ONLINE APPENDIX: Whose Truth is it Anyway? An Experiment on Annotation Bias in Times of Factual Opinion Polarization

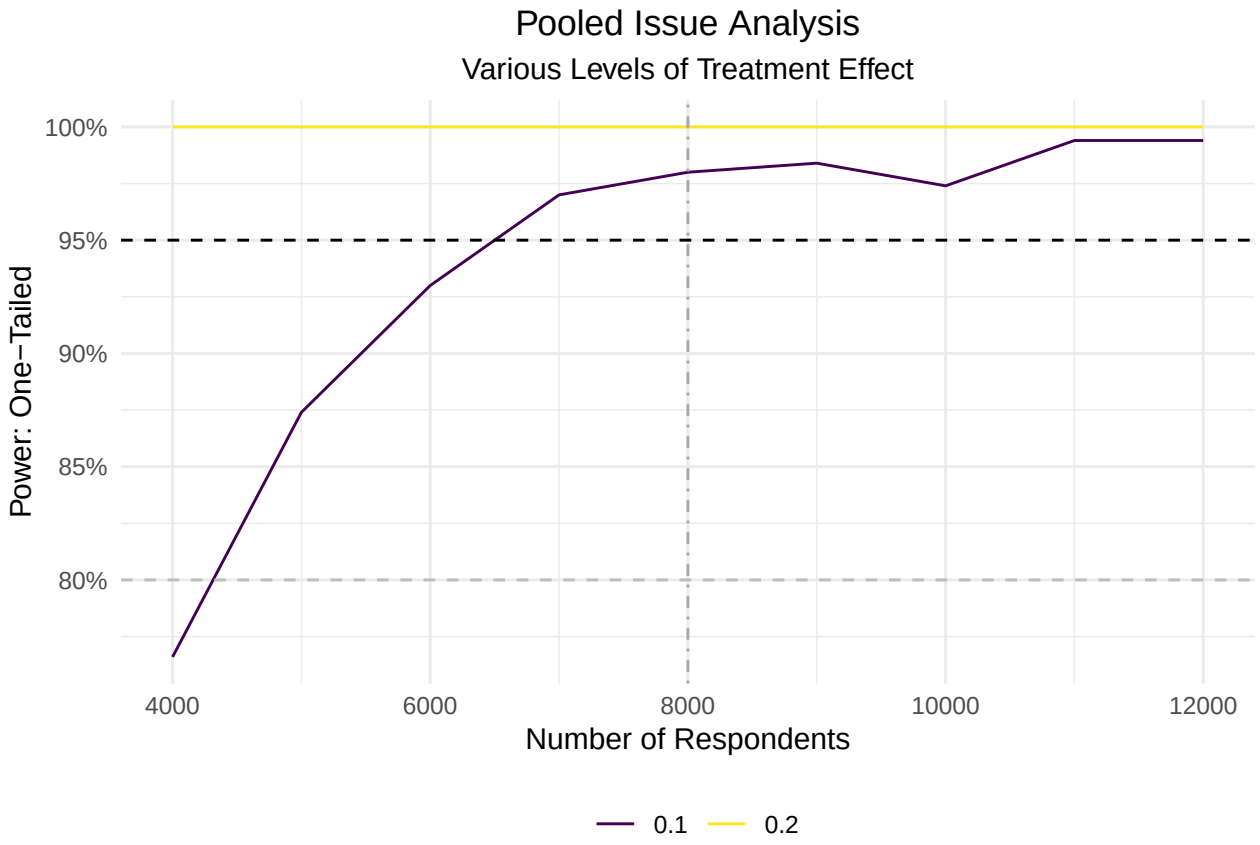
Contents

A-1. Power Analysis

As detailed in our pre-analysis plan, we conduct multi-level regressions for the four sentences. The manipulation, i.e. the masking condition vs. the political actor, as well as the specification of the sentence (i.e. *fully specified* or *underspecified*) as well as respondents ideological distance to the party and their levels of political knowledge are the main independent variables.

