AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

Local Immigration and Support for Anti-Immigration Parties: A Meta-Analysis

Sara Cools Institute for Social Research
Henning Finseraas Norwegian University of Science and Technology
Ole Rogeberg Frisch Centre for Economic Research

Abstract: *Does the share of immigrants in a community influence whether people vote for anti-immigration parties? We conduct a systematic review of the causal inference literature studying this question. We collect estimates from 20 studies and develop a new Bayesian meta-analysis framework to account for both between-study heterogeneity in effect sizes and the possibility of reporting bias. Although meta-analysis methods that do not adjust for reporting bias suggest a moderate effect of local immigration, our Bayesian model finds that the effect of local immigration on far-right voting is on average negligible once we account for reporting bias. However, the analysis also reveals a large heterogeneity in effects across contexts, suggesting that local immigration may be important for anti-immigration vote shares in certain settings.*

Verification Materials: The data and materials required to verify the computational reproducibility of the results, procedures and analyses in this article are available on the *American Journal of Political Science* Dataverse within the Harvard Dataverse Network, at: https://doi.org/10.7910/DVN/TEPAK4.

Immigration has become a prominent political issue in recent decades (Grande, Schwarzbözl, and Fatke 2019), with increasing public support for anti-immigration parties in many European countries (Arzheimer 2018). A rapidly growing research literature has addressed this political shift to the extent that even some researchers in the field view the attention as disproportionate (Arzheimer 2018; Mudde 2013). Despite this attention, there is a lack of consensus on whether immigration causally triggers anti-immigrant voting. Although Arzheimer's literature review (2018, p. 160) argues that immigration tends to increase the vote share of the radical right, Golder (2016 p. 485) notes that only a minority of studies have employed research designs that allow for causal inference.[1]

In this article, we conduct a meta-analysis of the causal inference literature. In a systematic search of the literature, we screen papers against a set of criteria, the first of which requires that a study employs a research design developed explicitly to address bias due to selection and reverse causality. We identify 20 studies with a total of 147 qualified estimates. From the 147 estimates, we choose one main estimate per estimation technique per study, resulting in a sample of 31 estimates that we use in the main analysis.

A simple unweighted average across the 31 estimates indicates that a 1 percentage point increase in immigrant share is associated with a 0.57 percentage point increase in the vote share of anti-immigration parties. This average masks a substantial heterogeneity, as individual

Sara Cools is Research Professor, Institute for Social Research, P. O. Box 3233, Elisenberg, 0208 Oslo, Norway (sara.cools@samfunnsforskning.no). Henning Finseraas is Associate Professor, Norwegian University of Science and Technology P. O. Box 8900, Torgarden, 0208 7491 Trondheim, Norway (henning.finseraas@ntnu.no). Ole Rogeberg is Research Professor, Frisch Centre for Economic Research Gaustadalleen 21, 0349 Oslo, Norway (ole.rogeberg@frisch.uio.no).

We would like to thank Don Green, Eirik Strømland, and seminar participants at the Frisch Centre, Hertie School of Governance, NTNU, University of Oslo, OsloMet, the 2019 EPSA conference, and the 2019 Meeting of Norwegian Political Scientists for useful comments. Grant number 270687 (Norwegian Research Council) is acknowledged.

[1]See also Amengay and Stockemer (2018), who count the share (38%) of estimations that yield a positive and statistically significant coefficient for immigration on radical right voting. Kaufmann and Goodwin's (2018) meta-analysis finds that increases in ethnic diversity tend to be associated with more negative views on immigration.

American Journal of Political Science, Vol. 65, No. 4, October 2021, Pp. 988–1006

DOI: 10.1111/ajps.12613

estimates range from −0.04 to +2.36 percentage points. The variability across estimates reflects a mix of statistical sampling variability, true effect heterogeneity, reporting and publication bias, and specification bias. We find a strong and systematic negative relationship between statistical precision and effect magnitudes in the reported estimates, which is, in our opinion, most likely the result of reporting bias. At the same time, we expect substantial effect heterogeneity across contexts, related, for instance, to a country's immigration history and its party system.

