Science Advances
AAAS

Supplementary Materials for

Ideological bias in the production of research findings

George J. Borjas and Nate Breznau

Corresponding author: George J. Borjas, george_borjas@hks.harvard.edu

Sci. Adv. **12**, eadz7173 (2026)
DOI: 10.1126/sciadv.adz7173

This PDF file includes:

Supplementary Text

Notes for Users

To allow for easy replication, this appendix describes the variables used in our study including further methodological explanations, and tables and figures. The experimental data used in this study were produced by BRW[1] and are available in the Github repository at https://github.com/nbreznau/CRI. The analysis uses the main data file in that repository, cri.csv. The file has 1,253 observations and the data are at the team-model level. A fully reproducible repository with the analysis run independently in two different programming languages, R and Stata are available in our reproducible repository at https://github.com/nbreznau/ideology_specification. The main analyses are run in Stata. Back up analyses and the production of visualizations were done in R.

Data

The data come from the experiment of Breznau, Rinke and Wuttke (BRW)[1]. In the BRW experiment, a 'many analysts' study, the research teams of 1-3 researchers were given data from the International Social Survey Program (ISSP). The ISSP records attitudes towards the government provision of social programs by asking (BRW's *S. Appendix*, p. 7): Do you think it should or should not be the government's responsibility to...

... provide a job for everyone who wants one
... provide health care for the sick
... provide a decent standard of living for the old
... provide a decent standard of living for the unemployed
... reduce income differences between the rich and the poor
... provide decent housing for those who can't afford it.

The answer to each item is on a 4-point scale with no midpoint, ranging from "Definitely should not be" to "Definitely should be". The Brady-Finnigan study related these responses to measures of the immigrant supply to estimate the impact of immigration on those attitudes.

The BRW experiment instructed the participating research teams to first computationally reproduce the findings of Brady and Finnigan (2014), which correlated public attitudes towards the government provision of social programs (as measured in the ISSP) with the level of immigration in 17 European countries. This preliminary reproduction was part of the experimental design to determine not only how well teams could computationally reproduce the findings, but also to bring all participating teams to a similar level of awareness of how a typical study in this area is conducted. The teams were then instructed to extend the empirical analysis in any way they determined would best reflect the data-generating process, and were provided additional waves of the ISSP and country-level data measuring immigrant shocks and other macroeconomic and social indicators.

In their efforts to replicate and extend Brady and Finnigan (2014), the research teams estimated some version of the regression model:

$$y_{rs} = \beta_{rs} m_{rs} + controls + error, \quad (1)$$

where y_{rs} is the variable measuring the ISSP respondent's stance on the government's responsibility to provide social programs used by research team *r* in regression specification *s*; and m_{rs} is the measure of the immigrant supply shock used in that specification.[1] When submitting their own ideal estimates of how they think they could best test the data generating model which would link immigration and public opinion, and thus best test the hypothesis that 'immigration reduces support for social policy', the research teams often reported estimates of β_{rs} from multiple regression models. In fact, the teams jointly estimated 1,253 alternative regressions. The median team estimated 12 models, and the 10th and 90th percentiles are 3 and 36, respectively.

BRW converted the submitted estimates of β_{rs} into a statistic that is comparable across teams and models. The "average marginal effect" (AME) gives the change in the probability that the government should be responsible for providing social programs resulting from a one-percentage-point increase in the immigrant population share. Figure *1B*, main text, shows the frequency distribution of the AMEs in the experimental data. Although the estimated AMEs cluster around zero (perhaps influenced by the fact that the researchers initially reproduced the Brady-Finnigan study, where the key result is that there is a zero correlation), many of the estimates suggest that immigration has a numerically important and significant effect on social cohesion. For example, the 10th percentile estimate is -0.071 (with a standard error of 0.019), and the 90th percentile estimate is 0.052 (0.011). In substantive terms, depending on who conducts the empirical analysis, a one-percentage-point increase in the share of the population that is foreign-born reduces or increases the probability that the public supports government provision of social programs by -7.1 or +5.2 percentage points, respectively.[2]

As noted above, the experimental data records researcher attitudes towards immigration using a 7-point scale, with a higher number indicating the researcher preferred a more relaxed immigration policy (see *Figure 1A*, main text). The publicly available data contains a measure of a *team's* pro-immigration sentiment, measured by the mean of this index across the (up to three) team members.

The summary statistics reported in *Table S1* show noticeable differences in the mean of the AME distribution (and in other relevant variables) across the three types of teams. The mean AME is slightly positive (0.014) for the pro-immigration teams, slightly negative (-0.008) for the moderate teams, and most negative (-0.019) for the anti-immigration teams. More striking differences appear if we focus on the tails of the AME distribution and in the significance of the estimates. Among the AME estimates produced by pro-immigration teams, 5.9 percent are positive and significant at the 5% level in a one-tailed test (i.e., $t > |1.645|$), and 2.8 percent are negative and significant. In

[1] The dependent variable could measure attitudes towards a particular social policy examined in the ISSP questionnaire (e.g., jobs or health care), or some combination thereof. The independent variable could measure the percent foreign born in the population or a measure based on net immigration flows.

[2] If we average across the various programs, the fraction of respondents in the 2006 wave of the ISSP who responded that the government "definitely should be" or "probably should be" responsible for the provision of government programs is 80.3 percent. The 10th and 90th percentile estimated effects of immigration on social cohesion, therefore, are sizable relative to the baseline.

contrast, among the estimates produced by the anti-immigration teams, 3.7 percent are positive and significant, and 11.9 percent are negative and significant.

Figure 1C, main text, illustrates the importance of introducing the underlying political ideology of teams for understanding some of the variation in the estimated impact of immigration. It plots density distributions of the AME for the three types of teams. The distribution has much more mass in the negative tail for the anti-immigration teams. In contrast, the AMEs estimated by pro-immigration teams have more positive values. In short, the raw data suggests that pro-immigration teams tend to adopt research strategies leading to the conclusion that immigration increases public support for social policies, while the opposite is true for anti-immigration teams.

Further Method Details

We examine the link between ideological bias and the AME by estimating regressions that relate the estimated AME to a vector of team-specific variables. The generic regression model is:

$$AME_{rs} = \alpha I_r + controls + error, \quad (2)$$

where I_r gives a measure of team *r*'s ideology towards immigration. The regressions are weighted by the inverse of the number of models estimated by the team and standard errors are clustered at the team level.

The controls in equation (2) include variables that summarize the team's pre-existing familiarity with empirical methods and immigration research. BRW collected information on each researcher's prior experience teaching or publishing research in statistical methods (and software skills). They used factor analysis to combine this information both at the researcher and team levels. Similarly, BRW collected information on prior experience in either teaching courses or publishing papers related to immigration and social policy, and on whether the researcher was familiar with the hypothesis being examined. They again used factor analysis to combine the various answers into an index of topic experience. These researcher characteristics likely shape modeling decisions. Table S1 documents the differences in these indices (measured in standardized units) across the ideologically defined teams. The statistical skills index of pro-immigration teams is about half a standard deviation higher than that of anti-immigration teams. Similarly, the topic experience index is lowest for moderate teams and highest for pro-immigration teams.

