Whose Truth is it Anyway? An Experiment on Annotation Bias *

Mariken A.C.G. van der Velden*
Dep. of Communciation Science, Vrije Universiteit Amsterdam
Myrthe Reuver
Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam
Wouter van Atteveldt
Dep. of Communciation Science, Vrije Universiteit Amsterdam
Antse Fokkens
Computational Linguistics & Text Mining Lab, Vrije Universiteit Amsterdam
Felicia Loecherbach
Center for Social Media & Politics, New York University
Kasper Welbers
Dep. of Communciation Science, Vrije Universiteit Amsterdam

Information is key to inform the behavior of citizens, and thereby for the social scientists studying them. The democratization of data has lead to numerous possibilities to gather and analyze textual data. These enourmous amounts of data are typically handled by machine learning techniques to classify into meaningful variables. The performance of these models is evaluated based on a gold standard, created by human annotators. Having a high level of agreement between these annotators is key, but some suggest personal characteristics of annotators, like political ideology or knowledge, interfere. We show in two pre-registered experiments that XXX. Thereby [contribution].

Keywords: Experiment, Annotation Bias, Ideology, Measurinig Political Position, Text-as-Data, Political Knowledge

---

Introduction

Information is key to inform the behavior of citizens, and thereby for the social scientists studying them. Classical theories of political communication, such as agenda setting or framing (e.g. Van Aelst and Walgrave 2016; Lecheler and De Vreese 2019), formulate that political information drives opinion formation and participation in politics – from voting to protests (for overviews hereof across countries, see Pfetsch and Esser 2013). The democratization of data and advent of computational social science has paved the way for new possibilities to gather and analyze textual data (for a recent overview, see Van Atteveldt, Trilling, and Calderon 2022). In particular, advances in Natural Language Processing (NLP) have made it possible to automatically analyze large quantities of data using machine learning (e.g., see Bender 2016; Wei et al. 2023). To determine the validity of such large scale, computational analyses, we rely on “gold standard” data, created by human annotators. If follows that the validity of these analyses hinges on the quality of these golden standards. If a gold standard contains biases, i.e., systematic errors, it can foster bias in any downstream analysis.

Most data collection efforts to create gold standards assume that there is only one correct interpretation for every input example, and that disagreement between the annotators is something that needs to be dealt with at all costs (Aroyo and Welty 2015a). A recent study (Van Atteveldt,

** = Corresponding author, Replication files are available on the author’s Github account (https://github.com/MarikenvdVelden/bias-experiment); Author contributions: a) designed the study: MACGvdV, WvA, AF, FL, MR, & KW; b) conducted the study: MACGvdV, FL & MR; c) data cleaning & analysis: MACGvdV; d) writing of the paper: MACGvdV

Van der Velden, and Boukes 2021) demonstrates that using crowd-coding platforms is a good way to collect such golden standards without too much disagreement. These platforms have also been used for experiments, and the quality of the respondents has been a center of attention (Coppock 2019; Clifford, Jewell, and Waggoner 2015; Huff and Tingley 2015; Berinsky, Margolis, and Sances 2014). While it has been argued that this data is of similar quality to data from a random sample of the population (Coppock 2019), others have demonstrated that these online platforms are populated by people that are "unlike" the general public, being younger and holding more liberal values (Clifford, Jewell, and Waggoner 2015; Huff and Tingley 2015). If the latter is true, this might cause a problem for the coding of political texts: Ennser-Jedenastik and Meyer (2018) report that coders of political texts incorporate prior beliefs about parties' issue stances into their coding decisions. The authors find that party labels cue coders to a stance. For example, coders are more likely to report a left-wing party to be pro-immigration and a populist right-wing party to be against based on the exact same sentence. *Is this actually bias or a diversity of view points? And how big of a problem does this bias/diversity of view points generate for scholars relying on golden standard data?*

There is a long history of text annotation in studies analyzing political text. While this yields lots of experiences as to how to train coders so that we get reliable hand-coded data, the procedure is expensive, protracted, and sometimes does not even get us the quality of data we need (Weber et al. 2018). While the crowd offers a solution to some of these issues (Van Atteveldt, Van der Velden, and Boukes 2021), the main underlying assumption that there is one correct interpretation for every input example remains untouched. At the same time, trends of polarization have shown that people do interpret information according to their ideological position. *Does this mean that some are right and others are wrong? Or is there an ideological difference in the ground truth?* These questions present a fundamental challenge to the main way of working when collecting gold standard data, as we operate from the baseline assumption that disagreement among the annotators should be avoided or reduced. Typically, when specific cases continuously cause disagreement, more instructions are added to limit interpretations Ying, Montgomery, and Stewart (2022). However, work in computational linguistics has shown that increased annotation instructions do not increase quality (Parmar et al. 2022). This leaves us between a rock and a hard place. Is there a potential bias in annotators that we should account for?

In this paper, we build upon the NLP literature on disagreement – or bias – in annotation (e.g., see Q. Shen and Rose 2021; Geva, Goldberg, and Berant 2019; Sommerauer 2020; Plank, Hovy, and Søgaard 2014a) and so-called perspectivism (Basile et al. 2021; Havens et al. 2022) – i.e. the adoption of methods that integrate the opinions and perspectives of the human subjects involved in the knowledge representation step of the machine learning processes (Basile et al. 2021). This literature puts forward that disagreement can occur because of differences in ideological position or political knowledge (Q. Shen and Rose 2021; Alkiek, Zhang, and Jurgens 2022; Joseph et al. 2021). This allows us to test the extent to which disagreement takes place, for what type of stances, as well as gives us directions on how to deal with the diversity in conceptions and the political heterogeneity that nowadays potentially occurs in our sample of annotators. To do so, we have fielded two high-powered pre-registered experiments (see here and here) in the Netherlands – a low-level polarized country – and in the U.S. – a high-level polarized country testing the effect of ideological distance between the annotator and the political actor in the text (H1), the effect of overinterpretation based on political knowledge or ideological engagement (H2), and a offered solution of masking the political actor to mitigate the effects of ideology and knowledge (H3).[1] In the experiments, we vary the level of specification with which a political actor takes a stance – a declarative sentence

[1]The data and research compendium is published on the main author's github page – annonymized for the review process.

versus a sentence where with some knowledge on politics, the stance might be inferred – as well as whether the political actor is shown or masked with putting `[ACTOR]` instead of the political party. We do this for four different political issues: *Environment*, *Immigration*, *Tax Policy*, and *EU* (for the Dutch case) or *Foreign Policy* (for the American case). This country selection does not only allow us to showcase the scope conditions of disagreement due to different levels of political heterogeneity, but also differentiates between languages. English is not only the most dominant language for computational text analysis in the social science [Baden et al. (2022); Dolinsky et al. 2023], crowd-coders do not need to be native speakers, given the dominance of English in our daily lives. This is different for Dutch, it is a language spoken by a smaller community, typically native speakers, yet still an often-enough researched case in computational text analysis in the social science [Baden et al. (2022); Dolinsky et al. 2023].

Our results demonstrate that overall sentence where with some knowledge on politics the stance might be inferred are really difficult for the crowd to annotate – people overinterpret the position using their own knowledge of the world. This is problematic as these sentence are very common in political text – like legislative debates or speeches – as well as media reports. Moreover, our results also demonstrate that for these disagreements in the crowd to occur, the level of polarization needs to be high. We do find support for our hypotheses in the American context, but not in the Dutch context – except for the situation where masking overcomes political knowledge, we do find support for that in the Dutch case, but not in the American one. Our findings thus underline the importance of taking disagreement seriously for the creation of gold standard text – the bread-and-butter of all machine learning endaveours. We should look beyond the majority vote and modeling it in the data, because if an algorithm is trained on biased data from disagreeing annotators, it will reproduce and often exacerbate that bias when it is applied to new data (e.g., see Prost, Thain, and Bolukbasi 2019). To be able to model these characteristic of annotators, we should survey the characteristics of annotators when using the crowd (see Webb-Williams et al. 2023 for a similar argument, yet different annotator characteristics).

Whose Truth is it Anyway? Disagreement & Perspectivism in Creating Gold Standard Data

Generating large data-sets has become one of the main drivers of progress in natural language understanding. In studies of political communication, the most familiar annotation tasks involve identifying basic concepts. This often involve noting the topic of the text, the position of the actor or the tone of the text. A recent study (Van Atteveldt, Van der Velden, and Boukes 2021) demonstrates that using crowd-coding platforms is a good way to collect such large data-sets. However, having only a few workers annotate the majority of text of interest has raised concerns about data diversity and models' ability to generalize beyond the crowd-workers: In a series of experiments, Geva, Goldberg, and Berant (2019) show that often models do not generalize well to annotations from annotators that did not contribute to the training set, suggesting that annotator bias should be monitored during data-set creation. One such potential bias, especially in times of increasing polarization (Iyengar et al. 2019; Gidron, Adams, and Horne 2019; Boxell, Gentzkow, and Shapiro 2022), is based on ideological position of the annotator. Given that annotators on crows-coding platforms tend to be younger and hold more liberal values than the general public (Clifford, Jewell, and Waggoner 2015; Huff and Tingley 2015), this could potentially hamper the data diversity and generalizability of the model. An additional reason to monitor the annotators ideological position as a potential source of annotator bias is that a recent study in NLP showed that experiential factors influence the consistency of how political ideologies are perceived (Q. Shen and Rose 2021). Their finding challenges the "ground-truth" assumption we as researcher make that a position for example is either left-leaning or right-wing leaning. People with different ideological backgrounds might experience that position differently. This challenges our way of data collection. We are interested in the effect of e.g. elite communication. To study this, we allow for heterogeneous treatment effects in experimental work. This indicates that we often do not assume that the treatment, often using text, has the same effect for different partisans. Yet, at the same time, we forget or ignore that knowledge when creating large data-sets for our machine learning models.

