NBER WORKING PAPER SERIES

IDEOLOGICAL BIAS IN ESTIMATES OF THE IMPACT OF IMMIGRATION

George J. Borjas
Nate Breznau

Working Paper 33274
http://www.nber.org/papers/w33274

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2024

We are grateful to Michael Amior, Catalina Amuedo Dorantes, Orley Ashenfelter, David Brady, Hugh Cassidy, Anthony Edo, Tom Emery, Daniel Hamermesh, Theodore Joyce, Joan Llull, Joan Monras, Jan Stuhler, Stephen Trejo, Robert VerBruggen, David Weakleim, and Christopher Wlezien for valuable suggestions and discussions. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2024 by George J. Borjas and Nate Breznau. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Ideological Bias in Estimates of the Impact of Immigration
George J. Borjas and Nate Breznau
NBER Working Paper No. 33274
December 2024
JEL No. C90, I38, J69

ABSTRACT

When studying policy-relevant topics, researchers' policy preferences may shape the design, execution, analysis, and interpretation of results. Detection of such bias is challenging because the research process itself is not normally part of a controlled experimental setting. Our analysis exploits a rare opportunity where 158 researchers working independently in 71 research teams participated in an experiment. After being surveyed about their position on immigration policy, they used the same data to answer the same well-defined empirical question: Does immigration affect the level of public support for social welfare programs? The researchers estimated 1,253 alternative regression models, producing a frequency distribution of the measured impact ranging from strongly negative to strongly positive. We find that research teams composed of pro-immigration researchers estimated more positive impacts of immigration on public support for social programs, while anti-immigration research teams reported more negative estimates. Moreover, the methods used by teams with strong pro- or anti- immigration priors received lower "referee scores" from their peers in the experiment. These lower-rated models helped produce the different effects estimated by the teams at the tails of the immigration sentiment distribution. The underlying research design decisions are the mechanism through which ideology enters the production function for parameter estimates.

George J. Borjas
Harvard Kennedy School
79 JFK Street
Cambridge, MA 02138
and NBER
george_borjas@hks.harvard.edu

Nate Breznau
German Institute for Adult Education/Leibniz Center for Lifelong Learning
breznau.nate@gmail.com

Ideological Bias in Estimates of the Impact of Immigration

George J. Borjas and Nate Breznau*

1. Introduction

Consumers of empirical research in the social and behavioral sciences know that there is often huge variation in existing estimates of important theoretical or policy-relevant parameters. Even in the very narrow context of labor market policy, for example, it is easy to identify important literatures that illustrate this point. Many empirical studies claim and document that minimum wage increases reduce the employment of affected workers, while other studies conclude that minimum wages either have no impact on employment or perhaps even increase it (Neumark, 2019; Manning, 2021). Similarly, some studies claim that the impact of immigration on the economic opportunities of native workers is negative, but others claim that it is zero or positive (Dustmann, Shönberg, and Stuhler, 2016; Blau and Mackie, 2017; Monras, 2021).

Metascience research aims to understand how biases in the research process can generate such large dispersions in results (Korbmacher *et al.*, 2023). This line of study has given us a better grasp of the role played by confirmation bias, where researchers attempt to hack their results to find what they wanted in the first place (Head *et al.*, 2015; Brodeur, Cook, and Heyes, 2020; Schneck, 2023); publication bias, where editors and reviewers filter out findings for reasons that are not purely scientific (Gerber and Malhotra, 2008; Blanco-Perez and Brodeur, 2020); the "file drawer problem," where some research findings never get submitted for publication (Mervis, 2014); the use of questionable data and faking of results (John, Loewenstein, and Prelec, 2012); and the noise and errors that enter the research production function because of human judgment and fallibility (Kahneman, Sibony, and Sunstein, 2021).

* Borjas: Harvard Kennedy School, National Bureau of Economic Research, and IZA; Breznau: German Institute for Adult Education/Leibniz Center for Lifelong Learning. We are grateful to Michael Amior, Catalina Amuedo Dorantes, Orley Ashenfelter, David Brady, Hugh Cassidy, Anthony Edo, Tom Emery, Daniel Hamermesh, Theodore Joyce, Joan Llull, Joan Monras, Jan Stuhler, Stephen Trejo, Robert VerBruggen, David Weakleim, and Christopher Wlezien for valuable suggestions and discussions.

There are also claims, particularly in discussions of policy-related research, that the political ideology of scientists biases both their results and the presentation of the evidence (Abramowitz *et al.*, 1975; Clark and Winegard, 2020; Honeycutt and Jussim, 2020; Jelveh, Kogut, and Naidu, 2024). One problem with investigating ideological bias, however, is the absence of controlled experiments that would allow observers to isolate the role played by this specific bias, and another is that ideological bias can enter the research production function in many ways, including the framing of a hypothesis and the design of the research methodology.

This study takes advantage of the unique and imaginative experiment designed, organized, and conducted by Nate Breznau, Eike Mark Rinke, and Alexander Wuttke (Breznau, Rinke, Wuttke *et al.*, 2022; henceforth BRW) to further understand how and why the dispersion in estimates of an important parameter arises. Specifically, BRW led an experiment in which 71 research teams (comprising 158 researchers) used publicly available International Social Survey Program (ISSP) data to conduct an empirical analysis of the proposition that "greater immigration reduces support for social policies among the public" (BRW, 2022, p. 1). The participating researchers had no control over the hypothesis to be tested or the data to be used. Moreover, there were enough participants in the experiment to establish statistical inference at the meta-level.

The work of Alesina and Glaeser (2004) represents an early and influential examination of this proposition. They argued that racial, cultural, and ethnic homogeneity might explain why European countries (prior to the immigrant shocks many of those countries received in recent decades) had developed more advanced and generous welfare systems than the United States. The hypothesis has been examined frequently over the past thirty years and studies across disciplines have arrived at different conclusions about the effect of immigration on public preferences. Collectively they suggest that the correlation between immigration and support for social programs may be negative (Schmidt-Catran and Spies, 2016; Eger and Breznau, 2017; and Alesina, Murard, and Rapoport, 2021), may be zero (Brady and Finnigan, 2014; and Auspurg, Brüderl, and Wöhler, 2020), or may be positive (Finseraas, 2009; Burgoon, 2014; and Garand, Xu, and Davis, 2017).

The BRW experiment instructed the participating research teams to first computationally reproduce the findings of a classic study of this hypothesis by Brady and

Finnigan (2014). The Brady-Finnigan study 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 task was part of the experimental design to determine not only how well teams could replicate 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. As part of the experiment, and *prior* to any empirical analysis being done, the participating researchers were also asked about their attitudes towards immigration; specifically, whether immigration laws "should be made tougher" or "should be relaxed".

