A Grip on the Mind: Stereotypes and Information Recognition Failure

August 19, 2026

Abstract

Why do narratives linking immigration and violence persist despite contrary evidence? We argue that biases in how information is registered and retained help explain their political durability. The bias we identify differs from standard confirmation bias because it does not require individual priors. Instead, it operates when information contradicts collectively held stereotypes. We demonstrate this mechanism using a survey experiment in four European countries. In a two-by-two design, respondents read descriptions of violent incidents in which perpetrators and victims were either migrants or natives. They were then asked to perform a minimal task: accurately identify the information presented. Across countries, respondents were more likely to recall stereotype-consistent events—migrants as perpetrators and natives as victims—while counter-stereotypical cases were forgotten or misattributed. These patterns hold across ideology and immigration attitudes and are robust to incentives for accuracy, indicating that the bias persists among those least likely to endorse the stereotype.

Keywords: Information processing; Stereotypes; Discrimination; Immigration; Violence

Why do false narratives linking immigration and violence endure despite contrary evidence? Research consistently finds no systematic relationship between immigration and violent crime in Western democracies (Ousey and Kubrin, 2018; Bell, Fasani, and Machin, 2013), yet public opinion remains attached to the idea that immigrants drive violence (Alizade, 2025)—fueling support for restrictive immigration policies and strengthening radical right parties (Arzheimer, 2009; Halla, Wagner, and Zweimüller, 2017). Perceived immigrant threat also shapes high-stakes institutional decisions, including asylum adjudications (Spirig, 2023). Understanding why such beliefs persist is key to explaining how stereotypes distort political judgment.

The dominant answer in public opinion research centers on motivated reasoning: individuals selectively process, reinterpret, or discount information that conflicts with their prior attitudes (Kahan, 2013; Taber and Lodge, 2006). On this account, the immigration-violence narrative endures because those predisposed to view immigrants as threatening actively resist disconfirming evidence. Recent extensions of this framework point to the role of interpretive wiggle room — exploiting ambiguity to reach preferred conclusions (Eyting, 2022), selective exposure to congenial information environments (Stroud, 2008), and asymmetric rejection of uncomfortable facts (Little, 2025). These accounts differ in their proposed mechanisms, but they share a common structural assumption: the bias operates through and upon prior individual attitudes. It is the person's own directional motivation — their personal stake in a particular conclusion — that drives the distortion. Consequently, individuals without strong anti-immigrant priors should be largely immune, and corrective information should work for everyone except the already-convinced.

We propose an alternative account, one that locates the bias earlier in the information processing chain and does not require individual-level motivation to sustain it. Rather than operating through personal endorsement of the stereotype or motivation to protect a prior belief, the mechanism we identify is rooted in collectively held cultural schemas that structure attention and memory prior to any evaluative process — affecting even those who consciously reject the underlying association. We call this stereotype reinforcement bias: a tendency to encode and retain information congruent with prevailing group stereotypes, while failing to register information that contradicts them.

We test this argument in a survey experiment conducted across four European countries (Germany, Hungary, Italy, Sweden; N = 7,975), in which participants read a brief vignette describing a violent attack, with the identities of the attacker and the victim — migrant or native — varied in a 2×2 design. After an intervening attitude battery, respondents were asked to identify, from the four possible scenarios, which one they had actually read. The task required no belief updating, only accurate recognition of an unambiguous fact they had just encountered. It offered no interpretive ambiguity, and the interval between exposure and recall was minimal. Yet over 80% correctly recalled the stereotype-consistent scenario — a migrant attacking natives — compared to only 52–60% in all other conditions. Importantly, recall errors were not random: nearly three-quarters of incorrect answers involved respondents misattributing the attack to a migrant perpetrator and a native victim. The robustness of this pattern is striking — the bias affects participants who place themselves at the far left of the ideological spectrum, those with clearly positive attitudes toward immigration, and even those who expressed willingness to donate part of their experimental earnings to the UNHCR. It persists, moreover, when respondents are offered a monetary incentive for accurate recall, ruling out expressive motives as drivers of provided responses. In other words, the bias persists among individuals for whom personal motivation to protect an anti-immigrant belief is highly implausible. Rather than reflecting how prejudiced individuals distort information to protect their worldview, this evidence points to a culturally ambient stereotype sufficiently pervasive to shape how virtually everyone — regardless of personal conviction — processes information about immigration and violence.[1]

These findings have direct policy implications. Consider, for instance, the debate over whether media outlets should systematically report the immigration status of criminal offenders. Empirical research suggests that the absence of clear reporting rules often leads to selective coverage, with crimes involving migrants receiving disproportionate attention relative to their frequency in official statistics (Keita, Renault, and Valette, 2024; Couttenier et al., 2024; Brill et al., 2021). Uniform reporting rules could reduce selective exposure, although they would not address biases arising earlier in policing, charging, or official recording. Our results show, however, that even a balanced set of

[1]For a similar account connecting fear of death to antisemitism through widely shared cultural schemas, see Kanol and Schaub (2026).

reported incidents may be encoded and retained asymmetrically. Incidents in which migrants attack natives align with a widely circulating stereotype and may therefore loom larger in public memory, while otherwise equivalent incidents with reversed roles are less likely to be retained.[2]

In this sense, the mechanism we identify complicates a class of policy approaches that seek to combat discrimination by correcting false beliefs through information provision. If counter-stereotypical information fails to be encoded in the first place, simply providing accurate data may have limited impact. The persistence of this bias, in turn, helps explain why immigration remains an unusually potent issue for political mobilization, particularly by right-wing actors (Abou-Chadi and Krause, 2020; Rooduijn, 2019; Halla, Wagner, and Zweimüller, 2017; Arzheimer, 2009). Because stereotype-consistent incidents are more likely to register and persist in public memory, even rare events involving migrants can loom disproportionately large in public consciousness. The result is a feedback loop in which rare but emotionally salient events shape perceptions of threat and sustain political agendas built around immigration.

Research Design

We conducted an online survey experiment in Germany, Hungary, Italy, and Sweden, with a sample of approximately 2,000 respondents per country ($N = 7{,}975$) representative in terms of age, gender, and education.[3] The experiment involves no deception—scenarios are presented as hypothetical, and we used START (2021) to confirm that conceptually similar events have occurred in all four countries. Participants are randomly assigned to one of four treatments, each presenting a vignette describing a

[2] To be sure, if stereotypes are indeed partly formed through repeated media exposure, sustained reductions in these upstream distortions could weaken them over time.

[3] The experimental design and data collection were preregistered with the OSF. The pre-analysis plan is available at: *(removed for anonymization)*. Both studies have been reviewed and approved by an ethics commission *(removed for anonymization)*. The study was initially designed to examine how varying attacker and victim identity in violent attacks affects immigration attitudes; the striking asymmetry in recall accuracy emerged as an unexpected finding. To guard against the possibility that this pattern reflected a false positive or was driven by expressive responding rather than genuine encoding failure, we pre-registered a follow-up study specifically designed to test the robustness of this result, introducing a monetary incentive for accurate recall in a random subset of the sample. The follow-up confirmed the original finding.

hypothetical bombing attack on a cafe that killed six people. Treatments vary attacker and victim identity (migrant or native) in a 2×2 design:

Treatment Migrant→Native: the attack is carried out by a migrant on a cafe frequented by natives

Treatment Migrant→Migrant: the attack is carried out by a migrant on a cafe frequented by other migrants

Treatment Native→Native: the attack is carried out by a native on a cafe frequented by natives

Treatment Native→Migrant: the attack is carried out by a native on a cafe frequented by migrants

We chose a terrorist attack specifically because, while such attacks occur with both migrant and native perpetrators, migrant-perpetrated terrorism is the archetypal frame linking immigration to violence in Western media and political discourse (Lajevardi, 2021; Klein and Gross, 2019). After the vignette, participants completed a battery of immigration attitude questions.[4] Our main outcome, measured after the attitude battery, asks participants to identify which of the four scenarios they read. To assess whether the bias reflects expressive responding, we conducted a pre-registered follow-up study in Germany ($N = 1{,}556$) focusing on the Migrant→Native and Native→Migrant conditions and orthogonally varying whether respondents received a monetary reward for a correct recall answer.[5]

Results

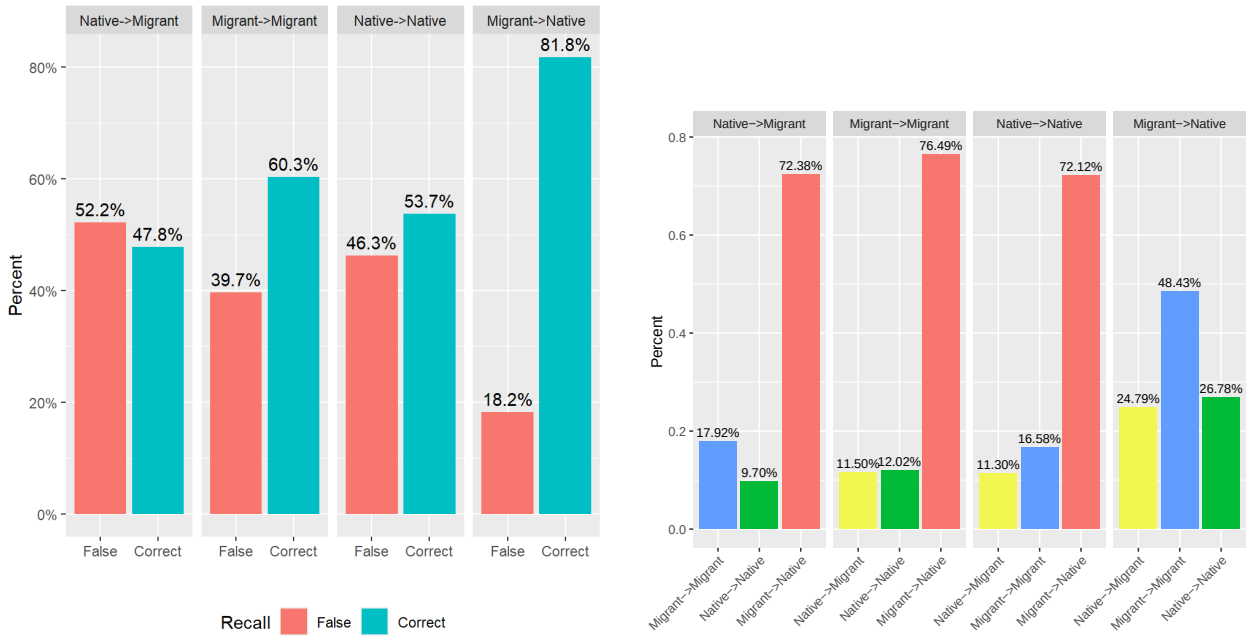
Figure 1a reports the share of participants correctly identifying their vignette. In the Migrant→Native condition, 81.8% answered correctly—roughly 20–30 percentage points higher than in all other conditions (60.3%, 53.7%, and 52.2%; Kruskal-Wallis $p < 0.01$).[6] Both attacker and victim identity independently shape recall: $P^{Correct}_{M\to N} > P^{Correct}_{M\to M} > P^{Correct}_{N\to N} > P^{Correct}_{N\to M}$, with all pairwise differences significant at the 1% level.

