When Conservatives See Red but Liberals Feel Blue: Labeler Characteristics and Variation in Content Annotation

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

Abstract

Human annotation of data, including texts and images, is a bedrock of political science research. Yet we often fail to consider how the identities of our labelers may systematically affect their annotations and our downstream applications. Collecting annotator demographic information, regardless of task type, can help us establish measurement validity and better appreciate variation in interrater reliability. We may also discover things about our topic that we did not previously appreciate. We demonstrate the benefits of collecting labeler characteristics with two annotation cases, one using images from the United States and the second using text from the Netherlands. For both cases on a range of tasks, we find that annotator gender and political identity are associated with significantly different annotations. We consider three main approaches to addressing labeler characteristic issues: adjusting labels based on labeler identity, weighting composite labels based on target population demographics, and intentionally modeling subgroup variation. (149 words)

Short title (49 characters): When Conservatives See Red but Liberals Feel Blue

Keywords: human annotation, content analysis, labeler characteristics, manual labeling, text/images as data

Replication files are available in the JOP Data Archive on Dataverse (https://dataverse.harvard.edu/dataverse/jop). The empirical analysis has been successfully replicated by the JOP replication analyst. Supplementary material for this article is available in the appendix in the online edition. The study was conducted in compliance with relevant laws and was deemed exempt by the appropriate institutional research ethics committees. Support for this research was provided by National Science Foundation grant #1727459, "The Power of Images: A Computational Investigation of Political Mobilization via Social Media."

[*] University of Illinois at Urbana-Champaign: nww3@illinois.edu

[†] Royal Holloway University of London: andreu.casas@rhul.ac.uk

[‡] Independent Scholar: kevin.aslett.30@gmail.com

[§] University of Washington: jwilker@uw.edu

1 Introduction

Human annotation of data (also referred to as human labeling or coding) 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. Jones et al. 2023; 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.

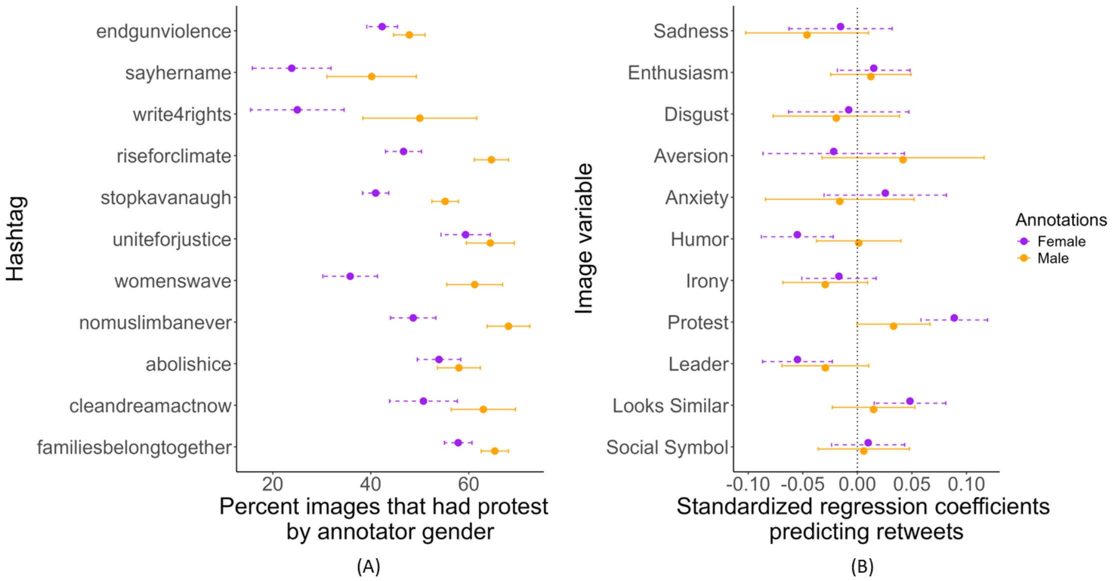
The strength of conclusions drawn from annotated data depends on our confidence in those annotations. For example, if we want to analyze news sources for partisan biases, we need to be confident in our measure of partisan slant. Yet if we are annotating news articles, it is not difficult to imagine that annotators might disagree about whether an article is partisan or not. Political scientists are attentive to concerns about the validity and reliability of their labels (see, e.g. Grimmer and Stewart 2013). We want to know that the tasks we set for our annotators (whether those annotators are ourselves, our research assistants, or crowdsourced labelers) produce "good" data and that the annotations reflect stable concepts. Aiming for strong inter-rater reliability (IRR) and validity, we design research procedures with attention to *how* labels are generated (especially for crowdsourced annotation on platforms like Mechanical Turk). We pilot our labeling forms; train our coders and test their performance; monitor and drop labels from annotators who speed through the tasks or fail attention checks, or those of annotators whose labels consistently diverge from a gold standard; we use anchoring vignettes to help define complex concepts, and so on (Barberá et al. 2021; Struthers, Hare, and Bakker 2020; Winter, Hughes, and Sanders 2020; DeBell 2017; Ying, Montgomery, and Stewart 2022; Benoit et al. 2016).

Political scientists have been less attentive to threats to validity and reliability that arise because of *who* our labelers are. The task of labeling news *for* partisanship might itself be subject *to* partisanship – a liberal might be less likely to identify an article with a liberal slant than a conservative. We conducted a meta-analysis of 97 articles from 1965 to 2022 in five top political science journals that relied on human annotation. Only 55% percent included any sort of IRR statistic.[1] Only 25% provided any information about who did the labeling. Two-thirds (62%) mentioned education level in some way (see online Appendix A, Table 3). Also relatively common were mentions of language abilities. Just 2 of the 97 articles provided information about annotator gender and only one included partisanship or party identification.

Importantly, variation in annotations based on labeler characteristics are not always easy to predict. The left panel (A) in Figure 1 shows clear differences in how often male and female respondents saw protest in the same set of images. Scholars recognize that biased labels can produce biased results (Hopkins and King 2010; Benoit, Laver, and Mikhaylov 2009) and propose correction approaches that assume that there is a single right answer from which annotators have deviated (Hopkins and King 2010; Grimmer, King, and Superti 2015; Bachl and Scharkow 2017). Our protest example suggests another concern. What if the task is not as objective as the researcher assumes? What if there is no single ground truth that all labelers would agree on given sufficient instruction? And how would we know that there are demographic differences in responses if we have not collected information about who is doing the labeling and tested for differences?

[1] Papers from the 1990s (roughly 9% of the total sample) reported some form of IRR statistic 66% of the time. Papers from the 2010s (roughly 28% of the total sample) reported IRR statistics 63% of the time. For details on the data collection and analysis, see online Appendix A.

Figure 1: **Image Annotations for "Protest" (Panel A) and Regression Results (Panel B) Vary with Annotator Gender.** Purple dashed lines for female annotations, orange solid lines for male. Error bars are 95% confidence intervals.