Traditional meta-analytic approaches are not designed to account for both effect heterogeneity, and reporting bias, which limits their usefulness for our purpose. To assess the evidence in light of the pattern we find in variability between studies, we therefore develop a Bayesian selection model that allows for both reporting bias and effect heterogeneity. It models how the observed estimates are selected from the underlying latent distribution of (published and unpublished) estimates, making it possible to recover the parameters of the latent effect distribution we are interested in.

Our results can be summarized as follows: Conventional random effects (REs) meta-analyses suggest that a 1 percentage point increase in immigration on average raises the vote share of anti-immigration parties by about 0.4 percentage points. We argue that this estimate is strongly inflated by publication bias, as the point estimate from the Bayesian bias model is close to 0 with a 95% credibility interval that does not include 0.4. All models agree, however, that there is substantial effect heterogeneity, limiting the extent to which the results of any single study can be generalized across contexts. In the conclusion, we discuss the implications of our results for the research on immigration and voting.

Immigration and Anti-Immigration Voting

We limit our examination to studies that estimate causal effects of immigration on anti-immigration voting. In practice, research designs appropriate for causal inference have used variation in immigration at the local level to assess how immigration influences voting patterns. A number of mechanisms have been suggested through which local immigration would influence how people vote. Most of them predict that an increase in the anti-immigration vote will follow from increased immigration.

The first line of argument concerns the labor market. Although most research in labor economics finds small effects, on average, of immigration on wages and employment (Dustmann, Schönberg, and Stuhler 2016), immigration could still have a significant impact in certain parts of a local labor market (see, e.g., Dustmann, Frattini, and Preston 2012). Those who suffer negative economic consequences may be mobilized to vote against immigration and, according to sociological conflict theory, coethnics that are not themselves personally affected by the shock may be moved in the same direction (Quillian 1995).

A second line of argument concerns the provision and quality of local services and welfare benefits (Cavaille and Ferwerda 2018). In contexts where local authorities have important responsibilities in welfare provision (e.g., health services, public education, social assistance, public housing), immigration—like other types of rapid demographic change—may put pressure on the quality of services and spur competition for welfare benefits, particularly if public budgets are under strain. These issues were prominent in the U.K. Brexit debate over immigration (Becker and Fetzer 2017).

A third line of argument involves problems of social integration. Much research finds that areas that have both high immigration and high levels of unemployment and poverty also tend to have higher support for anti-immigration parties (Arzheimer 2018). Often, crime is also higher in these areas. Anti-immigration parties tend to run on law-and-order platforms, arguing that social problems in such areas result from liberal immigration and crime policies (Dinas and van Spanje 2011).

The final type of argument is that natives might have direct preferences about the ethnic, linguistic, or religious composition of the neighborhood—what Card et al. (2012) label "compositional amenities." Changes in the share of immigrants can be viewed as a threat to cultural, demographic, and local identifications and could increase uncertainty about the future (Newman and Velez 2014; Kaufmann 2018). These types of cultural concerns are typically considered the most important ones in the sociological and political science literature on attitudes to immigration (Hainmueller and Hopkins 2014).

These mechanisms predict that anti-immigration parties would benefit in elections from an increase in immigration at the local level. However, for many voters, whether immigrants are present in ones local community might be less important than the regional or national level of immigration. This could be because immigration might affect labor markets outside where immigrants settle (Borjas 2003), concerns about the effect of immigration on central government budgets, or their notion of national culture (Sides and Citrin 2007). The same is true if anti-immigration voting is driven by, for example, regional (Alba and Foner 2017) or economic (Piketty

2018) inequality, distrust of established parties and elites (Akkerman, Mudde, and Zaslove 2014), media coverage of immigration debates (Hopkins 2010), or other antiglobalization views (Colantone and Stanig 2018). Such mechanisms would cause the effects of immigration in one region to spill over into other regions, diluting estimates of the effect of immigration at the local level.

Finally, it should be noted that a higher immigrant share might also be related to reduced vote shares for anti-immigration parties, as implied by the contact hypothesis. This holds that positive interethnic contact might increase with immigrant share and lead to reduced ethnic prejudice and anti-immigrant voting (see Paluck, Green, and Green 2019, for a review).