The field of Natural Language Processing often works on automatically classifying texts on labels of concepts such as stance, sentiment, and political orientation. These models are trained on data created by human annotators. Often, this process has a final step where disagreements and differences in annotations are leveled by aggregating, averaging, or in other ways coming to a consensus on one label for one example, which is then seen as "ground truth" (Aroyo and Welty 2015b). Differences from this ground truth label are seen as errors that need to be removed or sorted out. Models then learn to predict labels for new examples based on this ground truth. However, since several years there is some discussion on how realistic it is to have one label, especially for subjective or complex concepts and/or texts with multiple interpretations. Aroyo and Welty (2015b) describes succinctly how the predominant annotation procedures for classification models run into myths such as "disagreement is bad" and "one annotation is enough". Plank, Hovy, and Søgaard (2014a) adds to this that underlying ambiguity and linguistic complexity should be considered for disagreement in annotations: not all linguistic examples are created equal. Disagreement even occurs in seemingly objective tasks such as Part of Speech tagging Plank, Hovy, and Søgaard (2014a).

Another aspect to consider is that disagreement can be informative for the concept under measure - sometimes agreement can be used to validate hypotheses about how universal the perceptions of such concepts are (Sommerauer 2020). Additional doubts on smoothing out disagreement in annotation have focussed on the lack of diversity when only annotating with one label or annotator, leading to a homogeneity especially in subjective and social tasks (Geva, Goldberg, and Berant 2019) such as hatespeech detection or political affiliation classification. The question then is: *Whose perspective is being recorded in these datasets, and then later in the models trained on these datasets?* Framing arbitrary representations in data as "bias" misses the political character of datasets: There is no neutral data and no apolitical standpoint from where we can call out bias. Datasets are always "a worldview" [26] and, as such, data always remains biased." (Miceli et al., 2022, p. 5). This is key to social scientist in general, and those studying political text in particular, since several tasks of interest are intrinsically societal, with answers that differ based on the make-up of the worldview of annotators. The answer of the annotators in turn influence how machine learning examples classify new models. For instance, hate speech detection and abuse detection is one NLP task where the race and gender of annotators influences annotators and model performanceWaseem (2016). Language is inherently connected to society and culture: J. H. Shen et al. (n.d.) analyze sentiment analysis, and find that human annotators lead to certain perspectives on sentiment being recorded, and that notably that African American English dialects are often misunderstood by such models.

Most recently, new annotation paradigms have gone one step further by asking whether we modelling the task, or the annotator (Geva, Goldberg, and Berant 2019)? Pavlick and Kwiatkowski (2019) find that for the logical coherence task Natural Language Inference, annotators have several valid interpretations that are not reflected in one ground truth label. They call for new training paradigms that can reflect "the full range of possible human inferences"(Pavlick and Kwiatkowski 2019, 688). Recent approaches in NLP have sought to explicitly incorporate disagreement and diversity in training data annotations. (Röttger et al. 2021) introduces the idea of an explicitly subjective annotation paradigm existing in addition to one focussed on one label and "ground truth". Such a subjective annotation paradigm can be used for purposes where the goal is finding diverse perspectives on the task or concepts, and for models to model more accurately how humans interpret a task or text. Additionally, "perspectivism" (Basile et al. 2021) is a paradigm and research agenda where different perspectives are explicitly incorporated in the training data, and used by models to provide more human-like classifications. Another paradigm is "jury learning" (Gordon et al. 2022), in which machine learning models do not learn to replicate one specific ground truth, but are trained with different annotator juries to reflect the judgement of different populations. In both approaches, demographic and other individual aspects of the annotator are explicitly mentioned and highlighted as having an influence on classification performance, but is used as an asset rather than as an error.

Troiano, Padó, and Klinger (2021, 2) also note how for a complex concept such as the "emotion" of a text, the annotator can make several assumptions during the annotation process on what is wanted, that are all valid and may or may not be useful in different contexts: "It is possible to assess one's own emotion after reading the text, to reconstruct the affective state of the writers who produced it, to guess the reaction that they intended to elicit in the readers, and so on." Q. Shen and Rose (2021) find that political ideology is not an inherent concept in many texts, but rather dependent on who is asked to annotate and their perceptions and background. Extralinguistic factors, such as annotator's own political ideology and also knowledge, influenced annotation and in turn model performance. Thorn Jakobsen et al. (2022) specifically analyze a task related to stance detection, argumentative sentence detection, on these extralinguistic factors, and find an effect of gender and political leaning on annotations and also model performance. However, to our knowledge this phenomenon has not yet been tested in a controlled experimental setting with manipulations in the texts.

To test whether people with different ideological backgrounds as well as their political knowledge might experience that position differently, challenging our ground-truth assumption in data annotation, we propose the following hypotheses:

> **Ideological Bias hypothesis** (*H1a*): The larger the ideological distance between respondent and the party, the less likely respondents annotate statements according to the party's uttered position.

> **Ideological Bias hypothesis** (*H1b*): The effect of H1a is stronger for sentences in which the party's position can be infered (i.e. underspecified sentences).

As noted by Plank, Hovy, and Søgaard (2014b), some linguistic examples are ambiguous, and open to multiple interpretations even for seemingly objective tasks such as part of speech tagging, where annotators have to distinguish parts of speech such as nouns and verbs. A complex concept such as political ideology is much more likely to lead to multiple interpretations. We call such sentences with more possible interpretations and less explicit standpoints "underspecified". A lack of explicitness in the annotated text is one of the main causes of the disagreements in earlier literature. Thorn Jakobsen et al. (2022) deduce that annotator bias comes from a process known as the affect heuristic (Slovic et al. 2007): making a decision based on the emotional response related to your own personal attitude towards the discussed topic, especially when the text is relatively ambiguous. To test whether people with different ideological backgrounds as well as their political knowledge might experience that position differently, challenging our ground-truth assumption in data annotation, we propose the following hypotheses:

Ideological Overinterpretation hypothesis (*H2a*): The larger the ideological distance between respondent and the party, the more likely respondents interpret underspecified sentences as stance.

Political Knowledge Overinterpretation hypothesis (*H2b*): The more political knowledge, the more likely people interpret underspecified sentences as stance.

If these biases exists, how can we alleviate them? There have been several previous approaches to solve biased annotations, especially where it concerns political or societal aspects. Geva, Goldberg, and Berant (2019) introduce an approach where training set annotators are separated from annotators annotating datasets that evaluate the models, to ensure the evaluation is not simply accurate at replicating the original annotators, but can generalize to new annotators' judgements. However, these approaches are not aimed at reducing biases during the training set annotation procedure. Other approaches are aimed at leveraging multiple perspectives - but this is not useful when one wants one label to learn from. For the Austrian National Elections, Ennser-Jedenastik and Meyer (2018) already demonstrated that showing party labels impact annotators' assessment of the party position. So, one solution is masking:

Masking Solution hypothesis (*H3a*): Masking reduces the effect of respondents' ideological position for coding stances according to the party's position.

Masking Solution hypothesis (*H3b*): Masking reduces the effect of respondents' level of political knowledge for coding stances according to the party's position.

Data, Methods & Measurement

Data

We will conduct this survey experiment in the Netherlands in May 2022. The sample, recruited through KiesKompas, consists of 3,000 participants (based on the power analysis presented in our online compendium) of 18 years and older. Kieskompas works with non-random opt-in respondents. Therefore, we measure many demographic background variables, and balance checks have been conducted to demonstrate whether certain categories are over represented in a certain experimental group. The study has been approved by the Research Ethics Review Committee of the *Vrije Universiteit Amsterdam* (see the approval here). To ensure good quality of our data, one attention check (discussed in more detail in Section 3.3) is included (Berinsky, Margolis, and Sances 2014).

Measurement

Experimental Conditions. Respondents are randomly assigned to either view a political party as an actor, or a masked condition, where they see `X` as an actor; simultaneously, respondents see either a fully specified sentence or a underspecified sentence, in which one needs additional information to interpret the positon on an actor. Table 2 gives an overview of the variations in treatment in the survey as well as their English translations.

Dependent Variable. We rely on whether or not a party's (implied) stance is coded according to the party's position (H1 and H3) as well as whether or not the statement is coded as a stance at all (H3). For each issue, we ask the respondent `what is according to the sentence above the position of [ACTOR]?`, with the answer categories: `in favor`, `against`, `no stance`, `don't know`.