If we were to use these "protest or not?" annotations in a regression analysis predicting retweets, our conclusions might differ depending on the proportion of our labelers who are male versus female. The right panel (B) of Figure 1 shows that with an all female labeling pool, we would conclude that there is a positive, significant association between images of protests and retweets, but not if we had an all-male labeling pool. Other regression coefficients in Figure 1.B indicate that who is doing the labeling impacts conclusions. Images labeled as including humor or leaders by female annotators have negative, statistically significant associations with retweets, but not those labeled by male annotators. [2]

Comparing overall (e.g., across-group) IRR to IRR within particular groups of coders can help us identify potential systematic differences in annotations such as those found above. Consider a task that asks annotators to indicate whether policy documents address a

[2] Corrections for multiple hypothesis testing would result in different conclusions about the statistical significance of some of the coefficients (see online Appendix E, Figure 3 for Figure 1.B with 99% confidence intervals). The point about changing signs and magnitudes of coefficients stands despite changes to the chosen alpha for significance.

women's issue.[3] We might be disappointed to discover that IRR is low on the task. Absent demographic information, we can only guess at why. With demographic information, we might discover there is strong agreement within female coders and within male coders, but that agreement is in opposite directions between the two groups, pulling down overall IRR. Evaluating IRR both across groups (e.g., overall IRR) and within groups (e.g., IRR within sub-groups of coders) can therefore be helpful. We can choose to focus on some responses as more valid than others. We can be more discriminating in terms of who we choose to do future annotations. We might also decide to pursue a completely new project explaining the gender differences in perceiving gender issues.

In this paper we comprehensively address issues of labeler characteristics in large-scale annotation tasks. We suggest that arraying tasks on a spectrum from objective to subjective helps us gain purchase on the issues and solutions. Where labeler characteristics systematically shift annotations away from a researcher-assumed, objective ground truth, we have "labeler characteristic bias," or LCB. Where subjectivity in labeling is expected, differences in annotations based on identity are not bias, they are the focus of the research. Annotation data are often more similar to survey opinion data than has generally been recognized. Depending on the task type, variation in annotations can persist despite extensive training and clear annotation guides. For all task types, collecting information about our annotators strengthens our research by establishing that we are measuring what we assume we are measuring. We can both improve the validity of our annotations and open up new avenues for research and understanding. Here we explore labeler characteristics and their consequences in two research settings, one involving image labeling and the other text annotation. We also discuss how researchers can adapt their projects in light of systematic differences in annotations.

[3] This task is similar to one of the premier examples of large-scale annotation in political science, the Comparative Agendas Project (CAP) (Jones et al. 2023). To our knowledge, the CAP project has never tested for systematic coding differences based on labeler characteristics.

2 Tasks Types, Labeler Problems, and Solutions

Why should researchers collect demographic information about their labelers and test for systematic differences in annotations based on labeler characteristics? If we do find differences, what should we do about it? The answers emerge most clearly if we consider a subjective-objective spectrum of human annotation tasks. This section defines the spectrum and explains how labeler characteristics present problems for annotation on different tasks types, with reference to how these issues might be obscured or revealed by within- or across-group IRR. We also lay out the dangers of assuming a task is objective when in reality it tips towards subjective. Finally, we address how to approach and account for annotation variation across groups with three general classes of solutions: adjust, weight, or model.

What do we mean by a spectrum of subjective or objective tasks? At the most subjective end, tasks asks annotators to respond to prompts that require personal interpretation. Others have referred to these tasks as labeling for “projective” or more “latent” content (Potter and Levine-Donnerstein 1999). Another way to define a subjective task is to say that there is no single, factually true answer to the question. An example would be asking “How angry does this text make you feel?” At the objective end, in contrast, the goal is for annotators to label for a single true response that the researcher assumes is present in or identifiable from the material. For image labeling, an objective task might entail object recognition: “Is there a person in this picture?” For text, an objective task might ask: “Is the president mentioned in this news article?” Others have described these tasks as labeling for “manifest” content (Potter and Levine-Donnerstein 1999).[4]

[4] This spectrum should not be confused with a distinction between simple and complex tasks. Subjective tasks can be both simple (“How happy does this text make you feel?”) or complex (“Which emotions do you feel looking when reading this text?”). The complexity of objective tasks might similarly vary between simple (“Is there a person in this picture?”), and complex (“Name all the people in this picture.”). The subjective-objective spectrum should also not be confused with the clarity or ambiguity of a task. If tasks are not clearly explained to annotators, annotators may end up relying more on personal judgment than the researcher desires. In this discussion, we assume that the researcher has defined their task as clearly and as simply as possible for the annotators.

In the case of a subjective-type task, labeler characteristics and variation in labels are not a problem, they are the point. We fully expect that Democrats and Republicans will feel different emotions when they look at images from left-leaning protests. We must therefore collect and report data on who our annotators are to make the case that our annotations can be appropriately generalized. We may expect that there will be low IRR overall across groups. At the same time, we may expect that there will be strong IRR within a particular group. For example, we might expect that people under the age of 50 respond with the same degree of (presumably higher) anger compared to people over the age of 75 when reading a text about raising the minimum age for Social Security eligibility. Our labelers are conceptually analogous to respondents on a public opinion survey. We could therefore use a total survey error framework to understand annotation "error" as being composed of problems with measurement (e.g., poorly worded tasks) and representation (e.g., sample population does not match the population of interest.[5] Assuming that the tasks have already been written as clearly as possible and have been extensively piloted, the main issue is to whom you want to generalize with the annotations.

On an objective task, the primary goal for a researcher is to acquire labels that match the labels they would have assigned. The concern is that labelers with different demographic or identity characteristics might systematically perceive something different in the material than what the researcher assumes is the ground truth. We refer to this as "labeler-characteristic bias," or LCB.

Consider a task that asks annotators to indicate if there are any Republican politicians in a photo. There is a ground truth here – either a Republican politician is pictured or not. However, we might imagine that Democratic annotators may not be as good at recognizing Republican politicians as Republican annotators. A mixed-partisan labeling pool might

[5] We thank an anonymous reviewer for suggesting Salganik (2019, 89-91) for a helpful framework summary.

demonstrate weak across-group IRR, pulled down by the poor performance of the Democrats, but with strong within-Republican IRR. We would only see the differences within- and across-groups if we had a diverse pool of labelers, collected demographic data, and checked if there were systematic differences in labels between groups. In our meta-analysis of political science papers, the average reported IRR statistic was around 0.7, with many instances falling below this agreement level. By comparing within- and across-group IRR, researchers may be able to better diagnose sources of unreliability and improve performance.[6]

Some solutions to LCB come immediately to mind. For example, the researcher could invest in better training or codebooks for the annotators, perhaps providing politician reference photos in this case. Or they could retain only accurate labelers, based on a researcher-generated ground-truth. They could also take the findings from the diverse annotators as indication that they need to rescope their research goals. While these solutions may be practical and defensible, they risk smoothing over theoretically important variation.

The researcher may also be confronting a task they thought was objective that has been revealed to be subjective. Take our task of identifying "protest" in images, which we initially assumed was objective. Only after we compared responses by labeler characteristics did we discover the systematic variation. In short, tasks that researchers assume are fully objective may turn out to be closer to the subjective end of the spectrum. The silver lining of this perhaps disheartening discovery is new opportunities to explore interesting subgroup differences - opportunities that arise only because you have collected information about your annotators.

