need to be attentive to such potential labeler-characteristic biases. Unfortunately, the tests that political scientists currently rely on to assess coding performance - in particular overall IRR - may not address such concerns.[10] In two demonstration studies (one involving image labeling and one text), we showed how demographic variation in labelers can significantly impact label assignments and, potentially, downstream analyses.

[10] A similar point is made by Grimmer, King, and Superti (2015).

For example, it makes intuitive sense that political identities could lead labelers to interpret images or messages differently. But, perhaps surprisingly, we also find that labelers with different political identities disagree on basic questions such as what is in a picture or the target of a message. Republican and Democratic annotators were not equally likely to see a protest in an image, and Dutch students of different ideologies had different rates of stating that a message was directed at someone else. Less surprisingly, Republicans reported higher rates of disgust, resentment, and bitterness (on average) when viewing images associated with left-leaning causes, while Democrats reported higher rates of enthusiasm, pride, and hope. In the text study, political ideology also affected how coders responded emotionally to politicians' messages.

Some of our initial predictions were not supported or had findings that went in the opposite direction of what we originally expected. In our view these findings make it even more imperative that researchers collect demographic information as part of the labeling process and test for systematic labeling bias. We can not assume, as is common practice, that labeler biases are unimportant or that we can predict which tasks they might affect. Researchers need to prepare for the possibility of unexpected LCB. For example, we found important gender differences in image labeling for US social movement tweets, but very few gender differences in the labeling of Dutch politicians' messages. It seems unlikely that we can generalize about when gender differences will introduce bias, or many other demographic characteristics for that matter. The concern is that failing to account for LCB will bias our results, whether that is predicting the labels of other cases in machine learning, or drawing conclusions about the importance of variables in a regression analysis.

Addressing LCB starts with a consideration of who the relevant population is for a given study. If, for example, we were only interested in the impacts of social movement images on Democrats, then we would be justified in recruiting an all-Democratic pool of labelers. However, we could not then claim that the extracted image content reflected a universal

truth about the images. Instead we would need to be transparent about the annotator pool and our ability to draw conclusions about image effects only within that population.

Where systematic labeling differences are found, researchers have a choice. They can adjust labels to represent one ground truth, or they can see this as an opportunity to better understand the complex nature of their research questions. For example, this study found that male image labelers gave systematically higher ratings for both the evoked emotions and for seeing protest. We could respond to that difference by lowering the ratings for males and raising the ratings for females to get an "average" true response. Or, as we did in light of the partisan differences, we can separately model the heterogeneous image effects on Democrats/Republicans and men/women. An alternative solution is to reweight composite annotations to match the makeup of the target population, similar to reweighting survey responses.

Labeler-characteristic bias raises new and intriguing questions for future study, such as why the differences occur and which images or text factors drive the differences in reactions. Asking how and why content is perceived differently may help us to better understand polarization and social movement mobilization, for example. Figure 9 was annotated by two Republicans and two Democrats in the study. Both Republicans said it did not show a protest, while both Democrats said that it did. What is it about the image that created such different opinions? Future work will explore the divide that LCB has revealed.

In short, a serious treatment of LCB will lead to new studies and a reconsideration of established work. We suggest the following best practices to address LCB: first, recruit a diverse annotator pool, at a minimum for pilot testing. Researchers should collect data on a range of labeler characteristics, including subgroups that the researchers may not have thought about as relevant ex ante. Second, researchers should test if there are differences in annotations between labeler groups – a simple test of differences in mean responses between subgoups can be telling. Third, where differences are found, we should try to improve our

Figure 9: Democrats Saw Protest; Republicans Did Not

labeling practices to see if it improves IRR, model subgroup effects, or adjust labels for subsequent analyses.

6 References

Aarøe, Lene, Michael Bang Petersen, and Kevin Arceneaux (2017). "The Behavioral Immune System Shapes Political Intuitions: Why and How Individual Differences in Disgust Sensitivity Underlie Opposition to Immigration". *American Political Science Review* 111.2, pp. 277–294.

Agrawal, Amritanshu, Wei Fu, and Tim Menzies (2018). "What is wrong with topic modeling? And how to fix it using search-based software engineering". *Information and Software Technology* 98, pp. 74–88.

Ahn, Woo Young et al. (2014). "Nonpolitical images evoke neural predictors of political ideology". *Current Biology* 24.22, pp. 2693–2699.

Baerg, Nicole and Will Lowe (2020). "A textual Taylor rule: estimating central bank preferences combining topic and scaling methods". *Political Science Research and Methods* 8.1, pp. 106–122.

Barberá, Pablo et al. (2021). "Automated Text Classification of News Articles: A Practical Guide". *Political Analysis* 29.1, pp. 19–42.

Benoit, Kenneth, Drew Conway, et al. (2016). "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data". *American Political Science Review* 110.2, pp. 278–295.

Benoit, Kenneth, Michael Laver, and Slava Mikhaylov (Apr. 2009). "Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions". *American Journal of Political Science* 53 (2), pp. 495–513.

Benoit, Kenneth, Kevin Munger, and Arthur Spirling (2019). "Measuring and explaining political sophistication through textual complexity". *American Journal of Political Science* 63.2, pp. 491–508.

Bolukbasi, Tolga et al. (2016). "Debiasing Word Embedding". In: *30th Conference on Neural Information Processing Systems*. NIPS 2016, pp. 1–9.

Boussalis, Constantine et al. (2021). "Gender, candidate emotional expression, and voter reactions during televised debates". *American Political Science Review* 115.4, pp. 1242–1257.

Brown, Catherine C (2014). "Gender Difference in Emotional Reactions to Media : Examining Self-Report During Bittersweet Video Clips".

Budak, Ceren, Sharad Goel, and Justin M. Rao (Apr. 2016). "Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis". *Public Opinion Quarterly* 80.S1, pp. 250–271.

Bullock, John G. and Gabriel Lenz (2019). "Partisan bias in surveys". *Annual Review of Political Science* 22, pp. 325–342.

Cantú, Francisco (2019). "The fingerprints of fraud: Evidence from Mexico's 1988 presidential election". *American Political Science Review* 113.3, pp. 710–726.

Casas, Andreu and Nora Webb Williams (2018). "Images that Matter: Online Protests and the Mobilizing Role of Pictures". *Political Research Quarterly* 72.2, pp. 360–375.

Chen, Xiaohong, Han Hong, and Denis Nekipelov (2011). "Nonlinear Models of Measurement Errors". *Journal of Economic Literature* 49.4, pp. 901–937.

Chen, Yunliang and Jungseock Joo (Aug. 2021). "Understanding and Mitigating Annotation Bias in Facial Expression Recognition". In:

DeBell, Matthew (2017). "Harder Than It Looks: Coding Political Knowledge on the ANES". *Political Analysis* 21.4, pp. 393–406.

Deng, Yaling et al. (2016). "Gender differences in emotional response: Inconsistency between experience and expressivity". *PLoS ONE* 11.6, pp. 1–12.

Denny, Matthew J and Arthur Spirling (2018). "Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it". *Political Analysis* 26.2, pp. 168–189.

Dolezal, Martin et al. (2016). "Analyzing Manifestos in their Electoral Context A New Approach Applied to Austria, 2002–2008". *Political Science Research and Methods* 4.3, pp. 641–650.

Gamm, Gerald and Thad Kousser (2010). "Broad Bills or Particularistic Policy? Historical Patterns in American State Legislatures". *American Political Science Review* 104.1, pp. 151–170.

Gordon, Mitchell L et al. (2021). "The Disagreement Deconvolution: Bringing Machine Learning Performance Metrics In Line With Reality". In: *CHI Conference on Human Factors in Computing Systems*.

Grimmer, Justin, Gary King, and Chiara Superti (2015). *The Unreliability of Measures of Intercoder Reliability, and What to do About it*.

Grimmer, Justin and Brandon Stewart (July 2013). "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts". *Political Analysis* 21.3, pp. 267–297.