Not surprisingly, the research teams produced wildly different estimates of the impact of immigration. As BRW (2022, p. 5) note, "no two teams arrived at the same set of numerical results or took the same major decisions during data analysis." Our paper examines the experimental data to determine if the dispersion depends on researcher priors about immigration. The data reveal that such a correlation exists. Research teams that had strong pro-immigration sentiments were more likely to obtain positive parameter estimates, suggesting that immigration *increases* public support for social programs, which social and behavioral scientists consider to be an indicator of increased social cohesion (Dragolov et al, 2016). Conversely, strongly anti-immigration teams were more likely to obtain negative estimates, potentially suggesting that immigration reduces social cohesion.

The experimental data also shows dispersion in the quality of the model specifications adopted by different teams, as measured by a random and double-blind review of each team's modeling strategy(ies) by other researchers in the experiment. The research designs of teams that are strongly anti-immigration or strongly pro-immigration are more likely to receive lower referee scores than those received by moderate teams in the middle of the immigration sentiment distribution. The combined choices in research design such as sample selection, variable definitions, and included regressors are the mechanism that leads anti-immigration teams to produce more negative estimates and pro-immigration teams more positive estimates of the impact of immigration. In fact, five

specific research design decisions such as the use of specific survey waves, the construction of the dependent variable, and the measure of the immigrant supply shock account for about two-thirds of the difference in the impact reported by pro- and anti-immigration teams.

The evidence from our examination of a relatively small sample of researchers in a very specific experiment has limitations in generalizability and statistical power. This limitation, however, is outweighed by the rare prospect provided by the design of the experiment and the public availability of the experimental data to begin to understand how political ideology shifts the research production function. "Many-analysts" studies are the only known experimental setting that allows outsiders to observe this process and, to date, the BRW study is the only one that asked participants about their prior position on a policy directly affected by the outcomes of their research. Moreover, the BRW experiment is among the largest in terms of the number of participating teams.[1] Thus, our investigation exploits a unique opportunity to observe and document how the policy preferences of researchers influence their output.

2. Data

The call for researchers to participate in the crowdsourced experiment was published on the web in April 2018 (BRW, 2022, *Supplemental Information Appendix*, p. 41; bold in original):

> We seek researchers to participate in a crowdsourced replication project on a high-profile social science question: **How does immigration shape public opinion?...**Participating researchers will (a) **replicate** and (b) **expand** a previously published **cross-national quantitative study**. We plan to distribute the results…and prepare them for publication in a high visibility social science journal. **All participants who complete the analytical tasks will be co-authors on the final paper...**We invite teams of 1-3 researchers to **independently analyze the data**.

[1] Other than a recent study (with 164 teams) that examines statistical properties of financial trends in the Eurozone market (Menkveld et al, 2024), the other comparable experiments in the literature are far too small to conduct multivariate analysis of central tendencies.

Each research team was given five waves of the International Social Survey Program (ISSP) data (spanning the 1985-2016 period), measures of immigration levels including percent foreign-born in the population and net immigration flows (in-migration minus out-migration per 1,000 inhabitants), and an updated array of the macro-indicator data used in the Brady-Finnigan (2014) study. The teams were instructed to first replicate the Brady-Finnigan results, and then extend that analysis to test the hypothesis that *immigration reduces public support for social welfare policies*. By the end of the project, BRW had collected results from 161 researchers in 73 teams. Two teams, however, used models that did not generate numerical estimates of the impact of immigration and are excluded from our analysis. This leaves a working sample of 158 researchers in 71 teams.

The ISSP records attitude towards the government provision of social programs by asking (BRW, 2022, *SI 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.

As part of the experiment, BRW collected information about the researchers in four separate waves of interviews (conducted before, during, and after the research activity). These data include the researchers' educational background, familiarity with statistical methods, prior experience in immigration or social policy research, and general attitudes towards immigration and the hypothesis under study.

The key question that measures a researcher's stance towards immigration was asked in the first wave of interviews, *prior* to any data analysis (BRW *Supplemental Information Appendix*, 2022, p. 87): "Do you think that, in your current country of residence, laws on immigration of foreigners should be relaxed or made tougher?" The

researchers were given a 7-point scale to answer the question, with the extremes being "immigration laws should be made tougher" and "immigration laws should be relaxed."

Figure 1*A* illustrates the frequency distribution of the responses among the participating researchers.[2] The researchers skew heavily towards a pro-immigration stance: Nearly half strongly support the proposal that immigration laws should be relaxed, choosing a "5" or a "6" as responses to the question, and note that no researcher responded with a "0".

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

$$y_{rs} = \beta_{rs} m_{rs} + controls + error, \tag{1}$$

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.[3] When submitting their estimates, 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 share. Figure 2*A* shows the frequency distribution of the AMEs in the experimental data. Although the estimated AMEs cluster around zero, 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

[2] The Data Appendix gives a detailed description of all the variables used in our analysis.

[3] 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.

estimate is 0.052 (0.011). In substantive terms, depending on who conducts the empirical analysis, a 10-percentage-point increase in the fraction 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.[4]

The interesting question is whether the variation in the AMEs is strictly random or can partly be attributed to pre-existing and observable researcher characteristics. 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. The publicly available data contain a measure of a *team's* pro-immigration sentiment, given by the mean of this index across the (up to three) team members. It is obvious, however, that a simple averaging of this index may not completely capture the immigration ideology of a particular team. For example, two three-person teams could both have a mean sentiment of 4.0, but this outcome could describe a team where all researchers responded with a "4", or a team where one researcher responded with a "6" and the other two responded with a "3".

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

[4] 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.

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]

The summary statistics reported in Table 1 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 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 2*B* 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 suggest 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.

[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.

3. Regression results

The link between ideological bias and the AME can be established 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 1 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; fixed effects for two-person teams with researchers from different disciplines; and fixed effects for three-person

teams indicating if the majority of researchers are sociologists or political scientists, and if the lead author's discipline differs from that of the other two team members.[6] We show below that our results are robust to using alternative controls for the disciplinary composition of multi-person teams.

Columns 1-4 in the top panel of Table 2 report the basic set of regressions, using four alternative specifications for capturing the differences in ideology across teams.[7] The first column shows that the team's mean immigration sentiment index has a positive and significant impact on the estimated AME. In fact, going from one extreme of the sentiment distribution to the other (from a "1" to a "6" in the scale) increases the estimated AME by 0.054 points (0.029).

The second column relates the AME to the team's ideological composition, as measured by the percent of the team members that are either pro- or anti-immigration. The larger the representation of anti-immigration researchers in the team, the lower the estimated AME; and the larger the representation of pro-immigration researchers, the higher the estimated AME. We can use the coefficients to estimate the difference in the estimated AME between teams composed solely of pro- or anti-immigration researchers (given by Δ in the table). This difference is 0.074 (0.024) points.

The third column estimates the adjusted difference in the AME between pro-immigration teams (where more than half the members are pro-immigration) and all other teams, thus bypassing the need to rely on the small sample of anti-immigration teams to estimate the impact. The difference is again large and statistically significant; the coefficient is 0.040 (0.016).