To calculate power for the hypotheses, the R package `DeclareDesign` is used (Blair et al. 2019). The effect sizes are between $b = 0.2$ and $b = 0.1$ – i.e. a small effect visualized by the purple and blue lines in Figure OA.1. The hypothesis are directional, Figure OA.1 therefore displays one-tailed tests with $\alpha = 0.05$. The power analysis shows that testing the hypotheses requires a sample size of $2,000$ participants, who all annotate the four sentences, i.e. $8,000$ participants in total, (x-axis) to reach very high levels of power, i.e. $> 95\%$, (indicated by the black dashed line in Figure OA.1).

Figure OA.1: Power Analysis



A-2. Data Collection: Questionnaire & Stimulus Material

Questionnaire

Dependent Variables. 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`.

Control Variables. As control variables, the following *demographics* are measured post-treatment: gender, age, education, geographical region, level of urbanization, employment, and income. For the analysis, only variables that are unbalanced over the experimental conditions will be included.

- *Gender* is measured as `sex`. The answer categories are `Male` (value of `1`), `Female` (value of `0`), and `No answer` (value of `999`).
- *Age* is measured using 6 categories: `17 or younger`, `18--29`, `30--39`, `40--49`, `50--59`, `60--74`.
- *Education* is measured as the highest successfully completed level of education, recoded into four categories: `low`, `middle`, `high`, and `none`. We created dummy variables for each level of education with the lowest category as base category.
- *Employment* Respondents were asked which category of employment – `Full-time employed`, `Part-time employed`, `Entrepreneur`, `Unemployed and searching for a job`, `Unemployed and not searching for a job or incapacitated`, `Housewife/Househusband or else`, `Retired`, `Student or full-time education` – applied most to them.
- *Income* Respondents were questioned on their monthly income in bins of 500 Euro's or Dollars – `500 or less`, `501-1000`, `1001-1500`, `1501-2000`, `2001-2500`, `2501-3000`, `3001-3500`, `3501-4000`, `4501-7500`, `7501 or more` – as well as giving them the options of `won't say` and `don't know`.
- *Geographical region* is measured using the *Nielsen districs*, dividing the Netherlands into 1) the 3 major cities plus suburbs, Amsterdam (plus Diemen, Ouder-Amstel, Landsmeer, Amstelveen), Rotterdam (plus Schiedam, Capelle aan den IJssel, Krimpen aan den IJssel, Nederlek, Ridderkerk, Barendrecht, Albrandswaard) and The Hague (plus Leidschendam, Voorburg, Rijswijk, Wassenaar, Wateringen); 2) West (Noord-Holland, Zuid-Holland and Utrecht (excluding the major cities and their suburbs; 3) North (Groningen, Friesland and Drenthe), 4) East (Overijssel, Gelderland and Flevoland); and South (Zeeland, Noord-Brabant and Limburg). In the American case, we use the states.

In addition, pre-treatment, respondents' ideological position, political knowledge, vote recall, and position on the issues migration, climate, tax, and EU are measured. Those variables will only be included in the analyses if balance checks indicate they are necessary. Moreover, the variables will be used to explore heterogeneous relationships.

- *Ideological position* is measured using an 11-point scale ranging from left (`0`) to right (`10`) in the Dutch case, since this is common in the national election studies (DPES). In the American case, we use the ANES standard liberal-conservative 7 point scale from extremely conservative to extremely liberal.

- *Vote Recall* Respondents were asked which party they voted for in the 2021 parliamentary elections in the Dutch case, since party id is not a hellpful measure in this context. The options were 1) all parties that were elected into parliament – `Bij1`, `BoerBurgerBeweging`, `CDA`, `ChristenUnie`, `D66`, `Denk`, `Forum for Democracy`, `JA21`, `GroenLinks`, `PvdA`, `Animal Rights Party`, `PVV`, `SGP`, `SP`, `VOLT`,`VVD`, `50Plus Party` – 2) another party; 3) blanco vote; and 4) a `Don't know` option.

- *Party ID* For the American case, respondents where asked whether they think of themselves as a Democrat, a Republican, an independent, or something else.

- *Political knowledge* is measured with six items from the DPES for the Dutch study and four items from the ANES for the American study.

- *Position on migration* is measured by asking people whether or not there are too many immigrants in the Netherlands using a 5-point Likert-scale (`fully disagree`, `disagree`, `neutral`, `agree`, `fully agree`).