Homan, Maaike D., Gijs Schumacher, and Bert N. Bakker (n.d.). "Do We Mimic Politicians? Examining Voter Responses to the Emotional Displays of Politicians".

Hopkins, Daniel J and Gary King (2010). "A Method of Automated Nonparametric Content Analysis for Social Science". *American Journal of Political Science* 54 (1), pp. 229–247.

Hube, Christoph, Besnik Fetahu, and Ujwal Gadiraju (2019). "Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments". In: *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12.

Kahn, Kim Fridkin and Patrick J. Kenney (1999). "Do Negative Campaigns Mobilize or Suppress Turnout? Clarifying the Relationship between Negativity and Participation". *American Political Science Review* 93.4, pp. 877–889.

Kam, Cindy D. and Beth A. Estes (2016). "Disgust sensitivity and public demand for protection". *Journal of Politics* 78.2, pp. 481–496.

Kharroub, Tamara and Ozen Bas (Feb. 2015). "Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution". *New Media & Society*.

King, Gary, Jennifer Pan, and Margaret E. Roberts (2013). "How Censorship in China Allows Government Criticism but Silences Collective Expression". *American Political Science Review* 107.2, pp. 326–343.

Ksiazkiewicz, Aleksander, Window, and Amanda Friesen (2020). "Slimy worms or sticky kids How caregiving tasks and gender identity attenuate disgust response". *Politics and the Life Sciences* 39.2.

Marcus, George E., W. Russell Neuman, and Michael MacKuen (2000). *Affective Intelligence and Political Judgement*. Chicago and London: University of Chicago Press.

Peng, Yilang (2020). "What Makes Politicians' Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision". *The International Journal of Press/Politics* 26.1, pp. 143–166.

Pesigan, Ivan Jacob Agaloos, Rongwei Sun, and Shu Fai Cheung (2023). "betaDelta and betaSandwich: Confidence intervals for standardized regression coefficients in R". *Multivariate Behavioral Research*. R package version 1.0.1.

Peterson, Erik, Sharad Goel, and Shanto Iyengar (2021). "Partisan selective exposure in online news consumption: evidence from the 2016 presidential campaign". *Political Science Research and Methods* 9.2, pp. 242–258.

Putnam, Robert D. (1971). "Studying Elite Political Culture: The Case of "Ideology"". *American Political Science Review* 65.3, pp. 651–681.

Sap, Maarten et al. (2019). "The Risk of Racial Bias in Hate Speech Detection". In: *Proceedings ofthe 57th Annual Meeting ofthe Association for Computational Linguistics*, pp. 1668–1678.

Schaffner, Brian F. and Samantha Luks (2018). "Misinformation or Expressive Responding?" *Public Opinion Quarterly* 82.1, pp. 135–147.

Segal, Jeffrey A. and Harold J. Spaeth (1996). "The Influence of Stare Decisis on the Votes of United States Supreme Court Justices". *American Journal of Political Science* 40.4, pp. 971–1003.

Steephen, John Eric, Samyak Raj Mehta, and Raju Surampudi Bapi (2018). "Do We Expect Women to Look Happier Than They Are? A Test of Gender-Dependent Perceptual Correction". *Perception* 47 (2). PMID: 29199878, pp. 232–235.

Steinert-Threlkeld, Zachary C., Alexander M. Chan, and Jungseock Joo (2022). "How State and Protester Violence Affect Protest Dynamics". *The Journal of Politics* 84.2, pp. 798–813.

Struthers, Cory L., Christopher Hare, and Ryan Bakker (2020). "Bridging the pond: measuring policy positions in the United States and Europe". *Political Science Research and Methods* 8.4, pp. 677–691.

Todorov, Alexander et al. (2005). "Inferences of Competence from Faces Predict Election Outcomes". *Science* 308.5728, pp. 1623–1626.

Tufekci, Zeynep and Christopher Wilson (Apr. 2012). "Social Media and the Decision to Participate in Political Protest: Observations From Tahrir Square". *Journal of Communication* 62.2, pp. 363–379.

Winter, Nicholas J. G., Adam G. Hughes, and Lynn M. Sanders (2020). "Online coders, open codebooks: New opportunities for content analysis of political communication". *Political Science Research and Methods* 8.4, pp. 731–746.

Wojcik, Stefan and Adam Hughes (2019). *Sizing Up Twitter Users*. Tech. rep. Pew Research Center.

Yang, Yu et al. (July 2022). "Enhancing Fairness in Face Detection in Computer Vision Systems by Demographic Bias Mitigation". In: Association for Computing Machinery, Inc, pp. 813–822.

Ying, Luwei, Jacob Montgomery, and Brandon Stewart (2022). "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures". *Political Analysis* 30.4, pp. 570–589.

Appendix A Meta-Analysis

We ran multiple searches through Google Scholar to generate a list of high-impact political science articles that potentially used human-labeled data. We queried Google Scholar using six unique keywords that could indicate the use of human-labeled data for quantitative analyses: machine learning, text as data, inter rater reliability, audio as data, images as data, and inter coder reliability. A few notes on this search: (1) Our search was not case-sensitive. (2) For multiple word keywords we searched for the exact phrase. (3) For keywords with spaces we also searched for the same phrase, but replaced the space with a dash. For example, we did not just collect articles that contained "inter rater reliability." We also collected articles that contained "inter-rater-reliability." The same goes for "text as data" and "text-as-data."

This search was limited to five high-impact political science journals (listed in no particular order): American Political Science Review, Journal of Politics, American Journal of Political Science, Political Analysis, and Political Science Research Methods. Querying Google Scholar for articles published in these journals that contain at least one of our six keywords returned 393 hits, of which 378 were unique articles. The publication years ranged from 1965 to 2022.

The authors of this study plus one graduate research assistant read and annotated the 378 articles. We first developed an initial coding form and each coded 5 articles as a pilot. After meeting to discuss our annotations, we revised the form and assigned each article to at least one coder. We read the main paper and all available appendices for each paper. We excluded papers that used existing annotated data and papers where human annotation was conducted on the output of unsupervised machine learning (e.g. interpreting topics from text topic models).

A total of 77 articles were double-coded using the finalized labeling form to assess IRR. Our main indicator of IRR is agreement on the question of whether or not a given article was relevant to our study (a low-subjectivity, low-identity-sensitivity task, based on our typology). In order to be deemed relevant (and therefore subject to in-depth annotation) an article had to (a) contain original human labeling and (b) use that human labeling for either direct analysis or machine learning. Criterion (a) excluded papers that used existing labeled data. Criterion (b) primarily excluded papers where the human labeling was done to interpret the results from unsupervised machine learning (e.g. interpreting topics from a topic model). While we understand topic interpretation to be a form of human labeling where LCB could apply, this paper highlights the issues with LCB in direct analysis or in supervised machine learning. The impact of LCB on the results of unsupervised models is an area for future study (with acknowledgement that there are recent papers that directly tackle the instability of topic models and the challenges of deriving reliable topic labels ; see (Agrawal, Fu, and Menzies 2018; Denny and Spirling 2018)).

We first checked IRR for the decision to deem papers as relevant or not using Cohen's kappa and Krippendorff's alpha. We have two indicators of article relevance for our study, one that is simple (a "yes/no" question on if the study contains original human labeling) and the other that is more complex (a combination of the "yes/no" question, a question about

the labeling type to exclude unsupervised topic interpretation, and a question asking the coder to describe the labeling task). Table 5 contains the IRR statistic values between two pairs of coders – coder 1 double-coded articles from both coder 2 and coder 3. We see low values between coder 1 and 2, with much better agreement between coder 1 and 3. Coder 3 did the vast bulk of the labeling, so it is encouraging to see that their interpretation generally matched that of coder 1. There were a total of 2 articles with disagreements on relevance between coder 1 and 3 while there were 11 articles with disagreement between coder 1 and 2. To resolve conflicts on the double-coded articles, the remaining third coder acted as a tie-break (e.g. coder 3 broke ties between coder 1 and 2).