What should researchers do if they suspect prior to labeling or discover after labeling that there are differences in annotations based on labeler characteristics? It depends on the task type, the inferences the researcher wants to make, and whose interpretations they wish to privilege. If the researcher wants to accurately replicate what they themselves would

[6] Online Appendix A, Figure 1 summarizes all reported IRR statistics from the meta-analysis.

see, read, or hear in the material, then the solution is to fix wrong labels by accounting for systematic bias away from the researcher ground truth. We refer to this class of solutions as "adjust." If the researcher wants to infer what a specific type of person would see or read in the material, then the solution is to make sure that the annotators represent the person the researcher envisions. We refer to this class of solutions as "weight." Finally, if there are many potentially correct responses to the task, then the solution is to represent the diversity of responses. We refer to this class of solutions as "model." Table 1 summarizes the issues and solutions facing researchers with large-scale human annotation tasks.

Consider a situation where a researcher has collected a million images from a social media platform. They want to label the images for two things: the presence of a water fountain and the presence of a patriotic symbol. The first task is arguably more objective, while the second is more subjective. The researcher draws a sample of ten thousand images from the million images for human annotation. They then take a subsample of one thousand images from the larger sample and explore the images themselves to get a sense of the variation. They label the images for their concepts of interest to create a researcher ground-truth. They write a codebook, recruit diverse annotators, collect data on their annotators, train annotators and run a pilot on the subsample of images that the researcher labeled.

On the fountain task, which the researcher assumed was objective, the pilot annotations could reveal many scenarios with potential problems and solutions. For example, imagine that the diverse pool of annotators all sees fountains in the one thousand images at roughly the same rate. IRR is high within-group, across-group, and with the researcher-generated ground truth. This is the best-case scenario for tasks assumed to be objective – the researcher has evidence of a single, near-universally stable concept of a fountain. They would be on solid ground following the same annotation procedure for the rest of their sample and training a machine learning model to learn whether a picture contains a fountain or not.

Table 1: Threats to Annotation Reliability/Validity by Task Type

For all tasks		
Problem	*Indicator*	*Solutions*
Annotator inattention, fatigue, or boredom	Speeding through task; poor IRR (within or across groups) after training with other coders on double-coded subsamples	Attention checks; force breaks
Unclear task, questions, or codebook	Confusion expressed by labelers, poor IRR (within or across groups) both during training and after	Pilot task, questions, codebook, and instrument extensively; clarify task
For objective tasks		
Problem	*Indicator*	*Solutions*
Annotators perceive the concept differently than the researchers	Annotations do not match researcher ground truth on sample	Provide training; clarify instructions/definitions; only use annotators whose responses are correct based on ground truth sample
Labeler characteristic bias	Stubbornly low across-group IRR after exhausting above solutions; may have high within-group IRR	Only use annotators whose responses are correct based on ground truth sample; **adjust** labels to account for systematic differences from expected ground truth
For subjective tasks		
Problem	*Indicator*	*Solutions*
Annotator pool does not match underlying population of interest	Divergence in demographics between annotators and target population	**Weight** labels to match ideal population; **model** subgroups separately
Need to represent diversity of responses	Poor across-group IRR	**Model** subgroups separately

Another scenario that could emerge from the pilot, however, is poor IRR with the researcher-generated labels. Imagine a pool of 10 labelers, 5 of whom are women and 5 of whom are men. The researcher finds that one of the women and one of the men are consistently mislabeling fountains compared to the ground truth (for example, perhaps they have not identified drinking fountains as fountains). The researcher could try retraining the two poor annotators, ideally with a discussion about what the annotators are seeing and why they are misperceiving fountains. This discussion and training might lead to refinements to the codebook or instructions. The researcher could also institute attention checks or forced breaks to ensure that the mislabeling is not due to boredom or inattention. They could ultimately decide that these two annotators are simply not up to the task and not retain those coders past the pilot. In this scenario, the researcher is privileging their notion that "fountain" is an objective concept, and that error away from seeing fountain they way they do is due to incapable annotators.

What if the poor annotators were all from one demographic group, however? What if the researcher found strong IRR within the female and male labeling pools, but poor IRR across the groups? Imagine that within the all-female subset, there is stronger agreement with the researcher ground truth labels compared to the all-male subset. Again, our researcher could provide more training and discussion about the discrepancies. If the differences persist, however, what should the researcher do? They could retain only the female labelers beyond the pilot. They could release the task to all annotators and adjust the male labels to account for systematic bias away from the researcher's perceptions.[7] Or they could concede that maybe there are many different ways to define/see what a fountain is.

In this case, we enter a sub-branch of the researcher decision tree, one where the solutions become about addressing the subjectivity of the responses. We see that there are consistent differences in how men and women perceive a fountain. How can we honor the differences from the annotators as we move forward in the research?

One option is to *weight* the responses so that they match the ultimate subject of interest to the researcher. Let's say that the researcher wants to know how the average American

[7] In essence, this means modeling to correct for measurement error in the labels (Chen, Hong, and Nekipelov 2011; Hopkins and King 2010). More recently, Fong and Tyler (2020) and Egami et al. (2024) address the perils of measurement error from machine learning models in downstream analyses.

Facebook user would perceive the image to test how the image content translates into likes on the post. We could weight the responses from women and men (and other subgroups) according to their proportion of the Facebook user population in America to generate a single composite score. We could use this composite score for each picture in downstream analyses. If we use these data to train a ML model, we would need to clearly state that we are training the model to see like an average American Facebook user.

However, seeing like the average American may not feel like a satisfactory solution. We have smoothed over interesting variation in responses. A *modeling* solution would take that variation into account. Instead of assuming that there is one answer per image, we could allow there to be many correct answers. If we still want to train a ML classifier, it would be more appropriate to train separate classifiers for each subpopulation or use a more complicated ML architecture that can account for metadata. We would develop a fountain classifier that sees like the women annotators, a fountain classifier that sees like the male annotators, etc.

Note that the weighting and modeling solutions are the same solutions that the researcher would likely already be considering for their more subjective task of labeling for the presence of a patriotic symbol. Here the researcher assumed that what counts as patriotic would vary between groups. They would have thought about whose perceptions they were most interested in understanding – Women? Men? An average American? These questions would have informed not only their treatment of the data after it was collected but also their annotator recruitment. If we wish to speak to diverse interpretations, we require diverse annotations.

To address these issues, a pilot with a small subset of data and multiple labelers, as demonstrated in our example, can be sufficient. At a minimum, researchers should be transparent about their labelers and steps they take in the annotation process to account for potential LCB. Adjusting, weighting, and modeling solutions are potentially appropriate even with small groups of labelers, especially if there is sufficient diversity in a pilot to indicate that these solutions are warranted.

In the remainder of the paper, we explore labeler characteristics issues in two concrete examples. Both address two characteristics: gender and partisanship. To our knowledge, no prior research has explicitly examined whether partisanship impacts large scale image and text annotation tasks. 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. Boussalis et al. (2021, A5-A7) do discuss variation across gender in labels, finding that labels from female annotators more closely match model predictions. 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 among women (Brown 2014; Deng et al. 2016) and greater sensitivity to emotional nuance (Fischer, Kret, and Broekens 2018).