Finally, the link between ideology and the AME is strongest when we specifically compare pro- and anti-immigration teams in column 4 (where the anti-immigration teams have at least one researcher who is anti-immigration). The AME estimated by pro-immigration teams is

---

[6] Specifically, 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.

[7] The unit of observation in the regressions is a model estimated by a particular team. We weigh all regressions by the inverse of the number of models reported by the team. If all covariates are constant within a team, the regression coefficients would be numerically identical to those obtained when the regression is estimated using team-level data instead.

0.027 (0.015) points higher than the estimate of the moderate teams. In contrast, the AME estimated by anti-immigration teams is -0.057 points lower. The difference between the pro- and anti-immigration teams is 0.085 (0.031) points and statistically significant (with a *p* value of 0.008).

In sum, regardless of how we quantify the team's ideology, the analysis always shows a significant difference between the AME estimated by teams that can be generally considered pro-immigration and teams that can generally be considered anti-immigration. Further, the difference is numerically important. Holding constant all regressors at their mean values, the predicted AME in column 4 is -0.070 and 0.014 for the anti- and pro-immigration teams, respectively. This difference implies that going from one extreme to the other in the immigration sentiment distribution is equivalent to moving a team from the 16th percentile of the (adjusted) AME distribution to the 63rd percentile. In short, extreme differences in immigration ideology produce substantial dispersion in the estimate of the parameter of interest.[8]

The regressions also show that both the statistical skills and topic experience indices are important determinants of the size of the estimated AME. Teams that have better statistical skills produce more positive AMEs, while teams that have more topic experience produce more negative AMEs. Although it is interesting to speculate about what these effects represent, the data do not allow any inference about the mechanism. For example, the negative impact of topic experience could indicate that more informed researchers make research design decisions that are better suited for the question at hand, and those decisions happen to lead to negative estimated effects. But the negative effect could also reflect selection bias: the researchers who naturally gravitated to the immigration/social policy research arena in the past are not randomly chosen, and the negative coefficient could be reflecting part of that self-selection (which is not captured by the included immigration ideology variables).

[8] Although the regression in column 4 uses the moderate group as the baseline, it is important to emphasize that this expository choice does *not* imply that moderate teams estimated the "true" value of the parameter. We chose this baseline to simply illustrate how teams at the two ends of the immigration sentiment distribution tend to estimate parameters in the tails of the AME distribution.

Of course, many of the 1,253 estimates of the AME produced by the teams would never see the light of day in a peer-review research environment. They would likely be dismissed as technically clumsy or (too obviously) reflecting a researcher's priors.[9] The data collected in the experiment, however, allows us to partly account for this selection by re-estimating the regression model after adjusting for scientific community vetting.

As part of the experiment, BRW conducted a randomized double-blind refereeing exercise for each team's research design. Four or five randomly chosen researchers participating in the experiment were given the details of a regression model used by another team and were asked to score the research design (BRW, 2022, *SI Appendix*, p. 122):

> How confident [are you] that the respective research design is adequate for testing the hypothesis that 'immigration undermines social policy preferences' using ISSP data?

The reviewers responded using a 7-point scale ranging from "Unconfident" to "Confident". BRW averaged the votes across reviewers to construct a "referee score" for each specification. The bottom panel of Table 2 reports the coefficients produced by regressions that also weigh the observations by the referee score. The results are very similar to those reported in Panel A. For example, the difference in the estimated AME in column 4 between the pro- and anti-immigration teams increases slightly to 0.090 (0.034) points. The difference in referee scores across the different types of teams is of interest and will be discussed in detail in the next section.

It is important to emphasize that the regressions estimated in Table 2 do *not* include any variables that describe the actual specification of the regression model used to produce a particular estimate of the AME. After "playing around" with the data (perhaps in the initial replication phase of the experiment or based on previous topical experience), a researcher might be able to infer how particular variables and particular estimation techniques influence the sign and magnitude of the estimated AME. For example, the ISSP asked several questions about the government's responsibility to provide specific services, such as housing or health programs. It

[9] In fact, Breznau, Rinke, and Wuttke (2024) document that the participating researchers made errors in the replication phase of the experiment.

would not be surprising if the results are sensitive to which (sub)set of programs is used as the dependent variable. Similarly, the immigrant supply shock can be measured as a share of the population that is foreign-born or as net migration per year. The choice of variables and estimation techniques is one mechanism through which ideological bias might influence the estimate of the AME. In short, *model specification is endogenous*.[10] The link between ideology and research design decisions will be documented below.

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".

Figure 1*B* 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 the indices in Figures 1*A* and 1*B* is -0.08).

As a prior expectation might produce confirmation bias, we add a variable measuring the team's prior belief in the hypothesis to the 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 (see Table 1). Columns 5-6 of Table 2 adds the "hypothesis prior" variable to the most general specifications of the regression model. The inclusion of this variable does not

[10] BRW (2022, *SI Appendix*, pp. 27-28) show that regressions that relate the AME to various attitudinal measures *and* variables that describe the regression specification (e.g., logit or OLS, binary or categorical dependent variables, the definition of the immigrant supply shock, the set of countries used in the analysis, etc.) explain relatively little of the variance in the AME. Note, however, that those regressions were not designed to identify singular significant effects nor the causal impact of ideological priors on the AME, which is the objective of our analysis. As we argue and show below, the specification decision is endogenous and is one mechanism through which ideological bias might influence the estimate.

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.

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 3 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.

Up to this point, we have 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*).[11] 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.

Figure 2*C* shows that the frequency distribution of the AME in the dyad data again suggests that the distribution has more mass in the left tail for anti-immigration researchers and more mass in the right tail for pro-immigration researchers. We estimated the regression model in equation (2) using researcher-model dyads as the unit of observation, and the coefficients are reported in Table 4.[12]

The regression reported in column 2 shows that pro-immigration researchers estimate more positive AMEs (relative to all other researchers) and the difference is significant. The more general model in column 3 again reveals that anti-immigration researchers submit more negative AMEs than the moderate researchers, while pro-immigration researchers submit more positive AMEs. The difference between the average AME submitted by researchers at the tail ends of the immigration sentiment distribution is numerically large and statistically significant (a difference of 0.050 points, with a standard error of 0.017). Moreover, the relative effect resembles that found in the team-level analysis. In particular, the predicted AME for anti-immigration researchers (-0.041, with a standard error of 0.014) is in the 23rd percentile of the AME distribution in these data, while the prediction for pro-immigration researchers (0.009, with a standard error of 0.009) is in the 58th percentile.

Finally, the graphical analysis in Figures 2*B* and 2*C* suggests that the tails of the AME distribution are particularly sensitive to the immigration ideology of the team and the researchers. Table 5 exploits this insight by estimating a set of linear probability models (using both the team-model level data and the researcher-model dyads) that measure the impact of immigration ideology on the likelihood that the estimated AME is in the tails *and*

---

[11] 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.