Table 5: IRR of Relevant Articles from Meta-Analysis

Relevance measure	Coder pair	Cohen's kappa	Krippendorff's alpha
Complex	1-2	0.50	0.49
Complex	1-3	0.87	0.87
Simple	1-2	0.49	0.66
Simple	1-3	0.87	0.88

We also checked IRR on the measures of whether or not a paper reported any IRR statistics and of whether or not a paper reported any demographic information on coders. Those results are presented in Table 6

Table 6: IRR of Reporting IRR and Demographis in Meta-Analysis Articles

Measure	Coder pair	Cohen's kappa	Krippendorff's alpha
Reports IRR	1-2	0.76	0.77
Reports IRR	1-3	0.67	0.68
Reports Demographics	1-2	0.50	0.49
Reports Demographics	1-3	0.67	0.67

Appendix B Data Collection for Image Study

Starting on January 1st, 2018, we began collecting all tweets produced by 1,144 American public affairs organizations, 559 American political actors, and 30 American news media outlets. We then focused on the hashtags used in these tweets (hashtags are a means of organizing on Twitter). At the end of each day we pulled a list of hashtags that were used more than twice by the same tracked account. From this list, we removed all hashtags that did not have a capitalization in the middle or that were shorter than 12 characters. These requirements ensured that we tracked unique hashtags used prominently by at least one of these organizations. Once on our list, we immediately began collecting any tweet by any Twitter user that used that hashtag. Each hashtag was tracked for two days and then removed if it was not used more than twice by the same organization over the next two days. Due to the sheer number of posted tweets, we were unable to collect all the tweets using our tracked hashtags, but we were able collect around 1,000,000 tweets per day from an average of 600 hashtags.

In total, we have roughly 4 million tweets from our initial tracked accounts and around 400 million tweets collected by tracking hashtags. In addition to data from each tweet such as the tweet text, count of retweets/favorites, count of account followers, and count of friends, we also collected any pictures posted with the tweets. We did not collect videos, but we did collect the thumbnail image displayed tweets with videos.

The public affairs organizations we tracked came from the 56th edition of Encyclopedia of Associations – National Organizations of the United States (EoA), published in 2017. The EoA contains information on roughly 23,000 organizations, but after limiting our list to organizations in the "Public Affairs" subject category and manually removing inactive and removed Twitter accounts we settled on 1,144 Twitter accounts to track (74.7% of the total population of EoA Public Affairs associations). In addition, we supplemented the EoA list of Twitter accounts with the official accounts from the most prominent news organizations in the United States and from every member of the 115th United States Congress. For news media outlets, we referenced numerous lists of the most watched and read news organizations and the most Tweeted news organizations. In total we tracked 30 media accounts and 434 accounts from U.S. Representatives and 100 accounts from U.S. Senators (some U.S. Representatives did not have Twitter accounts). Full information on the Twitter accounts that we track are available in supplementary documents.

On Principles and Guidance for Human Subjects Research: our data collection and subsequent image labeling (via Qualtrics survey) was reviewed by IRBs at the University of [redacted for blind review] and the University of [redacted for blind review]. The study was granted "exempt" status by both bodies. The Twitter data collection involved no directed intervention with human subjects. For the survey, respondents were informed that they were participating in "a research project on images and social movement mobilization." Respondents were compensated via the Qualtrics panel service. We do not know what Qualtrics ultimately paid each respondent; our costs were between $4-6 per respondent. In general, image labeling or text annotation has not been treated as human subjects research, to our knowledge. Often these tasks are crowdsourced on platforms such as Mechanical Turk

with a rate per image or text annotated. Our survey took 10-15 minutes to complete for 8 images. We warned respondents that the images might contain adult/disturbing content and required that respondents be 18 or over. Due to privacy and copyright concerns, we are unable to share the raw images or tweets. However, we will make tweet ids available for replication purposes and will share intermediate data, including the image labels and demographic information of labelers.

Appendix C In-Depth Process of Determining Mobilizing Hashtags

We began by downloading a random sample of one percent of the tweets from our database of tweets from the social organizations, congresspeople, and senators that we were tracking (the sample came to 39,640 tweets). Using the text of these tweets, we created a frequency table of all unigrams used with in the text of these tweets (pulling out stop words and punctuation). All contributors parsed through the most frequently used 1,500 unigrams and pulled out 37 unigrams that are likely to promote an online or off-line mobilization (these unigrams are located in the appendix). This was our list of unigrams we would use to find mobilizing tweets. In addition to unigrams, we also wished to build a list of bigrams, we used this list of unigrams to pull a list of all bigrams from the random sample of tweets that have at least one unigram that we previously identified as mobilizing.

From this list of bigrams, we parsed through the 500 most frequently used bigrams in this random, we pull out the bigrams that are most likely to promote an online or offline mobilization. Utilizing this list of bigrams, we pulled all tweets that contain those bigrams and create a frequency table of bigrams in those tweets. We parse through the 1,500 most frequently used bigrams and pull out all bigrams that could be associated with promoting off-line mobilization. Using this list of mobilizing bigrams, we pull all tweets from the random sample of tweets this new list of mobilizing bigrams. This process of identifying mobilizing bigrams from the 1,500 most frequently used bigrams in the pulled sample of tweets and then using the new lit of mobilizing bigrams to pull another sample of possible mobilizing tweets is repeated until we reach convergence. This means that the list of mobilizing bigrams are no different from the previous list used to find tweets. Convergence occurs after three iterations of this process and we create a list of mobilizing bigrams totaling 160. From this list of bigrams, we find additional mobilizing unigrams not listed in our initial list of unigrams and them to our list of unigrams (support, action, speak, plan, effort).

In total, our active learning process identified in total 198 keywords (156 bigrams and 42 unigrams) that could be related to promoting online or offline mobilization. We use these keywords to pull tweets that are potentially mobilizing from the random sample of tweets that we pulled originally. We download a random sample of 50 tweets that contain each mobilizing bigram and unigram. As an example, we pulled 50 tweets that contain the unigram "join" (in our list of mobilizing unigrams), 50 tweets that contain the bigram "our voices" (in our list of mobilizing bigrams) and so forth for all 198 keywords we identified. In

total, we collect 9,900 tweets that are potentially and from this sample we take a random sample of 1,000 tweets that could be mobilizing.

Parsing through the text of these tweets, the authors of this manuscript individually coded these tweets as offline mobilizing, online mobilizing, both, or neither using a codebook that can be found in the supplementary materials. If the majority of the contributors viewed it as mobilizing, it was coded as a mobilizing tweet. From the 1,000 potentially mobilizing tweets we identified 252 tweets that were offline or online mobilizing. 426 hashtags were used in these 252 mobilization tweets.

We manually determine is each hashtag is mobilizing by looking up these hashtags and finding information. If the hashtag was geared towards getting people to act either online or offline, we code the hashtag as mobilizing. For this manuscript we only focus on eight of the hashtags that were promoting a specific off-line mobilization in the form of a protest are utilized. The eight hashtags are: riseforclimate, familiesbelongtogether, stopkavanaugh, uniteforjustice, womenswave, write4rights, cleandreamactnow, and abolishice.

Utilizing these hashtags, we pull all tweets that we had collected in our larger database of tweets that used these eight hashtags. We only pull tweets for these hashtags on days that were being tracked by our collection process. In addition to retweet count, follower count, and text data, we also collected any photos that was displayed with each tweet.