Data

In Table 1, we show the studies we found in our systematic search and remained after the application of a set of selection criteria, both of which are described below. These 20 studies and their 31 selected main estimates constitute the data for our meta-analysis.

Search Method

We first did a structured search in Google Scholar using the search string "(Immigration OR refugees OR asylum seekers) AND (vote share OR election result OR voting) AND (far right OR extreme right OR radical right OR populist right OR anti-immigration OR right-wing)." Next, we did a less structured search in Google Scholar using word phrases such as "immigration," "elections," and "causal."—or—"immigration, elections, causal." Next, we closely examined the reference lists in the identified studies. Finally, we used Google Scholar to search through all studies that cite any of the identified papers.

Selection Criteria

We established four criteria to select studies for inclusion in our meta-review. The first criterion requires that the study employs a *research design* that was explicitly developed to address bias due to selection and reverse causality.[2] As voters have direct influence over where they live and thus vote, studies of the effects of immigration face the difficult task of disentangling immigration effects from confounding factors like compositional effects and correlated contextual effects such as local unemployment. Moreover, in aggregated election data, there is a potential ecological fallacy problem because one cannot separate vote shares by natives and immigrants.

By limiting our review to studies with reasonably strong internal validity, we exclude a large and influential research tradition on far-right voting that aims to build comprehensive empirical models by simultaneously including "demand factors" such as immigration and unemployment and "supply factors" such as party system and electoral rules.[3] In this literature, which typically relies on cross-national data, the aim is to simultaneously model as many theoretically important factors as possible rather than to identify the causal effect of one particular variable. This research—reviewed in Arzheimer (2018)—has made important contributions to the understanding of far right voting (see, e.g., Arzheimer 2009), but we believe it is less appropriate for pinning down the specific causal impact(s) of one variable.[4] We therefore limit our review to studies with strong internal validity that use established causal inference designs to identify effects from spatial and longitudinal variation in immigrant and election shares within a single country.

The second criterion requires the use of *continuous measures of immigration and vote shares*, which enables cross-study comparison. The only study we identified that fulfills the other criteria but not this one is Steinmayr (2016), which estimates the relationship between vote shares for the far-right Fredom Party of Austria (FPÖ) and an indicator variable of whether the municipality housed migrants after the migrant crisis.

The third criterion restricts analysis to studies of European parties that are typically defined as *anti-immigration parties.* Most of these parties are defined as "populist radical right parties" by Mudde (2013, p. 3). We include Norway's Progress Party in this definition; it is the main anti-immigration party in Norway and is sometimes considered a "new radical right" party (Norris 2005).

The fourth and final criterion restricts our study to *national and local elections* and thus excluded European Parliament elections. Immigration policy is decided at the national level, and elections to the European Parliament have lower turnout and media coverage. This criterion excludes Becker and Fetzer's (2017) study of

[2]The research designs discussed in Angrist and Pischke (2009) can serve as a reference point.

[3]Demand factors are variables that influence the demand for restrictive immigration policies, whereas supply factors are variables that influence the formation of new parties and how easy it is for such parties to gain representation.

[4]Our view is similar to those expressed in Aronow and Samii (2016). One important point is that less emphasis on internal validity does not necessarily imply better external validity.