3 Example 1: Image Annotation

Image analysis has long been of interest to political scientists and has generated increasing attention with the rise of digitized media in politics, especially on social media and in social movement mobilization (Casas and Webb Williams 2018; Kharroub and Bas 2015). Our first illustration of the labeler characteristics issue in annotation relates to whether tweets promoting social movement political action are more likely to be retweeted based on their accompanying images. The image annotating tasks fall into two general categories – *content* and *reactions*. As initially conceived by the authors, we believed that the content labeling would be *objective*, while the reactions tasks were more *subjective*. The content questions asked respondents about things they saw in the images, specifically things that might predict political mobilization, such any celebrities or leaders, a protest, social symbols (e.g. a flag), or someone who "looks like me." We are also interested in potentially mobilizing reactions to images, including whether the image was humorous or ironic and the level of 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).

Table 2: 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

Do we see variation in annotation responses based on who our labelers were? If we do, how might our downstream conclusions about the associations between images and retweets be affected by the labeler characteristics? 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 and thus collected data on these dimensions about our annotators. We did not limit the demographic information collected to just these characteristics, however. Having more demographic information on hand (e.g. education, income, religion) allows for a more comprehensive consideration of potentially relevant identities.

Our images are drawn from tweets associated with left-leaning social movements in the United States. 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. 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 online 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. We used an unsupervised visual clustering method similar to Peng (2020) to construct a stratified sample of about 7,500 images that cover a wide array of topics and account popularity.

We then used the Qualtrics panel service to recruit self-identified Republicans and Democrats (2,140 total respondents). Annotators were over 18, English speakers, and based in the United States. Prior to labeling, they answered a set of demographic and media use questions and passed an attention check. Each annotator then answered questions about 8 images associated with a hashtag (see Table 2). At least one Republican and one Democrat annotated each image. We also aimed for equal representation for women and men in the labeling population.

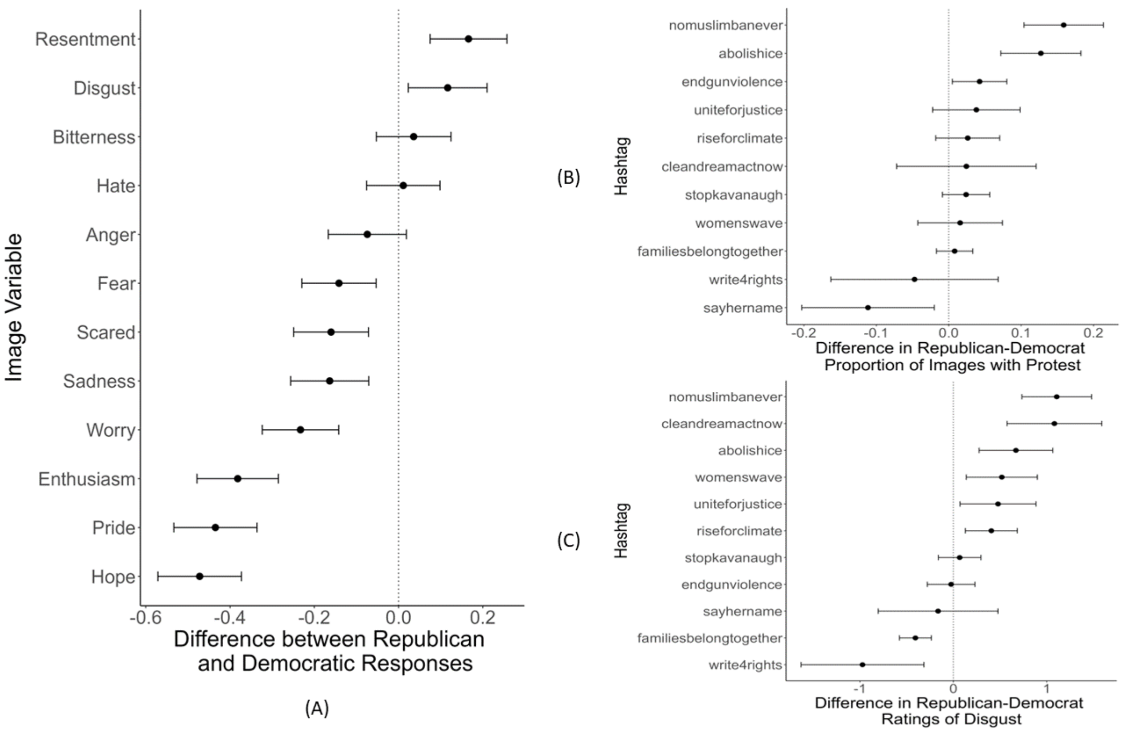
Because our responses were crowdsourced, we have many labelers but little to no overlap between pairs of coders – as such we do not have traditional IRR statistics to report. However, we can test how average labeling responses differ across demographic groups. We can also test the impact of different types of images on retweets varies depending on whose labels are used.

Figure 2.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 expect for these more subjective tasks on images from 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 fear, scared, sadness, worry, enthusiasm, pride, and hope. Republicans were more likely to respond that images elicited disgust and resentment. We see no significant differences in bitterness or hate. The largest difference is about 0.4 on an 11-point scale, which represents about 10% of a standard deviation for the emotions tasks.

More unexpected than the differences in reactions to the images are differences in content labels. These more objective tasks still showed systematic variation between the partisan labelers. For example, Figure 2.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 do not discuss differences for all of the possible annotations and hashtags. However, the results for one emotion, disgust, are illustrative. Disgust has been of particular interest to political scientists in recent years (Kam and Estes 2016; Aarøe, Petersen, and Arceneaux 2017; Ksiazkiewicz, Window, and Friesen 2020; Ahn et al. 2014). Figure 2.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 no movement type 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.

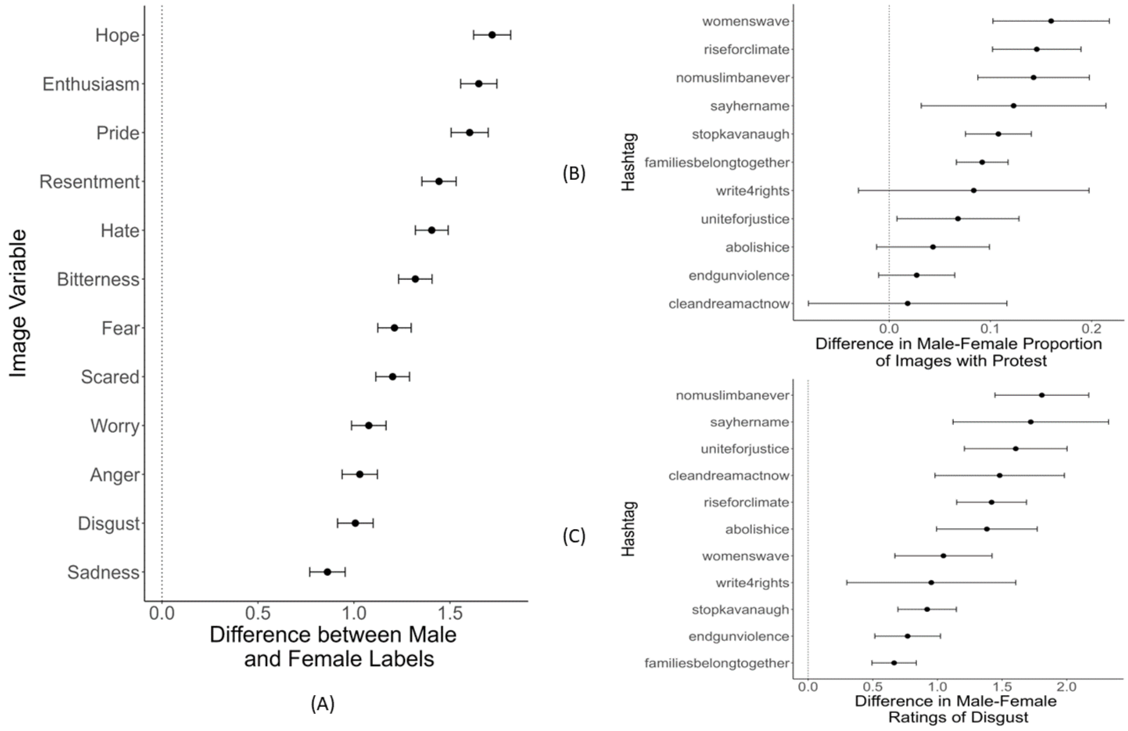
Figure 2: **Differences in Image Annotation by Partisanship:** The average differences between Republicans and Democrats in labeling images from different hashtags, with 95% confidence intervals around the differences in means. Panel A displays the average partisan difference of twelve emotional reactions 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.