[12] 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.

statistically significant. In other words, instead of having the AME as the dependent variable (implying that the regression coefficients are identifying the impact of immigration sentiments on the mean AME), we now estimate the impact of ideological bias on the probability of producing significant estimates in either tail of the AME distribution. Specifically, the regression models examine the probability that the estimated AME lies below the 10^{th} or above the 90^{th} percentile and is statistically significant at the 5 percent level in a one-tail test ($t > |1.645|$).

The marginal effects reported in Table 5 are striking. Anti-immigration teams are significantly less likely to estimate significant effects in the positive tail, and pro-immigration teams are significantly less likely to estimate significant effects in the negative tail. As a result, the probability that a pro-immigration team estimates a sizable positive AME is 8.7 (3.5) percentage points higher than that of an anti-immigration team. Similarly, the probability that an anti-immigration team estimates a sizable negative AME is 27.4 (12.5) percentage points higher than that of a pro-immigration team. The regressions using the researcher-model dyad data reveal a similar pattern: Pro-immigration researchers are 6.8 (3.5) percent more likely to estimate large and significant positive effects, while anti-immigration researchers are 16.7 (8.0) percent more likely to estimate large and significant negative effects. In short, as suggested by the visual differences in the AME distributions, anti-immigration teams (researchers) are relatively more likely to estimate significant effects in the left tail of the distribution and pro-immigration teams (researchers) are relatively more likely to estimate significant effects in the right tail.

4. Ideological Bias and Research Quality

Each team's research design was reviewed anonymously by 4 or 5 other researchers participating in the experiment (using a 7-point scale) and BRW constructed a "referee score" for each model by averaging these reviews. These data allow us to examine the possibility that immigration ideology not only biases the estimated AME, but that some of that bias arises because ideology affects research quality. If this conjecture is correct, the impact of immigration ideology on research quality may not be monotonic but might instead show up in both tails of the immigration sentiment distribution.

Figure 3 illustrates the distribution of standardized referee scores for each of the three types of teams. It shows that both the anti- and pro-immigration teams have raw distributions of referee scores that lie to the left of the distribution of the moderate teams. The moderate teams obtain the highest mean score of 0.35, as compared to 0.03 for the anti-immigration teams and -0.33 for the pro-immigration teams. In short, there is a sizable difference in referee scores between the moderate teams and the teams at the tails of the immigration sentiment distribution. The descriptive evidence is equally striking if instead of looking at the continuous measure of a referee score, we calculate the probability that the regression specification received a "high" grade, which we define as a referee score above the 75th percentile. As Table 1 reports, 31.3 percent of the models submitted by the moderate teams received a high grade, as compared to only about 16 percent for the models submitted by either anti- or pro-immigration teams.

It is of interest to determine if these differences in referee scores remain after controlling for the regressors introduced earlier. Table 6 summarizes the regression results. The first two columns of the table show the regressions when the dependent variable is the continuous referee score, while the last two columns report marginal effects from linear probability models where the dependent variable is set to unity if the model specification received a referee score above the 75th percentile. The qualitative findings are similar regardless of which dependent variable is used. To simplify the discussion, we focus on the results using the continuous standardized referee score variable.

The regressions reported in Table 6 shows that both tails of the immigration sentiment distribution receive a referee score that is at least one-half of a standard deviation below that received by the moderate teams (and the difference is statistically significant). In short, strong ideological biases in either direction produce regression models that are not well regarded by anonymous reviewers.

It is important to note that the documented differences in research quality across the three types of teams are *not* affected by potential ideological bias on the part of referees. The reviewing process was double-blind and random, so that even if referees preferred specifications that would confirm their own biases (Abramowitz *et al.*, 1975), the average referee score achieved by any model would not be affected.

5. Ideology and Research Design Decisions

One obvious inference from the evidence is that teams at the extremes of the immigration sentiment distribution use regression specifications that, although they received lower scores from anonymous reviewers, happen to produce a particular result. Teams aiming to produce a certain result might do so through specific combinations of modelling decisions. Therefore, it is of interest to determine if a relatively small number of design decisions explains the observed variation. More generally, how exactly do the types of teams differ in terms of defining variables, samples, and estimation methods?

The experimental data (2022, *SI Appendix*, pp. 54-63) record the outcome of 103 different specification decisions taken by more than one team—such as the exact definition of the dependent variable, the measure of the immigrant shock, the (sub)set of European countries used in the analysis, the waves of the ISSP included in the data, and the choice of statistical methods such as linear probability models, multinomial logit, random effects, etc. It turns out, however, that unique combinations of decisions made along *five* dimensions produce much of the observed variation in the mean AME estimated by the three types of teams. These key decisions are:

1. The ISSP records public attitudes towards the government provision of various types of programs (jobs, health, etc.). Are these responses aggregated to form a single dependent variable (e.g., by averaging or estimating a factor index)?
2. Is immigration measured as a stock or a flow?
3. Does the model include regressors controlling for variation at the country-year level (e.g., country-year fixed effects)?
4. Do the regressions use data for all countries in the ISSP?
5. Does the analysis use the 2016 wave of the ISSP (in addition to the 1996 and 2006 waves used in the original Brady-Finnigan study)?

The combination of these decisions produces 58 alternative regression specifications (with non-empty cells). As Table 7 shows, there are noticeable differences in design choices across the three types of teams. For example, only 10.4 percent of the anti-immigration teams, but over 15 percent of either the moderate or pro-immigration teams

use a composite dependent variable. Similarly, only a quarter of pro- and anti-immigration teams use the data for all available countries in the regression models, but almost half of the moderate teams used the entire sample. And anti-immigration teams were more likely to use data from the 2016 ISSP wave (73.1 percent of anti-immigration teams as compared to 60.2 percent of pro-immigration teams).

It is instructive to illustrate the impact of the decisions made by the different types of teams. For each of the 58 specifications, we calculated the "expected AME," defined as the mean AME in the subset of models using that specification.[13] Figure 4*A* ranks the 58 unique specifications from lowest to highest expected AME, and shows the frequency of adopting each unique specification for the three types of teams. It is notable that anti-immigration teams are the only teams that used the unique specifications that produce the lowest expected AMEs and that pro-immigration teams are the only teams that used the unique specifications that produce the highest expected AMEs. Figure 4*B* illustrates the data in an alternative way by showing the frequency distribution of the expected AME for each type of team. The density function for the anti-immigration teams has little mass for expected AMEs above 0.0, while the density function for pro-immigration teams has little mass below -0.1.

We re-estimated the regression models first reported in Table 2 to determine how the team's immigration ideology affects the mean of the expected AME distribution. The results reported in Table 8 are roughly similar to those in Table 2 that use the actual AME. In particular, the regression reported in column 4 shows that anti-immigration teams choose specifications that, on average, produce an expected AME that is -0.044 (0.024) points lower than the typical model estimated by a moderate team, while pro-immigration teams choose specifications that produce an AME that is 0.014 (0.008) higher. The difference in the AME estimated by the two extreme types of teams is -0.058 (0.025) points and significant with a *p*-value of 0.025. The regression reported in column 2 (which uses the fraction of the team that is either pro- or anti-immigration to capture the team's ideology) does not change the key insight. Teams composed exclusively of pro-immigration

[13] Equivalently, we estimated a regression of the AME on a fully interacted model that contains fixed effects for each of the 58 different specifications and the expected AME is the value of the fixed effect.

researchers adopt specifications that produce an expected AME that is 0.040 points (0.020) higher than the specifications adopted by teams composed exclusively of anti-immigration researchers.