Table 7: Final List of Unigrams that are Mobilizing

join	rights	justice	fight	protect	alert	protest	meeting
meet	stand	together	defend	fighting	attend	event	save
attack	forward	movement	march	prevent	demand	advocate	protest
ready	rally	event	voice	lets	forces	battle	resistance
celebrate	cause	happen	urge	resist			

Table 8: Final List of Bigrams that are Mobilizing

meeting with	to demand	we will	help us	fight to
to support	we should	to celebrate	meet the	to defend
to protect	we need	we have	we can	we must
to join	to fight	to meet	efforts to	plan to
fight for	joining the	join us	support for	joined by
join the	to address	to stand	working to	the fight
work with	support of	meet with	and support	we celebrate
together to	stand with	fighting for	we want	working with
stand up	action to	take action	attempt to	your voice
in support	celebrate the	join me	fight against	work together
participate in	standing up	to attend	advocate for	your support
out against	speak out	support our	where we	we stand
please join	get out	support this	will join	fight back
speaking out	come together	for joining	to protest	event on
the movement	we work	and join	supporting the	stand for
working together	an event	march for	joining us	to combat
fight the	stand by	attend the	keep fighting	protecting our
must act	together with	to advocate	movement to	be joining
and demand	calling on	event at	protests in	with us
i urge	standing with	demand for	the march	our fight
will fight	your support	join in	movement is	join a
a rally	join our	defend our	stand against	together we
must protect	are demanding	together for	joins the	event with
we demand	coming together	demand action	help prevent	working w/
call for	to speak	we fight	demand a	must stand
join my	support the	fighting to	meeting on	to end
act to	to promote	with me	protect our	protect the
need your	effort to the event	and fight	our voices	urge your
action on	attempts to	rally in	happening now	to act
a stand	defend the	for meeting	a meeting	demand that
events in	march in	we are	you need	for our
today we				

Appendix D Regression Tables

Figure 10: Results from linear regressions predicting retweets with data from labeler subgroups

	Annotations				
	Pooled	*Republican*	*Democrat*	*Women*	*Men*
Enthusiasm	0.0003 (0.009)	0.002 (0.007)	0.001 (0.006)	0.006 (0.007)	0.005 (0.008)
Anxiety	0.011 (0.017)	0.010 (0.012)	0.002 (0.013)	0.013 (0.014)	−0.007 (0.015)
Aversion	0.017 (0.020)	0.010 (0.015)	0.006 (0.014)	−0.011 (0.017)	0.019 (0.017)
Sadness	−0.020 (0.013)	−0.005 (0.009)	−0.015* (0.009)	−0.007 (0.010)	−0.018 (0.012)
Disgust	−0.023 (0.014)	−0.024** (0.010)	−0.002 (0.010)	−0.003 (0.012)	−0.008 (0.012)
Looks Like Me	0.062*** (0.020)	0.019 (0.015)	0.050*** (0.014)	0.048*** (0.017)	0.013 (0.017)
Protest	0.252*** (0.050)	0.152*** (0.039)	0.164*** (0.039)	0.242*** (0.043)	0.089* (0.046)
Humor	−0.112 (0.078)	−0.059 (0.058)	−0.072 (0.055)	−0.255*** (0.078)	0.004 (0.059)
Irony	−0.157** (0.075)	−0.074 (0.055)	−0.070 (0.051)	−0.068 (0.071)	−0.084 (0.057)
Leader	−0.228*** (0.068)	−0.144*** (0.053)	−0.144*** (0.050)	−0.202*** (0.06)	−0.085 (0.059)
Social Sumbol	0.082 (0.074)	0.025 (0.056)	0.047 (0.051)	0.036 (0.063)	0.017 (0.062)
start followers	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
type tweetQuoted	−0.535*** (0.040)	−0.542*** (0.041)	−0.539*** (0.040)	−0.520*** (0.044)	−0.529*** (0.044)
type tweetRetweet	0.450*** (0.040)	0.460*** (0.040)	0.452*** (0.040)	0.511*** (0.043)	0.458*** (0.044)
tweet hour	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)
as.character(tweet wday)Mon	0.127* (0.068)	0.140** (0.068)	0.122* (0.068)	0.076 (0.072)	0.129* (0.074)
as.character(tweet wday)Sat	0.136** (0.057)	0.152*** (0.057)	0.142** (0.057)	0.113* (0.061)	0.162*** (0.062)
as.character(tweet wday)Sun	0.174*** (0.060)	0.185*** (0.060)	0.176*** (0.060)	0.147** (0.064)	0.205*** (0.065)
as.character(tweet wday)Thu	0.062 (0.061)	0.059 (0.062)	0.064 (0.061)	0.056 (0.066)	0.066 (0.068)
as.character(tweet wday)Tue	0.145** (0.064)	0.146** (0.064)	0.150** (0.064)	0.093 (0.069)	0.157** (0.071)
as.character(tweet wday)Wed	0.131** (0.060)	0.133** (0.061)	0.137** (0.060)	0.114* (0.066)	0.128* (0.066)
hashtagcleandreamactnow	−0.336*** (0.105)	−0.309*** (0.106)	−0.310*** (0.105)	−0.293** (0.119)	−0.234** (0.119)
hashtagendgunviolence	−0.342*** (0.075)	−0.345*** (0.075)	−0.315*** (0.074)	−0.274*** (0.082)	−0.282*** (0.082)
hashtagfamiliesbelongtogether	−0.701*** (0.072)	−0.705*** (0.072)	−0.689*** (0.072)	−0.636*** (0.077)	−0.691*** (0.077)
hashtagnomuslimbanever	−0.156* (0.086)	−0.140 (0.086)	−0.144* (0.086)	−0.120 (0.095)	−0.162* (0.096)
hashtagriseforclimate	−0.393*** (0.081)	−0.381*** (0.081)	−0.363*** (0.081)	−0.325*** (0.090)	−0.363*** (0.089)
hashtagsayhername	−0.093 (0.130)	−0.102 (0.130)	−0.083 (0.129)	0.006 (0.150)	−0.206 (0.145)
hashtagstopkavanaugh	−0.547*** (0.073)	−0.562*** (0.073)	−0.547*** (0.073)	−0.492*** (0.080)	−0.532*** (0.079)
hashtaguniteforjustice	−0.392*** (0.101)	−0.373*** (0.101)	−0.381*** (0.101)	−0.383*** (0.113)	−0.348*** (0.110)
hashtagwomenswave	−0.506*** (0.107)	−0.508*** (0.107)	−0.491*** (0.106)	−0.398*** (0.114)	−0.508*** (0.114)
hashtagwrite4rights	−0.463* (0.280)	−0.442 (0.282)	−0.468* (0.280)	−0.411 (0.307)	−0.899*** (0.319)
Constant	1.244*** (0.096)	1.330*** (0.090)	1.230*** (0.090)	1.176*** (0.097)	1.317*** (0.103)
Observations	4,621	4,620	4,620	3,793	3,731
R2	0.174	0.167	0.171	0.187	0.183
Adjusted R2	0.168	0.161	0.165	0.18	0.176
Residual Std. Error	1.083 (df = 4589)	1.088 (df = 4588)	1.085 (df = 4588)	1.068 (df = 3761)	1.070 (df = 3699)
F Statistic	31.123*** (df = 31; 4589)	29.676*** (df = 31; 4588)	30.455*** (df = 31; 4588)	27.899*** (df = 31; 3761)	26.640*** (df = 31; 3699)