Whereas the partisan differences vary by hashtag and image content, the differences in labels assigned by men and women were clear and consistent. Across the board, the men report stronger emotional reactions (Figures 3.A for differences in all emotions pooled on all hashtags and 3.C for differences in disgust by hashtag). Men were also more likely to see protests in images (Figure 3.B). Given these consistent differences, it would be relatively straightforward to follow one of the solutions in Table 1 and adjust responses between men and women to produce one "true" response. We could systematically lower the scores from men or raise the scores from women. Doing so, however, paves over the interesting substantive finding of systematic variation across the annotators. Even our objective questions have differences across genders. Perhaps labeling for protest is not actually an objective task and nudges us towards future research questions: why do these labeling differences exist between men and women?

Ultimately, we are interested in whether image content and reactions are associated with more or less mobilization (as indicated by retweets). Would our conclusions differ depending on whose labels we used? 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. 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 2). 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 2. We imagine five scenarios where we recruited specific subsets of annotators. The five regressions use either (1) pooled labels (all labels), (2) Democrats' labels only, (3) Republicans' labels only, (4) men's labels only and (4) women's labels only. Do differences in who the labelers are lead to different conclusions about image content and retweets?

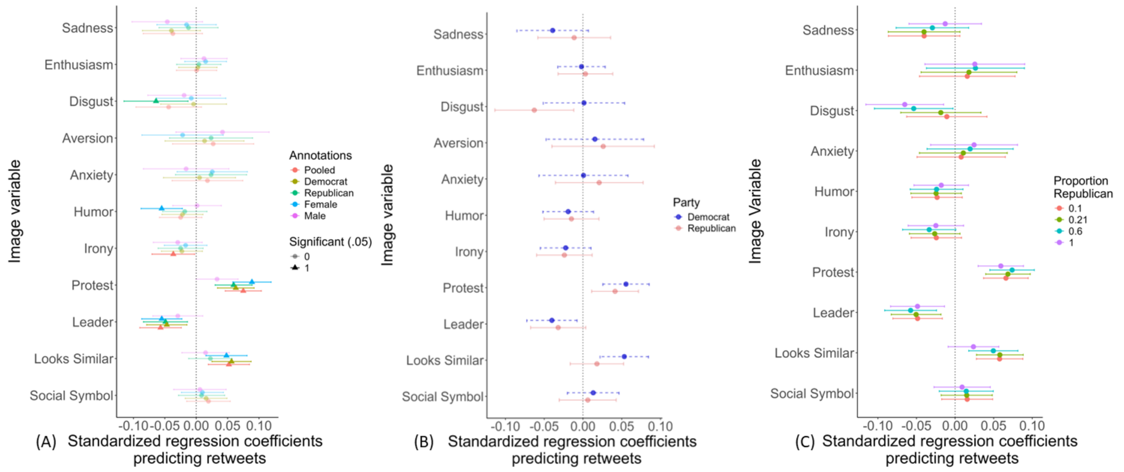
Figure 3: **Differences in Image Annotation by Gender:** The average differences between men and women in labeling images from different hashtags, with 95% confidence intervals around the differences in means. Panel A displays the average gender difference of twelve emotional reactions to all of the image from all hashtags. Panel B shows differences between men and women in the proportion of images where labelers saw a protest, by hashtag. Panel C shows 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.



To be clear, these models are intended to demonstrate that a scholar would potentially come to different conclusions based on whose image labels were collected. We are not trying to draw definitive conclusions. However, as a check that our models with all possible variables are not exaggerating coefficient sensitivity, we include in online Appendix E two alternative model specifications that separate the "content" and "reactions" image variables.[8]

[8] As with the fully-specified model in Figure 4.A, we see coefficients changing based on whose labels are included in these alternative specifications.

Figure 4: **Regression Results Vary by Annotator Demographics**: 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 ($\alpha < 0.05$) 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 where here each measure is a composite score combining the Republican and Democratic labels, with varying weights on the responses.



(A) Standardized regression coefficients predicting retweets

(B) Standardized regression coefficients predicting retweets

(C) Standardized regression coefficients predicting retweets

Figure 4.A reports the standardized regression coefficients with 95% confidence intervals[9] for the image-features in the models with all controls and all image variables. Coefficients that are statistically significant at the 0.05 level are represented as darker triangles.[10] If we were not attentive to who is doing the labeling, we might come to very different conclusions about the mobilizing effects of image features based on our labeling pool. The signs of some coefficients flip depending on whose labels are used (e.g., for Aversion), some coefficients that are statistically significant with one set of labelers lose their significance with a different set (e.g., Humor), and the positive and significant association between protest images and

[9] Full regression tables are available in online Appendix D. See Online Appendix Figure 4 for a replication of the figure with 99% confidence intervals to account for multiple hypothesis testing.

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

retweets disappears if we rely solely on male annotators.[11] In general, the coefficients for the variables where we observed larger differences across labelers are the most sensitive to whose labels are used in downstream analyses.

Figure 4.B follows the modeling strategy solution from Table 1. To account for the differences in content and reactions, we can report separate coefficients for Republicans versus Democrats or for women and men. That is, we can explicitly model the variable associations on subgroups within the labelers. In Figure 4.B increased levels of Republican disgust are associated with fewer retweets. 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 (see online Appendix D, Table 6 for regression table and online Appendix E, Figure 6 for 99% confidence intervals).