The rough similarity in the coefficients reported in Tables 2 and 8 suggests that inter-team differences in research design decisions along the five margins noted above account for much of the result that anti-immigration teams estimate more negative impacts and pro-immigration teams estimate more positive impacts. The regression in column 4 of Table 2 implies that the (adjusted) difference in *actual* AMEs between the two extreme types of teams is 0.085, while the corresponding regression in column 4 of Table 8 implies that the difference in *expected* AME is 0.058. Hence the design choices along those five margins alone account for 68 percent (or 0.058 ÷ 0.085) of this difference. The analogous calculation using the coefficients from the regressions reported in column 2 of Tables 2 and 8 that use the fraction of the team's members that are pro- or anti-immigration implies that those choices account for 54 percent of the difference.

In sum, the *endogenous* research design choices made by teams with immigration ideologies in either tail of the immigration sentiment distribution happen to produce parameter estimates that seem consistent with the teams' pre-existing ideological priors. The combination of modeling decisions along the five dimensions examined in this section produce unique regression specifications that account for much of the total ideology effect. Our analysis thus provides a straightforward depiction of the mechanism through which ideology influences research outcomes.

6. Conclusion

This paper analyzes data generated by a unique experiment, where 71 research teams (comprising 158 individual researchers) used the same publicly available surveys to answer the same question: Does immigration reduce the level of political support for the social programs that make up the welfare state?

To anyone familiar with the mechanics of empirical work in social science, it is not surprising that there were as many estimates of this impact as there were alternative regression models (1,253 to be exact). Each team had a unique way of selecting the sample

for analysis, of defining the key dependent and independent variables, and of specifying the statistical analysis.

The data produced by the various research teams, however, can provide useful information about whether the dispersion in the estimated effect of immigration is random or depends on the underlying ideology of the team members towards immigration policy. In short, do researchers who strongly favor tightening immigration laws produce different estimates of the impact of immigration on social cohesion than researchers who strongly favor relaxing immigration laws?

The experimental data suggest that the team's pre-existing preference towards immigration restrictions play a role in producing some of the observed differences. The measured effect of immigration on social cohesion is more positive if the researchers are pro-immigration, and more negative if the researchers are anti-immigration. Further, the regression specifications proposed by researchers who are either very pro-immigration or very anti-immigration get lower "grades" from their peers than the specifications adopted by researchers who have moderate immigration sentiments. In other words, immigration ideology influences research quality in a way that happens to produce evidence that seemingly reflects the team's pre-existing ideology.

One benign interpretation is that the time and effort devoted to research activities are limited resources, and researchers allocate their time and effort in the same way as everyone else. Namely, it is costly to develop an idea into a well-crafted empirical analysis, and researchers face many tradeoffs when making this labor supply decision. Given the opportunity cost, it would not be surprising that once a researcher starts examining the data and finds a particularly appealing result that can be easily assembled into a compelling narrative, the researcher stops the empirical search for alternative stories. Metascience research reveals that such behavior is self-reported by at least one-third of researchers across the behavioral and social sciences (John, Loewenstein and Prelec, 2012; Gopalakrishna *et al*., 2022). In this interpretation, researchers who are predisposed to favor a particular narrative about the impact of immigration begin to craft publishable results once the data seem to confirm that internally appealing story.

Even if this interpretation of the research production function were correct, it highlights a problem with empirical research in modern applied social science. As part of

the credibility revolution in empirical research, which emphasizes finding well-defined natural experiments to determine the causal impact of an exogenous shock on economic or social outcomes, the literature is increasingly dominated by policy evaluation studies. The shocks that are easy to find and examine empirically are typically shocks created by policy changes, and a major part of ongoing research activity essentially examines the impact of a particular policy shift on a particular set of outcomes.

This kind of policy-oriented research, however, may well attract an even more selected group of researchers who *really* care about the consequences of the specific policies they plan to examine. Combined with the tradeoffs in the labor supply decision, the strong policy focus that motivates much of current empirical research could easily amplify the role of ideological bias in estimates of relevant parameters.

We recognize that the conclusions drawn from any analysis of this specific experiment are based on a moderately sized sample of 71 teams. Given the sample size, skeptics may rightfully assert that ideology could simply reflect random chance. Nevertheless, comparing the research outcomes produced by teams who differ in their immigration sentiments across all our regression models, we reject the null hypothesis that ideology has a zero effect on the production of research findings in *all* cases.

We also recognize that our results may not be generalizable. After all, it is unclear how much effort the researchers put into the project because the key payoff to participating (i.e., promised co-authorship on a yet-to-be-written paper with uncertain prospects) may not have been sufficiently advantageous for researchers to reallocate their time from more promising projects.

Finally, it is likely that the ongoing revolution in Artificial Intelligence (AI) will influence the research process itself. The revolution introduces an additional source of ideological bias as the output of any AI algorithm may reflect the bias of its creators and certainly of the data that it is trained on (Breznau, 2021; Buyl *et al*, 2024). Moreover, the revolution may dramatically lower the costs incurred by ideologically motivated researchers as they search across many possible research designs for the model that confirms their priors. At the same time, however, just as an AI has already been trained to spot statistical errors (Nuijten and Wicherts, 2023), a future AI could theoretically be trained to identify ideological bias.

References

Abramowitz, S. I., Gomes, B. & Abramowitz, C. V. Publish or Politic: Referee Bias in Manuscript Review. *Journal of Applied Social Psychology* **5**, 187–200 (1975).

Alesina, Alberto F. and Edward L. Glaeser. 2004. *Fighting Poverty in the U.S. and Europe*. New York: Oxford University Press.

Alesina, Alberto, Elie Murard, and Hillel Rapoport. 2021. "Immigration and Preferences for Redistribution in Europe," *Journal of Economic Geography* 21, 925–954 (2021).

Auspurg, Katrin, Josef Brüderl, and Thomas Wöhler. 2020. "Does Immigration Reduce the Support for Welfare Spending? A Cautionary Tale on Spatial Panel Data Analysis." *American Sociological Review* 84 (4): 754–63.

Blanco-Perez, C. and Brodeur, A. (2020) 'Publication Bias and Editorial Statement on Negative Findings', *The Economic Journal*, 130(629), pp. 1226–1247. Available at: https://doi.org/10.1093/ej/ueaa011.

Blau, Francine D. and Christopher Mackie, eds. 2016. *The Economic and Fiscal Consequences of Immigration*. Washington, DC: National Academies Press.