Note: *p<0.1; **p<0.05; ***p<0.01

S

Table 9: Modeling Partisan Labels Separately

	Dependent variable:
	log_end_rts
avg_enthus_dem	−0.001 (0.006)
avg_anxiety_dem	0.0003 (0.013)
avg_avers_dem	0.007 (0.014)
avg_sad_dem	−0.015* (0.009)
avg_disgust_dem	0.0005 (0.010)
avg_socid_look_dem	0.047*** (0.014)
protest_avg_dem	0.145*** (0.040)
humor_avg_dem	−0.063 (0.055)
irony_avg_dem	−0.068 (0.051)
leader_avg_dem	−0.123** (0.051)
socsymb_avg_dem	0.040 (0.051)
avg_enthus_rep	0.001 (0.007)
avg_anxiety_rep	0.009 (0.012)
avg_avers_rep	0.011 (0.015)
avg_sad_rep	−0.004 (0.009)
avg_disgust_rep	−0.024** (0.010)
avg_socid_look_rep	0.015 (0.015)
protest_avg_rep	0.106*** (0.040)
humor_avg_rep	−0.048 (0.058)
irony_avg_rep	−0.072 (0.055)
leader_avg_rep	−0.094* (0.054)
socsymb_avg_rep	0.018 (0.056)
start_followers	0.00000*** (0.00000)
type_tweetQuoted	−0.536*** (0.040)
type_tweetRetweet	0.449*** (0.040)
tweet_hour	0.002 (0.002)
as.character(tweet_wday)Mon	0.127* (0.068)
as.character(tweet_wday)Sat	0.140** (0.057)
as.character(tweet_wday)Sun	0.175*** (0.060)
as.character(tweet_wday)Thu	0.061 (0.061)
as.character(tweet_wday)Tue	0.146** (0.064)
as.character(tweet_wday)Wed	0.134** (0.060)
hashtagcleandreamactnow	−0.335*** (0.106)
hashtagendgunviolence	−0.341*** (0.075)
hashtagfamiliesbelongtogether	−0.702*** (0.072)
hashtagnomuslimbanever	−0.153* (0.086)
hashtagriseforclimate	−0.396*** (0.082)
hashtagsayhername	−0.101 (0.130)
hashtagstopkavanaugh	−0.550*** (0.073)
hashtaguniteforjustice	−0.398*** (0.101)
hashtagwomenswave	−0.504*** (0.107)
hashtagwrite4rights	−0.473* (0.281)
Constant	1.246*** (0.096)
Observations	4,619
R^2	0.175
Adjusted R^2	0.167
Residual Std. Error	1.084 (df = 4576)
F Statistic	23.049*** (df = 42; 4576)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10: Regression Results with Reweighted Measures

	Dependent variable:			
	log_end_rts			
	Rep = 0.1	Rep = 0.21	Rep = 0.6	Rep = 1
	(1)	(2)	(3)	(4)
rw_enthus	0.008 (0.016)	0.010 (0.017)	0.016 (0.020)	0.011 (0.014)
rw_anxiety	0.004 (0.014)	0.006 (0.015)	0.011 (0.017)	0.010 (0.012)
rw_sad	−0.017* (0.010)	−0.019* (0.011)	−0.015 (0.013)	−0.005 (0.009)
rw_disgust	−0.005 (0.011)	−0.009 (0.012)	−0.028** (0.013)	−0.025** (0.010)
rw_socid_look	0.056*** (0.015)	0.062*** (0.017)	0.057*** (0.019)	0.020 (0.014)
rw_protest	0.185*** (0.042)	0.208*** (0.045)	0.243*** (0.049)	0.151*** (0.038)
rw_humor	−0.084 (0.060)	−0.098 (0.067)	−0.106 (0.077)	−0.058 (0.058)
rw_irony	−0.082 (0.056)	−0.098 (0.063)	−0.140* (0.074)	−0.075 (0.055)
rw_leader	−0.162*** (0.055)	−0.183*** (0.060)	−0.226*** (0.067)	−0.143*** (0.053)
rw_socsymb	0.052 (0.056)	0.055 (0.062)	0.059 (0.073)	0.027 (0.055)
start_followers	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)	0.00000*** (0.00000)
type_tweetQuoted	−0.538*** (0.040)	−0.537*** (0.040)	−0.537*** (0.040)	−0.542*** (0.041)
type_tweetRetweet	0.450*** (0.040)	0.449*** (0.040)	0.452*** (0.040)	0.460*** (0.040)
tweet_hour	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)	0.002 (0.002)
as.character(tweet_wday)Mon	0.122* (0.068)	0.122* (0.068)	0.131* (0.068)	0.140** (0.068)
as.character(tweet_wday)Sat	0.141** (0.057)	0.139** (0.057)	0.140** (0.057)	0.151*** (0.057)
as.character(tweet_wday)Sun	0.175*** (0.060)	0.175*** (0.060)	0.177*** (0.060)	0.185*** (0.060)
as.character(tweet_wday)Thu	0.064 (0.061)	0.064 (0.061)	0.062 (0.061)	0.061 (0.062)
as.character(tweet_wday)Tue	0.150** (0.064)	0.148** (0.064)	0.145** (0.064)	0.146** (0.064)
as.character(tweet_wday)Wed	0.136** (0.060)	0.134** (0.060)	0.131** (0.060)	0.133** (0.061)
hashtagcleandreamactnow	−0.315*** (0.105)	−0.322*** (0.105)	−0.334*** (0.105)	−0.309*** (0.106)
hashtagendgunviolence	−0.318*** (0.074)	−0.322*** (0.074)	−0.346*** (0.075)	−0.344*** (0.075)
hashtagfamiliesbelongtogether	−0.689*** (0.072)	−0.690*** (0.071)	−0.702*** (0.072)	−0.704*** (0.072)
hashtagnomuslimbanever	−0.147* (0.086)	−0.150* (0.086)	−0.154* (0.086)	−0.141 (0.086)
hashtagriseforclimate	−0.367*** (0.081)	−0.374*** (0.081)	−0.395*** (0.081)	−0.380*** (0.081)
hashtagsayhername	−0.083 (0.129)	−0.084 (0.129)	−0.098 (0.130)	−0.101 (0.130)
hashtagstopkavanaugh	−0.545*** (0.073)	−0.543*** (0.073)	−0.548*** (0.073)	−0.563*** (0.073)
hashtaguniteforjustice	−0.383*** (0.101)	−0.387*** (0.101)	−0.389*** (0.101)	−0.373*** (0.101)
hashtagwomenswave	−0.492*** (0.106)	−0.495*** (0.106)	−0.508*** (0.107)	−0.508*** (0.107)
hashtagwrite4rights	−0.467* (0.280)	−0.467* (0.280)	−0.461 (0.281)	−0.440 (0.282)
Constant	1.223*** (0.091)	1.218*** (0.093)	1.261*** (0.095)	1.331*** (0.090)
Observations	4,619	4,619	4,619	4,619
R^2	0.171	0.172	0.173	0.167
Adjusted R^2	0.166	0.167	0.167	0.161
Residual Std. Error (df = 4588)	1.085	1.084	1.084	1.088
F Statistic (df = 30; 4588)	31.648***	31.853***	31.901***	30.648***

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix E Alternative Regression Model Specifications

The first alternative specification includes the control variables listed above and only image variables deemed to be about "content": showing protest; a leader; someone who looks like the labeler; or a social symbol. The second includes the control variables and image variables deemed to be about "reactions": the emotions variables; humor; and irony. As with the fully-saturated model in Figure 5, we see coefficients changing based on whose labels are included in these alternative specifications.