- *Position on climate* is measured by asking people whether or not the climate crisis is exaggerated using a 5-point Likert-scale (`fully disagree`, `disagree`, `neutral`, `agree`, `fully agree`).

- *Position on tax* is measured by asking people whether or not the tax rate for the highest earners should go up using a 5-point Likert-scale (`fully disagree`, `disagree`, `neutral`, `agree`, `fully agree`).

- *Position on the EU* is measured by asking people whether or not membership in the European Union has been especially bad for the Netherlands so far using a 5-point Likert-scale (`fully disagree`, `disagree`, `neutral`, `agree`, `fully agree`).

- *Position on Foreign Policy* is measured by asking people whether or not say the U.S. needs to consider military build-up in the Pacific Ocean using a 5-point Likert-scale (`fully disagree`, `disagree`, `neutral`, `agree`, `fully agree`).

Attention Check. We include one attention checks in the survey. This is asked just before respondents enter the round of the experimental treatments. The attention checks are taken from Berinsky, Margolis, and Sances (2014) and adapted to the Dutch context by the authors – for the American data we use the original item. If a respondents fails the first attention check, a warning appears and the respondent can only continue with the survey once the respondent has correctly answered the question correctly.

When a big news story breaks people often go online to get up-to-the-minute details on what is going on. We want to know which websites people trust to get this information. We also want to know if people are paying attention to the question. To show that you have read this much, please ignore the question and select Volkskrant and Metro as your two answers. When there is a big news story, which is the one news website you would visit first? (Pleaseonly choose one). Eight (Dutch)

news outlets are provided to choose from. Respondents pass the attention check if they select **de Volkskrant** and **Metro**.

Experimental Protocol

The study is conducted online and in Dutch for the Dutch data as well as in English for the American data. Participants are told that they are taking part in a survey to get an overview of how people form their views on politics. After reading an informed consent message participants are forwarded to the main questionnaire (or the survey will be terminated if they do not agree to the consent form).

First, participants complete a set of pre-treatment variables (i.e. vote recall, issue positions of the issues used in the experiment (Climate, EU/ Foreign Policy, Immigration, and Tax), ideology, and political knowledge). This block ends with the attention check included in this survey. When participants fail this attention check, a warning appears asking them to read the question again carefully and to answer again. Therefter, participants see the experimental stimuli. The stimuli in the experiment are to show respondents fully specified or underspecified sentences on the issue position of a political actor or a masked political actor (see Section Treatment for more details).

Stimulus Material

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 OA.1 gives an overview of the variations in treatment in the survey as well as their English translations.