Brady, David and Ryan Finnigan. 2014. "Does Immigration Undermine Public Support for Social Policy," *American Sociological Review* 79 (February), pp. 17-42.

Breznau, N. Integrating Computer Prediction Methods in Social Science: A Comment on Hofman et al. (2021). *Social Science Computer Review* **40**, 844–853 (2022).

Breznau, Nate, Eike Mark Rinke, and Alexander Wuttke, et al. 2022. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Uncertainty," *Proceedings of the National Academy of Sciences* 119, No. 44. Available at: doi.org/10.1073/pnas.2203150119.

Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, et al. 2024. "The Reliability of Replications: A Study in Computational Reproductions." *MetaArXiv*. https://osf.io/preprints/socarxiv/j7qta/.

Brodeur, A., Cook, N. and Heyes, A. (2020) 'Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics', *American Economic Review*, 110(11), pp. 3634–3660. Available at: https://doi.org/10.1257/aer.20190687.

Burgoon, Brian. 2014. "Immigration, Integration, and Support for Redistribution in Europe." *World Politics* 66 (3): 365–405. https://doi.org/DOI: 10.1017/S0043887114000100.

Buyl, M. *et al.* Large Language Models Reflect the Ideology of their Creators. Preprint at https://doi.org/10.48550/arXiv.2410.18417 (2024).

Clark, C.J. and Winegard, B.M. (2020) 'Tribalism in War and Peace: The Nature and Evolution of Ideological Epistemology and Its Significance for Modern Social Science',*Psychological Inquiry*, 31(1), pp. 1–22. Available at: https://doi.org/10.1080/1047840X.2020.1721233.

Dragolov, G. *et al. Social Cohesion in the Western World: What Holds Societies Together: Insights from the Social Cohesion Radar*. (Springer, 2016).

Dustmann, Christian, Uta Schönberg, and Jan Stuhler. 2016. "The Impact of Immigration: Why Do Studies Reach Such Different Results?" *Journal of Economic Perspectives,* pp. 31-56.

Eger, Maureen A, and Nate Breznau. 2017. "Immigration and the Welfare State: A Cross-Regional Analysis of European Welfare Attitudes." *International Journal of Comparative Sociology* 58 (5): 440–63. https://doi.org/10.1177/0020715217690796.

Finseraas, Henning. 2009. "Income Inequality and Demand for Redistribution: A Multilevel Analysis of European Public Opinion." *Scandinavian Political Studies* 32 (1): 94–119. Available at https://doi.org/10.1111/j.1467-9477.2008.00211.x.

Garand, James C., Ping Xu, and Belinda C. Davis. 2017. "Immigration Attitudes and Support for the Welfare State in the American Mass Public," *American Journal of Political Science* 61, 146–162.

Gerber, A.S. and Malhotra, N. (2008) 'Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?', *Sociological Methods & Research*, 37(1), pp. 3–30. Available at: https://doi.org/10.1177/0049124108318973.

Gopalakrishna, G., Riet, G. ter, Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLOS ONE*, *17*(2), e0263023. Available at https://doi.org/10.1371/journal.pone.0263023.

Grossman, Jean Baldwin. 1982. "The Substitutability of Natives and Immigrants in Production," *Review of Economics and Statistics* 64 (November), pp. 596-603.

Head, M.L. *et al.* (2015) 'The Extent and Consequences of P-Hacking in Science', *PLoS Biology*, 13(3). Available at: https://doi.org/10.1371/journal.pbio.1002106.

Honeycutt, N. and Jussim, L. (2020) 'A Model of Political Bias in Social Science Research', *Psychological Inquiry*, 31(1), pp. 73–85. Available at: https://doi.org/10.1080/1047840X.2020.1722600.

Jelveh, Zubin, Kogut, Bruce, and Naidu, Surish. (2024). "Political Language in Economics," *Economic Journal* 134 (August), pp. 2439-2469.

John, L.K., Loewenstein, G. and Prelec, D. (2012) 'Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling', *Psychological Science*, 23(5), pp. 524–532. Available at: https://doi.org/10.1177/0956797611430953.

Kahneman, D., Sibony, O. and Sunstein, C.R. (2021) *Noise. A Flaw in Human Judgement*. New York, NY: Little, Brown Spark.

Korbmacher, M. *et al.* (2023) 'The replication crisis has led to positive structural, procedural, and community changes', *Communications Psychology* [Preprint]. Available at: https://gala.gre.ac.uk/id/eprint/42730/ (Accessed: 30 June 2023).

Lipset, Seymour M. and Gary W. Marks. 2000. *It Didn't Happen Here*. New York: Norton.

Manning, Alan. 2021. "The Elusive Employment Effect of the Minimum Wage," *Journal of Economic Perspectives*, pp. 3-26.

Menkveld, A. J. *et al.* Nonstandard Errors. *The Journal of Finance* **79**, 2339–2390 (2024).

Mervis, J. (2014) 'Why null results rarely see the light of day', *Science*, 345(6200), pp. 992–992. Available at: https://doi.org/10.1126/science.345.6200.992.

Monras, J. (2021). Local Adjustment to Immigrant-driven Labor Supply Shocks, Journal of Human Capital 15(1), 204-35.

Neumark, David. 2019. "The Econometrics and Economics of the Employment Effects of Minimum Wages: Getting from Known Unknowns to Known Knowns," *German Economic Review*, 293-329.

Nuijten, M. B. & Wicherts, J. The effectiveness of implementing statcheck in the peer review process to avoid statistical reporting errors. Preprint at https://doi.org/10.31234/osf.io/bxau9 (2023).

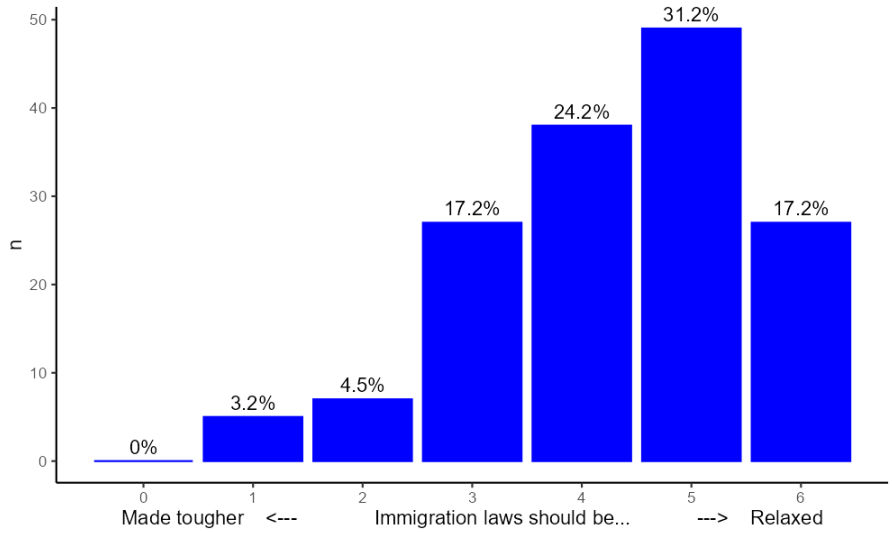
Schmidt-Catran, Alexander W, and Dennis C Spies. 2016. "Immigration and Welfare Support in Germany." *American Sociological Review* 81 (2): 1–20. https://doi.org/10.1177/0003122416633140.