Table 1: Survey Questions - Experimental Conditions

Condition	Wording ENG	Wording NL
Specified	[PVV/X] says immigration should be made harder.	[PVV/X] zegt dat immigratie moeilijker gemaakt moet worden.
Specified	[GreenLeft/X] says nitrogen emissions need to be reduced.	[GroenLinks/X] zegt dat stikstofuitstoot meer tegengegaan moet worden.
Specified	[Labour Party/X] says tax rate should go up for highest earners.	[PvdA/X] zegt dat het belastingtarief voor de hoogste inkomens omhoog moet.
Specified	[Forum for Democracy/X] says that membership in the European Union has been especially bad for the Netherlands so far.	[Forum voor Democratie/X] zegt dat het lidmaatschap van de Europese Unie tot nu toe vooral slecht geweest voor Nederland is.
Underspecified	[PVV/X] says many immigrants are coming this way.	[PVV/X] zegt dat veel immigranten deze kant op komen.
Underspecified	[GreenLeft/X] says nitrogen policy must be different.	[GroenLinks/X] zegt dat het stikstofbeleid anders moet.
Underspecified	[Labour Party/X] says tax system must be changed.	[PvdA/X] zegt dat het belastingstelsel moet worden aangepast.
Underspecified	[Forum for Democracy/X] says the Netherlands should have a different role in the European Union.	[Forum voor Democratie/X] zegt dat Nederland een andere rol in de Europese Unie moet hebben.

Moderating Covariates. *Ideological position* is measured using an 11-point scale ranging from left (`0`) to right (`10`). - *Political knowledge* is measured with six items from the DPES.

Controle Variables. In our analysis, we control for demographic information (gender, age, education, income, religion, job) as well as political background variables (trust in politics, ideological position on economic left-right scale and cultural progressive-conservative scale, and evaluations and prospects of the economy). Tables A.5 till A.17 in the OA demonstrate the descriptive information per country.

Method

To test our hypotheses, we will conduct a multilevel model, with respondents clustered in issues, see Equation 1. Using the pooled data we will estimate a within groups fixed effects model. We have conducted a balance test based on demographics (age, gender, education, geographical region, level of urbanness,employment, and income), vote choice in the 2021 parliamentary elections, ideological self-placement, political knowledge, and positions on the issues, using the `cobalt` R package (Greifer 2021). This balance test indicated that none of the variables are unbalanced over the experimental groups, and therefore, as pre-registered, will not be added to the regression formula. $Y\hat{Y}_{r,i,t}$ in Equation 1 denotes the evaluation of a stance by respondent r, during issue i and at experimental round t – ranging from round 1 to round 4. The standard errors are clustered at the individual level.

$$\hat{stancecorrect}_{r,i,t} = \beta_0 + \beta_1 masked_{r,i,t} + \beta_2 specification_{r,i,t} + \beta_3 ideologicaldistancetoparty_{r,i,t} + \beta_4 politicalknowledge_{r,i,t} + \alpha_i + \gamma_t + \varepsilon_{r,i,t} \quad (1)$$

Results

To answer whether there is an ideological or knowledge-basd annotation bias, we have conducted a two-by-two experiment. Table 3 demonstrates the average profile of respondents who annotate correctly and incorrectly. In terms of demographics, there is not much of a difference. Yet, people who are incorrectly identifying stances are more left-wing oriented compared to those who are correct – i.e. an average score of 4 for those who are incorrect vs. an average score of 5 for those who are correct. For other positions on issues or political knowledge, we do not see a difference in averages between those who are correctly and incorrectly identifying stances.

Table 2: Profile Dutch Stance Annotators

Incorrectly Identified Stance	Correctly Identified Stance
Male	Male
High-levels of education	High-levels of education
West of Netherlands	West of Netherlands
Fulltime Employed	Fulltime Employed
D66	D66
Age: 48	Age: 46
Income: 3250	Income: 3250
Position on Immigration: 3	Position on Immigration: 3
Position on Environment: 1	Position on Environment: 1
Position on Tax: 3	Position on Tax: 3
Position on EU: 1	Position on EU: 1
Ideological Position: 5	Ideological Position: 4
Ideological Distance: 2	Ideological Distance: 2
Issue Congruence: 0	Issue Congruence: 0
Political Knowledge: 2	Political Knowledge: 2

Looking at the effect of the experimental conditions on the two dependent variables – 1) correctly identifying a stance; and 2) over-interpreting a stance – Figure 1 visualizes the baseline. The left-hand panel demonstrates the effect of the two experimental conditions for correctly identifying the stance. On average, approximately half of the respondents where correct – as indicated by the intercept. When we mask the political actor – i.e. instead of mentioning the party, we put "X" – we see that this improves correctly interpreting the stance significantly. However, an effect of `0.06` (i.e. `6%`) is not a big improvement. We do see that the level of specification of a sentence has a significant effect. If a sentence is not fully specified, it has a substantive negative effect on the likelihood to correctly interpret the sentence. A coefficient of `-0.38` indicates that compared to a fully specified sentence, `38%` of the respondents are more likely to be incorrect when the sentence is under-specified – that is when the sentence does not state a clear position, but mentions the issue. For the left-hand panel, we look at a different dependend variable, so instead of seeing if

Table 3: Profile American Stance Annotators

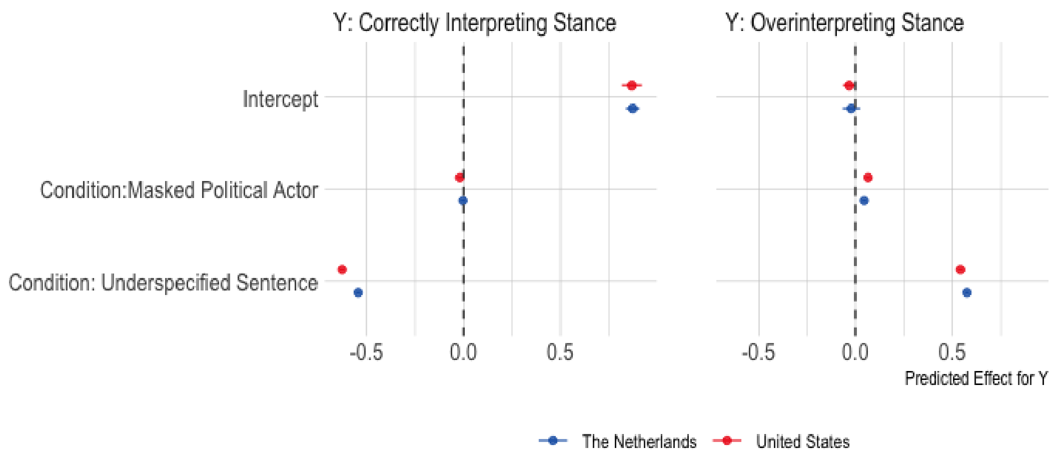
Incorrectly Identified Stance	Correctly Identified Stance
Male	Male
High-level of education	High-level of education
Southeast of the United States	Southeast of the United States
Working now	Working now
Democrat	Democrat
Age: 24	Age: 24
Income: 3250	Income: 3250
Position on Immigration: 3	Position on Immigration: 3
Position on Environment: 2	Position on Environment: 2
Position on Tax: 4	Position on Tax: 4
Position on Foreign Policy: 3	Position on Foreign Policy: 3
Ideological Position: 2	Ideological Position: 2
Ideological Distance: 2	Ideological Distance: 2
Issue Congruence: 1	Issue Congruence: 0
Political Knowledge: 1	Political Knowledge: 1
Bullshit Receptivity: 3	Bullshit Receptivity: 2

respondents where correct or not, we look at whether they interpreted the sentence as a stance or not. On average, almost nobody overinterprets a stance – as indicated by an intercept of `-0.01`. However, if people see an X compared to a political actor, they are statistically significantly more likely to interpret the sentence as a stance. Yet, a coefficient of `0.02` (i.e. `2%`) is a very small effect. For the condition of specification level, however, we see that compared to a fully specified sentence, people seeing an under-specified sentence are much more likely to interpret the sentence as a stance: an increase of `0.83`. This indicates that people do not excel in this task without any instruction. In the pre-registered section, we demonstrate the tests of the hypotheses, and afterwards, we visualize some explorations of the data to show the robustness of our findings.

Pre-registered Results

First, we test whether there is an ideological bias in interpreting stances (H1a), and if this bias increases for those that are further away from the ideological position of the political actor in the under-specified condition (H1b). Figure 2 demonstrates on the left-hand panel the regression coefficients and on the right-hand panel the average marginal effects of the interaction between ideological distance and under-specified sentences. The upper-left panel of Figure 2 demonstrates the coefficient of ideological distances for the likelihood of interpreting the stance correctly. There is a negligible positive effect – a coefficient of `0.002` – that is borderline significant. Substantially, this means that there is no effect of ideological distance for correctly interpreting the stance. Hence, no ideological bias found, thus no support for our H1a. Looking at the lower-left, and right-hand panel of Figure 2, we see a small but significant effect of the interaction between ideological distance and the under-specified condition. It is, however, going in the other direction than hypothesized. For those who are ideologically further away from the party, they are less likely to be wrong than those who are close too the party. Yet, the difference is about `3%` – from a coefficient of `-0.38` to `-0.35`. We thus find no evidence for H1b either. This demonstrates that even in times of hightened

Figure 1: Baseline Results of Experimental Conditions



polarization, ideological annotation bias is not a huge concern.