Misattributions are not random. As Figure 1b shows, 72–77% of wrong answers in each counter-

[4]Samples are balanced across treatments on all covariates. See Online Appendix for full balance checks (Table A.1) and exact question wording. Figure A.2 summarizes power analysis, which guided sample size selection.

[5]A correct recall answer earned respondents 100 panel points, equivalent to €1.

[6]The ≈20% failed-recall rate in the stereotype-consistent arm falls within the range reported for recall-based manipulation checks. Aronow, Baron, and Pinson (2019) report rates of 6.4%–7.5% among Mechanical Turk respondents, whom they note are more attentive than representative samples (2019, p. 578). Kao et al. (2024) report rates of 17%–35% in Jordan, Morocco, and Tunisia, and Getmansky (2025) reports rates of 23%–45% in a representative Israeli sample.



(a) Correct and false attributions per treatment

(b) Distribution of false attributions per treatment

Figure 1: **Recall accuracy and misattribution patterns by treatment condition.**
Note: Panel (a) shows the share of participants correctly (teal) and incorrectly (red) identifying their assigned vignette across the four treatment conditions. Panel (b) shows, for each treatment condition, how incorrect answers were distributed across the four possible scenarios. In all three counter-stereotypical conditions, roughly 72–77% of wrong answers involved respondents misattributing the attack to a migrant perpetrator and native victim — the stereotype-consistent scenario.

stereotypical condition identified the Migrant→Native scenario. The stereotype-consistent scenario acts as a near-universal attractor for recall errors: regardless of what participants read, mistakes almost always converge on the same answer.

Table 1 presents the regression analysis.[7] With Migrant→Native as baseline, all three alternative treatments significantly reduce correct attribution (probit coefficients of -0.65 to -0.97, all $p < 0.01$). Moving from the stereotype-consistent vignette (Migrant→Native, the baseline) to the counter-stereotypical Native→Migrant vignette lowers the predicted probability of correct attribution by 34.0 percentage points (from 81.8% to 47.8%).[8] The equivalent reduction in the probability of correct attribution for the other two treatment arms (vis-à-vis the baseline) is 28.1 and 21.5 percentage points for Native→Native and Migrant→Migrant respectively. Results are robust to controls for

[7]For full regression results displaying full co-variates list, see Tables A.2 and A.7 in the Online Appendix.

[8]Predicted probabilities computed as $\hat{p} = \Phi(\hat{\beta}_0 + \hat{\beta}_k)$, where $\hat{\beta}_0$ is the constant and $\hat{\beta}_k$ the coefficient on the relevant treatment indicator, using the column (1) estimates in Table 1.

demographics, country fixed effects, ideology, political participation, and vignette reading time. While older, more educated, and female respondents recall the vignette somewhat more accurately (full OLS specification in Table A.3; demographic correlates by treatment in Figure A.4), the treatment effects remain negative and highly significant. Furthermore, the bias is not driven by inattentive respondents: taking vignette reading time as a proxy for the attention to the treatment, results in Table A.4 show that the treatment effects remain negative and significant within every attention quintile.

Table 1: **Treatment effect: likelihood of correct attribution**

	(1)	(2)	(3)	(4)
	Correct attribution$_i$			
Baseline: Migrant → Native				
Native → Migrant	-0.962***	-0.973***	-0.971***	-0.973***
	(0.043)	(0.044)	(0.044)	(0.044)
Native → Native	-0.814***	-0.819***	-0.815***	-0.817***
	(0.043)	(0.043)	(0.043)	(0.043)
Migrant → Migrant	-0.646***	-0.653***	-0.650***	-0.651***
	(0.044)	(0.044)	(0.044)	(0.044)
Voted			0.081**	0.080**
			(0.039)	(0.039)
Left to right			-0.021***	-0.021***
			(0.006)	(0.006)
Demonstrated			-0.034	-0.041
			(0.059)	(0.059)
Reading time				0.001***
				(0.000)
Constant	0.907***	0.694***	0.763***	0.751***
	(0.033)	(0.076)	(0.089)	(0.089)
Demographics	No	Yes	Yes	Yes
Country FEs	No	Yes	Yes	Yes
Observations	7,975	7,975	7,964	7,964

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Probit regression of a dummy *Correct attribution*$_i$ set to one if participant i provided a correct response to the recall task and zero otherwise. Native → Migrant, Migrant → Migrant, and Native → Native are indicator variables for the corresponding treatments varying the identity of the attacker and the victim of the attack described in the vignette, with treatment Migrant → Native serving as the baseline. Demographic controls include age, gender, completion of higher education, employment status, and country of residence. Variables *Voted* and *Demonstrated* take value one if a participant stated to have voted in the last election and (respectively) participated in a rally in the last year, while left-right captures self-reported ideological leaning (on a scale between 1 (left) to 11 (right), asked prior to treatment). Reading time measures (in seconds) the time participants spent on the screen displaying the vignette. Table A.2 in Online Appendix presents the results with the full covariate set. Table A.3 in online Appendix reports the corresponding OLS estimates.

How deeply embedded is the bias? Figure 2 plots treatment effects across the ideological scale. This builds on findings of previous literature consistently showing ideology to be a strong predictor of immigration attitudes (Hainmueller and Hopkins, 2014; Sides and Citrin, 2007; De Vries, Hakhver-

dian, and Lancee, 2013). Right-leaning respondents show somewhat larger biases, consistent with confirmation bias. Crucially, however, treatment effects remain negative and statistically significant across the entire spectrum, including the most left-leaning participants: ideology moderates the magnitude of the bias but does not eliminate it. Pairwise comparisons show that even sizable differences in left-right placement (e.g. moving from the lower to the upper quartile) do not translate into significantly different treatment effects (Table A.5). A parallel analysis using post-treatment immigration attitude measures—willingness to grant refugee residency, Muslim feeling thermometer, and UNHCR donation behavior—yields the same conclusion: effects persist substantially and significantly even among those with the strongest pro-immigration orientations (Online Appendix Figure A.1).[9]

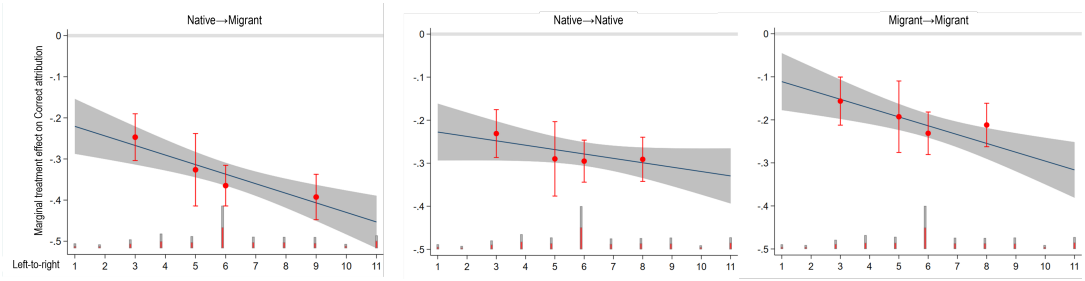


Figure 2: **Treatment effect on correct attribution by ideological orientation.**
Note: Treatment effects estimated by participants' left-to-right placement (1–11 scale) using the Interflex package (Hainmueller, Mummolo, and Xu, 2019) with a binning estimator; Migrant→Native is the baseline. Model follows column (3) of Table 1 (excluding country fixed effects).

Table 2 presents the incentivization study. Monetary rewards for correct recall have no significant effect on attribution accuracy in either treatment condition—the bias is equally large whether or not correct recall carries a monetary reward (coefficient on the interaction term is small and far from significant across all specifications). This rules out expressive responding as the primary driver and points to a genuine encoding failure.

[9]Unlike ideological self-placement, these immigration-attitude measures were recorded after exposure to the vignette. Conditioning on them directly may induce post-treatment selection bias. We therefore apply Lee's (2009) trimming-bounds procedure, under its monotonicity assumption, to bound the treatment effect among respondents who would report pro-immigration attitudes regardless of vignette assignment. Across all 21 vignette-comparison-by-measure cells, both the unadjusted and covariate-adjusted bounds remain negative and statistically significant. The results are insensitive to alternative random tie-breaking in the trimming procedure. See Online Appendix A5 and Tables A.8, A.9 and A.10.

Table 2: **Likelihood of correct attribution in Study 2 - the effect of incentivization**

	(1)	(2)	(3)	(4)
	Correct attribution$_i$			
Baseline: Migrant → Native				
Native→Migrant	-0.841***	-0.868***	-0.870***	-0.870***
	(0.107)	(0.109)	(0.110)	(0.110)
Incentive	0.077	0.072	0.081	0.078
	(0.122)	(0.124)	(0.125)	(0.125)
Native→Migrant*Incentive	0.107	0.127	0.110	0.111
	(0.154)	(0.157)	(0.158)	(0.158)
Voted			0.303***	0.304***
			(0.113)	(0.113)
Left to right			-0.015	-0.015
			(0.035)	(0.035)
Demonstrated			0.128	0.128
			(0.116)	(0.116)
Reading time				0.000
				(0.000)
Constant	1.234***	0.708***	0.522**	0.518**
	(0.085)	(0.192)	(0.248)	(0.248)
Demographic controls	No	Yes	Yes	Yes
Federal state of residence FEs	No	Yes	Yes	Yes
Observations	1,556	1,553	1,553	1,553

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Probit regression of a dummy *Correct attribution*$_i$ set to one if participant i provided a correct response to the recall task and zero otherwise, conducted among participants in the second study. Native → Migrant is an indicator variable for the treatment where the vignette described an attack of an immigrant against native majority. The roles are opposite in treatment Migrant → Native which serves as the baseline. Incentivized is an indicator for the case where a correct answer is rewarded with a monetary prize. Covariates as explained in Table 1. Table A.7 in online Appendix presents the results with the full covariate set.