TABLE 1 Descriptives of Assembled Studies

Paper	Country	Party	Election type & period	Immigration variable	Method	Main estimate	*S.E.*
Gerdes and Wadensjö (2008)	Denmark	FrP	National elections, 1989–2001	Log share non-Western	FE IV	0.16^{**} 0.44	(0.05) (0.32)
Mendez and Cutillas (2014)	Spain	Anti-immig	National elections, 1996–2011	Delta African pop share	IV	$0.04^{\dagger}$	(0.02)
Otto and Steinhardt (2014)	Germany	Extreme right	National elections, 1987–1998	Share of foreigners	FE	0.14^{***}	(0.04)
			Pooled elections, 1987–1998		IV	0.30^{**}	(0.11)
Barone et al. (2016)	Italy	Extreme right	National elections, 2001–2008	Share of immigrants	IV	0.25^{**}	(0.08)
Sekeris and Vasilakis (2016)	Greece	GD	National elections, 2012–2015	Share of refugees	FE IV	0.05^{***} 0.05^{**}	(0.01) (0.02)
Sørensen (2016)	Norway	FrP	National elections, 1977–2009	Share of non-Western	FE	0.30^{*}	(0.13)
Halla, Wagner, and Zweimüller (2017)	Austria	FPÖ	National elections, 1979–2013	Share of foreign citizens	FE	0.16^{***}	(0.04)
				Delta 20-year percent share of foreign citizens	IV	$0.08^{\dagger}$	(0.04)
Brunner and Kuhn (2018)	Switzer-land	SVP	National elections, 1970–2010	Share of culturally dissimilar immigrants	IV	1.66^{*}	(0.68)
Caselli, Fracasso, and Traverso (2018)	Italy	Far right	National elections, 1994–2008	Share of foreign born	FE	0.81^{***}	(0.20)
			National elections, 1994–2009		IV	2.00^{***}	(0.30)
Chasapopou-los, van Witteloost-uijn, and Boone (2018)	Nether-lands	Radical right	National elections, 2003–2012	Share of foreign-born non-Western	FE	0.43^{*}	(0.18)
Chletsos and Roupakias (2018)	Greece	Far right	National elections, 2004–2012	Share of immigrants	FE IV	$0.29^{\dagger}$ 0.32^{**}	(0.15) (0.11)

(Continued)

TABLE 1 (Continued)

Paper	Country	Party	Election type & period	Immigration variable	Method	Main estimate	*S.E.*
Dal Bó et al. (2018)	Sweden	SD	National elections, 2002–2014	Share of immigrants	FE	−0.04	(0.11)
Dinas et al. (2019)	Greece	GD	National elections, 2012–2015	Refugee arrivals per capita	FE	0.60**	(0.18)
			National elections, 2015		IV	0.74***	(0.17)
Harmon (2018)	Denmark	DF + FrP	Local elections, 1981–2000	Delta share non-Western	FE	0.70**	(0.20)
			National elections, 1980–2001		IV	1.06***	(0.28)
Vasilakis (2018)	Greece	GD	National elections, 2015	Inflow of refugees in 1 month	FE	0.06***	(0.01)
				Inflow of refugees in 3 months	IV	0.15***	(0.04)
Dustmann et al. (2019)	Denmark	DF	National elections, 1990–1997	Delta share of allocated refugees	IV	1.58***	(0.42)
Edo et al. (2018)	France	Far right	National elections, 1998–2012	Delta share of immigrants	FE	0.42**	(0.12)
			National elections, 1988–2012		IV	2.36***	(0.62)
Mehic (2019)	Sweden	SD	National elections, 2014–2020	Immigration rate	FE	0.12	(0.08)
			National elections, 2014–2019		IV	1.89*	(0.95)
Schaub, Gereke, and Baldassarri (2019)	Germany	AFD	National elections, 2017	Share of refugees	OLS	0.11	(0.10)
Tomberg, Stegen, and Vance (2019)	Germany	Far right	National elections, 1998–2017	Share of refugees	IV	0.40**	(0.15)

$^{\dagger}p < .10$, $*p < 0.05$; $**p < 0.01$; $***p < 0.001$.
Note: The table lists the studies included in our meta-analysis and their main estimates, standard errors (in parentheses), and *t*-values (in brackets). FE = fixed effects; IV = instrumental variable; OLS = ordinary least squares.

immigration and vote shares for the UK Independence Party (UKIP).

Overview of the Studies

The studies listed in Table 1 show that, although many types of research designs emphasize internal validity and causal inference (as required by our first criterion), the methods used in the set of studies that fit all four of our criteria are limited to *two-way fixed effects/first difference (FE/FD) models* and *instrumental variable (IV) regressions.* Schaub et al. (2019) analyze a cross-section of municipalities, but they rely on a seeming natural experiment (the rule-based allocation of refugees during the crisis of 2015), which makes the refugee share plausibly exogenous. They also control for the outcome in the previous election, implying that the interpretation of estimates approximate that of an FE model. Most papers present both FE/FD and IV estimates.

