Supplement:

Understanding Partisan Bias in Judgments of Misinformation:

Identity Protection vs. Differential Knowledge

Tyler J. Hubeny, Lea S. Nahon, Bertram Gawronski

University of Texas at Austin

Procedure for the Selection of Statements

Because the selection and pilot testing of statements followed the same procedure in Experiment 1 and Experiment 2, we describe this procedure for both experiments together.

Statement Search

To generate a list of statements, we first searched for factual information about either country (i.e., UK and France in Experiment 1; Spain and Greece in Experiment 2). Statements were drawn from sources such as the Pew Research Center, OECD, the World Bank, the UN, Statista, U.S. News and World Report, and the International Monetary Fund. Generally, the search for statements was guided by the following criteria: (1) the statements should have a clear slant towards one country; (2) the statements should be factual (as opposed to opinions); and (3) the statements should be unambiguously true or false and not misleading or having conflicting evidence. Statements were generated in the Summer of 2024 for Experiment 1 and in the Fall of 2024 for Experiment 2. For each statement, we recorded the following information: (1) veracity of the statement (i.e., true vs. false); (2) slant of the statement (i.e., pro-Country A vs. pro-Country-B); (3) valence of the statement as being either positive (i.e., "Country A is good"), negative (i.e., "Country B is bad"), or comparative (i.e., "Country A is better than Country B"); and (4) source(s) supporting the truth or falsity of the information. In this initial stage, a total of 98 statements were generated for Experiment 1 and a total of 91 statements were generated for Experiment 2.

Pilot Testing

To select the final set of statements, we conducted a pilot test for each set of statements. In this pilot test, participants were asked (1) *Assuming the above statement is true, how favorable is this statement towards [Country A]?* (2) *Assuming the above statement is true, how favorable*

is this statement towards [Country B]? and (3) *To the best of your knowledge, is the claim in this statement true or false?* Responses to the first two questions were measured on a 7-point scale ranging from *Very unfavorable to [Country]* (recorded as 1) to *Very favorable to [Country]* (recorded as 7) with a midpoint response option of *Neither favorable nor unfavorable* (recorded as 4). Responses to the third question were measured as either *True* or *False*.

For the pilot test for Experiment 1, a total of 201 U.S. participants were recruited from Prolific Academic. Of these, a total of 180 participants were included after excluding 21 participants who failed the same instructional attention check as in the main experiments. For the pilot test for Experiment 2, a total of 201 U.S. participants were recruited from Prolific Academic. Of these, a total of 183 participants were included after excluding 18 participants who failed the same attention check as in the main experiments.

Based on the collected pilot data, we calculated a statement's slant as the mean difference scores between favorability responses for Country A and Country B. We also calculated the probability that a statement was judged correctly (i.e., a "true" response for a true item, a "false" response for a false item). To select the final set of stimuli, we first eliminated statements that did not have a sufficiently strong slant by removing statements with a slant-difference score of less than one scale point. Next, to avoid the possibility of participants having limited knowledge about the statements, we selected the top 15 statements from each condition (i.e., pro-Country A true, pro-Country B true, pro-Country A false, and pro-Country B false) with the highest proportion of correct responses. The selection of these statements was balanced for valence (i.e., 5 statements per valence per condition). Next, to ensure that the statement slants were equivalent across conditions (i.e., pro-Country A vs. pro-Country B), we replaced some statements to achieve similar average slant-difference scores across conditions. To further ensure that there

were no differences in slant strength across conditions, we used participants' truth judgments to calculate partisan-bias scores using SDT (i.e., $c_{\text{pro-Country A}} - c_{\text{pro-Country B}}$) and then used the resulting scores to confirm that the selected statements were not biased in favor of one country over another. The data for both pilot tests as well as the final lists of 60 statements used in both experiments are available at https://osf.io/gj8wh/?view_only=5fb864782994462db44f71f00be61a83.