Finally, the regressions include fixed effects to control for team size and for the (sometimes mixed) disciplinary background of the team members (e.g., sociology, political science, economics, etc.). Most researchers are either sociologists (55.4 percent) or political scientists (27.4 percent). But 57 of the 71 teams have more than one researcher and 36 of those 57 combine myriad disciplines, making it difficult to construct a small vector of fixed effects that accurately reflects the team's expertise. The baseline regression specification includes fixed effects indicating the discipline of the lead author; a fixed effect indicating if two-person teams have researchers from different disciplines; fixed effects indicating if three- person teams are mainly composed of sociologists, or mainly composed of political scientists, or mainly composed of sociologists (political scientists) with a political scientist (sociologist) as the lead author, and a fixed effect indicating any other

type of three-person discipline combination. We show below that our results are robust to using alternative controls for the disciplinary composition of multi-person teams.

Tables S3-S5 present the entire set of regressions using three alternative dependent variables: the actual AME, the probability that the AME is below the 10th percentile and significant at the 5 percent level (($t > |1.645|$).; and the probability that the AME is above the 90th percentile and significant at the 5 percent level.

Up to this point, we documented the dispersion in the AME using the *team-model* as the unit of observation, and examined how this dispersion partly depends on differences in the immigration ideology of the researchers composing the various teams. We now show that our results would be similar if we instead examined the data at the *researcher-model* level (thus circumventing the need to specify either the *team's* ideology or the *team's* educational background).

Suppose a team has r researchers, and the team submitted s AME estimates. Each researcher in this team (implicitly or explicitly) participated in the calculation and submission of each of the s estimates, implying that the AME data describing this team consists of ($r \times s$) observations, one observation per researcher-model combination. By stacking these data across teams, we have created a dataset consisting of researcher-model dyads, where the unit of observation is a researcher-model pairing.

The classification of any given observation in this reformatting of the experimental data into anti-immigration or pro-immigration categories is trivial and follows directly from the response of each researcher to the immigration sentiment question (illustrated in Figure 1*A*).[3] Similarly, each researcher was asked for his/her specific field of study, and the responses can be used to easily construct discipline fixed effects at the researcher level. We estimated the regression model in equation (2) using researcher-model dyads as the unit of observation, and the coefficients are reported in Table S6.[4]

Figure S2 illustrates the AME frequency distribution in the researcher-model dyad data. It again suggests that the tails of the AME distribution are particularly sensitive to the immigration ideology of the team and the researchers. Table S5 estimates the regression models using the dyad data, again showing a significant relationship between ideology and the AME.

Table S6 shows that the results (at the team-model level) are also robust to using alternative controls for the disciplinary background of the team. The table reports the key regression coefficients using the various alternative measures of the team's ideology and two alternative sets of controls for the team's disciplinary background: (1) fixed effects simply indicating the discipline

[3] A researcher is classified as anti-immigration if he responds to the immigration attitude question with a "1" or a "2" and is classified as pro-immigration if he responds with a "5" or "6"; all other researchers are grouped into the moderate classification.

[4] The statistics skill and topic experience variables in the regressions reported in Table 4 are the individual-specific factor indices created in BRW (2022). The standard errors in the researcher-model dyad specifications, though clustered at the team-researcher-model level, would be identical if they were instead simply clustered at the team level.

of the lead researcher in the team (i.e., the researcher that corresponded with the principal investigators); and (2) a vector of 28 fixed effects that capture every possible combination of disciplines among the researchers in a team.

Variables

Pro-immigration sentiment & mean sentiment. Each researcher is asked whether immigration laws should be tightened or relaxed (using a 7-point scale) in the first wave questionnaire and the responses are recorded in *attitude_immigration_11*, *attitude_immigration_12*, and *attitude_immigration_13*, for the (up to three) researchers in each team. None of the researchers responded with the strongest anti-immigration sentiment (the "0" in *Figure 1A*), so that the publicly available data employs a 6-point scale for all these variables. The analysis also uses the team's mean immigration sentiment, *pro_immigrant*, which is a simple average of the index across the responses. The direction of the scale of the *attitude_immigrant_1j* variables is the reverse of the direction of the scale of the *pro_immigrant* variable. We standardized the responses so that a higher number always indicates a stronger pro-immigration sentiment. We also use a mean by team measure to test the robustness of our results.

Anti-, moderate or pro-immigration team. To allow for the possibility that some team members feel strongly about immigration (one way or the other) and to demonstrate the robustness of our results, we construct alternative measures of a team's ideology. In particular, we summarize the team's ideology in terms of two variables: the fraction of the team that is anti-immigration (a "1" or "2" in the distribution) and the fraction of the team that is pro-immigration (a "5" or "6" in the distribution), with the omitted variable indicating the fraction with moderate sentiments.

It is also convenient, particularly in terms of visualizing the impact of ideology, to classify teams into distinct categories. We define a pro-immigration team as a team where more than half the team has scores of "5" or "6" in the sentiment question. This definition classifies 31 of the 71 teams (or 43.7 percent) as pro-immigration. It is also sensible to separate out the remaining 40 teams into teams that have strong anti-immigration sentiments or are more moderate. Given the rarity of anti-immigration sentiments among the participating researchers, a simple approach is to classify a team that has at least one member responding with a "1" or a "2" to the sentiment question as anti-immigration. This definition classifies 9 teams (or 12.7 percent) as anti-immigration. The remaining 31 teams (or 43.7 percent) are then classified as "moderate."[5]

Field of highest degree. This question is asked in the first wave, and the variables for the team are: *backgr_degree1* (the discipline of the lead or corresponding author*)*, *backgr_degree2*, and *backgr_degree3*. The baseline regressions include fixed effects that indicate the discipline of the lead author (i.e., communications, economics, sociology, political science, psychology, and "other"); a fixed effect indicating if two- person teams have researchers from different disciplines; and fixed effects indicating if three- person teams are mainly composed of sociologists, or mainly

[5] In two three-person teams, one of the researchers responded with a "2" and the other two responded with a "5" or a "6". Because at least half of the team is strongly pro-immigration, those two teams are classified as pro-immigration teams. The regression results reported in Table 2 are almost identical if the 40 models estimated by those two "marginal" teams are excluded from the regressions.

composed of political scientists, or mainly composed of sociologists (political scientists) with a political scientist (sociologist) as the lead author, and a fixed effect indicating any other type of three-person discipline combination.

Topic knowledge. This variable measures the team's familiarity with research in immigration or social policy and is produced by a factor analysis of several questions exploring this background. The variable giving the factor index is *topic_ipred*. The topic knowledge information is missing for one team. The missing value was imputed using a hot-deck procedure based on the team size, the field of highest degree, and the gender composition of the team.