The selected studies use data from all geographic regions of Europe, but studies from Scandinavia and Greece are overrepresented. There is a mix of papers that estimate the effect of immigration on the success of a single party and papers where small anti-immigration parties are pooled together. A majority of the papers date from 2018 or 2019, and eight are working papers. Most papers were written by economists; among the 12 papers that have been published, only 2 appeared in political science journals. Eight were published in economic journals, whereas one was published in a migration journal and the other as a book chapter. 24 of 31 estimates (77%) are statistically significant at the 5% level.

Most of the studies measure immigrant share as the share of foreign-born citizens, but there is some variation. Some studies allow effects to differentiate for immigrants from different regions, whereas some distinguish between refugees and other immigrants.

Sixteen of the estimates are based on instrumental variable methods. Different types of instrumental variables are employed, with the chain migration instrument being the most common. This instrument uses historical settlement patterns to predict later immigration. Three studies, all from Greece during the refugee crisis, use distance from Turkey as an instrument for the share of refugees. Two studies, both from Denmark, use a placement policy to derive potential exogenous variation in immigration. Finally, one study relies on immigration in a broader area as the instrumental variable, whereas another uses public housing policy as an instrument.

Studies also differ with regard to what they control *for* (though most include some controls for labor market situation and demographic characteristics) and *how* (e.g., the number of controls and functional forms). Few studies discuss their choices on these issues in much detail. When collecting estimates from the studies, we were struck by the varied practices in reporting descriptive statistics and in the care taken when interpreting coefficients. Halla et al. (2017) is an exemplar for future studies to follow. For most specifications, the authors report specification-specific means and standard deviations for the dependent and independent variables and describe each specification in detail in table notes.

Main Estimates

Several of the identified papers present multiple estimates from different specifications that all fulfill the specified criteria. From the set of estimates, we select one for each research design or identification strategy used in the article. Thus, for papers presenting estimates using both FE/FD and IV research designs, we include two main estimates, one for each empirical approach. The main estimates and their standard errors are also reported in Table 1. The supporting information (SI) includes details on the selection of main estimates (SI:1). Although we identified main estimates, we collected all the relevant estimates in the papers and present robustness checks where we use the full set of coefficients.

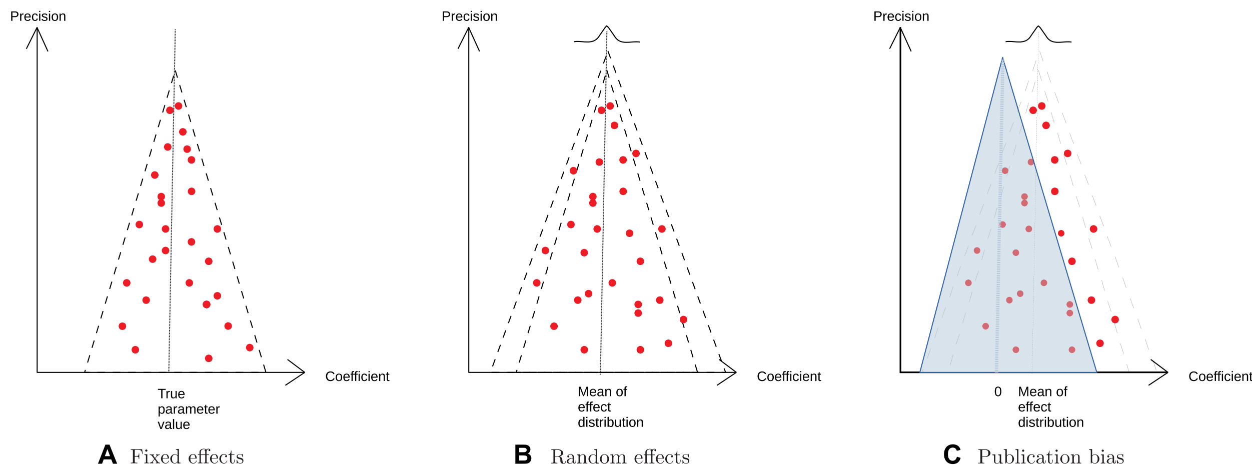
Analysis

The aim of a meta-analysis is to pool evidence across studies in order to increase statistical power and precision. The estimates reported in Table 1 vary considerably both in size and in precision. There are several a priori reasons why estimates of the same causal relationship may differ between studies, which we discuss in the next sections.