Figure 11: Standardized Regression Coefficients from "Content" Linear Regressions Showing Labeler-type Differences

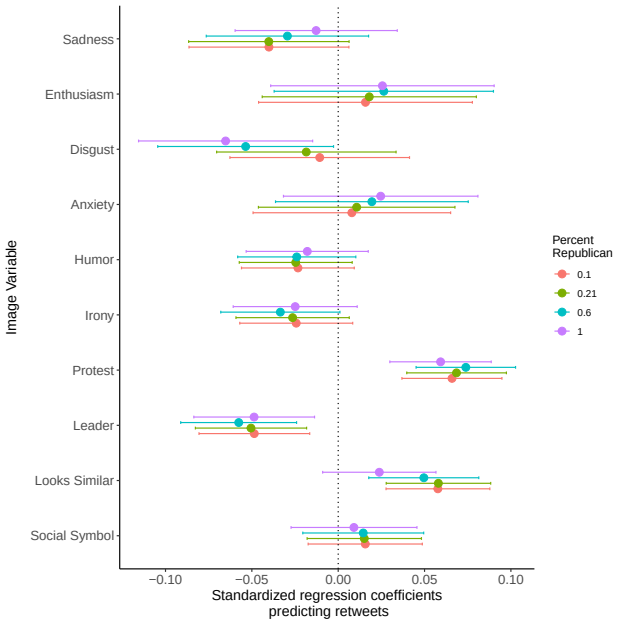
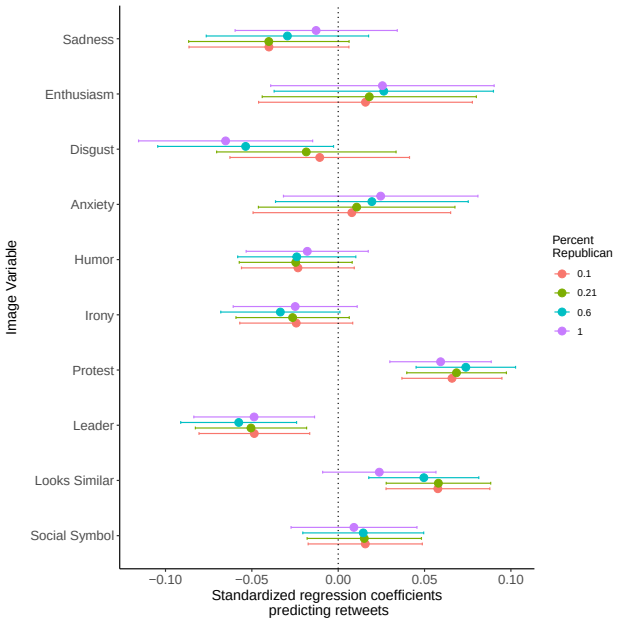
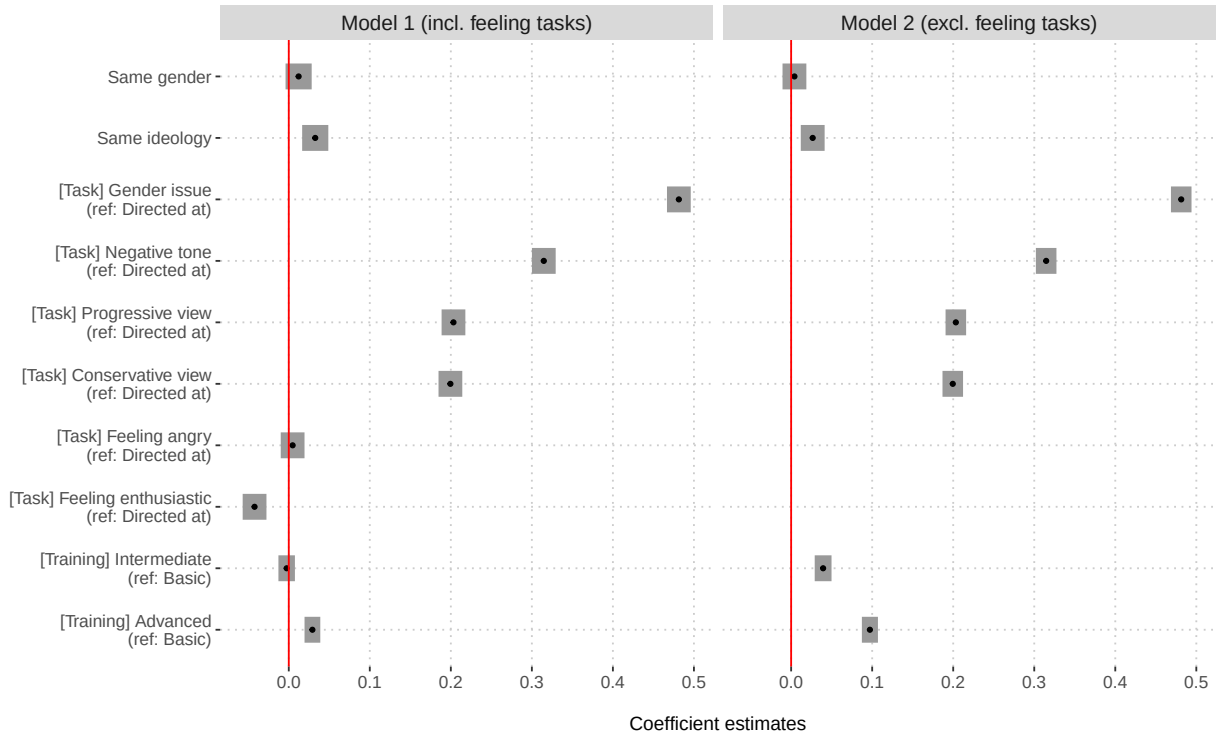


Figure 12: Standardized Regression Coefficients from "Reactions" Linear Regressions Showing Labeler-type Differences



Appendix F Modeling IRR in the text study

Figure 13: Linear regressions predicting IRR_{ijsz} (Cohen's Kappa) for each unique pair of coder ij, training session s, and annotation task z. These are multilevel models with random intercepts for each pair of coder.



To complement the descriptive results in Figure 7, in Figure 13 we report coefficients (+95% confidence intervals) for two linear regressions predicting IRR_{ijsz} as a function of whether a given pair ij is of the same gender and ideology, the training session s, and the annotation task z. Given the nested nature of the data, we run multilevel models with random intercepts for each pair ij. In Figure 13, IRR is not significantly higher for pairs of labelers of the same gender. IRR is significantly higher (+0.04) for labeler pairs of the same ideology. In Model 1 we also observe overall IRR improvement in the final "Advanced" annotation session (+0.03), but not the 'Intermediate' session, compared to the first "Basic" session. However, as shown in Model 2, this training effect is mainly driven by instructing coders to share how they personally felt when reading the message (and not whether they thought that the emotion was expressed in the text). If we exclude these "Feeling" tasks, IRR improves by about +0.04 and +0.1 in the "Intermediate" and "Advanced" sessions, respectively. The results for the other coefficients remain the same.

Appendix G Modeling IRR in the text study

In this Appendix we provide the results for all the models we specified in the pre-registration of the text study, available (blinded) **here**. In Figure 14 we present a visualization of the coefficients, and in Table 11 we present the coefficient tables. In these models we find support for most of (although not all) our expectations.

Gender In all the models that include *Same gender* as a predictor, we find a positive (about +0.01) effect, indicating that IRR is higher for pairs of coders of the same gender. However, although in the expected direction, these findings are not statistically significant.

Ideology In all the models that include *Same ideology* as a predictor, we find a positive (between +0.03 and +0.04) effect, indicating that IRR is indeed higher for pairs of coders of the same ideology. The findings are statistically significant in all but one model (Model 10), in which we included a very large number of interactions.

Training In Models 1, 5, 6, 9 and 10 we observe IRR to be higher in the third, and last, "Advanced" annotation session compared to the first "Basic" one; indicating that further training of the annotators contributed on average to improving IRR. In Models 1, 5, and 9 we do not observe the expected similar effect in the second "Intermediate" session (although we do observe a positive significant effect in Models 6 and 10). As discussed in the manuscript and shown in Figure 13, this is because by averaging effects across annotation tasks, the lower IRR for the most subjective and identity-dependent tasks (the two "Feeling" tasks) in the "Intermediate" and "Advanced" round of coding, we are unable to observe that IRR actually did also improve for the other tasks in the "Intermediate" round.

Gender × Task In Model 4 we test an interaction between pairs of coders being of the same gender (*Same gender*) and the particular annotation *Task* at hand. We find IRR to be particularly higher among coders of the same gender for the more subjective and identity-dependent tasks: Feeling angry and enthusiastic, and the finding for the latest is statistically significant.

Ideology × Task In line with the findings in Model 4, in Model 8 we also find IRR to be larger for pairs of coders of the same ideology for the most subjective and identity-dependent tasks (the Feeling tasks); however the findings are not statistically significant.