Secondly, we hypothesized that there is a risk of over-interpretation from those that are ideologically distant to the political actor (H2a) as well as those who have high levels of political knowledge (H2b). We measure this with an interaction between the experimental condition of specification and the variable of interest. Following Brambor, Clark, and Golder (2006), we visualize the average marginal effects for both interactions to enhance interpretation. Figure 3 demonstrates the average marginal effects for both interaction effects: In the left-hand panel the results for H2a, and in the right-hand panel, the results for H2b. The left-hand panel of Figure 3 shows that those who are more distant from the position of the political actor are more likely to over-interpret the sentence as a stance. The effect increases from about 80% to those that are closest to the party to 91% for those who are furthest away from the party. This supports our H2a. The right-hand panel of Figure 3 shows that political knowledge does statistically significantly effect over-interpretation of a sentence as a stance. The substantial effect, however, is small: From the least knowledgeable respondents to the most knowledgeable ones, the likelihood of over-interpretation increases with 2%. So, while we do find support for our H2b, we will in our recommendations section focus less on the political knowledge as a possible interfering feature, but focus on what to do with under-specified sentences when using crowd-coding.

Thirdly, we test whether masking of the political actor is a solution for potentially misinterpreting the stance. We hypothesized that masking should reduce the ideological bias (H3a) as well as the bias resulting from political knowledge (H3b). We test these hypotheses with an interaction between the condition masking and the variables of interest and Figure 4 demonstrates the average marginal effects. The left-hand panel of Figure 4 shows, against expectation, a slight increase: Those that are further away from the masked political actor are more likely to incorrectly interpret the stance. The decrease in slope is however negligible: From approximately 6% to 7% of the respondents being incorrect. The slope for political knowledge, on the right-hand panel of Figure 4, is so-possible even flatter. This means that masking of political actors does not help to correctly interpret a sentence as a stance – i.e. no support for H3a and H3b.

Exploratory Results

To check the robustness of our findings, Figure 5 demonstrates the analyses for each issue separately. On the left-hand panel of Figure 5, we show the results for correctly interpreting the stance. We see some variation between issues. For the issue *Tax* (lowest row), we see that almost everyone interprets the sentence correct – an intercept of 0.96. This increases statistically significantly still with about 1% when masking the political actor, and decreases statistically significantly with 68% when the sentence is under-specified. The regression results (in Online Appendix) show a small effect of ideological distance: the further away, the more likely to correctly interpret the stance – i.e. against H1a. Respondents found it most difficult to correctly interpret the issue *Environment* (top row of Figure 5), only 12% has this correct initially, as indicated by the intercept. This increases statistically significantly still with about 8% when masking the political actor, and decreases statistically significantly with 15% when the sentence is under-specified. The other two issues – *EU* and *Immigration*, in the middle rows of Figure 5 show a similar effect as the main base-line presented in Figure 1. The same pattern holds for the other dependent variable – overinterpretation of a stance – in the right-hand panel of Figure 5. Thus, while there are some differences in effect sizes between the issues, the overall findings are not driven by a single issue.

In addition to issue-specific analyses, we also explore an interaction between treatments, visualized in Figure 7 for both dependent variables. Figure 7 shows that masking is of help when sentences are under-specified. In the left-hand panel of Figure 7, it demonstrates that for under-specifief sentences, people are less likely to incorrectly identify a sentence as a stance when the actor is masked (coefficient of -0.30) than when an actor is revealed (coefficient of -0.45). That means there is a 15% increase in having it correct. The difference for over-interpreting is smaller between revealed and masked political actor – shown in the right-hand panel of Figure 7 – yet also statistically significant. Compared to 85% over-interpreting the sentence as a stance, in the masking solution "only" 80% over-interprets the sentence as a stance. In the recommendation section, we will reflect on the masking solution for under-specified sentences.

Lastly, we explore three different ways of measuring ideological distance. First, we measured ideological bias by looking at whether the respondent is congruent or not with the issue position in the sentence, visualized in Figure 8. Second, we measured ideological bias by looking at whether the person voted for the party displayed in the sentence, visualized in Figure 9. And thirdly, we measured ideological bias by looking at the ideology of the respondents – not in relation to the political actor revealed, visualized in Figure 10. Figures 8, 9, and 10 show that our null-finding regarding ideological bias is not conditional upon the measure we used. In none of the analyses, we find evidence for ideological bias.