FE Meta-Analysis

The simplest reason why estimates of the same causal relationship may differ between studies is sampling variability: Coefficient estimates based on repeated independent samples from the same population will be normally distributed around the true parameter value, with interstudy variation determined by the standard error of the estimator. If the studies vary in sample size, a scatter plot of the point estimate and standard error of each sample would show 95% of estimates falling within a funnel

FIGURE 1 Funnel Plots Under Different Scenarios



Note: The figure shows hypothetical funnel plots under three different scenarios.

extending 1.96 standard errors in each direction, as shown in Figure 1, panel (A). The correct meta-analytic method in such a setting would be an FE meta-analysis equal to the average of the estimates weighted by the inverse of their squared standard error.

The estimated pooled effect using an FE meta-analysis is 0.08 (see the top row in Table 2). This is substantially below the unweighted average of 0.57, which tells us that there is an inverse relationship between precision and estimated effect size.[5]

[5]See Figure A1 in the SI (p. 2) for details. The figure shows that 70% of the weights comes from three estimates from Greece with substantially higher precision than the rest of the studies (Sekeris and Vasilakis 2016).

TABLE 2 Meta-Analyses Without Correction for Reporting Bias

	Estimate	SE	τ
Constant effects (FE)	0.08	0.01	-
REML	0.39	0.09	0.35
Maximum-likelihood	0.38	0.08	0.34
Bayesian w/o priors	0.39	0.08	0.37
Hedges	0.45	0.10	0.59
Sidik–Jonkman	0.44	0.10	0.53
Empirical Bayes	0.43	0.09	0.48
Hunter–Schmidt	0.21	0.05	0.09
DerSimonian–Laird	0.22	0.05	0.10

Note: $N = 31$. The table reports meta-estimates with standard errors and heterogeneity estimates (τ) using different meta-analysis methods.