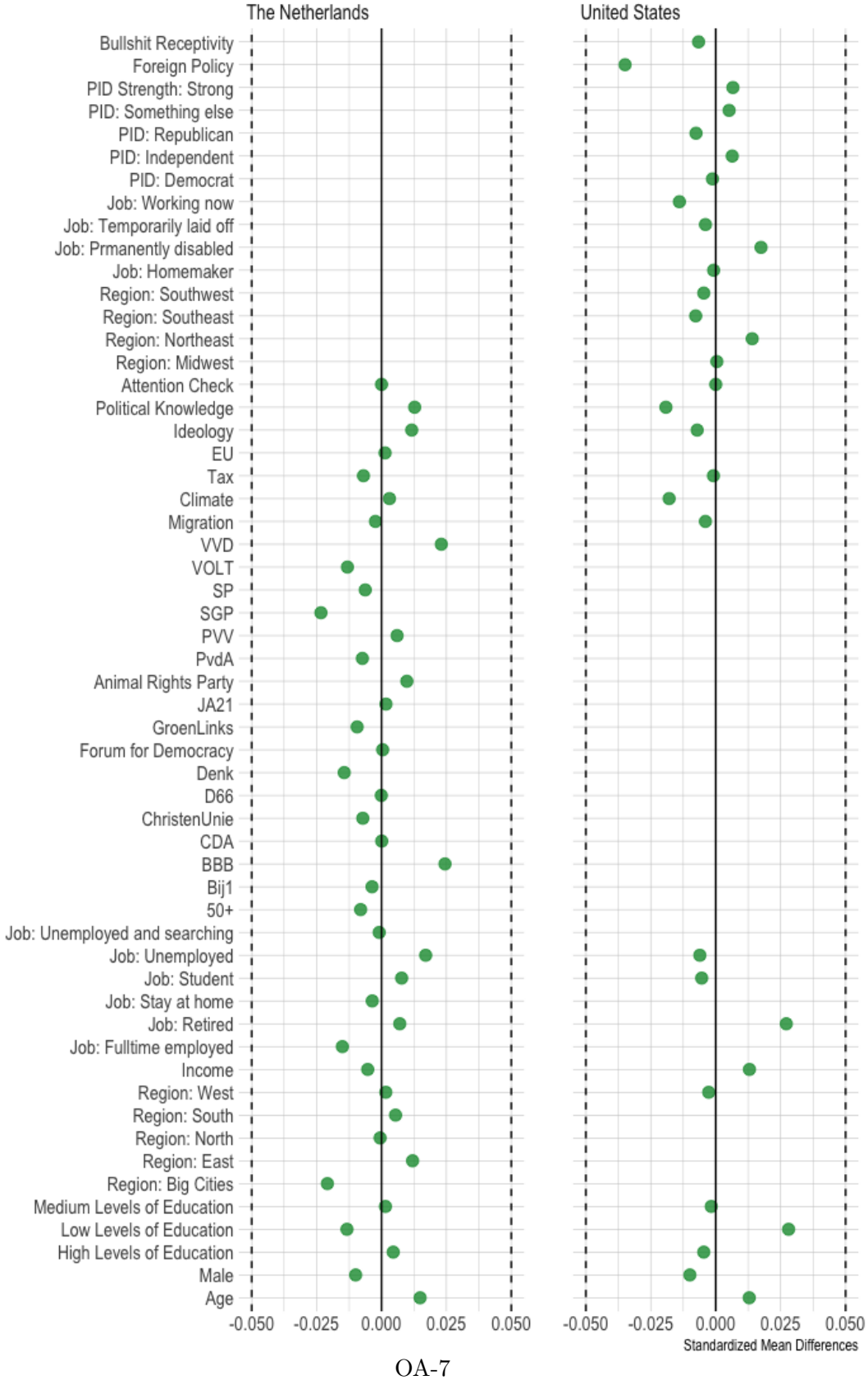
Table OA.1: Survey Questions - Experimental Conditions

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.

A-3. Balance Checkes

Figure OA.2 below shows that the data is balanced on all variables. As described in the Pre-Analysis Plan, we will therefore add no co-variates to the analyses as controls.

Figure OA.2: Balance Test



A-4. Methods of Analyses

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 urbanization, 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.

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

A-5. Descriptive Analyses

Tables OA.2 and OA.3 demonstrate the average profile of respondents who annotate correctly and incorrectly (where respondents who annotated some stances correctly and some incorrectly are weighted by proportion (in)correct). 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 four for those who are incorrect vs. an average score of five 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. This profile is quite similar for the lenient interpretation of what a stance is.

Table OA.2: Profile Dutch Stance Annotators

Incorrectly Identified Stance (Strict Interpretation)	Correctly Identified Stance (Strict Interpretation)
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

Table OA.3: Profile American Stance Annotators

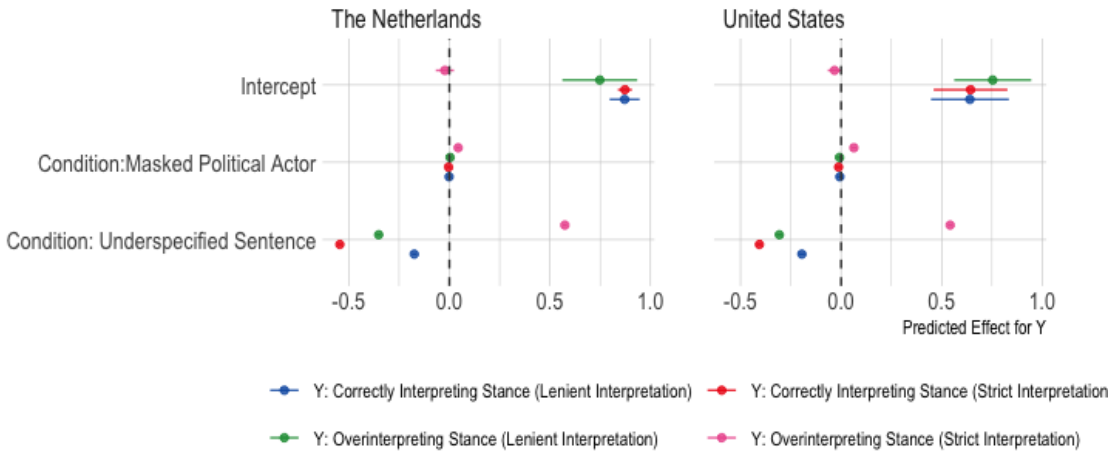
Incorrectly Identified Stance (Strict Interpretation)	Correctly Identified Stance (Strict Interpretation)
Male	Male
High-level of education	High-level of education
Southeast of the United States	Southeast of the United States
Working now	Working now
Democrat	Democrat
Age: 24	Age: 24
Income: 3250	Income: 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: 3	Ideological Distance: 2
Issue Congruence: 0	Issue Congruence: 1
Political Knowledge: 1	Political Knowledge: 1
Bullshit Receptivity: 3	Bullshit Receptivity: 2