Figure 2: Ideological Distance

Figure 3: Results Level of Sentence Specification

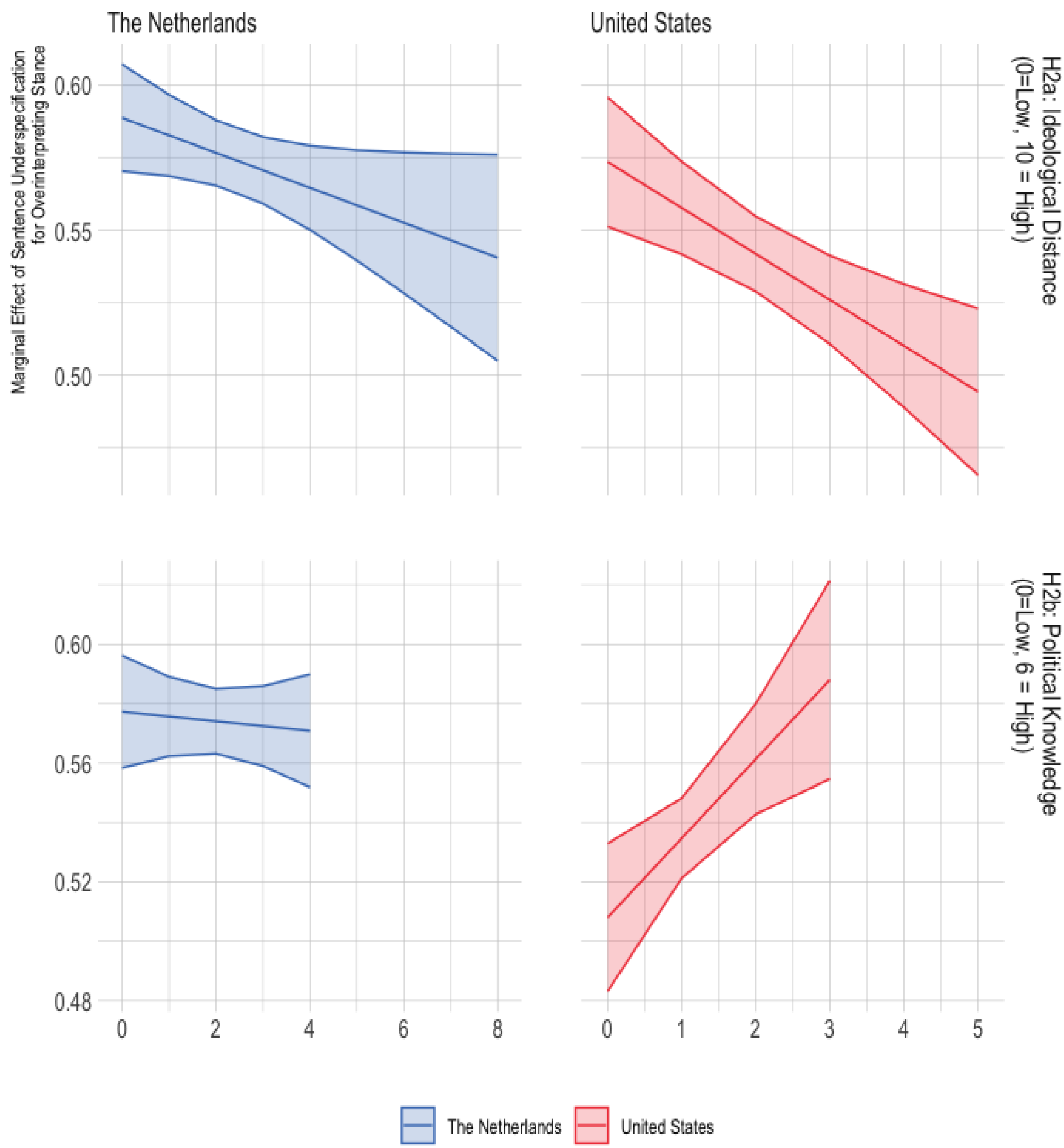


Figure 4: Results Masking Solution

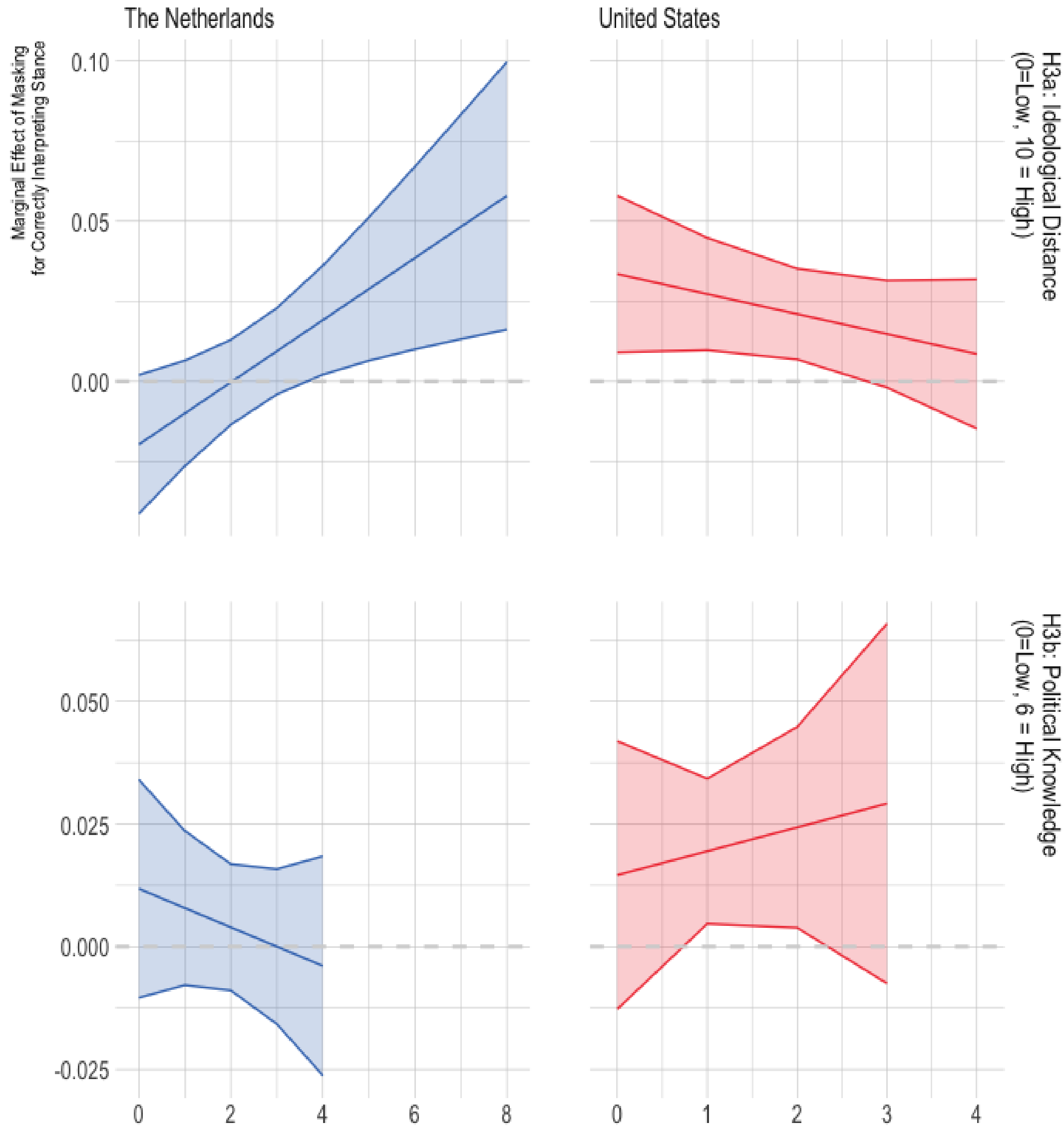


Figure 5: Exploration: Issue Speficific Analyses

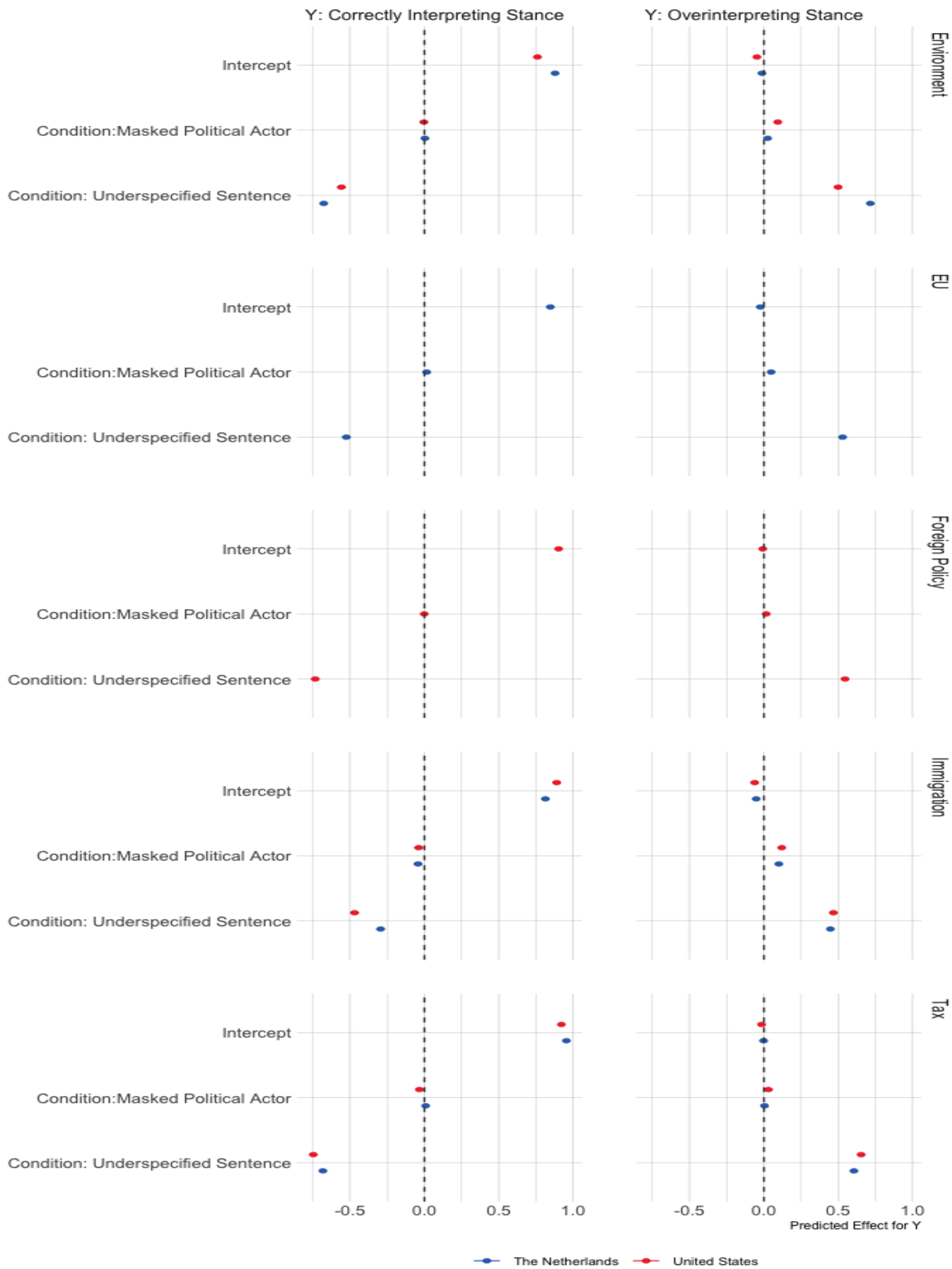


Figure 6: Exploration: Interaction with Treatments