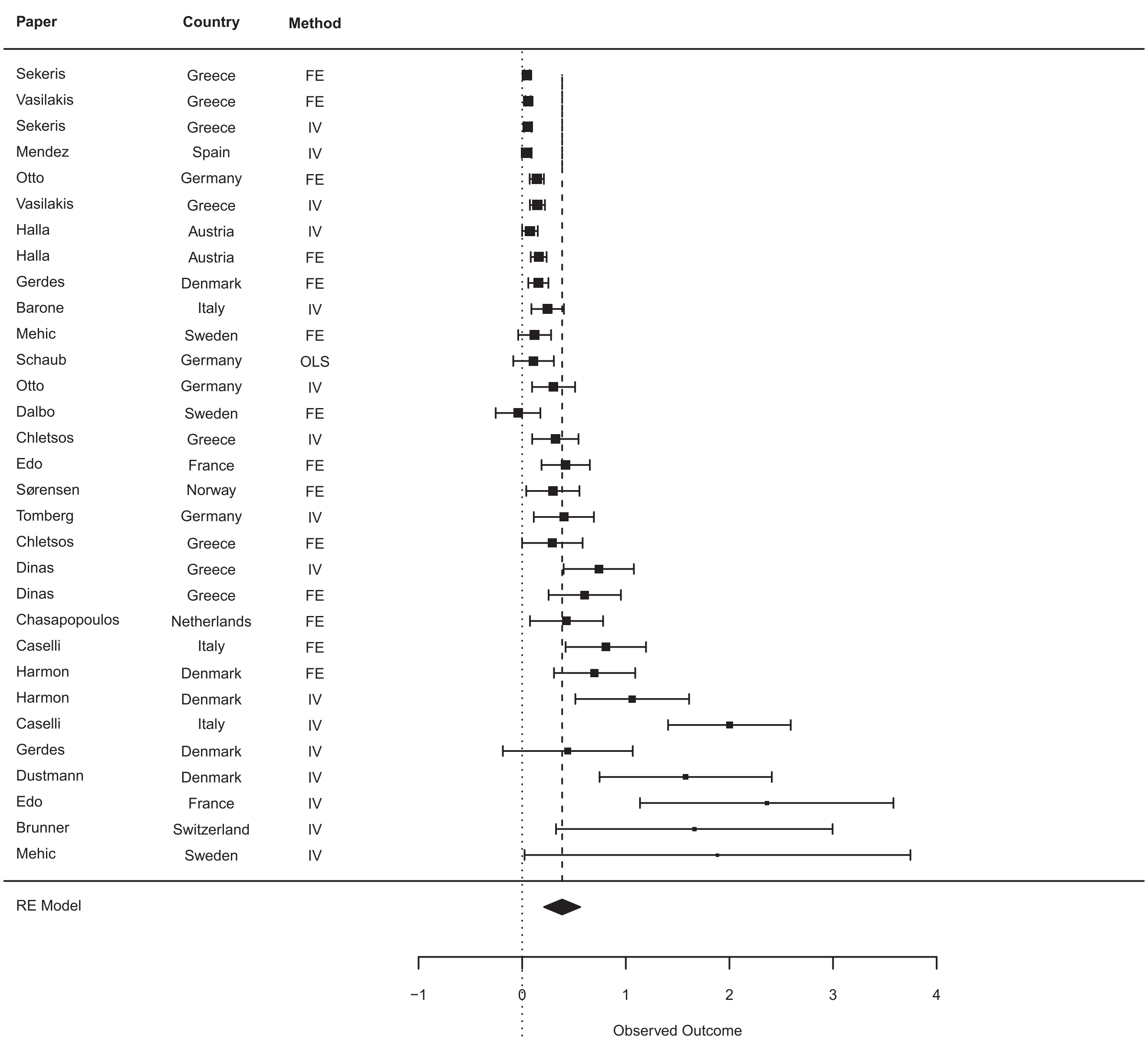
RE Meta-Analysis

The FE model is only appropriate if all studies estimate the same underlying effect. This is not plausible in our present context, as the estimates use data from different populations and time periods. There may be additional effect heterogeneity within populations, causing different IV estimates to identify local average treatment effects (LATEs) relevant for different subsets of the population. The impact of immigrant share on the anti-immigrant vote will also plausibly vary with, for example, electoral and party system, characteristics of the native or immigrant population, historical experiences, and institutional differences in areas like labor markets or benefit systems.

When the underlying parameter being estimated differs across studies, estimates will differ more than we would expect from sampling variability alone: Even studies with perfect precision would now yield different estimates because they estimate different (but related) parameters. In a point-estimate versus precision plot, this produces a broader funnel (see Figure 1, panel (B)), and the additional variation is used to infer the presence and size of parameter heterogeneity across studies. The appropriate model in such a case is called an RE meta-analysis. RE meta-estimates can be estimated under different assumptions and thus produce somewhat different results, especially when the number of included studies is low.[6]

[6]In an RE meta-estimate, each effect estimate is weighted by the inverse of the sum of its variance and the estimated variance in the distribution of true effects, but the latter can be estimated in different ways. Note that although the FE model restricts inference

FIGURE 2 Forest Plot, REML Model



Note: Dots refer to point estimates, lines to 95% CIs. The dotted line shows the meta-estimate. Studies are sorted by their standard errors.

Results from a range of RE models estimated on our data set are shown in Table 2, and Figure 2 shows the Forest plot for the restricted maximum likelihood (REML) model.[7] The estimates are fairly similar across approaches, typically around 0.4 percentage points, with the exception of DerSimonian–Laird and Hunter–Schmidt, two approaches with a known negative bias when the number of studies is low (Veroniki et al. 2016).

Correcting for Reporting Bias

The above models assume that all results—significant or not—have similar probabilities of being written up and

to the population in the included studies, RE models allow out-of-sample inference (see Borenstein et al. 2009, pp. 83–84).