Discussion

Our findings reveal a systematic bias in how information about immigration and violence is encoded and retained, favoring stereotype-consistent role attributions. Because the task involved no uncertainty and required only recognition—not belief updating—the pattern is difficult to reconcile with standard motivated reasoning accounts based on distorted priors or directional processing (Taber and Lodge, 2006; Kahan, 2013). Instead, the evidence points to a more basic mechanism: stereotype-incongruent information is neglected altogether, prior to any evaluative process.

Our findings echo Gilliam and Iyengar (2000)'s result that viewers exposed to a crime report featuring a white perpetrator, or none at all, often misremembered the suspect as non-white—a stereotype-congruent recall bias in race and crime. We extend this insight in four ways. First, we

show a structurally similar bias in a different domain (immigration-terrorism rather than race-crime) across four countries with markedly different immigration politics. Second, our 2×2 design shows that the bias cuts both ways: migrants are over-attributed the perpetrator role and under-recognized as victims. Third, an incentivized replication rules out expressive responding. Fourth, the bias persists among respondents least likely to hold the underlying stereotype—left-leaning, pro-immigration respondents who pledged part of their experimental earnings to a refugee-support organization. Taken together, these findings suggest that motivated reasoning alone cannot explain the bias. Widely shared stereotypes appear to shape information processing outside conscious deliberation (Little, 2025), even among respondents who neither endorse them nor have an evident motive to defend them.

One likely source of the stereotypes underlying this bias is cumulative media exposure, as repeated portrayals reinforce associations between migrants and violence. Prior research shows that repetition and familiarity contribute to collectively held schemas (Tversky and Kahneman, 1973; Fazio et al., 2015). Rather than explaining how such stereotypes arise or who endorses them, we examine their downstream consequences: once established, they can distort recognition of a single, unambiguous event, including among individuals who do not personally endorse them.

These findings carry implications for debates about information provision. Although the limited capacity of factual information to change attitudes and behavior is well established, information-based interventions remain central to migration politics, with competing political camps demanding changes in disclosure practices to counter what they regard as biased or incomplete information. In 2025, for example, the UK's National Police Chiefs' Council issued guidance encouraging police forces to disclose suspects' ethnicity and nationality, partly to counter misinformation and reduce community tensions. Numerous comparable policies replacing selective disclosure with more systematic reporting have been debated or implemented elsewhere (Keita, Renault, and Valette, 2024; Couttenier et al., 2024; Kanton Zürich, 2021). Our findings point to a limitation that precedes any attempt to persuade or correct beliefs: stereotype-incongruent information may fail to be registered accurately in the first place, raising doubts about the short- and mid-term effectiveness of such policies.

Our case involves terrorist violence, a domain selected for its stereotypical salience. The attribution

asymmetry may be weaker or absent for forms of violence that are less closely associated with migration, or for migrant groups that are not commonly portrayed through a threat-centered frame. We view this as a theoretically meaningful scope condition, because the immigration debates that motivate our study concern politically salient migrant flows involving groups for which strong stereotypes have already been documented in media coverage and public beliefs. Future work could extend our line of research by exploring such domain-specific variation.

References

Abou-Chadi, Tarik and Werner Krause (2020). "The causal effect of radical right success on mainstream parties' policy positions: A regression discontinuity approach". In: *British Journal of Political Science* 50.3, pp. 829–847.

Alizade, Jeyhun (2025). "The electoral politics of immigration and crime". In: *American Journal of Political Science*.

Aronow, Peter M., Jonathon Baron, and Lauren Pinson (2019). "A Note on Dropping Experimental Subjects who Fail a Manipulation Check". In: *Political Analysis* 27.4, 572–589. DOI: 10.1017/pan.2019.5.

Arzheimer, Kai (2009). "Contextual factors and the extreme right vote in Western Europe, 1980–2002". In: *American Journal of Political Science* 53.2, pp. 259–275.

Bell, Brian, Francesco Fasani, and Stephen Machin (Oct. 2013). "Crime and Immigration: Evidence from Large Immigrant Waves". In: *The Review of Economics and Statistics* 95.4, pp. 1278–1290.

Brill, Janine, Lars Guenther, Wibke Ehrhardt, and Georg Ruhrmann (2021). "Crime in Television News: Do News Factors Predict the Mentioning of a Criminal's Country of Origin?" In: *Mass Mediated Representations of Crime and Criminality*. Emerald Publishing Limited, pp. 31–48.

Couttenier, Mathieu, Sophie Hatte, Mathias Thoenig, and Stephanos Vlachos (2024). "Anti-muslim voting and media coverage of immigrant crimes". In: *Review of Economics and Statistics* 106.2, pp. 576–585.

De Vries, Catherine E, Armen Hakhverdian, and Bram Lancee (2013). "The dynamics of voters' left/right identification: The role of economic and cultural attitudes". In: *Political Science Research and Methods* 1.2, pp. 223–238.

Eyting, Markus (2022). "Why do we discriminate? the role of motivated reasoning". In.

Fazio, Lisa K, Nadia M Brashier, B Keith Payne, and Elizabeth J Marsh (2015). "Knowledge does not protect against illusory truth." In: *Journal of experimental psychology: general* 144.5, p. 993.

Getmansky, Anna (2025). "Trigger Happy: The Moral Hazard of Military Automation". Book manuscript in progress.

Gilliam, Franklin D and Shanto Iyengar (2000). "Prime suspects: The influence of local television news on the viewing public". In: *American Journal of Political Science* 44.3, pp. 560–573.

Hainmueller, Jens and Daniel J Hopkins (2014). "Public attitudes toward immigration". In: *Annual review of political science* 17.1, pp. 225–249.

Hainmueller, Jens, Jonathan Mummolo, and Yiqing Xu (2019). "How much should we trust estimates from multiplicative interaction models? Simple tools to improve empirical practice". In: *Political Analysis* 27.2, pp. 163–192.

Halla, Martin, Alexander F Wagner, and Josef Zweimüller (2017). "Immigration and voting for the far right". In: *Journal of the European economic association* 15.6, pp. 1341–1385.

Kahan, Dan M (2013). "Ideology, motivated reasoning, and cognitive reflection". In: *Judgment and Decision making* 8.4, pp. 407–424.

Kanol, Eylem and Max Schaub (2026). "Cultural Roots of Prejudice: Cultural Scripts and the Reactivation of Antisemitism in Germany". In: *Perspectives on Politics*, pp. 1–26.

Kanton Zürich (2021). *Nennung der Nationalität bei Polizeimeldungen: Einheitliche Regelung tritt am 1. Juli in Kraft*. https://www.zh.ch/de/news-uebersicht/medienmitteilungen/2021/04/nennung-der-nationalitaet-bei-polizeimeldungen--einheitliche-reg.html. Accessed July 19, 2026.

Kao, Kristen, Ellen Lust, Marwa Shalaby, and Chagai M. Weiss (2024). "Female Representation and Legitimacy: Evidence from a Harmonized Experiment in Jordan, Morocco, and Tunisia". In: *American Political Science Review* 118.1, 495–503. DOI: 10.1017/S0003055423000357.

Keita, Sekou, Thomas Renault, and Jérôme Valette (2024). "The usual suspects: Offender origin, media reporting and natives' attitudes towards immigration". In: *The Economic Journal* 134.657, pp. 322–362.

Klein, Ofir and Kimberly Gross (2019). "Does media coverage of terrorism exacerbate anti-Muslim sentiment? Evidence from a natural experiment in the United States". In: *Journal of Politics* 81.2, pp. 649–653.

Lajevardi, Nazita (2021). "The media matters: Muslim American portrayals and the effects on mass attitudes". In: *The Journal of Politics* 83.3, pp. 1060–1079.

Lee, David S (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". In: *The Review of Economic Studies* 76.3, pp. 1071–1102.

Little, Andrew T (2025). "How to distinguish motivated reasoning from Bayesian updating". In: *Political Behavior*, pp. 1–25.

Ousey, Graham C. and Charis E. Kubrin (2018). "Immigration and crime: Assessing a contentious issue". In: *Annual Review of Criminology* 1.1, pp. 63–84. DOI: 10.1146/annurev-criminol-032317-092026.

Rooduijn, Matthijs (2019). "State of the field: How to study populism and adjacent topics? A plea for both more and less focus". In: *European Journal of Political Research* 58.1, pp. 362–372.

Sides, John and Jack Citrin (2007). "European opinion about immigration: The role of identities, interests and information". In: *British journal of political science* 37.3, pp. 477–504.

Spirig, Judith (2023). "When issue salience affects adjudication: evidence from Swiss asylum appeal decisions". In: *American Journal of Political Science* 67.1, pp. 55–70.

START (National Consortium for the Study of Terrorism and Responses to Terrorism) (2021). *Global Terrorism Database 1970–2019 [data file]*. https://www.start.umd.edu/data-tools/GTD. Accessed March 2022.

Stroud, Natalie Jomini (2008). "Media use and political predispositions: Revisiting the concept of selective exposure". In: *Political behavior* 30.3, pp. 341–366.

Taber, Charles S and Milton Lodge (2006). "Motivated skepticism in the evaluation of political beliefs". In: *American journal of political science* 50.3, pp. 755–769.

Tversky, Amos and Daniel Kahneman (1973). "Availability: A heuristic for judging frequency and probability". In: *Cognitive psychology* 5.2, pp. 207–232.

ONLINE APPENDIX

A Grip on the Mind: Stereotypes and Information Recognition Failure

A1. Treatment Effect by Immigration Attitudes

Figure A.1 reports treatment effects on correct attribution across the distribution of respondents' immigration attitudes, with the Migrant→Native condition as the baseline. The top-left panel splits the sample by whether respondents donated any positive amount to UNHCR or not and plots the treatment effect separately for the two groups in each of the treatments. The remaining three panels use the Interflex package (Hainmueller, Mummolo, and Xu, 2019) with a kernel estimator to trace treatment effects continuously along willingness to grant residency to Afghan refugees, willingness to grant residency to Syrian refugees, and the first principal component constructed from four (standardized) immigration-attitude items — whether migration benefits the economy (reverse-coded), whether migration contributes to crime, and stated willingness to rally against and to vote against Muslims. Across all four measures, treatment effects remain negative and statistically significant even among respondents with the most favorable attitudes toward immigrants and refugees, indicating that the bias is not confined to those with anti-immigrant priors.