Statistical skill. This variable measures the team's familiarity with statistical methods and data analysis and is produced by a factor analysis of several questions documenting the background. The variable giving the factor index is *statistics_ipred*. This variable is missing for one team. The missing value was again imputed using the same hot deck procedure as the topic knowledge variable.

Team size. We discovered an error in the variable *team_size* constructed by BRW. The variable is incorrectly coded for team 93 (coded as 2 but is actually 3) and 94 (coded as 2 but is actually 1). After looking at the source files, this error was confirmed by Nate Breznau and therefore we recoded the variable for these two teams.

Referee score. The analysis uses the variable *peer_mean*, an enhanced version of the *total_score* variable in the repository. The enhanced version was added to the public data from the original study in the BRW Harvard Dataverse repository in November of 2024 as an improvement to the original data (Breznau, Rinke, and Wuttke, 2022). The teams pre-registered 79 different model specifications, which were then classified in terms of the sample used, the construction of the dependent variable, the definition of the immigrant supply shock, and many other details. The description of each model was then submitted to 4 or 5 reviewers and "refereed" in a double-blind setting, with each referee ranking the specification using a 1-7 scale. The original variable resulted from each of the 1,253 estimated models being compared to the pre-registered specifications and assigned the average referee score that matched at least 95 percent of the specification details. The enhanced version takes into account the full model specifications actually run by each team as many teams did not go into enough detail in their pre-registrations to cover all peer reviewed specifications. The *peer_mean* variable was missing for 30 of the models. In the regressions that use the referee score as a weight, we imputed those missing values using a regression of the enhanced measure of the referee score variable on the original measure and on a vector of variables describing specific details of the model specification (e.g., the definition of the dependent variable, the definition of the immigrant supply shock, the waves of the ISSP used, etc.).

Belief in hypothesis. In addition to the immigration sentiment question, BRW collected information on researchers' priors about whether "higher levels of [immigration]…reduces public support of social welfare policies" (BRW, 2022, *SI Appendix*, p. 84). The researchers could express their priors using a 5-point scale, ranging from "strongly disagree" to "strongly agree".

The question is asked in the first wave questionnaire for each researcher in the experiment. It asks whether the researcher believes the hypothesis that immigration reduces support for social

programs. The belief response was measured on a 1-4 scale. The information is reported in the variables *belief_H1_11*, *belief_H1_12*, and *belief_H1_1*3 for the up-to-three researchers in the team.

Figure S1 illustrates the distribution of the responses. Very few researchers either strongly agreed or disagreed with the hypothesis. Instead, 55.4 percent responded with a "4", indicating that they believed immigration "somewhat reduces" support, and 36.9 percent responded with a "3", indicating that they believed immigration has no effect on support. It is worth noting that the correlation between the anti- or pro-immigration sentiments and the hypothesis prior is near zero (i.e., the Pearson correlation between belief and ideology is -0.08).

As a prior expectation might produce confirmation bias, we added a variable measuring the team's prior belief in the hypothesis in our robustness regressions. Specifically, we used the first wave of the questionnaire to calculate the fraction of the team that moderately or strongly agrees with the claim that immigration reduces political support for social programs (i.e., the fraction of the team that answered with a "4" or a "5" in Figure 1*B*). The mean of this variable across teams is 0.58, with the fraction being lowest for the moderate teams. Adding the "hypothesis prior" variable to the regression models does not change the results linking immigration sentiments and the estimated impact of immigration. Moreover, the hypothesis prior variable itself does not have a significant effect on the AME. See columns 5-6 in Tables S2, S3, S4.

***Expected AME*:** The specification decisions used to calculate the expected AME are: using a dependent variable that aggregates attitudes towards government provision of specific programs (variable: *scale*); measures of the immigrant shock (variables: *shock* and *flow*); adjustments for units nested in country-year levels (variable: *level_cyear*); using all countries available in the ISSP data (variable: *allcountries*); and the ISSP waves used (variables: *w1996*, *w2006*, and *w2016*). All these variables are binary indicators.

***Disciplinary background*:** We have shown that our results are robust when we use alternative definitions of the ideological composition of the team. We now show that they are equally robust to using alternative controls for the disciplinary background of the team. *Table S6*[6] reports the key regression coefficients using the various alternative measures of the team's ideology and two alternative sets of controls for the team's disciplinary background: (1) fixed effects simply indicating the discipline of the lead researcher in the team (i.e., the researcher that corresponded with the principal investigators); and (2) a vector of 28 fixed effects that capture every possible combination of disciplines among the researchers in a team.

The key lesson from Table 3 is that the results are robust regardless of how the team's ideology is defined or which set of controls is used for the team's disciplinary background. For example, the baseline difference of 0.085 (0.031) between pro- and anti-immigration teams declines slightly to 0.080 (0.032) if we use a set of fixed effects that allows for every possible combination of disciplines in multi-person teams.

[6] The key lesson from Table 3 is that the results are robust regardless of how the team's ideology is defined or which set of controls is used for the team's disciplinary background. For example, the baseline difference of 0.085 (0.031) between pro- and anti-immigration teams declines slightly to 0.080 (0.032) if we use a set of fixed effects that allows for every possible combination of disciplines in multi-person teams.