Procedure for the Selection of Countries in Experiment 2

To select a pair of countries for the two teams in Experiment 2, we conducted a pilot test. In this pilot test, we included the following 23 countries: Argentina, Brazil, Chile, Spain, Italy, Portugal, Greece, Germany, Austria, Switzerland, England, Ireland, Scotland, New Zealand, Australia, Canada, France, Norway, Sweden, Denmark, Finland, Singapore, and Thailand. Participants were asked to rate their agreement with the following two statements for each country, *I identify with [Country]* and *I like [Country]*. Responses for both items were measured on 7-point rating scales ranging from *Strongly disagree* (recorded as 1) to *Strongly agree* (recorded as 7) with the midpoint of *Neither agree nor disagree* (recorded as 4).

A total of 200 U.S. participants were recruited from Prolific Academic. After following the same preregistered exclusion criteria as in the main experiment, a total of 180 participants were included in the final sample. Based on participants' responses, we calculated a preference score for each country as the mean score of the two items for that country.

To select the pair of countries for the two teams in Experiment 2, we aimed to meet the following criteria: (1) no difference in preference scores between the two countries; (2) sufficient similarity between countries to ensure the plausibility of the described rivalry between countries; and (3) countries about which U.S. participants are likely to have sufficient knowledge. Based on

these criteria, we selected Spain and Greece to use for the two teams in Experiment 2. The pilot data demonstrated that the preference for Spain versus Greece did not significantly differ, $t(179) = 0.12$, $p = .903$, $d = 0.01$.

Experiment 1 Re-Analysis

After completion of the main experiment, we were alerted that one pro-UK statement in our stimulus was incorrectly classified as true, despite being false (*The majority of French people find infidelity morally acceptable*). As a result, we dropped this item from analyses reported in the main text. Although there were no meaningful differences in the results after dropping this item, below we report the results of all analyses following the original preregistered data aggregation plan that included the incorrectly classified item.

Acceptance Thresholds

For the ANOVA comparing acceptance thresholds, the main effects of Group Assignment, $F(2, 560) = 0.21$, $p = .813$, $\eta_p^2 < .01$, 95% CI [.00, .01], and Statement Slant, $F(1, 560) = 0.45$, $p = .501$, $\eta_p^2 < .01$, 95% CI [.00, .01] were not significant. However, as predicted, there was a significant interaction between Group Assignment and Statement Slant, $F(2, 560) = 8.35$, $p < .001$, $\eta_p^2 = .03$, 95% CI [.01, .06]. Following our preregistered analysis plan, follow-up *t*-tests found that acceptance thresholds in the Team UK condition were significantly lower for pro-UK statements compared to pro-France statements, $t(189) = -3.34$, $p < .001$, $d = -0.32$, 95 CI [-0.51, -0.13]. While acceptance thresholds in the Team France condition were descriptively lower for pro-France statements compared to pro-UK statements, the observed difference did not reach our preregistered alpha level, $t(184) = -1.92$, $p = .056$, $d = -0.17$, 95% CI [-0.35, 0.00]. Finally, as predicted, acceptance thresholds in the No Team condition did not significantly differ for pro-UK statements and pro-France statements, $t(187) = 0.58$, $p = .565$, $d = 0.05$, 95% CI [-0.13, 0.23].

Non-preregistered follow-up *t*-tests comparing acceptance thresholds for a given statement type across experimental and control conditions revealed that, compared to participants in the No Team condition, participants in the Team UK condition showed significantly lower acceptance thresholds for pro-UK statements, $t(376) = -2.34$, $p = .020$, $d = -0.24$, 95% CI [-0.44, -0.04]. However, acceptance thresholds for pro-France statements did not significantly differ between the Team UK and No Team conditions, $t(376) = 1.41$, $p = .158$, $d = .15$, 95% CI [-.06, .35]. Acceptance thresholds in the Team France condition did not significantly differ from acceptance thresholds in the No Team condition for pro-UK statements, $t(371) = 0.52$, $p = .600$, $d = 0.05$, 95% CI [-0.15, 0.26] and pro-France statements, $t(371) = -0.77$, $p = .444$, $d = -0.08$, 95% CI [-0.28, 0.12].