Gender × Training As expected, in Model 5 the coefficients for the interaction terms are negative, indicating that IRR between pairs of coders of the same gender was lower in the second ("Intermediate") and third ("Advanced") annotation rounds, compared to the first "Basic" one. This indicates that on average there was some degree of criteria-harmonization between pairs of coders of different gender. However, the findings are not statistically significant.

Ideo. × Training As expected, and in line with Model 5, in Model 9 the coefficients for the interaction terms are negative, indicating that IRR between pairs of coders of the same ideology was lower in the Intermediate and Advanced annotation rounds, compared to the first Basic one; indicating that on average there was some degree of criteria-harmonization between pairs of coders of different ideology. In this case, the coefficient for the last "Advanced" round is statistically significant.

Figure 14: Linear regressions predicting IRR_{ijsz} (Cohen's Kappa) for each unique pair of coder ij, training session s, and annotation task z. These are multilevel models with random intercepts for each pair of coder ij.

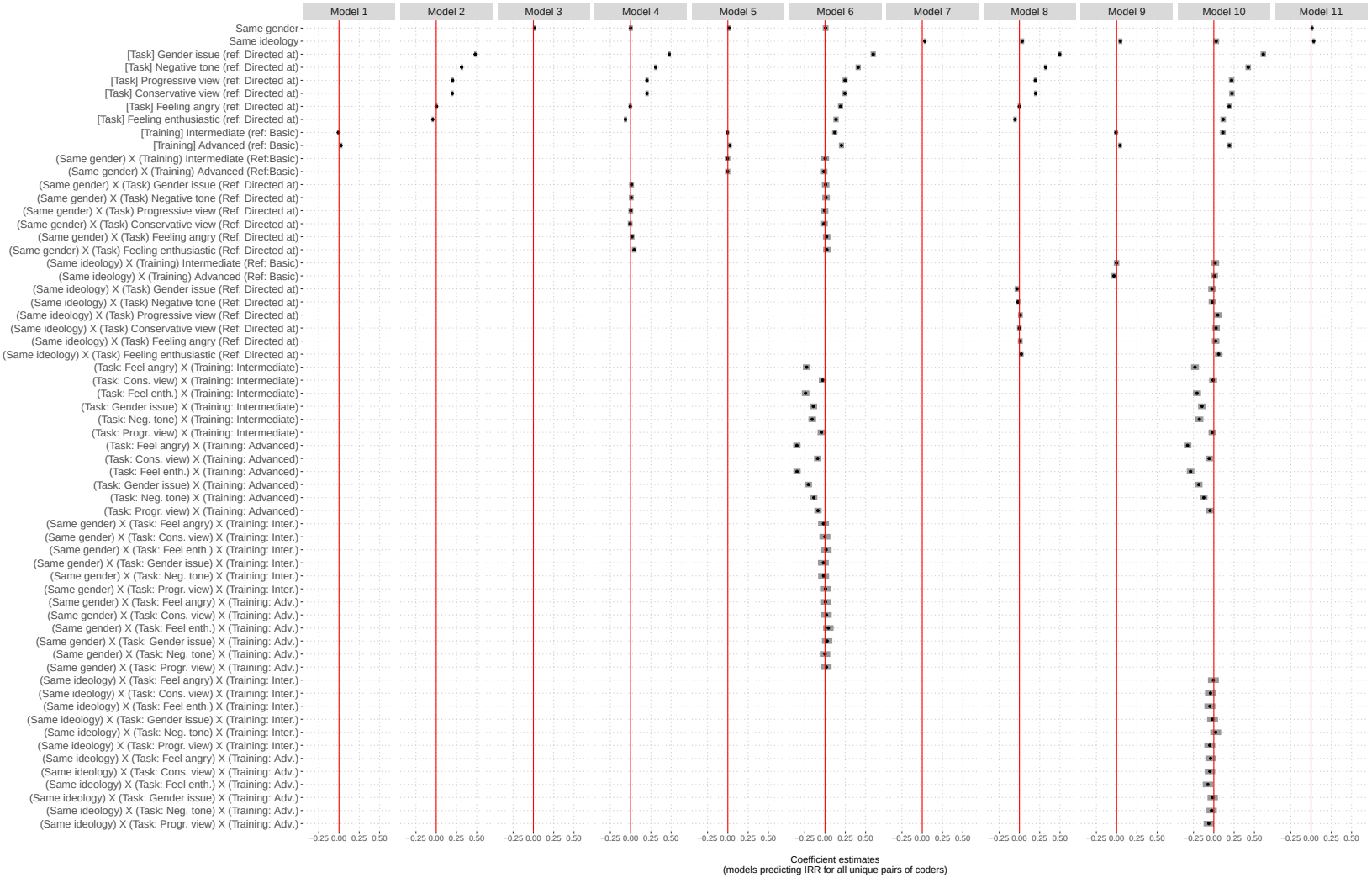


Table 11: Coefficient Tables for Models in Figure 14

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
Same gender	–	–	0.013 (0.009)	0.002 (0.013)	0.015 (0.013)	0.007 (0.018)	–
Same ideology	–	–	–	–	–	–	0.033 (0.008)*
[Task] Gender issue (ref: Directed at)	–	0.481 (0.008)*	–	0.476 (0.01)*	–	0.595 (0.017)*	–
[Task] Negative tone (ref: Directed at)	–	0.315 (0.008)*	–	0.311 (0.01)*	–	0.409 (0.017)*	–
[Task] Progressive view (ref: Directed at)	–	0.203 (0.008)*	–	0.202 (0.01)*	–	0.248 (0.017)*	–
[Task] Conservative view (ref: Directed at)	–	0.199 (0.008)*	–	0.202 (0.01)*	–	0.245 (0.017)*	–
[Task] Feeling angry (ref: Directed at)	–	0.005 (0.008)	–	-0.004 (0.01)	–	0.192 (0.017)*	–
[Task] Feeling enthusiastic (ref: Directed at)	–	-0.042 (0.008)*	–	-0.063 (0.01)*	–	0.136 (0.017)*	–
[Training] Intermediate (ref: Basic)	-0.008 (0.009)	–	–	–	-0.006 (0.012)	0.12 (0.017)*	–
[Training] Advanced (ref: Basic)	0.025 (0.008)*	–	–	–	0.026 (0.011)*	0.203 (0.016)*	–
(Same gender) X (Training) Intermediate (Ref:Basic)	–	–	–	–	-0.004 (0.017)	0.002 (0.024)	–
(Same gender) X (Training) Advanced (Ref:Basic)	–	–	–	–	-0.002 (0.016)	-0.015 (0.023)	–
(Same gender) X (Task) Gender issue (Ref: Directed at)	–	–	–	0.012 (0.015)	–	0.008 (0.024)	–
(Same gender) X (Task) Negative tone (Ref: Directed at)	–	–	–	0.009 (0.015)	–	0.013 (0.024)	–
(Same gender) X (Task) Progressive view (Ref: Directed at)	–	–	–	0.003 (0.015)	–	-0.005 (0.024)	–
(Same gender) X (Task) Conservative view (Ref: Directed at)	–	–	–	-0.006 (0.015)	–	-0.012 (0.024)	–
(Same gender) X (Task) Feeling angry (Ref: Directed at)	–	–	–	0.018 (0.015)	–	0.02 (0.024)	–
(Same gender) X (Task) Feeling enthusiastic (Ref: Directed at)	–	–	–	0.044 (0.015)*	–	0.023 (0.024)	–
(Task: Feel angry) X (Training: Intermediate)	–	–	–	–	–	-0.226 (0.024)*	–
(Task: Cons. view) X (Training: Intermediate)	–	–	–	–	–	-0.03 (0.024)	–
(Task: Feel enth.) X (Training: Intermediate)	–	–	–	–	–	-0.24 (0.024)*	–
(Task: Gender issue) X (Training: Intermediate)	–	–	–	–	–	-0.144 (0.024)*	–
(Task: Neg. tone) X (Training: Intermediate)	–	–	–	–	–	-0.157 (0.024)*	–
(Task: Progr. view) X (Training: Intermediate)	–	–	–	–	–	-0.045 (0.024)	–
(Task: Feel angry) X (Training: Advanced)	–	–	–	–	–	-0.347 (0.023)*	–
(Task: Cons. view) X (Training: Advanced)	–	–	–	–	–	-0.091 (0.023)*	–
(Task: Feel enth.) X (Training: Advanced)	–	–	–	–	–	-0.345 (0.023)*	–
(Task: Gender issue) X (Training: Advanced)	–	–	–	–	–	-0.205 (0.023)*	–
(Task: Neg. tone) X (Training: Advanced)	–	–	–	–	–	-0.138 (0.023)*	–
(Task: Progr. view) X (Training: Advanced)	–	–	–	–	–	-0.087 (0.023)*	–
(Same gender) X (Task: Feel angry) X (Training: Inter.)	–	–	–	–	–	-0.018 (0.034)	–
(Same gender) X (Task: Cons. view) X (Training: Inter.)	–	–	–	–	–	-0.001 (0.034)	–
(Same gender) X (Task: Feel enth.) X (Training: Inter.)	–	–	–	–	–	0.014 (0.034)	–
(Same gender) X (Task: Gender issue) X (Training: Inter.)	–	–	–	–	–	-0.02 (0.034)	–
(Same gender) X (Task: Neg. tone) X (Training: Inter.)	–	–	–	–	–	-0.017 (0.034)	–
(Same gender) X (Task: Progr. view) X (Training: Inter.)	–	–	–	–	–	0.006 (0.034)	–
(Same gender) X (Task: Feel angry) X (Training: Adv.)	–	–	–	–	–	0.005 (0.033)	–
(Same gender) X (Task: Cons. view) X (Training: Adv.)	–	–	–	–	–	0.018 (0.033)	–
(Same gender) X (Task: Feel enth.) X (Training: Adv.)	–	–	–	–	–	0.041 (0.033)	–
(Same gender) X (Task: Gender issue) X (Training: Adv.)	–	–	–	–	–	0.026 (0.033)	–
(Same gender) X (Task: Neg. tone) X (Training: Adv.)	–	–	–	–	–	0 (0.033)	–
(Same gender) X (Task: Progr. view) X (Training: Adv.)	–	–	–	–	–	0.017 (0.033)	–
N	4,415	4,415	4,415	4,415	4,415	4,415	4,415
Log Likelihood	239.052	2,413.775	235.424	2,397.460	229.990	2,718.613	241.904
Akaike Inf. Crit.	−468.103	−4,809.549	−462.848	−4,762.919	−443.981	−5,349.226	−475.808
Bayesian Inf. Crit.	−436.139	−4,752.015	−437.277	−4,660.635	−392.839	−5,067.945	−450.237