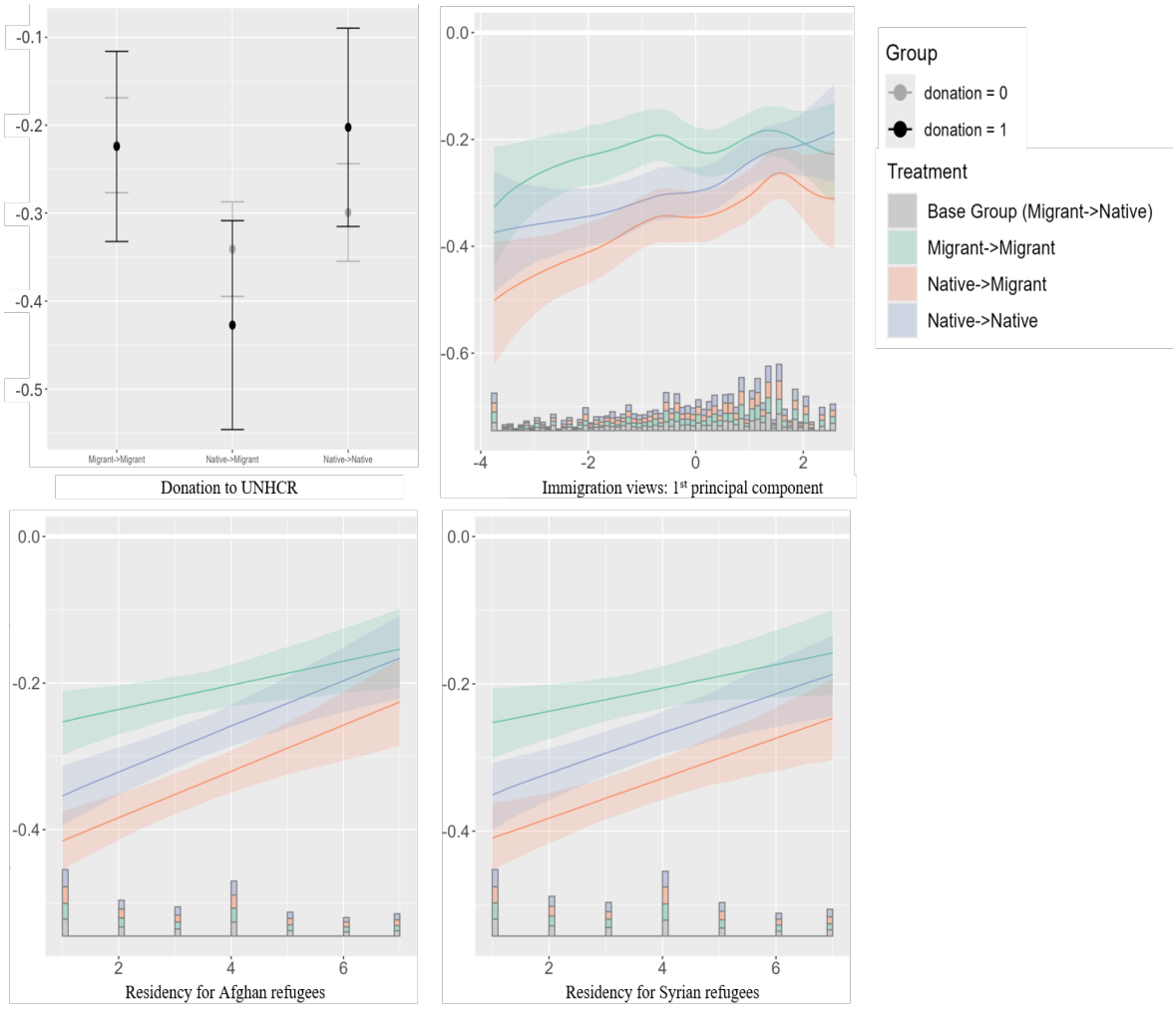


Figure A.1: Treatment effect on correct attribution by immigration attitudes

Notes: Treatment effects estimated along the distribution of four measures of participants' expressed immigration preferences (all recoded so that higher values indicate more positive attitudes). Vertical axes show the treatment effect—the (negative) difference in probability of correct attribution in the indicated treatment relative to the Migrant→Native baseline. The first panel shows effects estimated separately among UNHCR donors and non-donors. The remaining panels use the Interflex package (Hainmueller, Mummolo, and Xu, 2019) with a kernel estimator. Model follows column (3) of Table 1.

A2. Power Analysis

Figure A.2 reports the minimum sample size per treatment group required to detect a given treatment effect on correct recall, for power $(1-\beta) = 0.80$ and significance level $\alpha = 0.05$, using a standard two-proportion power calculation. The horizontal axis varies the assumed baseline rate of correct recall in the Migrant→Native condition, and each curve corresponds to a hypothesized reduction in accuracy (5, 7, or 10 percentage points) in a counter-stereotypical treatment relative to this baseline. Our realized sample of approximately 2,000 respondents per treatment comfortably exceeds the requirement even under relatively conservative predictions of the baseline accuracy and predicted effect size.

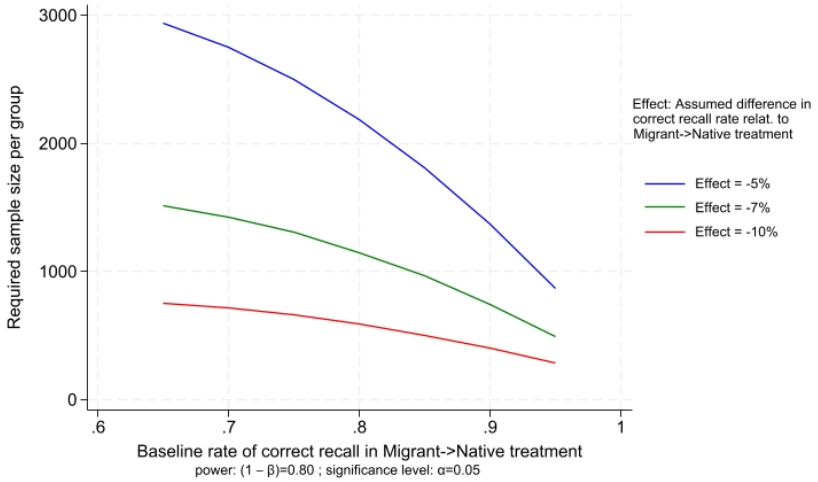


Figure A.2: Power analysis: required sample size

Notes: The figure depicts the calculated minimal number of participants per treatment group to detect a treatment effect on the rate of correct recall, for varying hypothesized effect sizes and baseline recall rates.

A3. Chance-adjusted Accuracy

Figure A.3 reports accuracy adjusted for the probability of a correct guess by chance (i.e. absent any genuine recall). Since respondents chose among four possible scenarios, random guessing yields a baseline accuracy of 25%; we rescale raw accuracy as:

$$\text{Adjusted Accuracy} = \frac{\text{Observed Accuracy} - 0.25}{1 - 0.25}, \tag{1}$$

which recovers the implied share of respondents who correctly retained the vignette, bounded between 0 (chance-level performance) and 1 (perfect recall).

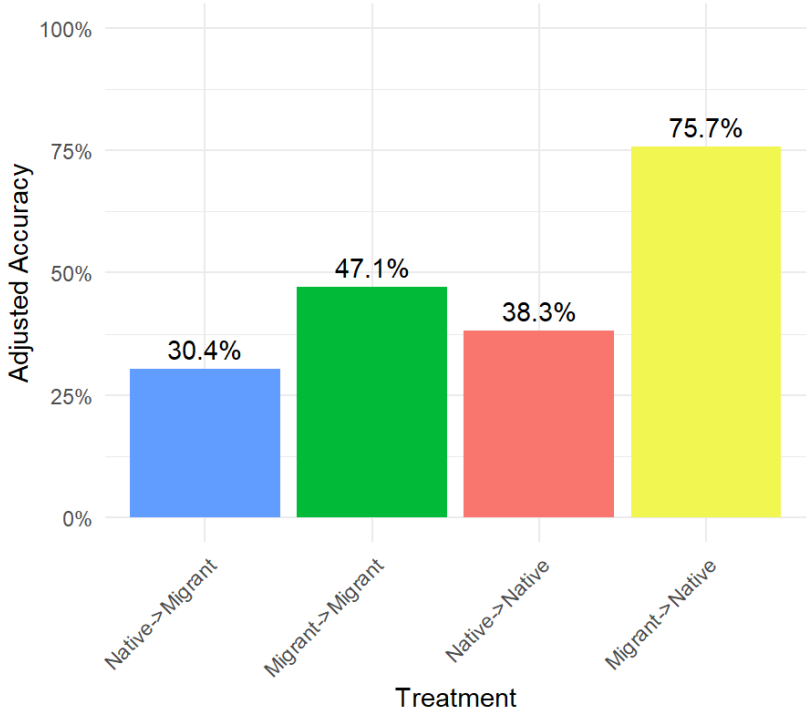


Figure A.3: Chance-adjusted accuracy of attribution

Notes: The figure depicts the accuracy of attribution adjusted for the baseline probability of providing a correct answer by chance (25%).

A4. Balance Checks and Additional Analysis

Table A.1 presents balance checks across the four treatment conditions, reporting the share of respondents with each demographic and political characteristic by treatment arm along with Pearson's χ^2 test of independence; no characteristic differs significantly across treatments, confirming successful randomization. Table A.2 reports the full specification underlying Table 1, including coefficients on all demographic and country controls omitted from the main text. Table A.3 reports the corresponding OLS estimates with the full specification, allowing examination of the role of demographic characteristics in predicting recall accuracy. To gain further descriptive insight into which demographic groups recall the vignette more accurately, Figure A.4 presents the same specification estimated separately by treatment. Table A.4 re-estimates the main treatment effects separately within quintiles of vignette reading time, to assess whether the recall bias is concentrated among respondents who spent little time on the vignette; the treatment effects remain negative and significant across all five quintiles. Table A.5 reports pairwise differences in treatment effects at the 25th, 50th, and 75th percentiles of the left-right ideology scale, estimated using the Interflex package with a kernel estimator, complementing the visual results in Figure A.1. Table A.6 reports results from the non-incentivized arm of Study 2 in isolation, and Table A.7 reports the full specification of the incentivization analysis summarized in Table 2, including all demographic and federal-state controls.