Schneck, A. (2023) 'Are most published research findings false? Trends in statistical power, publication selection bias, and the false discovery rate in psychology (1975–2017)', *PLOS ONE*, 18(10), p. e0292717. Available at: https://doi.org/10.1371/journal.pone.0292717

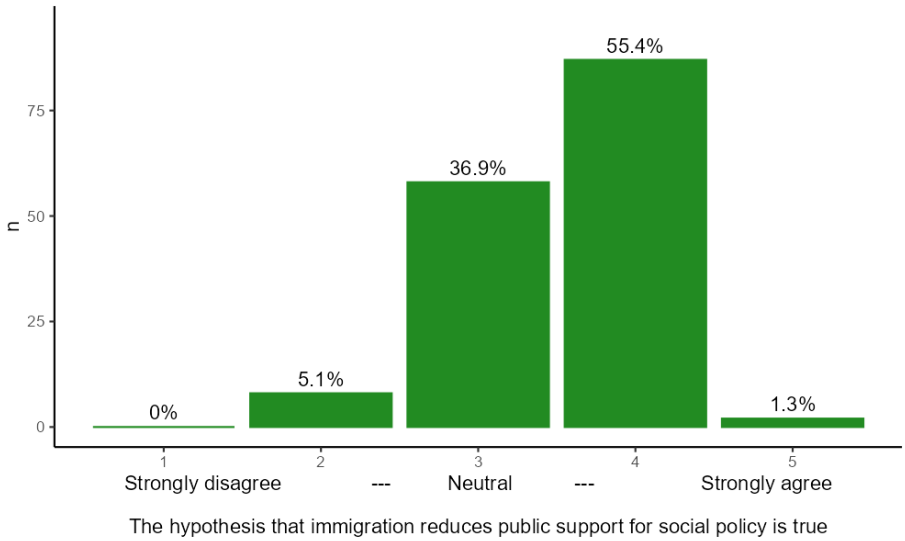
Silberzahn, R. & Uhlmann, E. L. Crowdsourced Research: Many Hands make Light Work. *Nature* **526**, 189–191 (2015).

Figure 1. Distribution of attitudes towards immigration

A. Immigration sentiment index



B. Prior belief on hypothesis

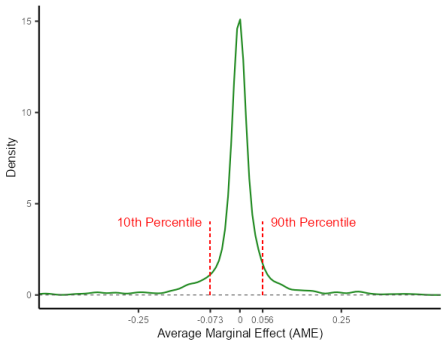


Notes: The frequency distributions are based on the survey responses of 158 researchers, with 3 missing cases due to non-response of either of the indices. Pearson correlation between the two measures in Panels A and B is -0.080.

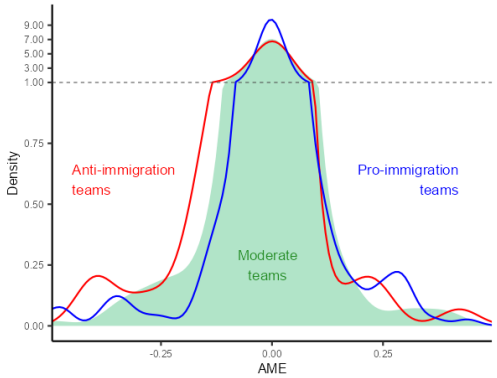
.

Figure 2. Distribution of AME across teams, models, and researchers

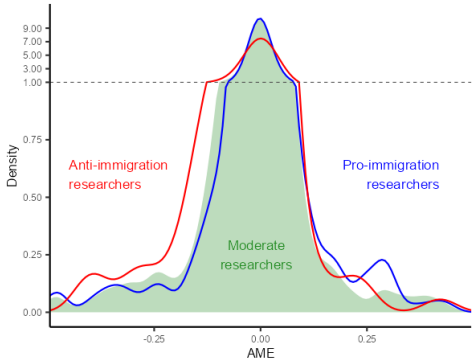
A. All teams and researchers



B. By team's immigration ideology

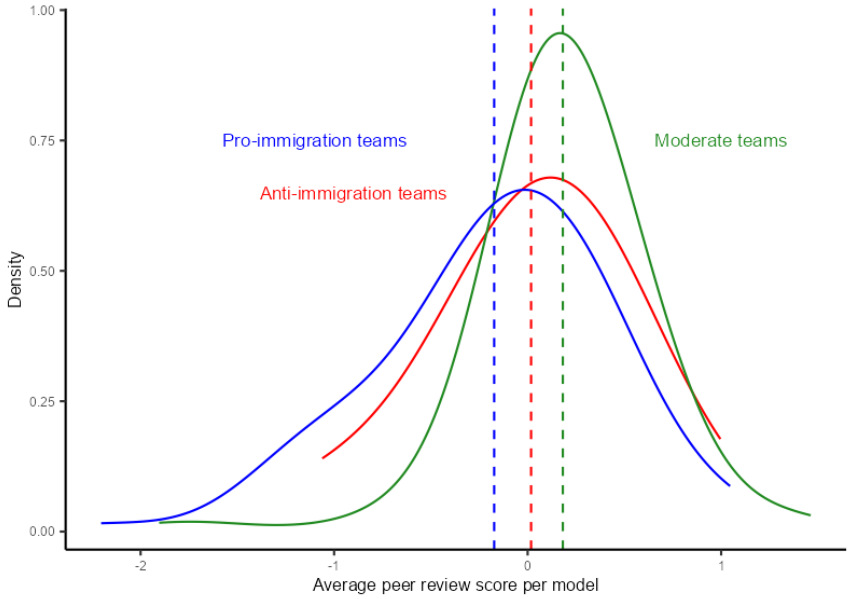


C. By researcher's immigration ideology



Notes: For ease of viewing, the *x*-axis in all panels is trimmed to include 98% of the distribution and the *y*-axis in Panels B and C is compressed above 1. The distributions in Panels A and B have 1,253 statistical models in 71 teams, and the distribution in Panel C has 2,680 researcher-model dyads.