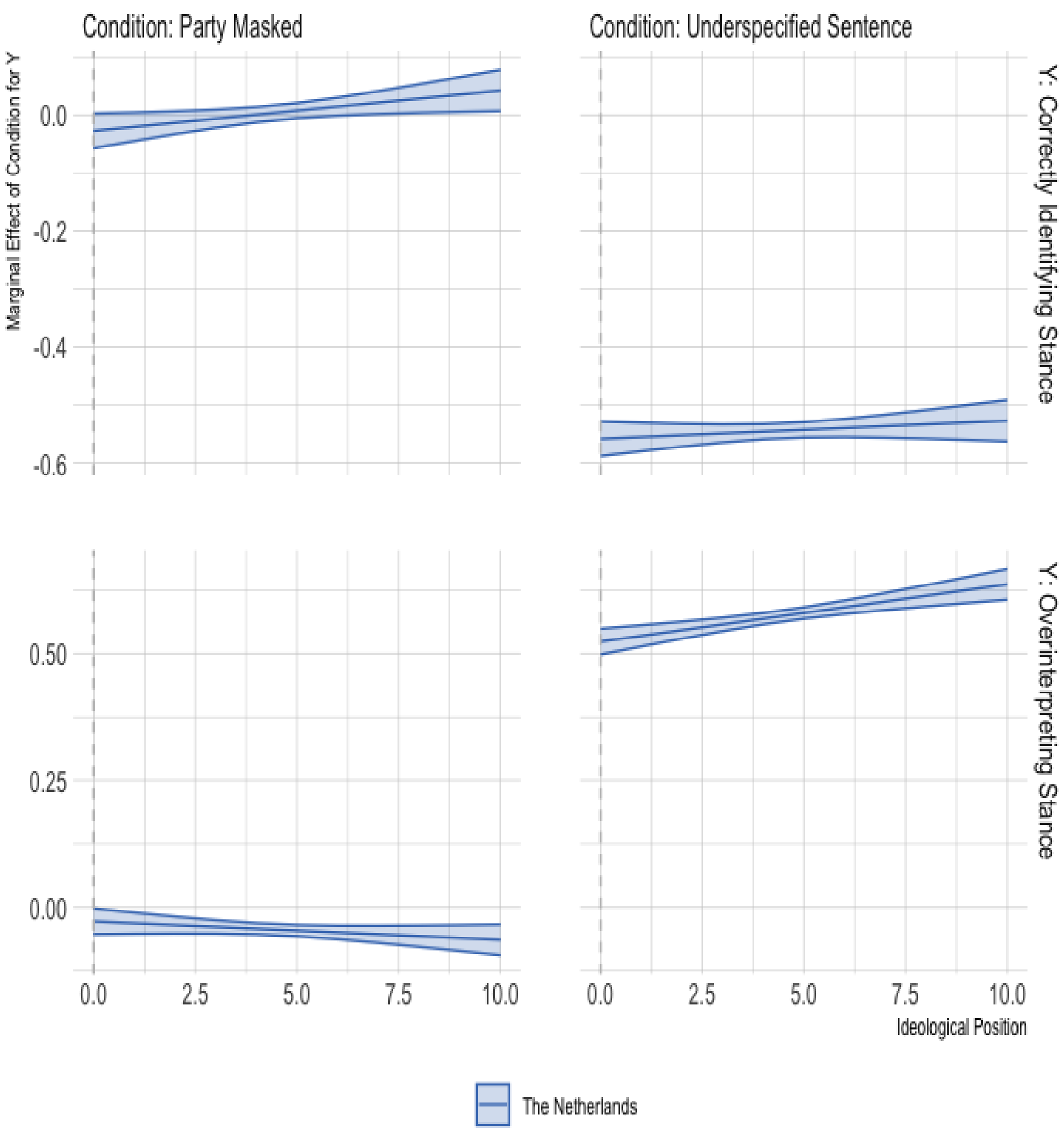


Figure 7: Exploration: Interaction with Treatments

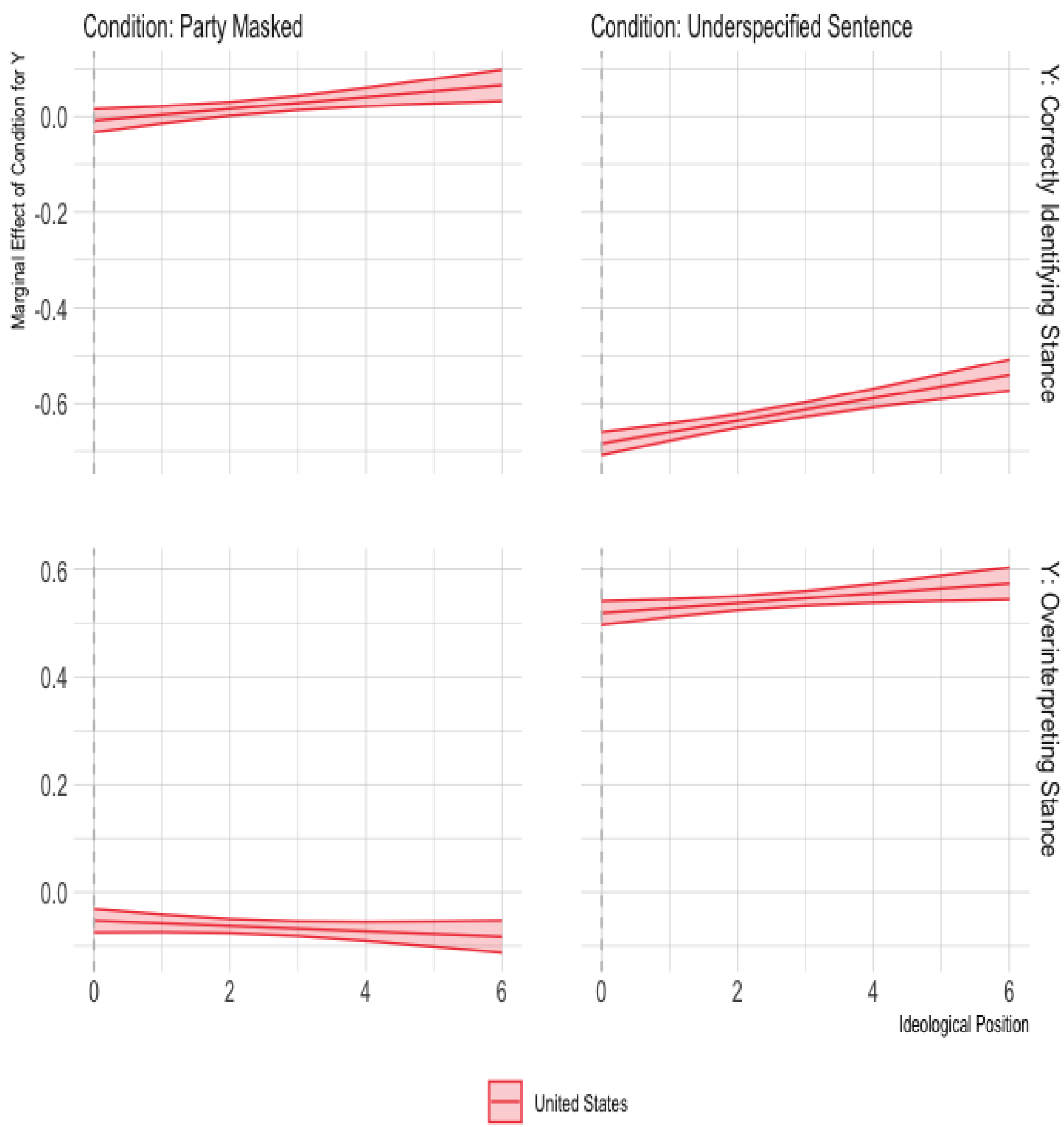


Figure 8: Exploration: Different Indicators of Ideological Distance (1)

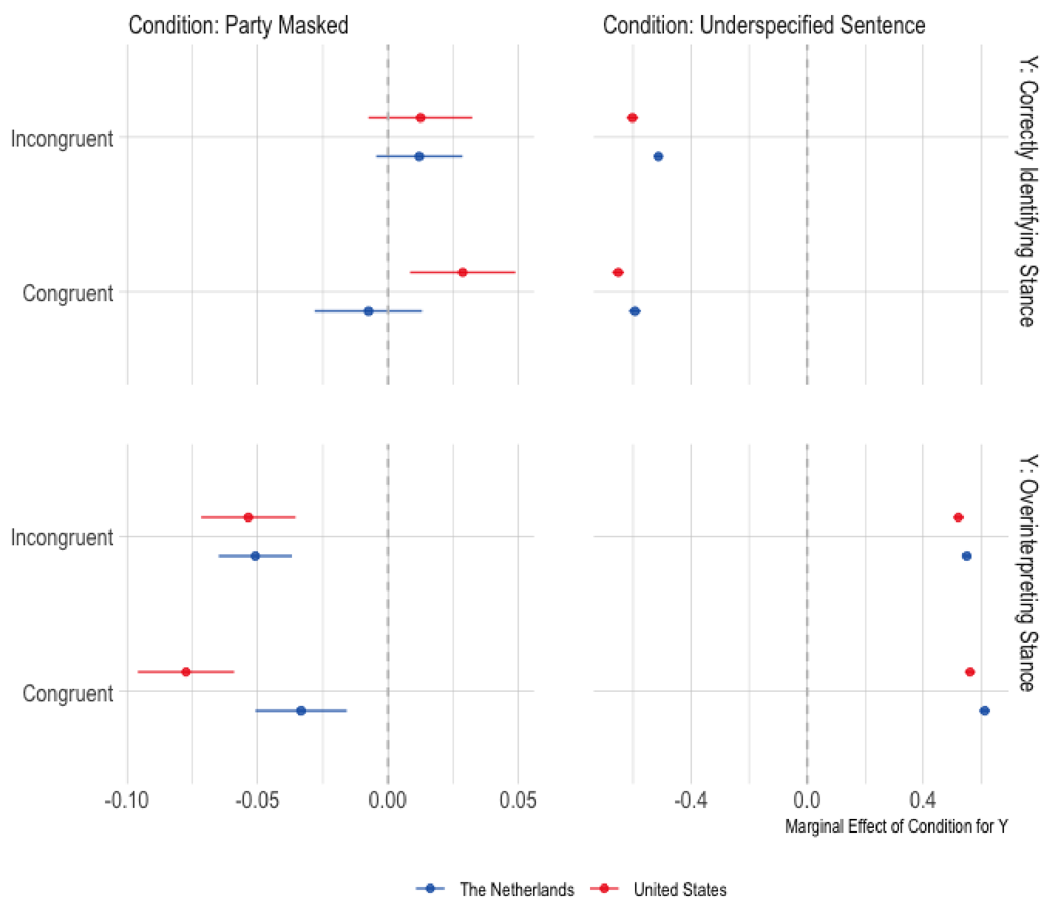


Figure 9: Exploration: Different Indicators of Ideological Distance (2)