Table A.1: Balance table

Treatment:	Native→Migrant	Native→Native	Migrant→Migrant	Migrant→Native	χ^2	p
Gender						
Male	0.481	0.495	0.501	0.487	1.809	0.613
Female	0.516	0.503	0.497	0.510	1.680	0.641
Other	0.003	0.001	0.003	0.004	1.775	0.620
Age						
18-29	0.203	0.206	0.201	0.201	0.166	0.983
30-39	0.192	0.190	0.208	0.180	5.210	0.157
40-49	0.216	0.206	0.197	0.213	2.515	0.473
50-59	0.216	0.215	0.215	0.205	0.934	0.817
60-70	0.173	0.182	0.179	0.201	5.870	0.118
University degree	0.272	0.270	0.283	0.259	3.029	0.387
Employment						
Employed	0.604	0.599	0.604	0.582	2.678	0.444
Self-employed	0.070	0.066	0.053	0.059	5.797	0.122
Unemployed	0.067	0.068	0.067	0.074	0.932	0.818
Retired	0.117	0.131	0.140	0.139	5.523	0.137
Student	0.076	0.065	0.069	0.073	1.811	0.613
House person	0.044	0.050	0.046	0.050	1.255	0.740
Long sick	0.022	0.021	0.021	0.024	0.479	0.923
Political behavior						
Left-to-right	6.145	6.094	6.095	6.010	2.613	0.455
Voted	0.815	0.796	0.811	0.815	3.016	0.389
Demonstrated	0.067	0.061	0.075	0.074	4.002	0.261
Observations	1,979	2,063	1,962	1,971		

Notes: The table reports the share of the sample with a corresponding characteristic, in each of the four treatments. The last two columns report the Pearson's χ^2 test statistic and the associated p values.

Table A.2: Treatment effect: likelihood of correct attribution (full specification)

	(1)	(2)	(3)	(4)
	Correct attribution$_i$			
Baseline: Migrant → Native				
Native → Migrant	-0.962***	-0.973***	-0.971***	-0.973***
	(0.043)	(0.044)	(0.044)	(0.044)
Native → Native	-0.814***	-0.819***	-0.815***	-0.817***
	(0.043)	(0.043)	(0.043)	(0.043)
Migrant → Migrant	-0.646***	-0.653***	-0.650***	-0.651***
	(0.044)	(0.044)	(0.044)	(0.044)
Age		0.00573***	0.00539***	0.00510***
		(0.00134)	(0.00136)	(0.00136)
University degree		0.165***	0.155***	0.158***
		(0.0342)	(0.0345)	(0.0345)
Gender (baseline: Male)				
Female		0.0808***	0.0746**	0.0720**
		(0.0304)	(0.0306)	(0.0306)
Other		-0.173	-0.125	-0.124
		(0.301)	(0.311)	(0.311)
Employment status (baseline: Employed)				
Self-employed		0.0170	0.0271	0.0273
		(0.0626)	(0.0627)	(0.0628)
Unemployed		0.00804	0.0113	0.00951
		(0.0594)	(0.0599)	(0.0599)
Retired		-0.110**	-0.119**	-0.120**
		(0.0527)	(0.0529)	(0.0530)
Student		0.131**	0.122*	0.120*
		(0.0639)	(0.0642)	(0.0643)
Household duties		-0.177**	-0.163**	-0.159**
		(0.0725)	(0.0728)	(0.0729)
Inactive		0.0735	0.0540	0.0439
		(0.103)	(0.103)	(0.104)
Country of residence (baseline: Germany)				
Hungary		-0.157***	-0.141***	-0.142***
		(0.0418)	(0.0421)	(0.0422)
Italy		-0.0878**	-0.0589	-0.0572
		(0.0429)	(0.0441)	(0.0441)
Sweden		-0.187***	-0.177***	-0.180***
		(0.0424)	(0.0427)	(0.0428)
Voted			0.081**	0.080**
			(0.039)	(0.039)
Left to right			-0.021***	-0.021***
			(0.006)	(0.006)
Demonstrated			-0.034	-0.041
			(0.059)	(0.059)
Reading time				0.001***
				(0.000)
Constant	0.907***	0.694***	0.763***	0.751***
	(0.033)	(0.076)	(0.089)	(0.089)
Observations	7,975	7,975	7,964	7,964

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Probit regression of a dummy *Correct attribution*$_i$ set to one if participant *i* provided a correct response to the recall task and zero otherwise. Native → Migrant, Migrant → Migrant, and Native → Native are indicator variables for the corresponding treatments varying the identity of the attacker and the victim of the attack described in the vignette, with treatment Migrant → Native serving as the baseline. Demographic controls include age, gender, completion of higher education, employment status, and country of residence. Variables Voted and Demonstrated take value one if a participant stated to have voted in the last election and (respectively) participated in a rally in the last year, while Left to right captures their self-reported ideological leaning (on a scale between 1 (left) to 11 (right)). Reading time measures (in seconds) the time participants spent on the screen displaying the vignette.

Table A.3: Treatment effect: likelihood of correct attribution OLS analysis

	(1)	(2)	(3)	(4)
		Correct attribution$_i$		
Baseline: Migrant → Native				
Native → Migrant	-0.340***	-0.341***	-0.340***	-0.340***
	(0.0150)	(0.0149)	(0.0149)	(0.0149)
Native → Native	-0.281***	-0.280***	-0.279***	-0.280***
	(0.0148)	(0.0148)	(0.0148)	(0.0148)
Migrant → Migrant	-0.215***	-0.216***	-0.214***	-0.214***
	(0.0150)	(0.0150)	(0.0150)	(0.0150)
Age		0.00208***	0.00196***	0.00187***
		(0.000484)	(0.000489)	(0.000488)
University degree		0.0602***	0.0564***	0.0569***
		(0.0122)	(0.0123)	(0.0123)
Gender (baseline: Male)				
Female		0.0293***	0.0270**	0.0263**
		(0.0109)	(0.0110)	(0.0109)
Other		-0.0662	-0.0489	-0.0482
		(0.105)	(0.108)	(0.108)
Employment status (baseline: Employed)				
Self-employed		0.00497	0.00885	0.00932
		(0.0224)	(0.0225)	(0.0225)
Unemployed		0.00415	0.00526	0.00468
		(0.0214)	(0.0215)	(0.0215)
Retired		-0.0398**	-0.0429**	-0.0430**
		(0.0189)	(0.0190)	(0.0189)
Student		0.0465**	0.0429*	0.0423*
		(0.0229)	(0.0230)	(0.0229)
Household duties		-0.0640**	-0.0587**	-0.0574**
		(0.0265)	(0.0265)	(0.0265)
Inactive		0.0283	0.0204	0.0156
		(0.0371)	(0.0372)	(0.0371)
Country of residence (baseline: Germany)				
Hungary		-0.0554***	-0.0492***	-0.0496***
		(0.0150)	(0.0151)	(0.0150)
Italy		-0.0285*	-0.0180	-0.0172
		(0.0153)	(0.0157)	(0.0157)
Sweden		-0.0662***	-0.0622***	-0.0629***
		(0.0152)	(0.0153)	(0.0153)
Voted			0.0291**	0.0288**
			(0.0142)	(0.0142)
Left to right			-0.00774***	-0.00772***
			(0.00232)	(0.00231)
Demonstrated			-0.0121	-0.0138
			(0.0211)	(0.0211)
Reading time				0.000391***
				(0.0000818)
Constant	0.818***	0.736***	0.763***	0.757***
	(0.0106)	(0.0269)	(0.0317)	(0.0317)
Observations	7975	7975	7964	7964
R-squared	0.0687	0.0772	0.0788	0.0815

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

OLS regression of a dummy *Correct attribution*$_i$ set to one if participant *i* provided a correct response to the recall task and zero otherwise. Native → Migrant, Migrant → Migrant, and Native → Native are indicator variables for the corresponding treatments varying the identity of the attacker and the victim of the attack described in the vignette, with treatment Migrant → Native serving as the baseline. Demographic controls include age, gender, completion of higher education, employment status, and country of residence. Variables Voted and Demonstrated take value one if a participant stated to have voted in the last election and (respectively) participated in a rally in the last year, while Left to right captures their self-reported ideological leaning (on a scale between 1 (left) to 11 (right)). Reading time measures (in seconds) the time participants spent on the screen displaying the vignette.

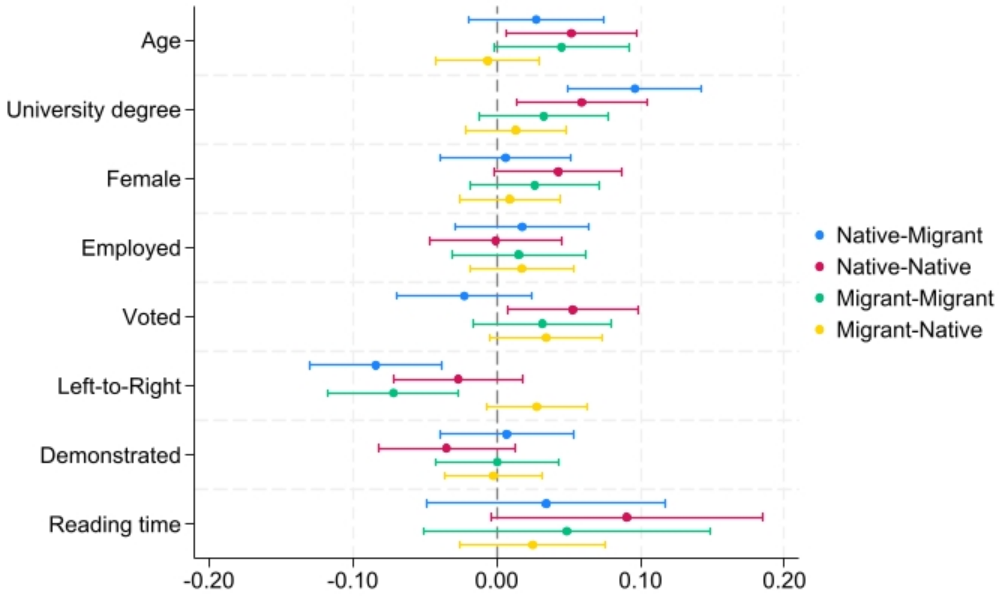


Figure A.4: Treatment-specific correlates of correct recall

Notes: The figure plots OLS coefficients, with 95% confidence intervals, from four separate regressions of accuracy of attribution on respondent characteristics, estimated separately within each treatment condition; both the outcome and all predictors are standardized to mean zero and unit standard deviation. All regressions additionally control for country fixed effects, omitted from the figure for legibility.