A-6. Baseline Analyses

Looking at the effect of the experimental conditions on the four dependent variables – correctly identifying a stance and over-interpreting a stance for both a strict and a lenient interpretation of stance – Figure OA.3 visualizes the baseline. The left-hand panel demonstrates the effect of the two experimental conditions for correctly identifying the stance in the Dutch case. The right-hand panel does so for the American case. On average, many respondents in both cases (respectively `85%` in the Dutch case and `65%` in the American case) correctly interpreted the stance using either the lenient or strict interpretation (respectively in blue and red) – 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 does on average not improve correctly interpreting the stance significantly in neither the Dutch or the American case. Additionally, 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 stance in both the lenient and strict interpretation of a stance. These effects are substantial in both the American and Dutch case, with coefficients varying between `-0.2` and `-0.5`. This indicates that compared to a fully specified sentence, between `20%` and `50%` 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. Looking at the other dependent variable, whether they interpreted the sentence as a stance or not, we see that almost nobody overinterprets a stance in the strict interpretation in either the Dutch or American case. Yet, they do overinterpret a stance in the lenient interpretation. Moreover, if people see an X compared to a political actor, they are statistically significantly more likely to interpret the sentence as a stance in its strict interpretation. 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 in the strict interpretation: an increase of `0.83`. This indicates that people do not excel in this task without any instruction. Using the lenient interpretation, however, people seem less likely to annotate the sentence as a stance. In the pre-registered section, we demonstrate the tests of the hypotheses, and afterwards, we discuss some explorations of the data to show the robustness of our findings, the visualizations thereof are displayed in OA A-3.