Using the weighting solution from Table 1 shifts our focus to making sure the annotations are relevant to the population we care about. The target population in this analysis is Twitter users – we want to know how Twitter users respond to social movement images. Because we have demographic information, we can weight our annotations to create a single composite label score that approximates the population of interest. Of course, 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 Hughes 2019). We weight our image labels based on that proportion, assuming that that the remaining Twitter population is all Democrats. After generating the composite score (e.g., *composite enthusiasm* = .21**Republican enthusiasm score* + .79**Democratic enthusiasm score*), we rerun the main model with only the composite measure for each of the variables of interest. As a sensitivity analysis, we reweight

[11] 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.

the responses thrice more to generate alternative composite measures, with the Republican proportion set to 0.1, 0.6 and 1.0. In Figure 4.C we show the results of the regression analysis that now uses the composite measures (see online Appendix D, Table 7 for regression table and online Appendix E, Figure 6 for 99% confidence intervals). We see different regression coefficients in terms of statistical significance for disgust and "looks similar" depending on the weighting choices.

4 Example 2: Text Annotation

The image data were not initially collected to test for labeler characteristic issues. The following text example was designed to more systematically assess problems arising from labeler characteristics and annotation tasks. Building on the objective-subjective framework we test: (a) for systematic annotation differences for labelers of different identities (specifically ideology and gender); and (b) whether identity-based differences can be mitigated with further training, again depending on annotation task.

Table 3: Seven Text Annotation Tasks, From Most to Least Objective as Initially Assumed

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?
(3a) Conservative view	Does the message reflect or contain a conservative (i.e. right-leaning) view?
(3b) Progressive view	Does the message reflect or contain a progressive (i.e. left-leaning) view?
(4) Gender issue	Does the message discuss a gender issue?
(5a) Feeling angry	Do you feel some anger when reading this message?
(5b) Feeling enthusiastic	Do you feel some enthusiasm when reading this message?

The data for this example consist of 150 social media messages (either a Twitter, Facebook, or Instagram post) sent by Dutch politicians during the 2021 electoral 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 hand-picked 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 annotations.

We identified 5 annotation tasks we believed would vary on the objectivity/subjectivity spectrum (listed in Table 3). First, the most objective task ("Directed at") asked annotators to indicate whether the message was directed at someone. Second, a less objective task "Negative tone," asked participants to indicate whether the message had a negative tone. The third task asked respondents to indicate whether a message contained a "Conservative view" and/or a "Progressive view" (non-mutually exclusive). Fourth, a similarly subjective task asked whether a message discussed a "Gender issue." Finally, we assigned two highly subjective tasks - whether annotators felt "angry" and/or "enthusiastic" (non-mutually exclusive) when reading a progressive message.

We recruited 23 undergraduate students from a Dutch university. In a pre-survey, they provided information about their ideology (15 progressive, 8 conservative students) and gender (13 female, 10 male). To assess whether additional training might improve IRR across these tasks and labeler characteristics, we provided escalating instruction and training in each of three annotating sessions. For each session, all 23 coders annotated the same set of 150 messages, randomly sorted and annotated in a different order. At the beginning of the first "Basic" coding session, participants were only given the questions/prompts described in Table 3. 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 (none of the messages to be coded were discussed) and answering outstanding questions about the codebook.

Additionally, two authors annotated the 150 messages for all subjective tasks (i.e., all tasks except the emotional reactions) to create a ground-truth set of labels against which we could compare the student annotations. We discussed the reasons for our decisions and came to agreement on all labels.

We expected IRR to be higher within groups of respondents of the same ideology and gender, especially for the more subjective 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 tasks.[12]. Here we focus on annotation differences by ideology because they were the most stark (results by gender are available in Appendix G).

We used Cohen's Kappa to measure IRR_{ijsz} for each unique pair ij of coders, annotation session s, and annotation task z. Figure 5 sorts the tasks by their level of expected objectivity, so that on the left we have the task we expected to be most objective. In Figure 5.A, with two notable exceptions ("Directed at" and "Gender issue"), we observe IRR to be higher on average for more objective tasks ("Negative tone" 0.57 in the first session; "Conservative view" 0.39; "Progressive view" 0.4), than for more subjective ("Feel angry" 0.35, "Feel enthusastic" 0.3). Contrary to our expectations, IRR for what we assumed was the most objective 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. This is a reminder of the importance of training to clarify concepts. 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.

[12] A blinded pre-registration for this study is available at the following **link** In response to anonymous reviewer comments, we have flattened the typology of tasks to fall on a single dimension instead of the two initial dimensions discussed in the pre-registration. The analysis remains the same.

Figure 5: **Results from the Text-Annotation Exercise:** (A) Average Cohen's Kappa between unique pairs of coders at each training level. (B) Overall improvement in Cohen's Kappa measure between the basic and advanced training sessions. (C) Difference in the average Cohen's Kappa for pairs of coders with the same ideology and pairs of coders with different ideologies.