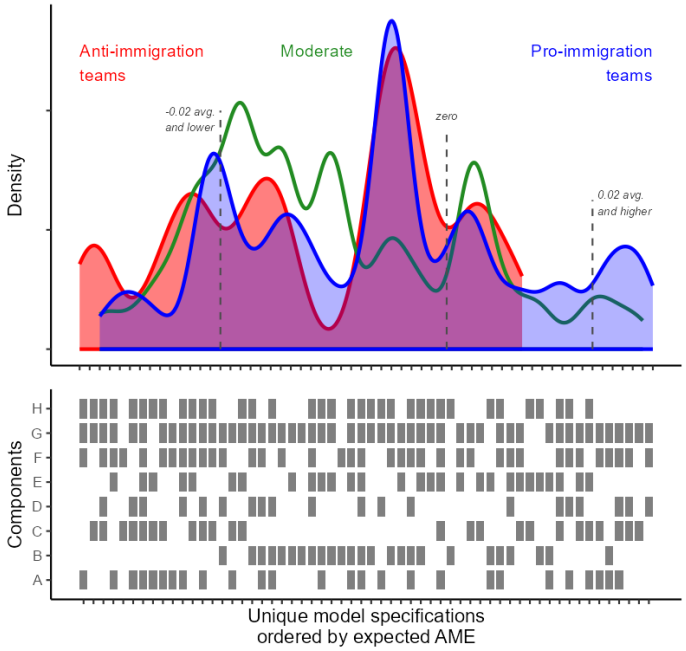
Figure 3. Distribution of standardized referee score, by type of team



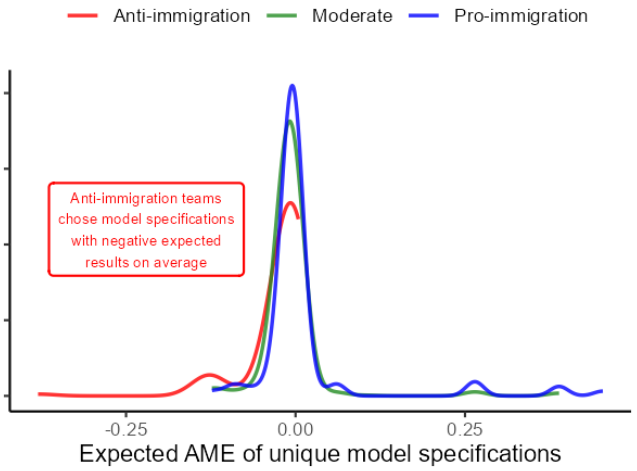
Notes: The figures present the raw distribution of standardized peer review scores for 1,215 models across 71 teams, by team ideology; the vertical dotted lines represent the raw means for each type of team.

Figure 4. Expected outcomes of AME based on five central modeling decisions

A. Specification density



B. Expected AME density



Notes: Panel A distribution of unique model specifications by team ideology ranked by order of expected AME, calculated as the mean AME for these models. The decisions are: 1. Scaling: A = Latent scale for dependent variable; 2. Test variable: B = Immigration measured as stock (% foreign-born), C = Immigration measured as flow (net migration); 3. Model structure: D = Regression includes country-year fixed effects; 4. Countries: E = All available countries in the data included; 5. Data waves: F = Data from the 1996 wave included, G = Data from the 2006 wave included, H = Data from the 2016 wave included. Panel B distribution with unique specifications plotted by expected AME.

Table 1. Summary statistics

Variable	Sample			
	All teams	Anti-imm.	Moderate	Pro-imm.
AME	0.001	-0.019	-0.008	0.014
% AME < 10th percentile and significant	0.063	0.119	0.086	0.028
% AME > 90th 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

Notes. 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.

Table 2. Determinants of AME, model level regressions

	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)
Δ: 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)
Δ: 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

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. All regressions in Panel B also include the statistics skill and topic experience indices. All regressions have 1,253 observations, and include fixed effects for team size, and fixed effects indicating the team's field of highest degree.

Table 3. Robustness of results to alternative controls for discipline composition of team

	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)
Δ: 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)
Δ: Pro – Anti	0.068**	0.080**	0.073**	0.088**
	(0.028)	(0.032)	(0.030)	(0.035)

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.

Table 4. Determinants of AME, researcher-model dyads

	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)
Δ: 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

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.

Table 5. Impact of ideology on probability of obtaining extreme and significant AMEs

	Team-model level regressions		Researcher-model level regressions	
Variable:	AME < 10th pct	AME > 90th pct	AME < 10th pct	AME > 90th pct
A. Main regressions				
1. % of team that is anti-immigration	0.081	-0.130**	---	---
	(0.103)	(0.057)		
% of team that is pro-immigration	-0.161**	-0.016	---	---
	(0.064)	(0.041)		
Δ: Pro – Anti	-0.242**	0.114**	---	---
	(0.107)	(0.046)		
2. Anti-immigration team (or researcher)	0.150	-0.070*	0.053	-0.072*
	(0.123)	(0.038)	(0.068)	(0.041)
Pro-immigration team (or researcher)	-0.124**	0.017	-0.113**	-0.005
	(0.044)	(0.034)	(0.042)	(0.025)
Δ: Pro – Anti	-0.274**	0.087**	-0.167**	0.068*
	(0.125)	(0.035)	(0.080)	(0.035)
B. Also weighted by referee score				
1. % of team that is anti-immigration	0.089	-0.132**	---	---
	(0.105)	(0.056)		
% of team that is pro-immigration	-0.175**	-0.021	---	---
	(0.067)	(0.039)		
Δ: Pro – Anti	-0.264**	0.111**	---	---
	(0.111)	(0.045)		
2. Anti-immigration team (or researcher)	0.163	-0.066*	0.056	-0.072*
	(0.132)	(0.036)	(0.069)	(0.039)
Pro-immigration team (or researcher)	-0.139**	0.016	-0.124**	-0.006
	(0.046)	(0.033)	(0.045)	(0.024)
Δ: Pro – Anti	-0.302**	0.081**	-0.180**	0.066*
	(0.134)	(0.032)	(0.084)	(0.034)

Notes: * $p < .1$; ** $p < .05$. All regressions are linear probability models. Standard errors reported in parentheses and are clustered at the team level in the team-level regressions, and at the researcher-team-model level in the researcher-model level regressions. The binary dependent variable is set to unity if the AME estimate is below (above) the 10th (90th) percentile and has a *t*-value greater than |1.645|. The regressions in Panel A are weighted by the inverse of the number of models in the team-model level regressions, and by the inverse of the product of the number of models and researchers in the team in the researcher-model level regressions. The regressions in Panel B are also weighted by the mean referee score awarded to the model. The team-model level regressions have 1,253 observations; the researcher-model level regressions have 2,680 observations. 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.

Table 6. Ideological bias and research quality

	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

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.

Table 7. Specification choices made by different types of teams

		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 fixed effects	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

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).

Table 8. Determinants of expected AME

	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)
Δ: 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

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.

Data Appendix: Variables

To allow for easy replication, this appendix describes the variables used in this study. The experimental data produced by BRW (2022) 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.

Pro-immigration 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.

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 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. 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.

***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*); controlling for variation at the country-year level (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.