Figure OA.3: Baseline Results of Experimental Conditions



A-7. Robustness Analyses

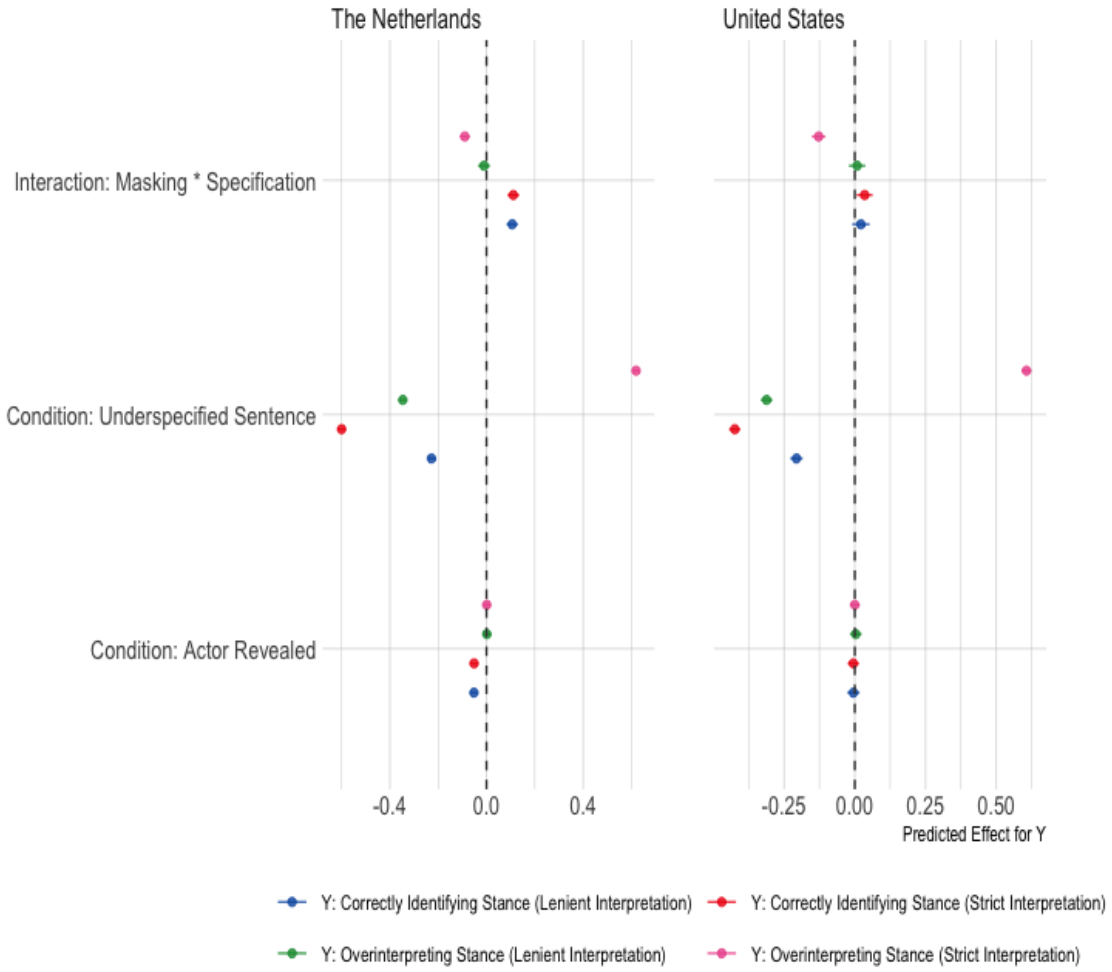
To check the robustness of our findings, Figure OA.4 demonstrates the analyses for each issue separately. The different colors visualize the different dependent variables. We do not see much variation between issues *Tax*, *EU/Foreign Policy*, and *Environment*. For those issues, we see that almost everyone interprets the sentence correctly (in blue and red). We also see that for a lenient interpretation of stances, people are quite likely to overinterpret a position as a stance. Being correct about the stance does not decrease when masking the political actor in both cases. Yet, the chance of being correct decreases statistically significantly when the sentence is underspecified. The same holds for overinterpreting for the lenient interpretation, but the opposite is true for the strict interpretation; there overinterpretation is more likely with underspecified sentences. Looking at *Immigration*, we see a different pattern. We see that masking does not increase the likelihood of being correct, but does increase the likelihood of overinterpretation regardless of how one defines a stance. Underspecified sentences are less likely to be correctly identified and more likely to be overinterpreted regardless of the definition of a stance. So, while there are some differences in effect sizes between the issues, the overall findings are not driven by a single issue.

Figure OA.4: Exploration: Issue Specific Analyses

In addition to issue-specific analyses, we also explore an interaction between treatments, visualized in Figure OA.5 for both dependent variables. This shows that masking is of help when sentences are under-specified. In the left-hand panel of Figure OA.5, it demonstrates that for under-specified sentences, people are less likely to incorrectly identify a sentence as a stance when