Table A.4: Likelihood of correct recall - sample split into quintiles by vignette reading time

	(1)	(2)	(3)	(4)	(5)
	Correct attribution$_i$				
Reading time:	quint.1	quint.2	quint.3	quint.4	quint.5
Baseline: Migrant → Native					
Native→Migrant	-1.484***	-1.034***	-1.170***	-0.739***	-0.876***
	(0.0999)	(0.101)	(0.108)	(0.104)	(0.109)
Native→Native	-1.503***	-1.064***	-0.903***	-0.511***	-0.689***
	(0.101)	(0.103)	(0.108)	(0.102)	(0.106)
Migrant→Migrant	-1.133***	-0.690***	-0.722***	-0.287***	-0.654***
	(0.0950)	(0.102)	(0.110)	(0.108)	(0.109)
Age	0.00243	0.000662	-0.000395	-0.00180	-0.00655**
	(0.00330)	(0.00330)	(0.00318)	(0.00338)	(0.00333)
University degree	0.0312	0.227***	0.210***	0.355***	0.186**
	(0.0820)	(0.0786)	(0.0808)	(0.0883)	(0.0894)
Gender (baseline: Male)					
Female	0.0249	-0.0621	0.0629	-0.0251	0.00623
	(0.0760)	(0.0724)	(0.0715)	(0.0746)	(0.0718)
Other	0.623	-0.156	-0.144	-0.765	-0.909
	(0.914)	(0.679)	(0.704)	(1.009)	(0.588)
Employment status (baseline: Employed)					
Self employed	-0.0127	-0.0457	0.231	-0.152	0.249
	(0.151)	(0.138)	(0.144)	(0.149)	(0.184)
Unemployed	0.0156	-0.0348	-0.380***	0.119	-0.00938
	(0.151)	(0.149)	(0.134)	(0.143)	(0.138)
Retired	-0.113	-0.0740	-0.182	-0.232**	-0.127
	(0.156)	(0.135)	(0.121)	(0.117)	(0.113)
Student	0.205	0.0603	0.197	0.0125	-0.229
	(0.137)	(0.148)	(0.162)	(0.173)	(0.162)
Household duties	-0.0329	-0.193	0.00497	-0.132	-0.277
	(0.168)	(0.169)	(0.177)	(0.163)	(0.178)
Inactive	-0.242	0.0639	-0.128	0.00403	-0.137
	(0.345)	(0.316)	(0.221)	(0.200)	(0.220)
Country of residence (baseline: Germany)					
Hungary	-0.124	-0.348***	-0.217**	-0.142	0.107
	(0.102)	(0.108)	(0.102)	(0.107)	(0.0945)
Italy	0.00328	0.303***	-0.117	-0.233**	-0.277**
	(0.108)	(0.0985)	(0.102)	(0.110)	(0.110)
Sweden	-0.124	-0.0365	-0.171*	-0.422***	-0.157
	(0.107)	(0.0993)	(0.0995)	(0.102)	(0.104)
Voted	0.0308	-0.0404	0.0422	0.212**	0.0443
	(0.0896)	(0.0954)	(0.0948)	(0.0979)	(0.0944)
Left to right	0.0220	-0.0380**	-0.0434***	-0.0223	-0.0310**
	(0.0157)	(0.0152)	(0.0152)	(0.0152)	(0.0155)
Demonstrated	-0.113	0.0494	-0.220	0.301*	-0.0861
	(0.136)	(0.143)	(0.140)	(0.156)	(0.149)
Reading time	0.0504**	0.0942***	0.0218	0.00225	-0.000630**
	(0.0228)	(0.0156)	(0.0134)	(0.0105)	(0.000262)
Constant	0.0353	0.190	1.240***	1.196***	1.713***
	(0.225)	(0.261)	(0.329)	(0.375)	(0.216)
Observations	1656	1530	1661	1574	1543

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Probit regression of a dummy *True recall*$_i$ set to one if participant *i* provided a correct response to the recall task and zero otherwise. Each column shows the treatments effects for an individual quintile of the sample according to the time spent in reading the vignette. Native → Migrant, Native → Native, and Migrant → Migrant are indicators for the corresponding treatments (relative to the Migrant → Native treatment which serves as the baseline). Demographic controls include age, gender, completion of higher education, employment status.

Table A.5: Treatment effects along left-to-right placement

Baseline: Migrant → Native Moderator: Left-to-Right placement						
	diff.estimate	sd	z-value	p-value	lower CI(95%)	upper CI(95%)
Migrant → Migrant						
50% vs 25%	-0.029	0.023	-1.251	0.211	-0.073	0.015
75% vs 50%	0.015	0.033	0.444	0.657	-0.048	0.076
75% vs 25%	-0.014	0.030	-0.458	0.647	-0.073	0.046
Native → Migrant						
50% vs 25%	-0.044	0.022	-2.016	0.044	-0.087	-0.003
75% vs 50%	-0.003	0.033	-0.099	0.921	-0.069	0.058
75% vs 25%	-0.047	0.031	-1.527	0.127	-0.107	0.015
Native → Native						
50% vs 25%	-0.019	0.022	-0.840	0.401	-0.063	0.024
75% vs 50%	0.008	0.033	0.246	0.806	-0.059	0.072
75% vs 25%	-0.011	0.031	-0.342	0.733	-0.072	0.049

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
The table reports the differences in estimated effects of treatments at 25, 50 and 75 percentile of in-sample placement on the 7-point left-to-right scale of political orientation. The differences are estimated using the Interflex package (Hainmueller, Mummolo, and Xu, 2019) employing a kernel estimator. The effect of a treatment is captured by the likelihood of a participant correctly recalling the type of attack described by the vignette shown to them, with Migrant → Native treatment serving as the baseline in all pairwise comparisons.

Table A.6: Likelihood of correct recall in Study 2 - without incentivization

	(1)	(2)	(3)	(4)
		Correct attribution$_i$		
Baseline: Migrant → Native				
Native → Migrant	-0.841***	-0.873***	-0.881***	-0.880***
	(0.107)	(0.111)	(0.112)	(0.112)
Age		0.00542	0.00533	0.00555
		(0.00463)	(0.00469)	(0.00469)
University Degree		0.316**	0.280**	0.267**
		(0.130)	(0.132)	(0.133)
Female		0.170	0.130	0.134
		(0.107)	(0.110)	(0.110)
Employment status (baseline: Employed)				
Self-employed		-0.222	-0.137	-0.140
		(0.236)	(0.241)	(0.241)
Unemployed		-0.140	-0.0737	-0.0589
		(0.252)	(0.252)	(0.252)
Retired		-0.226	-0.212	-0.204
		(0.169)	(0.171)	(0.171)
Student		0.449*	0.479*	0.476*
		(0.249)	(0.257)	(0.256)
Voted			0.394**	0.394**
			(0.161)	(0.161)
Demonstrated			-0.00446	-0.00618
			(0.0499)	(0.0500)
Left to right			0.167	0.158
			(0.169)	(0.169)
Reading time				-0.00213
				(0.00208)
FEs for federal state of residence	No	Yes	Yes	Yes
Constant	1.234***	0.871***	0.559	0.600*
	(0.0849)	(0.257)	(0.344)	(0.347)
Observations	776	773	773	773

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
The table shows the replication of the results depicted in Table 1 run among participants in the second study who *were not* offered a monetary reward for providing a correct answer. Reported are the results of a Probit regression of a dummy *True recall*$_i$ set to one if participant *i* provided a correct response to the recall task and zero otherwise. Native → Migrant is an indicator variable for the treatment where the vignette described an attack of an immigrant against native majority. The roles are opposite in treatment Migrant → Native which serves as the baseline. Demographic controls include age, gender, completion of higher education, employment status. Variables Voted and Demonstrated take value one if a participant stated to have voted in the last election and (respectively) participated in a rally in the last year, while Left to right captures the standardized value of self-reported ideological leaning (on a scale between 1 (left) to 7 (right)). Reading time measures (in seconds) the time participants spent on the screen displaying the vignette.

Table A.7: Likelihood of correct attribution in Study 2 - the effect of incentivization (full specification)

	(1)	(2)	(3)	(4)
	Correct attribution$_i$			
Baseline: Migrant → Native				
Native→Migrant	-0.841***	-0.868***	-0.870***	-0.870***
	(0.107)	(0.109)	(0.110)	(0.110)
Incentive	0.077	0.072	0.081	0.078
	(0.122)	(0.124)	(0.125)	(0.125)
Native→Migrant×Incentive	0.107	0.127	0.110	0.111
	(0.154)	(0.157)	(0.158)	(0.158)
Age		0.00950***	0.00914***	0.00913***
		(0.00333)	(0.00336)	(0.00336)
University Degree		0.205**	0.173*	0.172*
		(0.0885)	(0.0897)	(0.0897)
Female		0.0404	0.0216	0.0224
		(0.0761)	(0.0773)	(0.0773)
Employment status (baseline: Employed)				
Self-employed		-0.0680	-0.0240	-0.0234
		(0.188)	(0.190)	(0.190)
Unemployed		-0.128	-0.0678	-0.0696
		(0.153)	(0.155)	(0.155)
Retired		-0.190	-0.170	-0.170
		(0.126)	(0.127)	(0.127)
Student		0.369**	0.372**	0.373**
		(0.174)	(0.177)	(0.177)
Voted			0.303***	0.304***
			(0.113)	(0.113)
Left to right			-0.015	-0.015
			(0.035)	(0.035)
Demonstrated			0.128	0.128
			(0.116)	(0.116)
Reading time				0.000
				(0.000)
Constant	1.234***	0.708***	0.522**	0.518**
	(0.085)	(0.192)	(0.248)	(0.248)
Federal state of residence FEs	No	Yes	Yes	Yes
Observations	1,556	1,553	1,553	1,553

Notes: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Probit regression of a dummy *Correct attribution*$_i$ set to one if participant i provided a correct response to the recall task and zero otherwise, conducted among participants in the second study. Native → Migrant is an indicator variable for the treatment where the vignette described an attack of an immigrant against native majority. The roles are opposite in treatment Migrant → Native which serves as the baseline. Incentivized is an indicator for the case where a correct answer is rewarded with a monetary prize. Demographic controls include age, gender, completion of higher education, employment status. Variables Voted and Demonstrated take value one if a participant stated to have voted in the last election and (respectively) participated in a rally in the last year, while Left to right captures the self-reported ideological leaning (on a scale between 1 (left) to 7 (right)). Reading time measures (in seconds) the time participants spent on the screen displaying the vignette.