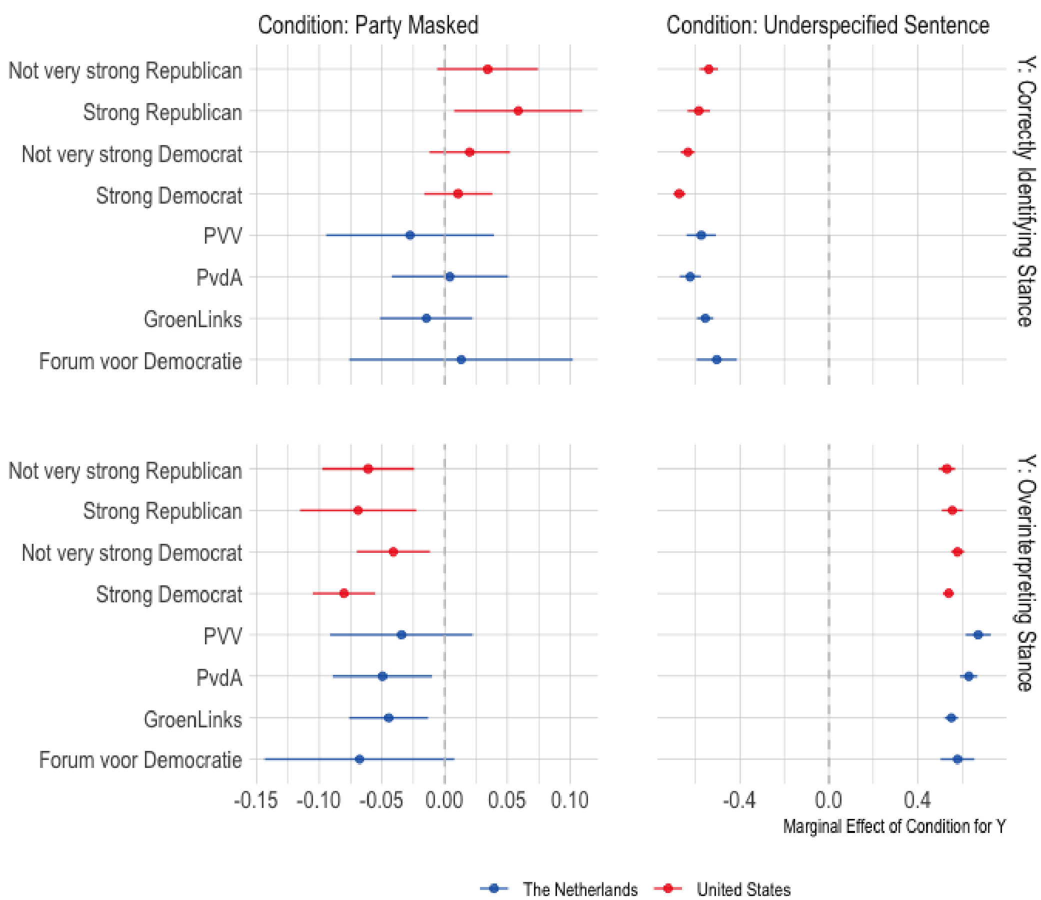


Figure 10: Exploration: Different Indicators of Ideological Distance (3)

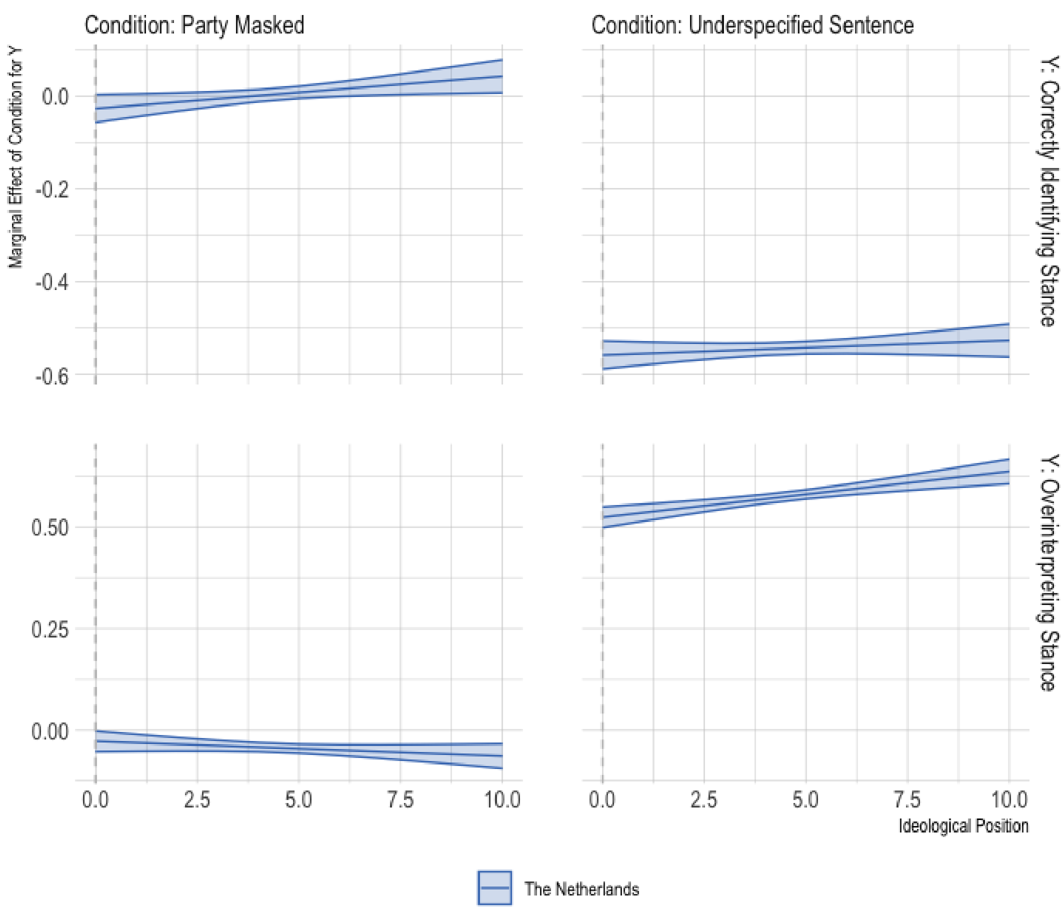
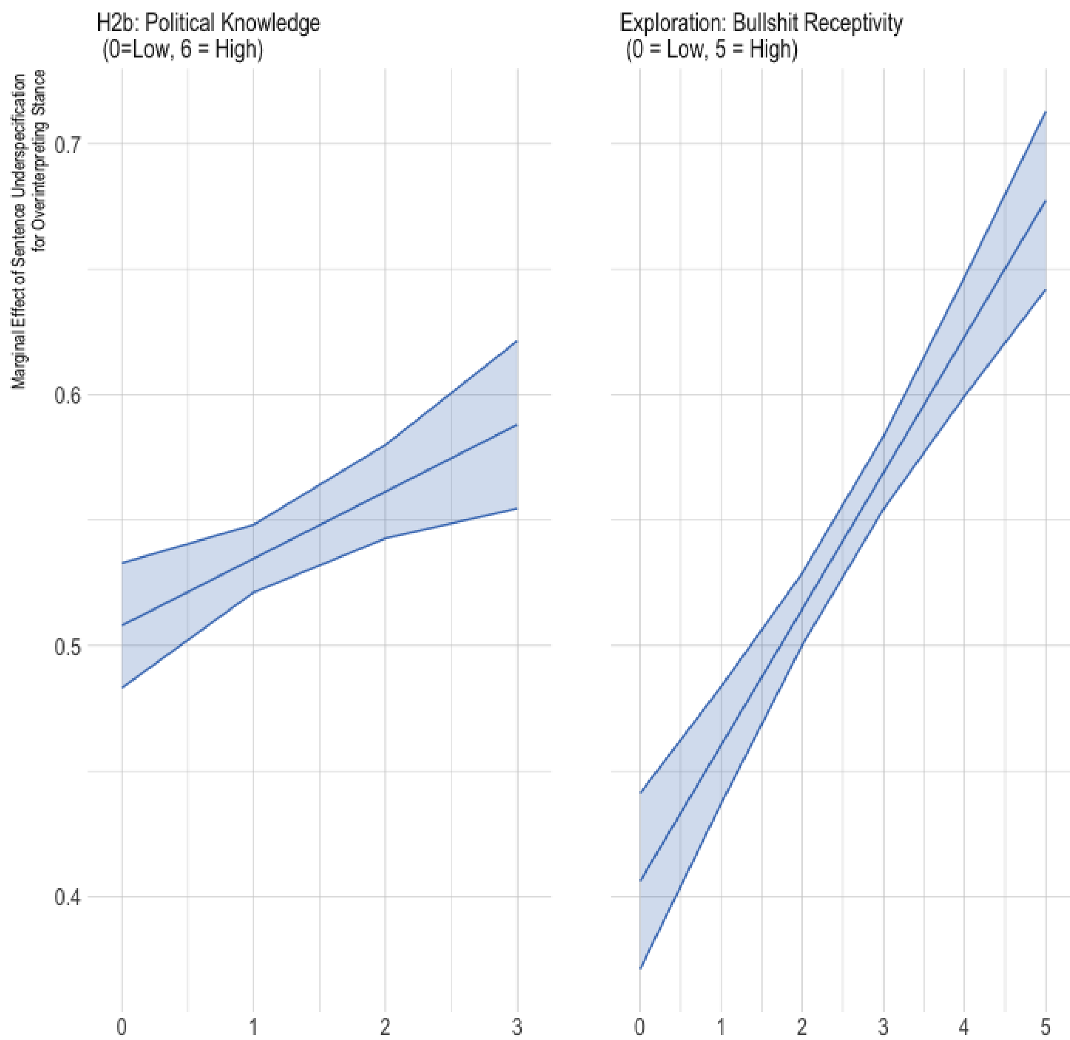


Figure 11: Results Level of Sentence Specification



Discussion

TBA

Recommendations

TBA

References

Alkiek, Kenan, Bohan Zhang, and David Jurgens. 2022. "Classification Without (Proper) Representation: Political Heterogeneity in Social Media and Its Implications for Classification and Behavioral Analysis." In *Findings of the Association for Computational Linguistics: ACL 2022*, 504–22. https://doi.org/10.18653/v1/2022.findings-acl.43.

Aroyo, Lora, and Chris Welty. 2015a. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." *AI Magazine* 36 (1): 15–24. https://doi.org/10.1609/aimag.v36i1.2564 .

———. 2015b. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." *AI Magazine* 36 (1): 15–24. https://doi.org/10.1609/aimag.v36i1.2564.