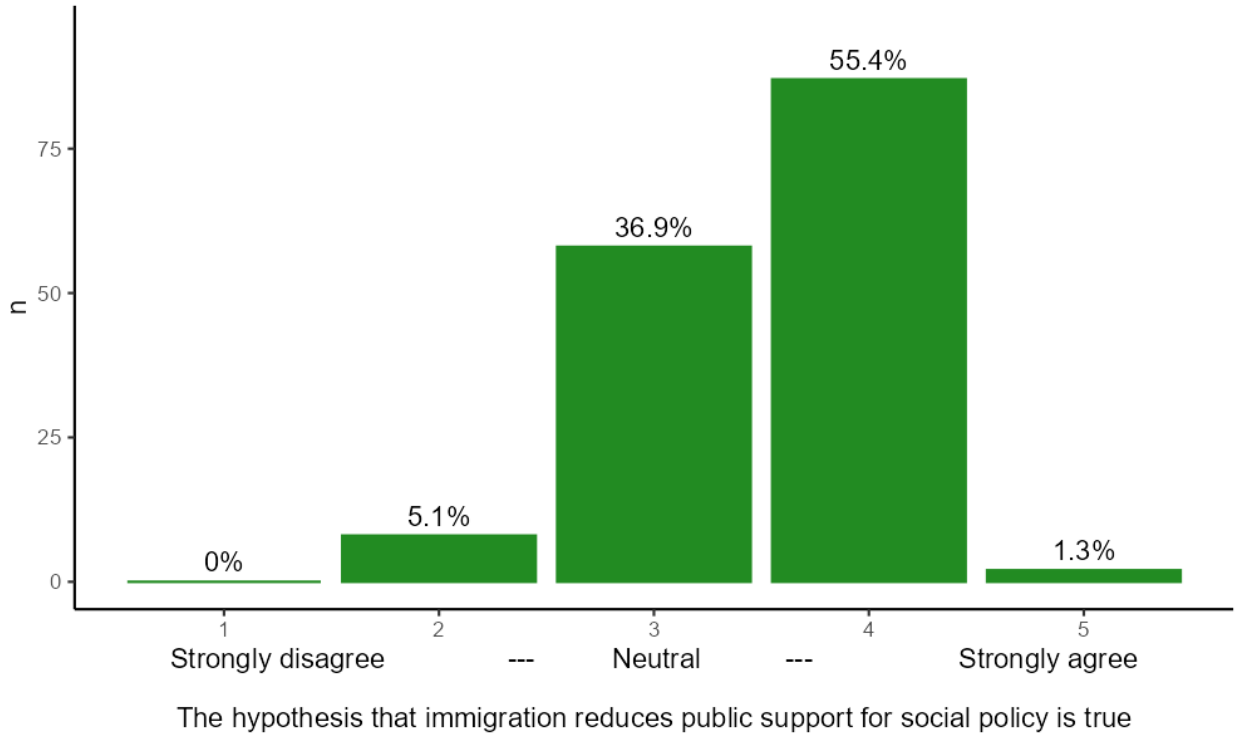

Fig. S1. Belief in Hypothesis

Raw survey results from the original experiment.

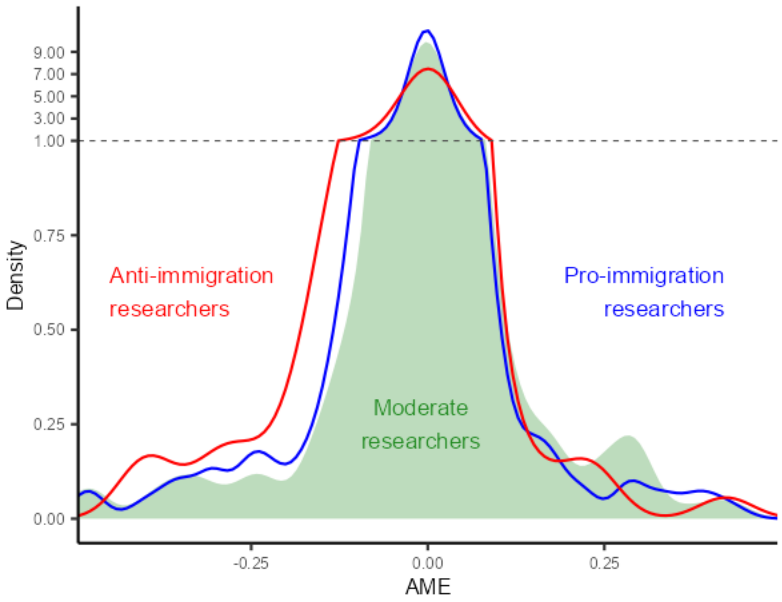

Fig. S2. Density of AME by Group, Researcher-Model Dyads

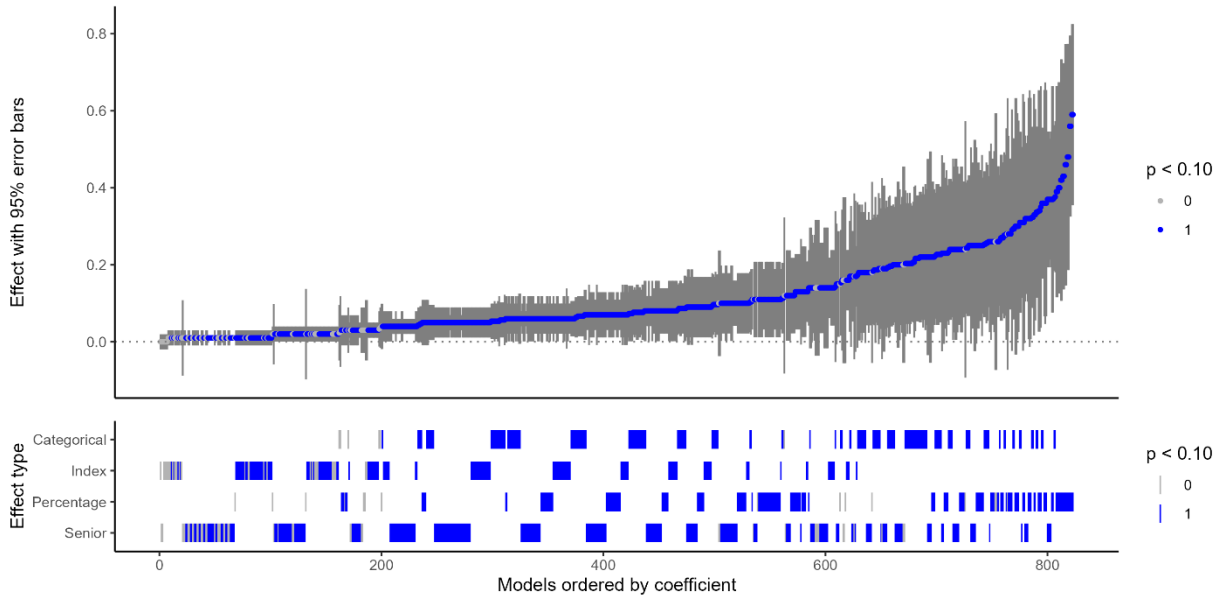

Fig. S3. Specification Curve of All Tables and Their Derivative Model Combinations

Notes: The effects are estimated in regressions using four alternative dependent variables: the AME, the likelihood of 10th and 90th (negative and positive) significant results, and the positive/negative directional results. They introduce independent variables stepwise, starting with none and then adding discipline, statistical skills and topic experience, team size, and hypothesis belief. We estimated all potential model specification for each of our four different measures of ideology (mean sentiment ("Index"), single category comparing results between pro-immigration teams and all other teams ("Categorical"), comparing results between the two categories of anti- and pro-immigration teams ("Categorical"), and ideology of the senior team member ("Senior"), and for each of our three definitions of disciplinary fixed effects. We also ran all models with weights given by the number of models per team, and then additionally weighted by the peer review scores. We used a similar procedure for the dyadic data with AME as the outcome (without the alternative discipline variables because they only make sense at the team level). See 03_Robustness.do in our Online Repository. Points represent the absolute values of coefficients or linear combination test coefficients for models with categorical groups. Effect type refers to the measurement method of ideology. Blue indicates an effect from that model is statistically significant at $p < 0.10$.