A5. Lee Bounds for the Post-Treatment Attitude Moderators

The recall item asks respondents to identify which of the four vignettes they were shown; a response is coded correct if it matches the assigned condition. As reported in the main text, the Migrant→Native vignette—the stereotype-consistent pairing of perpetrator and victim—is identified correctly far more often than the other three vignettes, and it is treated below, as in the main text, as the reference category against which each of the other three arms is compared.

We then ask whether this gap in correct attribution is itself larger or smaller among respondents with more pro-immigration attitudes, as measured via our four proxies: their pledge to donate to the UNHCR, their stated willingness to grant residency to Afghan and to Syrian refugees (each on a 1–7 scale), and a Pro-immigration index (the first principal component of five immigration-attitude items – residency for Syrians, Nigerians, and Afghans, and reversed agreement with anti-Muslim rallying and voting – oriented so that higher values indicate more pro-immigration views). All four measures are recorded *after* the experimental vignette, so any of them could in principle have been shifted by treatment itself; conditioning on them directly would then risk the standard form of post-treatment bias. To address this problem, we apply Lee's (2009) trimming-bounds method.

Relation to Lee's (2009 estimand. Lee bounds a treatment effect when a post-treatment binary variable S determines a subpopulation of interest, and the composition of that subpopulation may itself depend on treatment. In Lee's application, S is employment and Y is the wage, which is structurally unobserved for the unemployed; here, by contrast, "Correct Attribution" is recorded for every respondent regardless of their attitudes, so nothing is literally missing. The two problems are nonetheless formally equivalent: whether we condition on $S = 1$ because Y is unobserved otherwise or because $S = 1$ defines the substantive subpopulation we want to speak to, comparing Y across treatment arms *within* $S = 1$ is valid only if the $S = 1$ group is composed the same way in both arms—and random assignment alone does not guarantee this once S is itself measured after treatment. Thus, Lee bounds can be helpful in conditioning on a post-treatment variable more generally, not only for literal attrition.

Formally, let $D \in \{0,1\}$ denote assignment to the non-reference vignette (Arm $= 1$) versus the migrant→native reference (Arm $= 0$), let $S(d)$ denote the potential value of the selection indicator ("pro-immigration") under assignment d, and let $Y(d)$ denote the potential recall outcome. Under random assignment and a monotonicity assumption that D shifts S in one direction only (discussed below), Lee's method sharply bounds

$$\text{ATE}_{\text{always}} \;=\; E\big[\,Y(1)-Y(0)\;\big|\;S(1)=S(0)=1\,\big],$$

the treatment effect on recall accuracy *among respondents who would report pro-immigration attitudes regardless of which vignette they had been shown*.

Procedure. For each of the three pairwise comparisons (each alternative vignette against the migrant→native reference) and each of seven selection specifications – donation (already binary), and each of the three continuous attitude measures split at the sample median (primary specification) and, for robustness, at the 75th percentile – with thresholds fixed once on the full pooled sample so that "pro-immigration" means the same thing across all three comparisons for a given measure:

1. *Size the trim.* Compute $\hat{p}_1 = \Pr(S = 1 \mid \text{Arm} = 1)$ and $\hat{p}_0 = \Pr(S = 1 \mid \text{Arm} = 0)$ among respondents with non-missing selection status. Whichever arm has the larger share is the "over-selected" arm; the trimming proportion is $\hat{p}_0^{\text{trim}} = (\hat{p}_{\text{over}} - \hat{p}_{\text{under}})/\hat{p}_{\text{over}}$, following Lee's (2009) formula.

2. *Trim within the selected subsample.* Restrict to respondents with $S = 1$ in both arms. Within the over-selected arm's $S = 1$ respondents, drop the top $\hat{p}_0^{\text{trim}}$ share of the recall item—and, separately, the bottom $\hat{p}_0^{\text{trim}}$ share—leaving the under-selected arm's $S = 1$ respondents untouched in both cases.

3. *Compare means.* On each trimmed sample, compute the simple difference in mean correct recall between arms. The smaller and larger of the two differences are the lower and upper Lee bound on $\text{ATE}_{\text{always}}$.

For reference, we also report the naive, untrimmed difference in means among respondents observed with $S = 1$ in both arms.

Point estimates and standard errors are computed with Stata's `leebounds` command.

Why monotonicity is likely to hold. Lee's bound requires that assignment to a given vignette can only move respondents' probability of reporting pro-immigration attitudes in one direction, never both—it would be violated if the vignette pushed some respondents toward more pro-immigration views and other respondents toward more anti-immigration views. This is plausible here for two reasons, specific to the design. First, the vignette is a light-touch, single-exposure treatment—a short paragraph read for a median of a few seconds before an intervening attitude battery—with no explicit persuasive argument about immigration policy; any effect it has on general attitudes is far more likely to be a uniform, same-direction nudge (or no effect at all) than a treatment that moves different respondents' attitudes in opposite directions. Second, each vignette consistently assigns migrants to the same structural role (perpetrator or victim) throughout, so any priming effect it induces should operate in a single direction across the sample rather than depending on unobserved respondent characteristics in a way that reverses sign. Consistent with this, the estimated imbalance $|\hat{p}_1 - \hat{p}_0|$ is modest throughout (Table A.8, column "Trim %"), as expected under one small, monotonic shift rather than larger offsetting movements that happen to net out.

Results. Table A.8 reports the naive comparison and the two Lee bounds for all 21 comparison-by-selection cells. In every cell, the vignette effect on recall remains large, negative, and highly significant even among respondents who would report pro-immigration attitudes regardless of treatment: the lower and upper bounds range from about -0.09 to -0.46, and the naive (uncorrected) estimate falls within or very close to that range in every case. Accounting for the post-treatment selection problem therefore does not change the substantive conclusion: the recall bias persists even among holders of attitudes opposite to the stereotype, exactly as documented in the main text, and this does not appear to be an artifact of conditioning on attitudes that treatment may itself have shifted.

Table A.8: Lee bounds on the vignette effect on correct attribution, among respondents who would report pro-immigration attitudes regardless of treatment.

Comparison	Selection measure	Naive (S=1)	Lower bound	Upper bound	Over-sel. arm	Trim (%)	*N* trimmed	*N* (S=1)
Tr1 vs. Tr4	Donated to UNHCR	−0.417*** (0.060)	−0.457*** (0.056)	−0.416*** (0.060)	Tr4	4.2	3	157
Tr1 vs. Tr4	Residency: Afghans (> median)	−0.281*** (0.020)	−0.301*** (0.021)	−0.258*** (0.021)	Comp.	4.1	39	1,868
Tr1 vs. Tr4	Residency: Afghans (> p75)	−0.236*** (0.034)	−0.285*** (0.036)	−0.164*** (0.035)	Comp.	10.8	36	631
Tr1 vs. Tr4	Residency: Syrians (> median)	−0.279*** (0.026)	−0.326*** (0.027)	−0.221*** (0.027)	Comp.	9.5	56	1,118
Tr1 vs. Tr4	Residency: Syrians (> p75)	−0.238*** (0.033)	−0.293*** (0.035)	−0.154*** (0.034)	Comp.	12.2	43	662
Tr1 vs. Tr4	Pro-immigration index (> median)	−0.296*** (0.020)	−0.314*** (0.020)	−0.274*** (0.020)	Comp.	3.8	39	1,971
Tr1 vs. Tr4	Pro-immigration index (> p75)	−0.283*** (0.027)	−0.351*** (0.029)	−0.188*** (0.027)	Comp.	13.9	73	971
Tr2 vs. Tr4	Donated to UNHCR	−0.226*** (0.054)	−0.232*** (0.055)	−0.207*** (0.053)	Comp.	2.1	2	159
Tr2 vs. Tr4	Residency: Afghans (> median)	−0.184*** (0.020)	−0.198*** (0.020)	−0.158*** (0.020)	Comp.	3.9	37	1,863
Tr2 vs. Tr4	Residency: Afghans (> p75)	−0.156*** (0.034)	−0.185*** (0.035)	−0.094*** (0.033)	Comp.	8.4	27	621
Tr2 vs. Tr4	Residency: Syrians (> median)	−0.173*** (0.026)	−0.194*** (0.026)	−0.133*** (0.026)	Comp.	5.8	32	1,093
Tr2 vs. Tr4	Residency: Syrians (> p75)	−0.154*** (0.033)	−0.178*** (0.034)	−0.104*** (0.033)	Comp.	6.8	23	641
Tr2 vs. Tr4	Pro-immigration index (> median)	−0.187*** (0.019)	−0.196*** (0.019)	−0.170*** (0.019)	Comp.	2.5	25	1,953
Tr2 vs. Tr4	Pro-immigration index (> p75)	−0.193*** (0.026)	−0.234*** (0.028)	−0.107*** (0.026)	Comp.	11.3	57	953
Tr3 vs. Tr4	Donated to UNHCR	−0.203*** (0.056)	−0.243*** (0.052)	−0.201*** (0.057)	Tr4	7.1	5	146
Tr3 vs. Tr4	Residency: Afghans (> median)	−0.228*** (0.020)	−0.232*** (0.020)	−0.221*** (0.020)	Comp.	1.1	11	1,881
Tr3 vs. Tr4	Residency: Afghans (> p75)	−0.193*** (0.034)	−0.218*** (0.035)	−0.150*** (0.034)	Comp.	6.2	21	629
Tr3 vs. Tr4	Residency: Syrians (> median)	−0.229*** (0.026)	−0.273*** (0.027)	−0.161*** (0.026)	Comp.	10.1	62	1,148
Tr3 vs. Tr4	Residency: Syrians (> p75)	−0.227*** (0.033)	−0.292*** (0.035)	−0.125*** (0.033)	Comp.	14.3	54	687
Tr3 vs. Tr4	Pro-immigration index (> median)	−0.227*** (0.019)	−0.232*** (0.019)	−0.226*** (0.019)	Tr4	0.6	6	1,969
Tr3 vs. Tr4	Pro-immigration index (> p75)	−0.228*** (0.027)	−0.260*** (0.028)	−0.172*** (0.027)	Comp.	8.0	41	959

Note: Naive is the untrimmed difference in mean pass_mani between arms among respondents observed with the selection indicator $S = 1$ in both arms; robust standard errors in parentheses. Lower/Upper bound estimates and standard errors are from Stata's `leebounds` command (Lee, 2009). Tr4 (migrant → native) is the reference arm. Over-sel. arm identifies which arm has the higher share scoring $S = 1$ and is therefore trimmed ("Comp." = the non-reference vignette; "Tr4" = the reference). Trim (%) is the resulting trimming proportion; *N* trimmed is respondents dropped from the over-selected arm; *N* (S=1) is the total analytic sample (both arms, untrimmed). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Robustness: covariate adjustment. Table A.8 reports the unadjusted Lee bound – a simple difference in means, matching Stata's `leebounds` command. As a check, Table A.9 re-estimates all 21 cells replacing the raw mean comparison with the covariate-adjusted specification used in the main text – age, gender, education, employment, and country fixed effects – while leaving the trim itself (which arm is over-selected, and how large the trim is) unchanged, since Lee's trimming logic depends on the raw selection rates and outcome ranking, not on covariates. The covariate-adjusted bounds sit close to their unadjusted counterparts throughout, and the substantive conclusion is unchanged: the vignette effect on recall remains large, negative, and highly significant among respondents who would report pro-immigration attitudes regardless of treatment.