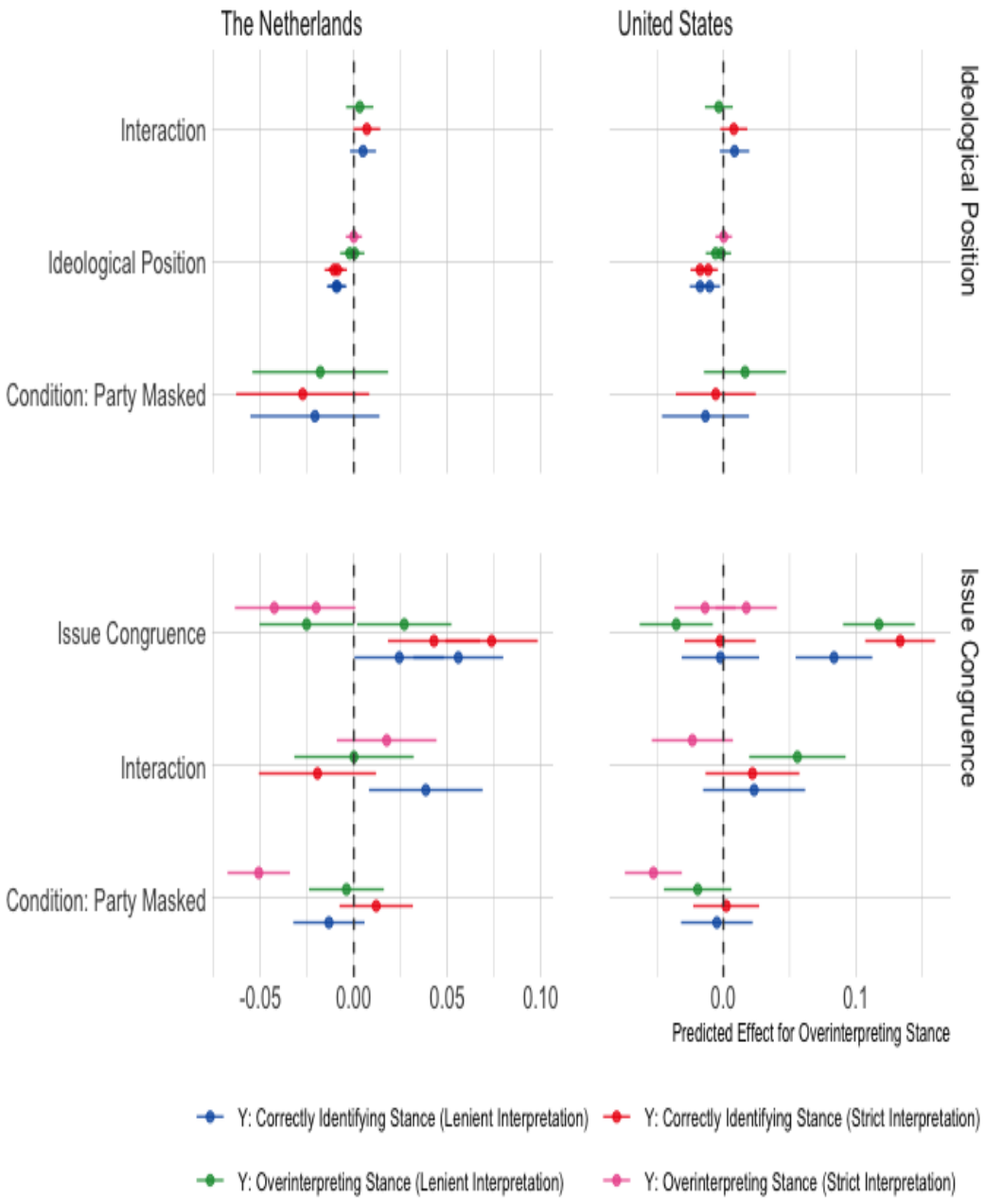
the actor is masked (coefficient of -0.30) than when an actor is revealed (coefficient of -0.45). That means there is a 15% increase in having it correct. The difference for overinterpreting is smaller between revealed and masked political actors – shown in the right-hand panel of Figure OA.5 – yet also statistically significant. Compared to 85% overinterpreting 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.