Variable	Sample			
	All teams	Anti-imm.	Moderate	Pro-imm.
AME	0.001	-0.019	-0.008	0.014
% AME < 10^{th} percentile and significant	0.063	0.119	0.086	0.028
% AME > 90^{th} percentile and significant	0.053	0.037	0.050	0.059
Mean immigration sentiment	4.462	2.423	3.992	5.383
% team members that are anti-immigration	0.078	0.632	0.000	0.023
% team members that are pro-immigration	0.541	0.104	0.225	0.940
% Believe imm. reduces social cohesion	0.581	0.726	0.469	0.653
% Statistical skills index (z)	0.000	-0.375	-0.176	0.254
% Topic experience index (z)	0.000	-0.074	-0.094	0.107
% Peer score (z)	0.000	0.034	0.345	-0.328
% High quality research design	0.227	0.164	0.313	0.162
Number of models	39.145	18.835	43.603	39.661
Team size	2.246	2.194	2.426	2.087
Number of models	1253	134	544	575
Number of teams	71	9	31	31

Table S1. Descriptive Statistics

Notes. In rows 2 and 3, the estimated AME is statistically significant if $|t| > 1.645$. All summary statistics are calculated at the model level. An anti-immigrant team consists of a team that has at least one team member who is anti-immigration (a "1" or "2" in the immigrant sentiment scale). A pro-immigration team consists of teams where more than 50 percent of the members are considered pro-immigrant (a "5" or "6" in the immigrant sentiment scale). All other teams are classified as moderate teams.

	Specification					
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
A. Basic regressions						
Mean immigration index	0.011*	---	---	---	---	---
	(0.006)					
% of team that is anti-imm.	---	-0.050**	---	---	-0.048*	---
		(0.025)			(0.025)	
% of team that is pro-imm.	---	0.023	---	---	0.025	---
		(0.019)			(0.019)	
Anti-immigration team	---	---	---	-0.057*	---	-0.055*
				(0.031)		(0.031)
Pro-immigration team	---	---	0.040**	0.027*	---	0.028*
			(0.016)	(0.015)		(0.015)
Hypothesis prior	---	---	---	---	-0.012	-0.011
					(0.016)	(0.016)
Statistical skills (*z*)	0.038**	0.037**	0.037**	0.034**	0.037**	0.034**
	(0.018)	(0.017)	(0.016)	(0.016)	(0.016)	(0.015)
Topic experience (*z*)	-0.041**	-0.042**	-0.040**	-0.041**	-0.042**	-0.041**
	(0.017)	(0.017)	(0.015)	(0.016)	(0.016)	(0.015)
D: Pro - Anti	---	0.074**	---	0.085**	0.073**	0.084**
		(0.024)		(0.031)	(0.024)	(0.031)
R-squared	0.062	0.065	0.066	0.071	0.065	0.071
B. Also weighted by referee score						
Mean immigration index	0.011**	---	---	---	---	---
	(0.006)					
% of team that is anti-imm.	---	-0.052*	---	---	-0.049*	---
		(0.026)			(0.025)	
% of team that is pro-imm.	---	0.026	---	---	0.028	---
		(0.019)			(0.019)	
Anti-immigration team	---	---	---	-0.061*	---	-0.059*
				(0.034)		(0.034)
Pro-immigration team	---	---	0.043**	0.029*	---	0.030**
			(0.016)	(0.015)		(0.015)
Hypothesis prior	---	---	---	---	-0.014	-0.012
					(0.016)	(0.016)
Difference: Pro - Anti	---	0.078**	---	0.090**	0.077**	0.089**
		(0.026)		(0.034)	(0.026)	(0.034)
R-squared	0.063	0.067	0.068	0.074	0.067	0.074

Table S2. Full Regression Results Predicting AME

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses and are clustered at the team level. The regressions in Panel A are weighted by the inverse of the number of models, and the regressions in Panel B are also weighted by the referee score awarded to the model. The team-model level regressions have 1,253 observations, and the final two with belief have 1,178 due to missing values. All regressions include the statistical skills factor index, the topic experience factor index, fixed effects for team size, and fixed effects indicating the team's or the researcher's field of highest degree.

	Specification					
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
A. Basic regressions						
Mean immigration index	-0.051**	---	---	---	---	---
	(0.021)					
% of team that is anti-imm.	---	0.081	---	---	0.097	---
		(0.103)			(0.095)	
% of team that is pro-imm.	---	-0.161**	---	---	-0.151**	---
		(0.064)			(0.065)	
Anti-immigration team	---	---	---	0.150	---	0.166
				(0.123)		(0.118)
Pro-immigration team	---	---	-0.157**	-0.124**	---	-0.115**
			(0.051)	(0.044)		(0.045)
Hypothesis prior	---	---	---	---	-0.082*	-0.090*
					(0.046)	(0.050)
Statistical skills (*z*)	-0.071**	-0.066**	-0.065**	-0.059**	-0.068**	-0.060**
	(0.033)	(0.017)	(0.030)	(0.029)	(0.030)	(0.027)
Topic experience (*z*)	0.070**	0.067**	0.067**	0.069**	0.066**	0.068**
	(0.027)	(0.024)	(0.025)	(0.027)	(0.024)	(0.025)
Difference: Pro - Anti	---	0.242*	---	0.274*	0.249*	0.280*
		(0.107)		(0.125)	(0.097)	(0.119)
R-squared	0.096	0.115	0.118	0.133	0.123	0.123
B. Also weighted by referee score						
Mean immigration index	-0.047*	---	---	---	---	---
	(0.021)					
% of team that is anti-imm.	---	0.052	---	---	0.087	---
		(0.102)			(0.091)	
% of team that is pro-imm.	---	-0.136*	---	---	-0.124*	---
		(0.054)			(0.053)	
Anti-immigration team	---	---	---	0.114	---	0.140
				(0.095)		(0.089)
Pro-immigration team	---	---	-0.117**	-0.089*	---	-0.080*
			(0.042)	(0.038)		(0.039)
Hypothesis prior	---	---	---	---	-0.092*	-0.096*
					(0.040)	(0.043)
Difference: Pro - Anti	---	0.187*	---	0.204*	0.211**	0.219*
		(0.101)		(0.096)	(0.086)	(0.088)
R-squared	0.099	0.111	0.108	0.117	0.123	0.130

Table S3. Full Regression Results Predicting Extreme Negative and Significant AME

Notes: * $p < .1$; ** $p < .05$. Linear probability models. Standard errors in parentheses and clustered at the team level. The binary dependent variable is set to unity if the AME estimate is below the 10th percentile and has a *t*-value greater than |1.645|. Panel A are weighted by the inverse of the number of models per team. Panel B are also weighted by the mean peer referee score awarded to each model. The team-model level regressions have 1,253 observations, and the final two with belief have 1,178 due to missing values. All regressions include the statistical skills factor index, the topic experience factor index, fixed effects for team size, and fixed effects indicating the team's or the researcher's field of highest degree.