One methodological point follows directly from this comparison. In the unadjusted specification, the point estimate is probably invariant to how the trim's random tie-break is resolved (Table A.8): since the outcome is binary and the comparison is a simple difference in means, which specific tied respondents are trimmed cannot affect the resulting means. Once covariates are added, this guarantee

no longer holds – the covariate profile of the dropped respondents can differ across draws – so the tie-break stability check becomes a genuine, non-tautological exercise for the covariate-adjusted specification. Table A.10 repeats it for the covariate-adjusted bounds: the resulting variation is real but small, with a standard deviation across 200 draws below 0.007 in every cell – an order of magnitude smaller than the gap between each cell's own lower and upper bound. The covariate-adjusted bounds are therefore not sensitive to the arbitrary tie-break either, though for a substantive rather than a mechanical reason.

Table A.9: Lee bounds on the vignette effect on correct attribution, with covariate adjustment (age, gender, education, employment, and country fixed effects)

Comparison	Selection measure	Naive (S=1)	Lower bound	Upper bound	Over-sel. arm	Trim (%)	*N* trimmed	*N* (S=1)
Tr1 vs. Tr4	Donated to UNHCR	-0.440^{***} (0.068)	-0.486^{***} (0.063)	-0.441^{***} (0.069)	Tr4	4.2	3	157
Tr1 vs. Tr4	Residency: Afghans (> median)	-0.284^{***} (0.020)	-0.303^{***} (0.021)	-0.261^{***} (0.021)	Comp.	4.1	39	1,868
Tr1 vs. Tr4	Residency: Afghans (> p75)	-0.234^{***} (0.035)	-0.284^{***} (0.037)	-0.160^{***} (0.036)	Comp.	10.8	36	631
Tr1 vs. Tr4	Residency: Syrians (> median)	-0.281^{***} (0.026)	-0.326^{***} (0.027)	-0.223^{***} (0.027)	Comp.	9.5	56	1,118
Tr1 vs. Tr4	Residency: Syrians (> p75)	-0.238^{***} (0.034)	-0.296^{***} (0.036)	-0.152^{***} (0.034)	Comp.	12.2	43	662
Tr1 vs. Tr4	Pro-immigration index (> median)	-0.297^{***} (0.020)	-0.315^{***} (0.020)	-0.275^{***} (0.020)	Comp.	3.8	39	1,971
Tr1 vs. Tr4	Pro-immigration index (> p75)	-0.288^{***} (0.027)	-0.356^{***} (0.029)	-0.193^{***} (0.028)	Comp.	13.9	73	971
Tr2 vs. Tr4	Donated to UNHCR	-0.178^{***} (0.057)	-0.187^{***} (0.059)	-0.161^{***} (0.057)	Comp.	2.1	2	159
Tr2 vs. Tr4	Residency: Afghans (> median)	-0.183^{***} (0.020)	-0.197^{***} (0.020)	-0.157^{***} (0.020)	Comp.	3.9	37	1,863
Tr2 vs. Tr4	Residency: Afghans (> p75)	-0.157^{***} (0.034)	-0.186^{***} (0.036)	-0.095^{***} (0.034)	Comp.	8.4	27	621
Tr2 vs. Tr4	Residency: Syrians (> median)	-0.172^{***} (0.026)	-0.191^{***} (0.027)	-0.131^{***} (0.026)	Comp.	5.8	32	1,093
Tr2 vs. Tr4	Residency: Syrians (> p75)	-0.151^{***} (0.034)	-0.176^{***} (0.035)	-0.100^{***} (0.033)	Comp.	6.8	23	641
Tr2 vs. Tr4	Pro-immigration index (> median)	-0.188^{***} (0.019)	-0.197^{***} (0.020)	-0.171^{***} (0.019)	Comp.	2.5	25	1,953
Tr2 vs. Tr4	Pro-immigration index (> p75)	-0.197^{***} (0.026)	-0.239^{***} (0.028)	-0.112^{***} (0.026)	Comp.	11.3	57	953
Tr3 vs. Tr4	Donated to UNHCR	-0.179^{***} (0.060)	-0.221^{***} (0.054)	-0.180^{***} (0.061)	Tr4	7.1	5	146
Tr3 vs. Tr4	Residency: Afghans (> median)	-0.223^{***} (0.020)	-0.228^{***} (0.020)	-0.217^{***} (0.020)	Comp.	1.1	11	1,881
Tr3 vs. Tr4	Residency: Afghans (> p75)	-0.191^{***} (0.035)	-0.214^{***} (0.035)	-0.147^{***} (0.035)	Comp.	6.2	21	629
Tr3 vs. Tr4	Residency: Syrians (> median)	-0.221^{***} (0.026)	-0.263^{***} (0.027)	-0.154^{***} (0.026)	Comp.	10.1	62	1,148
Tr3 vs. Tr4	Residency: Syrians (> p75)	-0.219^{***} (0.033)	-0.284^{***} (0.035)	-0.118^{***} (0.033)	Comp.	14.3	54	687
Tr3 vs. Tr4	Pro-immigration index (> median)	-0.225^{***} (0.020)	-0.231^{***} (0.019)	-0.224^{***} (0.020)	Tr4	0.6	6	1,969
Tr3 vs. Tr4	Pro-immigration index (> p75)	-0.228^{***} (0.027)	-0.259^{***} (0.028)	-0.170^{***} (0.027)	Comp.	8.0	41	959

Note: As Table A.8, except Naive/Lower/Upper are now estimated via OLS with age, gender, education, employment, and country fixed effects as additional covariates (robust standard errors in parentheses), rather than a raw difference in means. The trim itself (over-selected arm, trim proportion, *N* trimmed) is unchanged from Table A.8, since it depends only on the raw selection rates and outcome ranking, not on covariates. Because `leebounds` does not support covariate adjustment, Lower/Upper here are computed directly as the smaller/larger of the two covariate-adjusted `arm` coefficients from the top-trimmed and bottom-trimmed samples. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table A.10: Stability of the covariate-adjusted Lee bounds across 200 random tie-break draws

Comparison	Selection measure	Lower bound (mean)	Lower bound (SD)	Upper bound (mean)	Upper bound (SD)
Tr1 vs. Tr4	Donated to UNHCR	−0.486	0.0000	−0.439	0.0027
Tr1 vs. Tr4	Residency: Afghans (> median)	−0.303	0.0005	−0.260	0.0004
Tr1 vs. Tr4	Residency: Afghans (> p75)	−0.283	0.0013	−0.160	0.0020
Tr1 vs. Tr4	Residency: Syrians (> median)	−0.328	0.0010	−0.223	0.0011
Tr1 vs. Tr4	Residency: Syrians (> p75)	−0.294	0.0015	−0.153	0.0020
Tr1 vs. Tr4	Pro-immigration index (> median)	−0.315	0.0005	−0.275	0.0004
Tr1 vs. Tr4	Pro-immigration index (> p75)	−0.356	0.0017	−0.193	0.0014
Tr2 vs. Tr4	Donated to UNHCR	−0.184	0.0023	−0.164	0.0063
Tr2 vs. Tr4	Residency: Afghans (> median)	−0.197	0.0004	−0.157	0.0006
Tr2 vs. Tr4	Residency: Afghans (> p75)	−0.187	0.0012	−0.094	0.0019
Tr2 vs. Tr4	Residency: Syrians (> median)	−0.192	0.0007	−0.132	0.0009
Tr2 vs. Tr4	Residency: Syrians (> p75)	−0.175	0.0010	−0.100	0.0018
Tr2 vs. Tr4	Pro-immigration index (> median)	−0.197	0.0003	−0.171	0.0003
Tr2 vs. Tr4	Pro-immigration index (> p75)	−0.239	0.0012	−0.112	0.0011
Tr3 vs. Tr4	Donated to UNHCR	−0.221	0.0017	−0.175	0.0041
Tr3 vs. Tr4	Residency: Afghans (> median)	−0.228	0.0002	−0.216	0.0004
Tr3 vs. Tr4	Residency: Afghans (> p75)	−0.215	0.0011	−0.148	0.0017
Tr3 vs. Tr4	Residency: Syrians (> median)	−0.264	0.0009	−0.155	0.0016
Tr3 vs. Tr4	Residency: Syrians (> p75)	−0.283	0.0012	−0.119	0.0025
Tr3 vs. Tr4	Pro-immigration index (> median)	−0.230	0.0002	−0.224	0.0002
Tr3 vs. Tr4	Pro-immigration index (> p75)	−0.260	0.0012	−0.173	0.0013

Note: Mean and standard deviation of the covariate-adjusted lower- and upper-bound estimate across 200 independent re-draws of the trim's random tie-break; compare to Table A.9. Unlike the unadjusted case (Table A.8), variation across draws here is real, since the covariate profile of the respondents dropped at the trim margin can differ from draw to draw – but it is small throughout (maximum SD 0.0063), an order of magnitude below the gap between each cell's own lower and upper bound.