	Model 8	Model 9	Model 10	Model 11
Same gender	–	–	–	0.013 (0.008)
Same ideology	0.033 (0.013)*	0.046 (0.013)*	0.029 (0.018)	0.033 (0.008)*
[Task] Gender issue (ref: Directed at)	0.497 (0.011)*	–	0.61 (0.017)*	–
[Task] Negative tone (ref: Directed at)	0.324 (0.011)*	–	0.424 (0.017)*	–
[Task] Progressive view (ref: Directed at)	0.197 (0.011)*	–	0.219 (0.017)*	–
[Task] Conservative view (ref: Directed at)	0.2 (0.011)*	–	0.224 (0.017)*	–
[Task] Feeling angry (ref: Directed at)	-0.001 (0.011)	–	0.19 (0.017)*	–
[Task] Feeling enthusiastic (ref: Directed at)	-0.054 (0.011)*	–	0.115 (0.017)*	–
[Training] Intermediate (ref: Basic)	–	-0.008 (0.013)	0.112 (0.018)*	–
[Training] Advanced (ref: Basic)	–	0.043 (0.012)*	0.192 (0.017)*	–
(Same ideology) X (Training) Intermediate (Ref: Basic)	–	-0.001 (0.017)	0.017 (0.024)	–
(Same ideology) X (Training) Advanced (Ref: Basic)	–	-0.033 (0.016)*	0.007 (0.023)	–
(Same ideology) X (Task) Gender issue (Ref: Directed at)	-0.03 (0.015)*	–	-0.023 (0.024)	–
(Same ideology) X (Task) Negative tone (Ref: Directed at)	-0.018 (0.015)	–	-0.017 (0.024)	–
(Same ideology) X (Task) Progressive view (Ref: Directed at)	0.012 (0.015)	–	0.049 (0.024)*	–
(Same ideology) X (Task) Conservative view (Ref: Directed at)	-0.001 (0.015)	–	0.028 (0.024)	–
(Same ideology) X (Task) Feeling angry (Ref: Directed at)	0.01 (0.015)	–	0.022 (0.024)	–
(Same ideology) X (Task) Feeling enthusiastic (Ref: Directed at)	0.023 (0.015)	–	0.06 (0.024)*	–
(Task: Feel angry) X (Training: Intermediate)	–	–	-0.232 (0.025)*	–
(Task: Cons. view) X (Training: Intermediate)	–	–	-0.009 (0.025)	–
(Task: Feel enth.) X (Training: Intermediate)	–	–	-0.207 (0.025)*	–
(Task: Gender issue) X (Training: Intermediate)	–	–	-0.146 (0.025)*	–
(Task: Neg. tone) X (Training: Intermediate)	–	–	-0.178 (0.025)*	–
(Task: Progr. view) X (Training: Intermediate)	–	–	-0.016 (0.025)	–
(Task: Feel angry) X (Training: Advanced)	–	–	-0.324 (0.024)*	–
(Task: Cons. view) X (Training: Advanced)	–	–	-0.057 (0.024)*	–
(Task: Feel enth.) X (Training: Advanced)	–	–	-0.287 (0.024)*	–
(Task: Gender issue) X (Training: Advanced)	–	–	-0.186 (0.024)*	–
(Task: Neg. tone) X (Training: Advanced)	–	–	-0.124 (0.024)*	–
(Task: Progr. view) X (Training: Advanced)	–	–	-0.046 (0.024)	–
(Same ideology) X (Task: Feel angry) X (Training: Inter.)	–	–	-0.005 (0.034)	–
(Same ideology) X (Task: Cons. view) X (Training: Inter.)	–	–	-0.041 (0.034)	–
(Same ideology) X (Task: Feel enth.) X (Training: Inter.)	–	–	-0.049 (0.034)	–
(Same ideology) X (Task: Gender issue) X (Training: Inter.)	–	–	-0.016 (0.034)	–
(Same ideology) X (Task: Neg. tone) X (Training: Inter.)	–	–	0.023 (0.034)	–
(Same ideology) X (Task: Progr. view) X (Training: Inter.)	–	–	-0.05 (0.034)	–
(Same ideology) X (Task: Feel angry) X (Training: Adv.)	–	–	-0.041 (0.033)	–
(Same ideology) X (Task: Cons. view) X (Training: Adv.)	–	–	-0.047 (0.033)	–
(Same ideology) X (Task: Feel enth.) X (Training: Adv.)	–	–	-0.073 (0.033)*	–
(Same ideology) X (Task: Gender issue) X (Training: Adv.)	–	–	-0.014 (0.033)	–
(Same ideology) X (Task: Neg. tone) X (Training: Adv.)	–	–	-0.027 (0.033)	–
(Same ideology) X (Task: Progr. view) X (Training: Adv.)	–	–	-0.062 (0.033)	–
N	4,415	4,415	4,415	4,415
Log Likelihood	2,405.569	239.224	2,738.560	239.218
Akaike Inf. Crit.	−4,779.138	−462.447	−5,389.120	−468.436
Bayesian Inf. Crit.	−4,676.853	−411.305	−5,107.839	−436.472