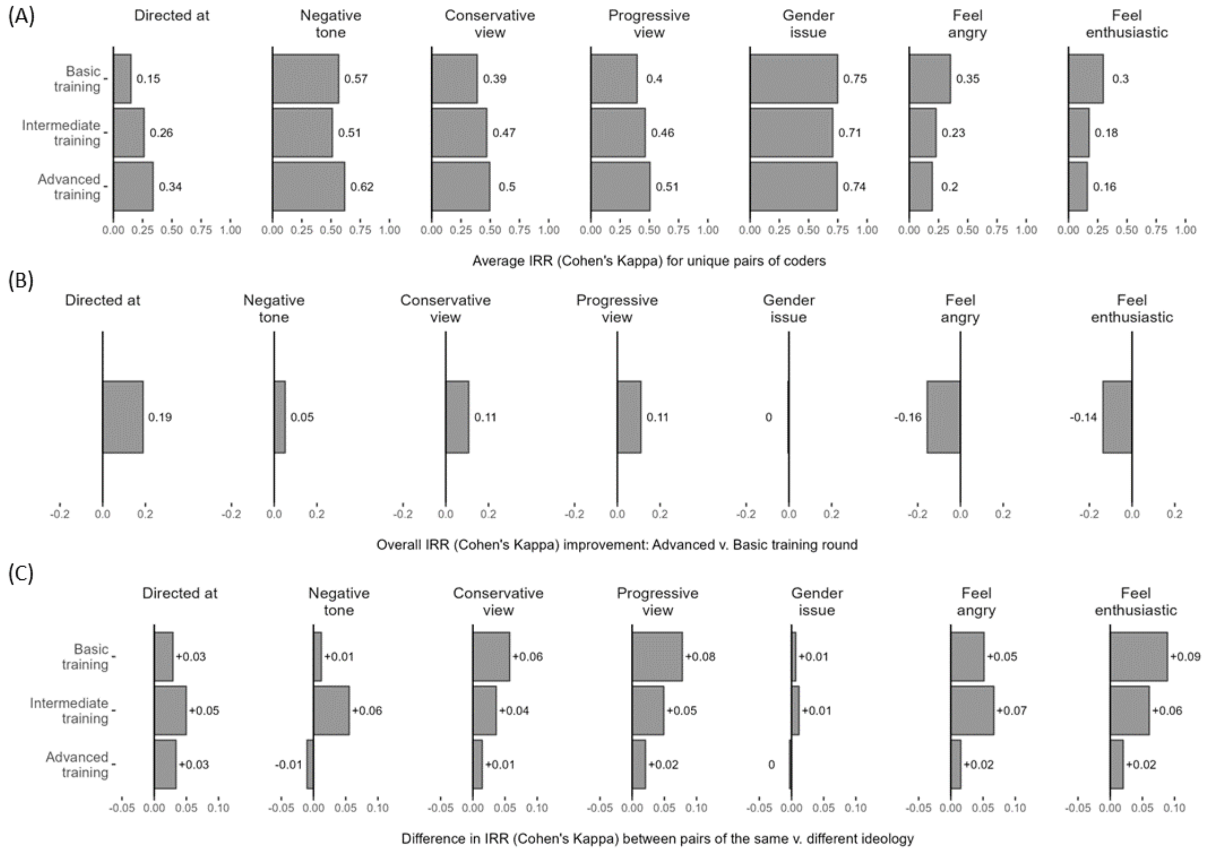


Figure 5.B displays changes in IRR between the "Advanced" and "Basic" training rounds (the difference between the last and first bars in Figure 5.A). As expected, IRR improvement is substantially larger for the most objective task ("Directed at", +0.19). Also as expected, the more subjective the task, the smaller the improvement in IRR from additional training ("Negative tone", +0.05; "Conservative view" +0.11; "Progressive view" + 0.11). IRR actually got worse for the most subjective tasks after training ("Feel angry" -0.16, "Feel enthusiastic" -0.14). Training on these tasks clarified that annotators should be leaning in to their personal interpretations. Online Appendix H (Figure 11) demonstrates how IRR

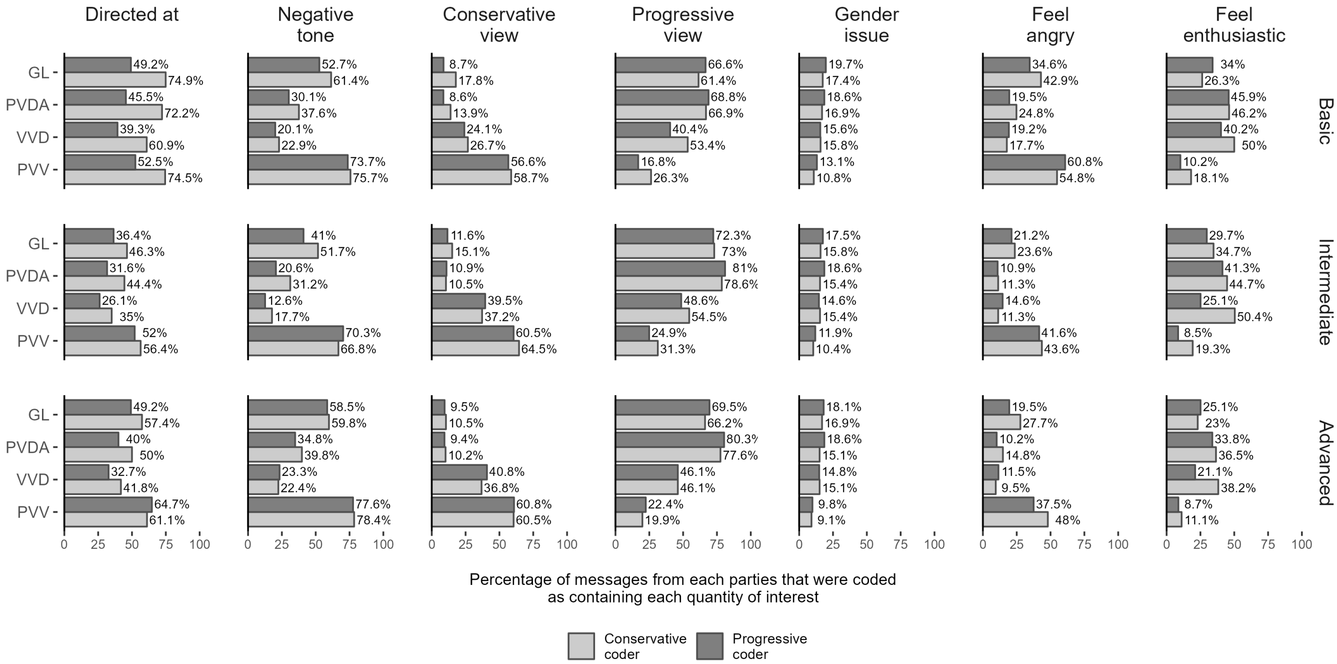
improved relative to the researcher generated-ground truth on all of the non-emotions tasks.

Finally, we compared IRR_{ijsz} for pairs of coders of the same versus different ideologies. In Figure 5.C the values shown are the within-group IRR minus the across-group IRR. Higher values indicate 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 more subjective tasks ("Conservative view" +0.06, "Progressive view" +0.06, "Feel angry" +0.05, "Feel enthusastic" +0.09), compared to the most objective task ("Directed at" +0.03). Figure 5.B indicates that additional training improved the overall IRR for most annotation tasks, with the exception of the two most subjective tasks. Figure 5.C shows that training also reduced the within- and between- IRR gap between coders of the same and different ideologies. In the final "Advanced" round of coding, the time spent training coders helped to improve IRR even in more identity-dependent tasks ("Conservative view" (+0.01) and "Progressive view" (+0.02)). In Appendix F (Figure 9) we show that these descriptive findings hold in a regression framework.[13]

Mirroring what we did for the images study, in Figure 6 we briefly discuss some effects the labeler characteristics 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 6 the parties are sorted by ideology, with the most left-leaning one at the top. The goal is to use this dataset to discuss potential effects on hypothetical downstream analyses, so no meaningful substantive conclusions can be drawn from this analysis.

[13] 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).

Figure 6: **Downstream Analysis Based on whose Annotations are used:** % of messages from each party coded as containing each quantity of interest, by labeler ideology.



In Figure 6, we observe in the first "Basic" round of coding some meaningful differences between the ratings of conservative and progressive participants. 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) more often rate the messages as being directed at someone (e.g. 74.9% v. 49.2%). Finally, we observe these ideological labeling differences mostly washing out in the "Advanced" coding round. Contrary to our initial expectations, as they received further training, participants harmonized their criteria even in the more subjective tasks. In regards to the previous examples, conservative (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 mentioned a gender issue in 16.9% (versus 18.1%). Yet, in line with our expectations, we still observed meaningful ideological differences in the annotations of the “Advanced” coding round when it comes to the most subjective tasks (feeling angry/enthusiastic).

This text experiment illuminates two points about labeler characteristics in human annotation. First, it emphasizes the difficulty of knowing ex ante which coding tasks will be susceptible to LCB. A second point is that coder training can make a difference in improving IRR, even if it cannot entirely remove inter-group subjectivity. 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 within-group 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 rather than getting quick labels online.

5 Conclusion

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), leading to downstream models that carry measurement error (Fong and Tyler 2020).

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). Unfortunately, the tests that political scientists currently rely on to assess coding performance - in particular overall IRR - may not address such concerns.[14] In two demonstration studies (one involving image labeling and one text), we showed how demographic variation in labelers can correlate with significant differences in annotations and different conclusions from downstream analyses. Perhaps surprisingly, we find that labelers with different political identities disagree on basic questions such as what is in a picture or who the target is 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 about which tasks were more objective did not obtain in the annotation data. These findings make it even more imperative that researchers collect demographic information as part of the labeling process and test for systematic labeling differences. We can not assume, as is common practice, that labeler biases are random. We also cannot generalize about which specific characteristics will introduce labeling differences. For example, we found gender differences in image labeling for US social movement tweets, but very few gender differences in the labeling of Dutch politicians' messages.