We also conducted non-preregistered follow-up *t*-tests to compare partisan bias in the Team UK and Team France conditions to the No Team condition, respectively. To do so, we calculated partisan-bias scores by subtracting acceptance thresholds for pro-France statements from acceptance thresholds for pro-UK statements ($c_{pro-UK} - c_{pro-France}$). Positive scores on this index reflect partisan bias favoring France over the UK, whereas negative scores reflect partisan bias favoring the UK over France. Analyses revealed that partisan bias favoring the UK over France was significantly greater in the Team UK condition compared to the No Team condition, $t(376) = -2.94$, $p = .004$, $d = -0.30$, 95% CI [-0.50, -0.10]. However, partisan-bias scores did not significantly differ between the Team France condition and the No Team condition, $t(371) = 1.05$, $p = .295$, $d = 0.11$, 95% CI [-0.09, 0.31].

Truth Sensitivity

For the ANOVA comparing truth sensitivity scores, results confirmed all preregistered hypotheses, in that the main effect of Group Assignment was not significant, $F(2, 560) = 0.84$, p

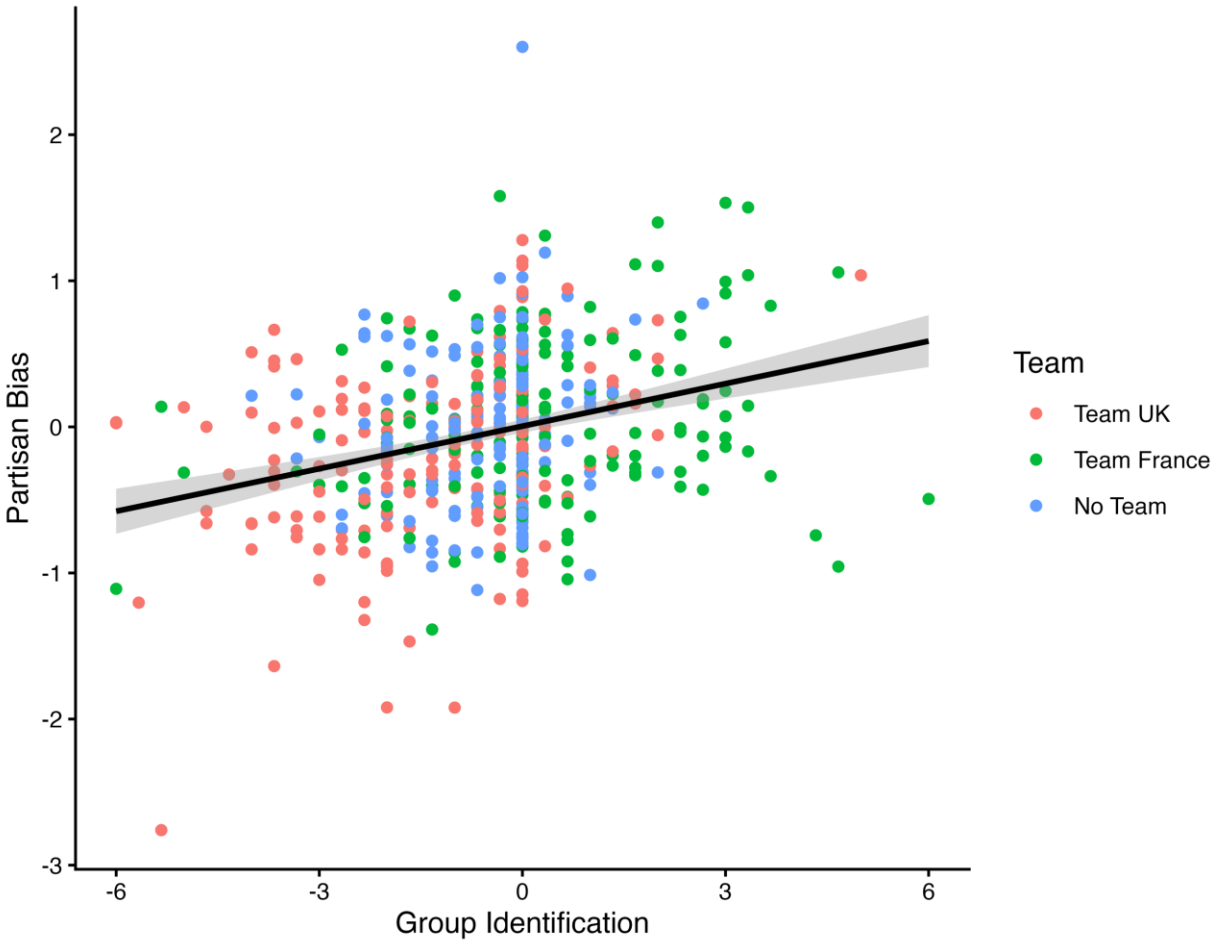
= .431, η_p^2 < .01, 95% CI [.00, .02], nor was the interaction between Group Assignment and Statement Slant, $F(2, 560) = 1.17$, $p = .311$, η_p^2< .01, 95% CI [.00, .02].