	Specification					
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
A. Basic regressions						
Mean immigration index	0.019*	---	---	---	---	---
	(0.012)					
% of team that is anti-imm.	---	-0.130*	---	---	-0.127*	---
		(0.057)			(0.057)	
% of team that is pro-imm.	---	-0.016	---	---	-0.014	---
		(0.041)			(0.040)	
Anti-immigration team	---	---	-0.079**	-0.070*	---	-0.066*
			(0.031)	(0.037)		(0.039)
Pro-immigration team	---	---	---	0.017	---	0.019
				(0.034)		(0.034)
Hypothesis prior	---	---	---	---	-0.016	-0.019
					(0.030)	(0.029)
Statistical skills (*z*)	0.020	0.021	0.020	0.019	0.021	0.018
	(0.017)	(0.016)	(0.017)	(0.017)	(0.016)	(0.017)
Topic experience (*z*)	-0.021	-0.029*	-0.024*	-0.023	-0.029*	-0.023
	(0.013)	(0.015)	(0.014)	(0.038)	(0.015)	(0.015)
D: Pro - Anti	---	0.114*	---	0.087*	0.113*	0.085*
		(0.045)		(0.034)	(0.047)	(0.031)
R-squared	0.087	0.090	0.088	0.089	0.091	0.090
B. Also weighted by referee score						
Mean immigration index	0.035*	---	---	---	---	---
	(0.016)					
% of team that is anti-imm.	---	-0.115	---	---	-0.117	---
		(0.074)			(0.071)	
% of team that is pro-imm.	---	0.008	---	---	0.008	---
		(0.062)			(0.060)	
Anti-immigration team	---	---	-0.111*	-0.126*	---	-0.128**
			(0.050)	(0.063)		(0.062)
Pro-immigration team	---	---	---	-0.020	---	-0.020
				(0.063)		(0.062)
Hypothesis prior	---	---	---	---	0.004	0.006
					(0.051)	(0.050)
Difference: Pro - Anti	---	0.124*	---	0.105*	0.125*	0.107*
		(0.063)		(0.054)	(0.063)	(0.054)
R-squared	0.119	0.116	0.117	0.118	0.116	0.117

Table S4. Full Regression Results Predicting Extreme Positive and Significant AME

Notes: * $p < .1$; ** $p < .05$. Linear probability models. Standard errors in parentheses and clustered at the team level. The binary dependent variable is set to unity if the AME estimate is above the 90th percentile and has a *t*-value greater than |1.645|. Panel A are weighted by the inverse of the number of models per team. Panel B are also weighted by the mean peer referee score awarded to each model. The team-model level regressions have 1,253 observations, and the final two with belief have 1,178 due to missing values. All regressions include the statistical skills factor index, the topic experience factor index, fixed effects for team size, and fixed effects indicating the team's or the researcher's field of highest degree.

	Basic regressions				Also weighted by peer score	
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
Immigration index	0.086**	---	---	---	---	---
	(0.039)					
Anti-immigration	---	---	-0.208*	-0.204*	-0.215**	-0.211**
			(0.110)	(0.110)	(0.104)	(0.105)
Pro-immigration	---	0.153*	0.113	0.106	0.130	0.125
		(0.084)	(0.090)	(0.091)	(0.088)	(0.090)
Hypothesis prior	---	---	---	0.050	---	0.039
				(0.097)		(0.100)
Statistical skills (*z*)	0.071*	0.076**	0.055	0.062	0.046	0.051
	(0.035)	(0.037)	(0.035)	(0.039)	(0.034)	(0.038)
Topic experience (*z*)	-0.081*	-0.090*	-0.087*	-0.089*	-0.084*	-0.086**
	(0.046)	(0.050)	(0.048)	(0.049)	(0.047)	(0.048)
D: Pro - Anti	---	---	0.321**	0.309**	0.345**	0.336**
			(0.108)	(0.111)	(0.103)	(0.107)
R-squared	0.132	0.127	0.135	0.136	0.138	0.139

Table S5. Regressions Predicting Positive or Negative Direction of AME

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses are clustered at the researcher-team-model level. The regressions are weighted by the inverse of the product of the number of models and researchers in the team; the regressions in columns 5-6 are also weighted by the mean referee score awarded to the model. Dependent variable = 1 if AME >= 0.015 and = 0 if AME <= 0.015, with values in between considered roughly 'zero' and coded to missing. All regressions have 601 observations with 66 remaining team clusters, and include fixed effects for team size and for the researcher's field of highest degree.

	Weighted by inverse number of models		Also weighted by peer score	
Alternative definitions of ideology:	Lead author discipline	All discipline combinations	Lead author discipline	All discipline combinations
1. Mean immigration index	0.010*	0.012*	0.010*	0.013*
	(0.006)	(0.007)	(0.005)	(0.007)
2. Team composition variables:				
% of team that is anti-immigration	-0.038	-0.019	-0.037	-0.017
	(0.025)	(0.030)	(0.026)	(0.030)
% of team that is pro-immigration	0.025	0.044*	0.029	0.049*
	(0.020)	(0.024)	(0.020)	(0.025)
D: Pro – Anti	0.063**	0.063**	0.065**	0.067**
	(0.025)	(0.025)	(0.026)	(0.026)
3. Pro-imm. team relative to all others:				
Pro-immigration team	0.035**	0.048**	0.039**	0.052**
	(0.017)	(0.017)	(0.018)	(0.017)
4. Baseline definition of teams:				
Anti-immigration team	-0.043	-0.042	-0.046	-0.047
	(0.028)	(0.033)	(0.030)	(0.036)
Pro-immigration team	0.024	0.038**	0.027	0.041**
	(0.018)	(0.016)	(0.018)	(0.016)
D: Pro – Anti	0.068**	0.080**	0.073**	0.088**
	(0.028)	(0.032)	(0.030)	(0.035)
5. Without controls:				
a. Mean-immigration index	0.016**	0.013**	0.017**	0.013*
	(0.007)	(0.006)	(0.007)	(0.007)
b. D: Pro – Anti (% of team)	0.084**	0.068**	0.089**	0.072**
	(0.033)	(0.032)	(0.036)	(0.035)
c. Pro-immigration team	0.048**	0.040*	0.050**	0.082**
	(0.020)	(0.020)	(0.014)	(0.031)
d. D: Pro – Anti (team categories)	0.099**	0.078**	0.105**	0.043**
	(0.037)	(0.034)	(0.040)	(0.021)

Table S6. Robustness of Results to Alternative Coding of Researcher Disciplines

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses and are clustered at the team level. The regressions in Panel A are weighted by the inverse of the number of models; the regressions in Panel B are also weighted by the mean referee score awarded to the model. All regressions have 1,253 observations and include the statistical skills factor index, the topic experience factor index, fixed effects for team size, and fixed effects indicating the team's or researcher's field of highest degree except for '5. Without controls' which are without the statistical skills, topic experience and team-size variables.