Baden, Christian, Christian Pipal, Martijn Schoonvelde, and Mariken AC G van der Velden. 2022. "Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda." *Communication Methods and Measures* 16 (1): 1–18. https://doi.org/10.1080/19312458.2021.2015574.

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. https://doi.org/10.1017/pan.2020.8 .

Basile, Valerio, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. "Toward a Perspectivist Turn in Ground Truthing for Predictive Computing." *arXiv Preprint arXiv:2109.04270*. https://doi.org/10.48550/arXiv.2109.04270.

Bender, Emily M. 2016. "Linguistic Typology in Natural Language Processing." *Linguistic Typology* 20 (3): 645–60. https://doi.org/10.1515/lingty-2016-0035.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver, and Slava Mikhaylov. 2016. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review* 110 (2): 278–95. https://doi.org/10.1017/s0003055416000058.

Berinsky, Adam J, Michele F Margolis, and Michael W Sances. 2014. "Separating the Shirkers from the Workers? Making Sure Respondents Pay Attention on Self-Administered Surveys." *American Journal of Political Science* 58 (3): 739–53. https://doi.org/10.1111/ajps.12081.

Boxell, Levi, Matthew Gentzkow, and Jesse M Shapiro. 2022. "Cross-Country Trends in Affective Polarization." *Review of Economics and Statistics*, 1–60. https://doi.org/10.1162/rest_a_01160.

Brambor, Thomas, William Roberts Clark, and Matt Golder. 2006. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14 (1): 63–82. https://doi.org/10.1093/pan/mpi014.

Clifford, Scott, Ryan M Jewell, and Philip D Waggoner. 2015. "Are Samples Drawn from Mechanical Turk Valid for Research on Political Ideology?" *Research & Politics* 2 (4): 2053168015622072. https://doi.org/10.1177/2053168015622072.

Coppock, Alexander. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7 (3): 613–28. https://doi.org/10.1017/psrm.2018.10.

DeBell, Matthew. 2013. "Harder Than It Looks: Coding Political Knowledge on the ANES." *Political Analysis* 21 (4). https://doi.org/10.1093/pan/mpt010 .

Ennser-Jedenastik, Laurenz, and Thomas M Meyer. 2018. "The Impact of Party Cues on Manual Coding of Political Texts." *Political Science Research and Methods* 6 (3): 625–33. https://doi.org/10.1017/psrm.2017.29.

Fornaciari, Tommaso, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. "NAACL-HLT 2021." In, 25912597. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.204.

Geva, Mor, Yoav Goldberg, and Jonathan Berant. 2019. "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets." *arXiv Preprint arXiv:1908.07898*. https://doi.org/10.48550/arXiv.1908.07898.

Gidron, Noam, James Adams, and Will Horne. 2019. "Toward a Comparative Research Agenda on Affective Polarization in Mass Publics." *APSA Comparative Politics Newsletter* 29: 30–36.

Gordon, Mitchell L., Michelle S. Lam, Joon Sung Park, Kayur Patel, Jeffrey T. Hancock, Tatsunori Hashimoto, and Michael S. Bernstein. 2022. "Jury Learning: Integrating Dissenting Voices into Machine Learning Models." *arXiv:2202.02950 [Cs]*, February. https://doi.org/10.1145/3491102.3502004.

Greifer, Noah. 2021. "Cobalt: Covariate Balance Tables and Plots. R Package Version 4.3.1." https://cran.r-project.org/web/packages/cobalt/index.html.

Havens, Lucy, Benjamin Bach, Melissa Terras, and Beatrice Alex. 2022. "Beyond Explanation: A Case for Exploratory Text Visualizations of Non-Aggregated, Annotated Datasets." In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP@ LREC2022*, 73–82.

Huff, Connor, and Dustin Tingley. 2015. "'Who Are These People?' Evaluating the Demographic Characteristics and Political Preferences of MTurk Survey Respondents." *Research & Politics* 2 (3): 2053168015604648. https://doi.org/10.1177/2053168015604648.

Iyengar, Shanto, Yphtach Lelkes, Matthew Levendusky, Neil Malhotra, and Sean J Westwood. 2019. "The Origins and Consequences of Affective Polarization in the United States." *Annual Review of Political Science* 22: 129–46. https://doi.org/10.1146/annurev-polisci-051117-073034.

Joseph, Kenneth, Sarah Shugars, Ryan Gallagher, Jon Green, Alexi Quintana Mathé, Zijian An, and David Lazer. 2021. "(Mis) Alignment Between Stance Expressed in Social Media Data and Public Opinion Surveys." *arXiv Preprint arXiv:2109.01762*. https://doi.org/10.48550/arXiv.2109.01762.

Lecheler, Sophie, and Claes H De Vreese. 2019. *News Framing Effects: Theory and Practice*. Taylor & Francis.

Parmar, Mihir, Swaroop Mishra, Mor Geva, and Chitta Baral. 2022. "Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions." *arXiv Preprint arXiv:2205.00415*. https://doi.org/10.48550/arXiv.2205.00415.

Pavlick, Ellie, and Tom Kwiatkowski. 2019. "Inherent Disagreements in Human Textual Inferences." *Transactions of the Association for Computational Linguistics* 7 (0): 677–94. https://transacl.org/ojs/index.php/tacl/article/view/1780.

Pfetsch, Barbara, and Frank Esser. 2013. "Comparing Political Communication." In *The Handbook of Comparative Communication Research*, 47–69. Routledge.

Plank, Barbara, Dirk Hovy, and Anders Søgaard. 2014a. "Learning Part-of-Speech Taggers with Inter-Annotator Agreement Loss." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 742–51.

———. 2014b. "ACL 2014." In, 507511. Baltimore, Maryland: Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-2083.

Prost, Flavien, Nithum Thain, and Tolga Bolukbasi. 2019. "Debiasing Embeddings for Reduced Gender Bias in Text Classification." *arXiv Preprint arXiv:1908.02810*. https://doi.org/10.48550/arXiv.1908.02810.

Röttger, Paul, Bertie Vidgen, Dirk Hovy, and Janet B. Pierrehumbert. 2021. "Two Contrasting Data Annotation Paradigms for Subjective NLP Tasks." *arXiv:2112.07475 [Cs]*, December. http://arxiv.org/abs/2112.07475.

Shen, Judy Hanwen, Lauren Fratamico, Iyad Rahwan, and Alexander M Rush. n.d. "Darling or Babygirl? Investigating Stylistic Bias in Sentiment Analysis," 5.

Shen, Qinlan, and Carolyn Rose. 2021. “What Sounds ‘Right’ to Me? Experiential Factors in the Perception of Political Ideology.” In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1762–71. https://doi.org/10.18653/v1/2021.eacl-main.152.

Slovic, Paul, Melissa L Finucane, Ellen Peters, and Donald G MacGregor. 2007. “The Affect Heuristic.” *European Journal of Operational Research* 177 (3): 1333–52.

Sommerauer, Pia. 2020. “Why Is Penguin More Similar to Polar Bear Than to Sea Gull? Analyzing Conceptual Knowledge in Distributional Models.” In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 134–42. https://doi.org/10.18653/v1/2020.acl-srw.18.

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): 677–91. https://doi.org/10.1017/psrm.2019.22.

Thorn Jakobsen, Terne Sasha, Maria Barrett, Anders Søgaard, and David Lassen. 2022. “LAW-LREC 2022.” In, 4461. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.law-1.6.

Troiano, Enrica, Sebastian Padó, and Roman Klinger. 2021. “Emotion Ratings: How Intensity, Annotation Confidence and Agreements Are Entangled.” *arXiv:2103.01667 [Cs]*, March. http://arxiv.org/abs/2103.01667.

Van Aelst, Peter, and Stefaan Walgrave. 2016. “Political Agenda Setting by the Mass Media: Ten Years of Research, 2005–2015.” *Handbook of Public Policy Agenda Setting*, 157–79. https://doi.org/10.4337/9781784715922.00018.

Van Atteveldt, Wouter, Damian Trilling, and Carlos Arcila Calderon. 2022. *Computational Analysis of Communication*. John Wiley & Sons.

Van Atteveldt, Wouter, Mariken ACG Van der Velden, and Mark Boukes. 2021. “The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms.” *Communication Methods and Measures* 15 (2): 121–40. https://doi.org/10.1080/19312458.2020.1869198.

Waseem, Zeerak. 2016. “Are You a Racist or Am i Seeing Things? Annotator Influence on Hate Speech Detection on Twitter.” In, 138142. Austin, Texas: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-5618.

Weber, René, J Michael Mangus, Richard Huskey, Frederic R Hopp, Ori Amir, Reid Swanson, Andrew Gordon, Peter Khooshabeh, Lindsay Hahn, and Ron Tamborini. 2018. “Extracting Latent Moral Information from Text Narratives: Relevance, Challenges, and Solutions.” *Communication Methods and Measures* 12 (2-3): 119–39.

Wei, Chengwei, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. 2023. “An Overview on Language Models: Recent Developments and Outlook.” *arXiv Preprint arXiv:2303.05759*.

Winter, Nicholas JG, 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): 731–46. https://doi.org/10.1017/psrm.2019.4.

Ying, Luwei, Jacob M Montgomery, and Brandon M Stewart. 2022. “Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures.” *Political Analysis* 30 (4): 570–89. https://doi.org/10.1017/pan.2021.33.