Group Identification and Partisan Bias

As an additional non-preregistered exploratory analysis, we analyzed correlations between group identification and partisan bias to test if the strength of participants' identification with a randomly assigned identity is related to their levels of partisan bias. Group-identification scores were calculated as the difference between the level of identification with, attachment to, and attitude towards the respective country, such that positive group-identification scores indicate a stronger identification with France than the UK (Exp. 1) and a stronger identification with Greece than Spain (Exp. 2) and negative scores indicate a stronger identification with the UK than France (Exp. 1) and a stronger identification with Spain than Greece (Exp. 2), respectively. Partisan-bias scores were calculated as the difference in acceptance thresholds, such that positive partisan-bias scores represent a partisan bias favoring France over the UK (Exp. 1) and Greece over Spain (Exp. 2) and negative partisan bias scores represent a partisan bias favoring the UK over France (Exp. 1) and Spain over Greece (Exp. 2), respectively. In both Experiment 1, $r(561) = .29$, $p < .001$, 95% CI [.21, .36], and Experiment 2, $r(846) = .23$, $p <$ .001, 95% CI [.17, .29], group identification showed significant positive correlations with partisan bias (see Supplemental Figures 1 and 2).

Supplemental Figure 1

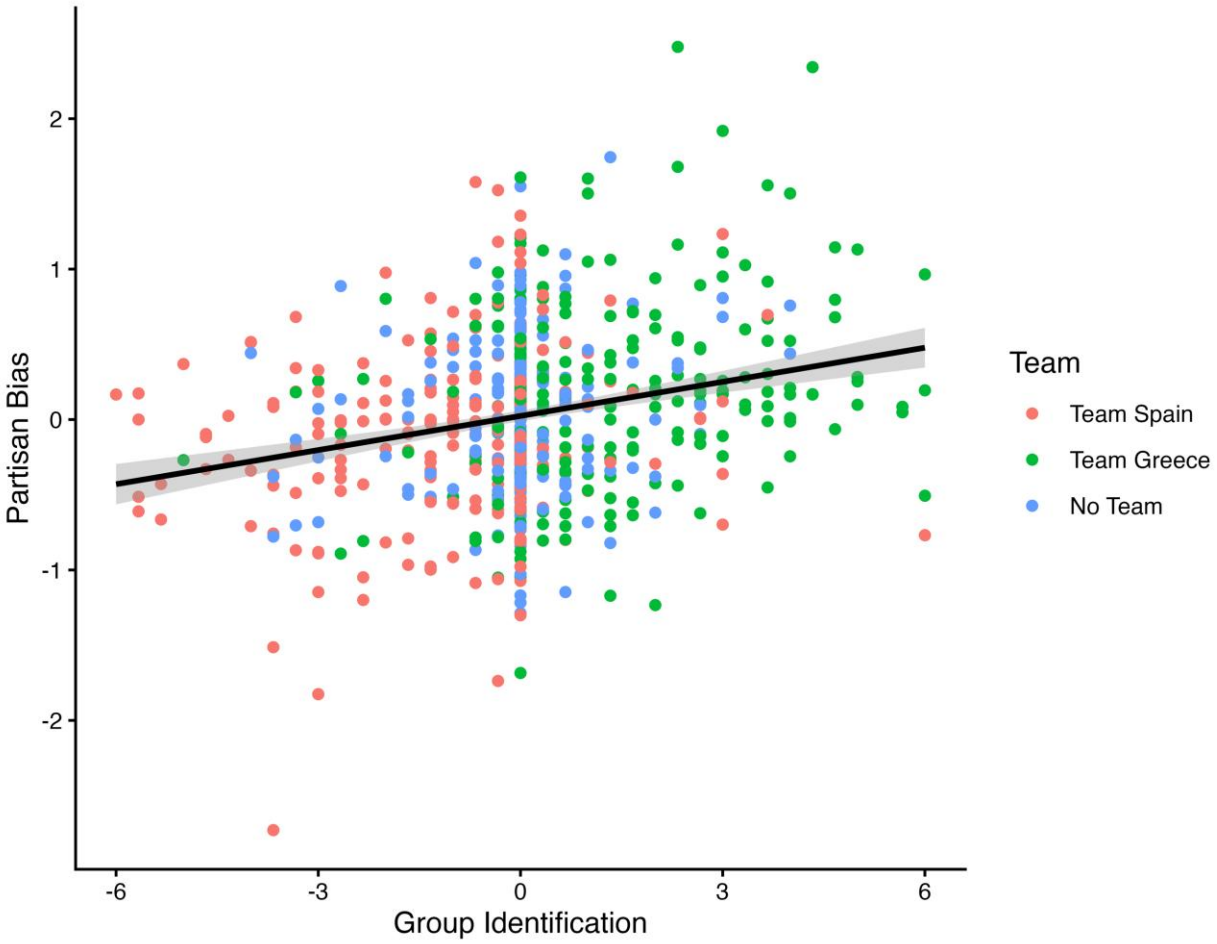
Partisan Bias as a Function of Group Identification, Experiment 1



Note. Partisan bias is calculated as the difference between pro-UK and pro-France acceptance thresholds ($c_{pro-UK} - c_{pro-France}$) such that positive partisan-bias scores indicate a partisan bias favoring France over the UK and negative partisan-bias scores indicate a partisan bias favoring the UK over France. Group identification is calculated as the difference between the identification with, attachment to, and attitude towards the UK/France, such that positive group-identification scores indicate a stronger identification with France than the UK and negative group-identification scores indicate a stronger identification with the UK than France.

Supplemental Figure 2

Partisan Bias as a Function of Group Identification, Experiment 2



Note. Partisan bias is calculated as the difference between pro-Spain and pro-Greece acceptance thresholds ($c_{pro-Spain} - c_{pro-Greece}$) such that positive partisan-bias scores indicate a partisan bias favoring Greece over Spain and negative partisan-bias scores indicate a partisan bias favoring Spain over Greece. Group identification is calculated as the difference between the identification with, attachment to, and attitude towards the Spain/Greece, such that positive group-identification scores indicate a stronger identification with Greece than Spain and negative group-identification scores indicate a stronger identification with Spain than Greece.