Addressing the issues raised in this paper starts with a consideration of the task type and the relevant population 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 a universal truth about the images. Instead we would need to draw conclusions about image effects only within that population.

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

Where systematic labeling differences are found, researchers can adjust or weight labels to represent one ground truth, or model differences in their downstream analysis. The image 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 adjusting the ratings for males and raising the ratings for females to get an "average" true response. An alternative solution is to weight composite annotations to match the make-up of the target population. Or, as we did in light of the partisan differences, we can separately model the heterogeneous image-retweet associations with Democrats/Republicans and men/women labelers. Labeler characteristics also raise new and intriguing questions for future study, such as why the differences occur and which images or text factors drive the differences in reactions.

We suggest the following best practices: first, collect data on a range of labeler characteristics, including subgroups that the researchers may not have thought about as relevant ex ante. Second, test if there are differences in annotations across labeler groups – a simple test of differences in mean responses across subgroups can be telling, as can tests of within- and across-group IRR. Third, where differences are found, try to improve labeling practices (e.g., training) to see if they improve IRR (both within- and across-group); model subgroup effects; or adjust labels and labeler populations for subsequent analyses.

Until proven otherwise, we are often on firmer research ground if we think of annotations as survey data instead of as uncovering a single truth. If we recruit a variety of labelers and find that they generally agree with our ground truth baseline annotations, then we can make a stronger case for the objectivity of our task, and can argue that after an initial pilot we do not need to be overly concerned about who our labelers are. If we simply assume that we have an objective task with answers that are obvious to all, we risk LCB and the opportunity to learn more about subpopulations.

6 Acknowledgments

We thank the reviewers and all those who offered suggestsion along the way. For exemplary research assistance, we thank Lucie Lu, Tyler Pichette and Trevor Joireman.

7 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): 277–294.

Ahn, Woo Young, Kenneth T. Kishida, Xiaosi Gu, Terry Lohrenz, Ann Harvey, John R. Alford, Kevin B. Smith, et al. 2014. "Nonpolitical images evoke neural predictors of political ideology." *Current Biology* 24 (22): 2693–2699.

Bachl, Marko, and Michael Scharkow. 2017. "Correcting Measurement Error in Content Analysis." *Communication Methods and Measures* 11 (2): 87–104.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42.

Benoit, Kenneth, Drew Conway, Benjamin Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–295.

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

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. “Debiasing Word Embedding.” In *30th Conference on Neural Information Processing Systems,* 1–9. NIPS 2016.

Boussalis, Constantine, Travis G Coan, Mirya R Holman, and Stefan Müller. 2021. “Gender, candidate emotional expression, and voter reactions during televised debates.” *American Political Science Review* 115 (4): 1242–1257.

Brown, Catherine C. 2014. “Gender Difference in Emotional Reactions to Media : Examining Self-Report During Bittersweet Video Clips.” *Tennessee Research and Creative Exchange: Chancellor’s Honors Program Projects.*

Budak, Ceren, Sharad Goel, and Justin M. Rao. 2016. “Fair and Balanced? Quantifying Media Bias through Crowdsourced Content Analysis.” *Public Opinion Quarterly* 80, no. S1 (April): 250–271.

Casas, Andreu, and Nora Webb Williams. 2018. “Images that Matter: Online Protests and the Mobilizing Role of Pictures.” *Political Research Quarterly* 72 (2): 360–375.

Chen, Xiaohong, Han Hong, and Denis Nekipelov. 2011. “Nonlinear Models of Measurement Errors.” *Journal of Economic Literature* 49 (4): 901–937.

DeBell, Matthew. 2017. “Harder Than It Looks: Coding Political Knowledge on the ANES.” *Political Analysis* 21 (4): 393–406.

Deng, Yaling, Lei Chang, Meng Yang, Meng Huo, and Renlai Zhou. 2016. “Gender differences in emotional response: Inconsistency between experience and expressivity.” *PLoS ONE* 11 (6): 1–12.

Dolezal, Martin, Laurenz Ennser-Jedenastik, Wolfgang C. Müller, and Anna Katharina Winkler. 2016. "Analyzing Manifestos in their Electoral Context A New Approach Applied to Austria, 2002–2008." *Political Science Research and Methods* 4 (3): 641–650.

Egami, Naoki, Musashi Hinck, Brandon M. Stewart, and Hanying Wei. 2024. *Using Imperfect Surrogates for Downstream Inference: Design-based Supervised Learning for Social Science Applications of Large Language Models.*

Fischer, Agneta H., Mariska E. Kret, and Joost Broekens. 2018. "Gender Differences in Emotion Perception and Self-reported Emotional Intelligence: A Test of the Emotion Sensitivity Hypothesis." *PLoS ONE* 13 (1): e0190712.

Fong, Christian, and Matthew Tyler. 2020. "Machine Learning Predictions as Regression Covariates." *Political Analysis* 29 (4): 467–484.

Gamm, Gerald, and Thad Kousser. 2010. "Broad Bills or Particularistic Policy? Historical Patterns in American State Legislatures." *American Political Science Review* 104 (1): 151–170.

Gordon, Mitchell L, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S Bernstein. 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." *Semantic Scholar.*

Grimmer, Justin, and Brandon Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21 (3): 267–297.

Hopkins, Daniel J, and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 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,* 1–12.

Jones, Bryan D., Frank R. Baumgartner, Sean M. Theriault, Derek A. Epp, Cheyenne Lee, and Miranda E. Sullivan. 2023. *Policy Agendas Project: Codebook.*

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): 877–889.

Kam, Cindy D., and Beth A. Estes. 2016. "Disgust sensitivity and public demand for protection." *Journal of Politics* 78 (2): 481–496.

Kharroub, Tamara, and Ozen Bas. 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): 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): 143–166.

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

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): 242–258.

Potter, W. James, and Deborah Levine-Donnerstein. 1999. "Rethinking validity and reliability in content analysis." *Journal of Applied Communication Research* 27 (3): 258–284.

Putnam, Robert D. 1971. "Studying Elite Political Culture: The Case of "Ideology"." *American Political Science Review* 65 (3): 651–681.

Salganik, Mathew J. 2019. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press.

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): 971–1003.

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): 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): 677–691.

Todorov, Alexander, Anesu N Mandisodza, Amir Goren, and Crystal C Hall. 2005. “Inferences of Competence from Faces Predict Election Outcomes.” *Science* 308 (5728): 1623–1626.

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): 731–746.

Wojcik, Stefan, and Adam Hughes. 2019. *Sizing Up Twitter Users.* Technical report. Pew Research Center.

Ying, Luwei, Jacob Montgomery, and Brandon Stewart. 2022. “Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures.” *Political Analysis* 30 (4): 570–589.

8 Biographical Statements

Nora Webb Williams is an Assistant Professor at the University of Illinois Urbana-Champaign, Urbana, IL, 61801. Andreu Casas is an Assistant Professor at the Royal Holloway University of London, London, UK. Kevin Aslett is an Independent Scholar, Bellevue, WA, 98005. John Wilkerson is a Professor at the University of Washington, Seattle, WA, 98195.