Model	ame (1)	ame (2)	ame (3)	neg10s (1)	neg10s (2)	neg10s (3)	pos10s (1)	pos10s (2)	pos10s (3)
0. No controls									
	0.00	0.03	0.05	-0.04*	-0.16	-0.19*	0.00	0.01	0.03
	(0.01)	(0.03)	(0.03)	(0.02)	(0.11)	(0.11)	(0.01)	(0.07)	(0.04)
Baseline degree discipline vector									
1. Degree only	0.02**	0.08**	0.10***	-0.06**	-0.26*	-0.30**	0.02*	0.12**	0.10***
	(0.01)	(0.03)	(0.04)	(0.03)	(0.13)	(0.14)	(0.01)	(0.05)	(0.04)
2. Model 1 +	0.01*	0.07***	0.09***	-0.05**	-0.24**	-0.27**	0.02*	0.11**	0.09**
skills and experience	(0.01)	(0.02)	(0.03)	(0.02)	(0.10)	(0.12)	(0.01)	(0.05)	(0.04)
3. Model 2 +	0.01*	0.07***	0.08***	-0.05**	-0.24**	-0.27**	0.02*	0.11**	0.09**
team-size	(0.01)	(0.02)	(0.03)	(0.02)	(0.11)	(0.13)	(0.01)	(0.05)	(0.03)
Degree discipline defined by lead researcher in team									
1. Degree only	0.01**	0.07**	0.08**	-0.05**	-0.23*	-0.24**	0.02	0.11**	0.07**
	(0.01)	(0.03)	(0.03)	(0.02)	(0.12)	(0.12)	(0.01)	(0.04)	(0.03)
2. Model 1 +	0.01*	0.06**	0.07**	-0.05**	-0.22**	-0.23**	0.02	0.10***	0.06*
skills and experience	(0.01)	(0.03)	(0.03)	(0.02)	(0.10)	(0.11)	(0.01)	(0.04)	(0.03)
3. Model 2 +	0.01*	0.06**	0.07**	-0.05**	-0.22**	-0.23**	0.01	0.10**	0.06*
team-size	(0.01)	(0.02)	(0.03)	(0.02)	(0.10)	(0.11)	(0.01)	(0.04)	(0.03)
Degree discipline defined by all possible combinations									
1. Degree only	0.02**	0.09**	0.11**	-0.06**	-0.30*	-0.35**	0.03**	0.13**	0.11**
	(0.01)	(0.04)	(0.05)	(0.03)	(0.15)	(0.16)	(0.01)	(0.06)	(0.04)
2. Model 1 +	0.01*	0.06**	0.08**	-0.06**	-0.25**	-0.30**	0.03**	0.11**	0.08**
skills and experience	(0.01)	(0.02)	(0.03)	(0.02)	(0.11)	(0.14)	(0.01)	(0.05)	(0.03)
3. Model 2 +	0.01*	0.06**	0.08**	-0.06**	-0.25**	-0.30**	0.03**	0.11**	0.08**
team-size	(0.01)	(0.02)	(0.03)	(0.02)	(0.11)	(0.14)	(0.01)	(0.05)	(0.03)

Table S7. Robustness of Results to Fewer Controls Across Alternative Ideology Measurements

Notes: * $p < .1$; ** $p < .05$; *** $p < 0.01$. Standard errors reported in parentheses are clustered at the researcher-team-model level. "ame" are models predicting the AME; "neg10s" are models predicting results that are negative and statistically significant at the $p < 0.05$ level; and "pos10s" are results that are positive and statistically significant. Models (1) measure ideology as the mean, models (2) measure the difference in coefficients between percentage of the team that are anti- and percentage pro-immigration; and model (3) measure the difference in coefficients between categorical definitions of teams as anti- or pro-immigration. The regressions are weighted by the inverse of the product of the number of models and researchers in the team. All regressions have 1,253 observations.

Model	ame (1)	ame (2)	neg10s (1)	neg10s (2)	pos10s (1)	pos10s (2)
0. No controls						
	0.02	0.03	-0.15**	-0.12***	-0.02	-0.00
	(0.02)	(0.02)	(0.06)	(0.04)	(0.04)	(0.03)
Baseline degree discipline vector						
1. Degree only	0.05**	0.05**	-0.20***	-0.16***	0.02	0.03
	(0.02)	(0.02)	(0.07)	(0.05)	(0.03)	(0.03)
2. Model 1 + skills and experience	0.03*	0.04**	-0.18***	-0.14***	0.01	0.03
	(0.02)	(0.02)	(0.07)	(0.05)	(0.03)	(0.03)
3. Model 2 + team-size	0.03*	0.04**	-0.18***	-0.16***	0.01	0.03
	(0.02)	(0.02)	(0.06)	(0.05)	(0.03)	(0.03)
Degree discipline defined by lead researcher in team						
1. Degree only	0.04*	0.04**	-0.19***	-0.15***	0.01	0.02
	(0.02)	(0.02)	(0.07)	(0.05)	(0.04)	(0.03)
2. Model 1 + skills and experience	0.03*	0.03**	-0.18***	-0.14***	0.01	0.01
	(0.02)	(0.02)	(0.06)	(0.04)	(0.03)	(0.03)
3. Model 2 + team-size	0.03*	0.04**	-0.18***	-0.15***	0.01	0.01
	(0.02)	(0.02)	(0.06)	(0.05)	(0.04)	(0.03)
Degree discipline defined by all possible combinations						
1. Degree only	0.06**	0.06***	-0.23**	-0.19***	0.04	0.05
	(0.02)	(0.02)	(0.09)	(0.07)	(0.03)	(0.03)
2. Model 1 + skills and experience	0.05**	0.05***	-0.21***	-0.17***	0.03	0.04
	(0.02)	(0.02)	(0.08)	(0.06)	(0.03)	(0.03)
3. Model 2 + team-size	0.05**	0.05***	-0.21***	-0.17***	0.03	0.04
	(0.02)	(0.02)	(0.08)	(0.06)	(0.03)	(0.03)

Table S8. Robustness of Results: Comparing Pro-Immigration Teams with All Other Teams

Notes: * $p < .1$; ** $p < .05$; *** $p < 0.01$. Standard errors reported in parentheses are clustered at the researcher-team-model level. "ame" are models predicting the AME; "neg10s" are models predicting results that are negative and statistically significant at the $p < 0.05$ level; and "pos10s" are results that are positive and statistically significant. Models (1) measure percentage of the team that is pro-immigrant as the coefficient (2) measure pro-immigrant teams compared to everyone else. The regressions are weighted by the inverse of the product of the number of models and researchers in the team. All regressions have 1,253 observations.