Figure OA.5: Exploration: Interaction bewtween Treatments



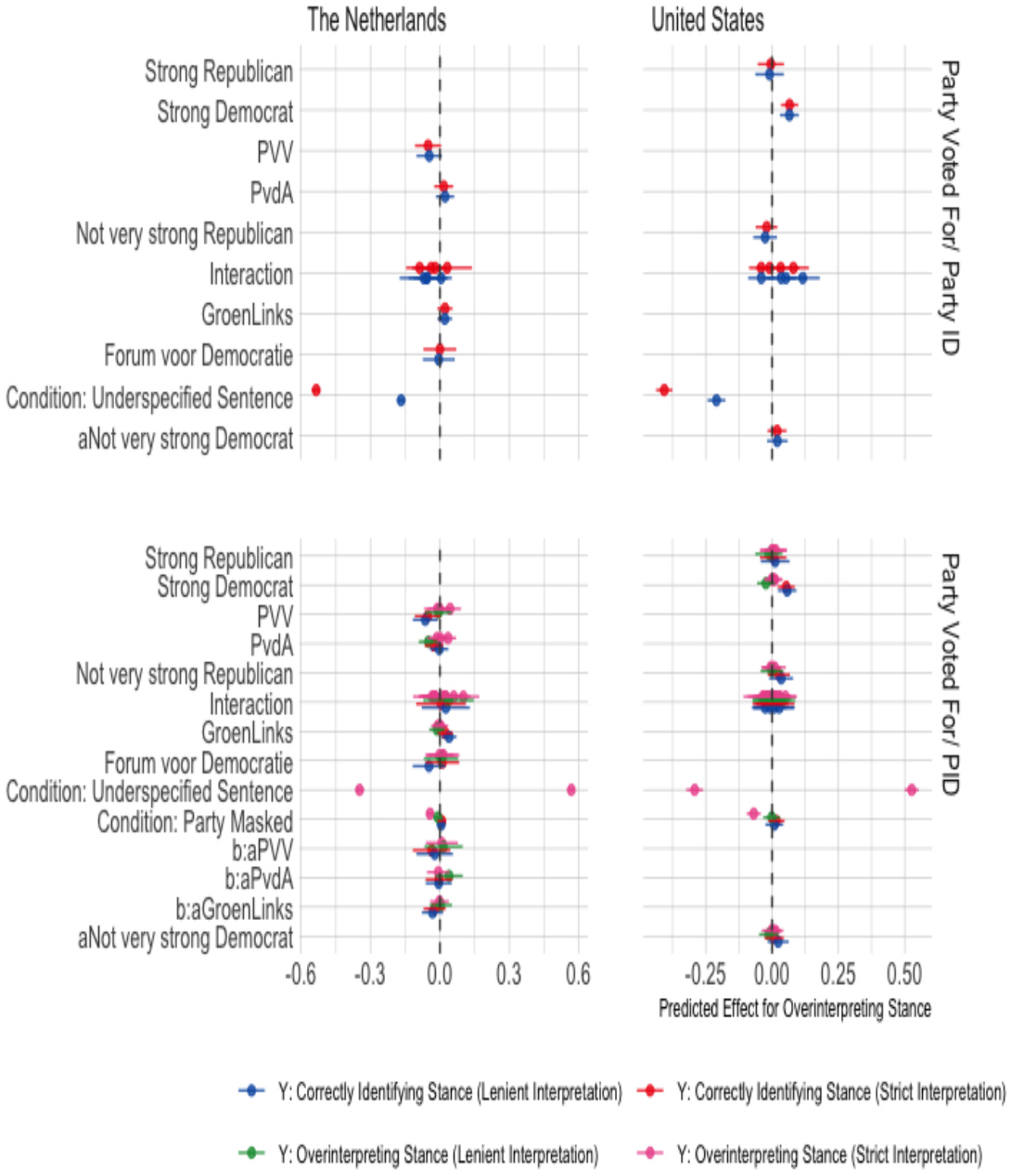
Lastly, we explore three different ways of measuring ideological distance and an alternative for political knowledge in the American case. First, we measured ideological bias by looking at the ideology of the respondents – not in relation to the political actor revealed, visualized in the upper panels of Figure OA.6. Secondly, we measured ideological bias by looking at whether the respondent is congruent or not with the issue position in the sentence, visualized in the lower panels of Figure OA.6.

Figure OA.6: Exploration: Interaction bewtween Treatments



And thirdly, we measured ideological bias by looking at whether the person voted for the party displayed in the sentence, visualized in the upper-rows of Figure OA.7. In none of the analyses, we find evidence for ideological bias. Also for the alternative measurement of political knowledge in the US, we find the same results as reported in the main analyses.

Figure OA.7: Exploration: Interaction bewtween Treatments



References

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.

Blair, Graeme, Jasper Cooper, Alexander Coppock, and Macartan Humphreys. 2019. "Declaring and Diagnosing Research Designs." *American Political Science Review* 113 (3): 838–59. https://doi.org/10.1017/S0003055419000194.

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.