	Basic regressions					Also weighted by peer score	
Variable:	(1)	(2)	(3)	(4)		(5)	(6)
Senior imm. index	0.013**	---	---	---		---	---
	(0.004)						
Senior anti-imm.	---	---	-0.023	-0.022		-0.022	-0.020
			(0.021)	(0.021)		(0.023)	(0.023)
Senior pro-imm.	---	0.042**	0.038**	0.040**		0.039**	0.040**
		(0.015)	(0.018)	(0.015)		(0.018)	(0.018)
Hypothesis prior	---	---	---	-0.016		---	-0.018
				(0.016)			(0.016)
Statistical skills (*z*)	0.038**	0.037**	0.036**	0.036**		0.037**	0.037**
	(0.018)	(0.016)	(0.016)	(0.016)		(0.017)	(0.017)
Topic experience (*z*)	-0.040**	-0.038**	-0.039**	-0.039**		-0.039**	-0.040**
	(0.016)	(0.015)	(0.015)	(0.015)		(0.015)	(0.015)
D: Pro - Anti	---	---	0.061***	0.061***		0.060***	0.061***
			(0.015)	(0.015)		(0.015)	(0.015)
R-squared	0.064	0.067	0.067	0.068		0.069	0.069

Table S9. Regressions Predicting AME Using Senior Team Member Ideology

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses are clustered at the team level. The regressions are weighted by the inverse of the product of the number of models and researchers in the team; the regressions in columns 5-6 are also weighted by the mean referee score awarded to the model. All regressions have 1,253 observations and include fixed effects for team size and for the researcher's field of highest degree.

	Basic regressions				Also weighted by peer score	
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
Immigration index	0.008**	---	---	---	---	---
	(0.004)					
Anti-immigration	---	---	-0.029*	-0.028**	-0.029**	-0.028*
			(0.014)	(0.014)	(0.014)	(0.014)
Pro-immigration	---	0.026**	0.021	0.021	0.023*	0.023*
		(0.013)	(0.013)	(0.013)	(0.014)	(0.014)
Hypothesis prior	---	---	---	-0.002	---	-0.002
				(0.010)		(0.010)
Statistical skills (*z*)	0.025**	0.026**	0.025**	0.025**	0.025**	0.025**
	(0.013)	(0.012)	(0.012)	(0.012)	(0.012)	(0.012)
Topic experience (*z*)	-0.022**	-0.022**	-0.022**	-0.022**	-0.021**	-0.022**
	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)	(0.009)
D: Pro - Anti	---	---	0.050**	0.049**	0.052**	0.052**
			(0.017)	(0.017)	(0.018)	(0.018)
R-squared	0.021	0.023	0.024	0.024	0.024	0.024

Table S10. Researcher-Model Dyad Regressions Predicting AME

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses are clustered at the researcher-team-model level. The regressions are weighted by the inverse of the product of the number of models and researchers in the team; the regressions in columns 5-6 are also weighted by the mean referee score awarded to the model. All regressions have 2,680 observations and include fixed effects for team size and for the researcher's field of highest degree.

		Team ideology		
Design decision:	All teams	Anti-immigration	Moderate	Pro-immigration
Composite dependent variable	0.156	0.104	0.169	0.155
Stock immigrant measure	0.496	0.582	0.471	0.499
Flow immigrant measure	0.470	0.410	0.513	0.443
Country-year level adjustment	0.134	0.134	0.129	0.188
All available countries	0.341	0.239	0.467	0.245
1996 wave	0.764	0.910	0.640	0.847
2006 wave	0.946	1.000	0.956	0.923
2016 wave	0.639	0.731	0.656	0.602
Number of models	1,253	134	544	575

Table S11. Specification choices made by different types of teams

Notes: The statistics give the fraction of the estimated models that employ the particular research design decision. A "composite" dependent indicates that the research team somehow aggregated the separate responses to whether the government should be responsible for specific types of programs (e.g., jobs, housing, health).

	Specification					
Variable:	(1)	(2)	(3)	(4)	(5)	(6)
Mean immigration sentiment	0.007*	---	---	---	---	---
	(0.004)					
% of team that is anti-immigration	---	-0.026	---	---	-0.027	---
		(0.019)			(0.020)	
% of team that is pro-immigration	---	0.014	---	---	0.017	---
		(0.014)			(0.015)	
Anti-immigration team	---	---		-0.044*	---	-0.049*
				(0.024)		(0.027)
Pro-immigration team	---	---	0.024**	0.014*	---	0.014*
			(0.010)	(0.008)		(0.008)
Statistical skills (*z*)	0.021	0.020	0.020	0.018	0.022	0.019
	(0.014)	(0.013)	(0.013)	(0.012)	(0.013)	(0.012)
Topic experience (*z*)	-0.018	-0.018*	-0.017*	-0.018*	-0.019*	-0.018*
	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)
D: Pro - Anti	---	0.040*	---	0.058**	0.044*	0.063**
		(0.020)		(0.025)	(0.023)	(0.028)
R-squared	0.104	0.110	0.116	0.140	0.119	0.156
Weighted by:						
Inverse number of models	Yes	Yes	Yes	Yes	Yes	Yes
Peer score	No	No	No	No	Yes	Yes

Table S12. Determinants of expected AME

Notes: * $p < .1$; ** $p < .05$. Standard errors reported in parentheses and are clustered at the team level. The dependent variable in the regressions is the expected AME implied by the research design decisions that characterize the model in terms of the definition of the dependent variable, the stock/flow immigrant measures, the inclusion of country-year fixed effects, the use of all available countries, and the addition of the 2016 panel of the ISSP. All regressions have 1,253 observations, include the statistical skills factor index, the topic experience factor index, fixed effects indicating the team's size, and a vector of fixed effects indicating the team's field of highest degree.

	Dependent variable			
Variable:	Mean referee score (z)		High quality indicator	
% of team that is anti-immigration	-0.971**	---	-0.497*	---
	(0.374)		(0.262)	
% of team that is pro-immigration	-1.162**	---	-0.467**	---
	(0.333)		(0.163)	
Anti-immigration team	---	-0.435*	---	-0.303
		(0.266)		(0.197)
Pro-immigration team	---	-0.648**	---	-0.295**
		(0.256)		(0.119)
Statistical skills (z)	-0.284**	-0.286**	-0.100*	-0.103*
	(0.104)	(0.102)	(0.059)	(0.060)
Topic experience (z)	0.235**	0.295**	0.025	0.050
	(0.115)	(0.120)	(0.052)	(0.048)
R-squared	0.325	0.266	0.304	0.266

Table S13. Ideological bias and research quality

Notes: * $p < .1$; ** $p < .05$. Standard errors in parentheses and are clustered at the team level. The regressions are weighted by the number of peer reviews per model and the inverse number of models per team. The regressions in the last two columns are linear probability models. All regressions have 1,215 observations and include fixed effects indicating the team's size and a vector of fixed effects indicating the team's discipline of highest degree. The "high quality indicator" is a binary variable set to unity if the peer score is the top quartile (and zero otherwise).