[7]We use the implementation in Viechtbauer (2010) for the frequentist estimates and (Röver 2020) for the Bayesian model. We use the Knapp–Hartung modification of the frequentist confidence interval (CI) to account for potential bias due to a small number of studies (Guolo and Varin 2017). The Bayesian model is estimated using so-called improper, uninformative priors for the estimate, its standard deviation, and between-study heterogeneity.

published. If results are more likely to be published when they are statistically significant and align with the field's expectations regarding coefficient sign, this will skew the published record and create a systematic relationship between precision and coefficients in the published literature (Figure 1, panel (C)). Applying standard meta-analytic techniques to the selection-distorted sample of observed estimates will give biased estimates of the true pooled effect.

To recover the true mean and variance of the effect distribution requires a technique that corrects for selection into publication while allowing for effect heterogeneity. To our knowledge, there are no off-the-shelf methods available for doing this. Two common approaches to adjusting for publication bias, WAAP (Stanley, Doucouliagos, and Ioannidis 2017) and FAT-PET-PEESE (Stanley and Doucouliagos 2014), ignore effect heterogeneity and are better suited to an FE setting (see SI, p. 5, for details and results when using these methods on our sample).

To address both effect heterogeneity and publication bias, we propose a Bayesian inference model where results may have different probabilities of being written up and/or accepted for publication depending on their statistical significance and coefficient sign.[8]

We start with a simple data-generating process consistent with the standard assumptions of an RE model: Researchers estimate a true effect θ_i drawn from a normally distributed effect distribution, $N(\mu, \sigma)$. Their estimate, $\hat{\theta}_i$, is drawn from a normal distribution centered on the true parameter value, $N(\theta_i, se_i)$, where se_i denotes the study's standard error. Conditional on a standard error, estimated coefficients will now be normally distributed, with variation that reflects both effect heterogeneity and sampling variability:

$$\hat{\theta}|se_i \sim N(\mu, \sqrt{\sigma^2 + se_i^2}).$$

This baseline model has no publication bias, and serves as a Bayesian analogue to the earlier RE models. We use f to denote the probability density function.

Using a 5% significance level, let s be a rule that assigns a "significance type" $k \in K = \{s-, ns, s+\}$ to any estimate $\hat{\theta}_i$ with standard error se_i, with negative and statistically significant coefficients denoted as $s-$, non-significant coefficients denoted as ns, and positive and statistically significant coefficients denoted as $s+$.

We can now allow for publication bias by letting the publication probability vary by significance type. The probability that a new result will be observed in the bias model is then:

$$P(observed|se_i) = \sum_{k \in K} P(observed|k, se_i)P(k|se_i).$$

Because the significance thresholds are determined by the standard errors, the probability of drawing different result types can be found using the cumulative distribution function.

Given parameter values, our model can be used to express the probability that an observation will take a specific value. This probability is the likelihood contribution of such an observation. The probability of observing the full data set—the likelihood of the data—is the product of these probabilities. Using the definition of a conditional probability, the likelihood contribution of any single observation $\hat{\theta}_i$ can be written as

$$\begin{aligned} P(\hat{\theta}_i|observed, se_i) &= \frac{P(\hat{\theta}_i \wedge observed|se_i)}{P(observed|se_i)} \\ &= \frac{P(observed|\hat{\theta}_i, se_i)\, f(\hat{\theta}_i|se_i)}{P(observed|se_i)} \\ &= \frac{P(observed|k = s(\hat{\theta}_i, se_i))\, f(\hat{\theta}_i|se_i)}{P(observed|se_i)} \\ &= \frac{\tau_{k=s(\hat{\theta}_i, se_i)}\, f(\hat{\theta}_i|se_i)}{\sum_{k \in K} \tau_k P(k|se_i)}. \end{aligned}$$

Here, τ_k is the *relative* publication probability of results of significance type k, using $s+$ as the reference (i.e., with $\tau_{s+} = 1$). Using the relative publication probabilities is needed to identify the parameters, as we lack information on the total number of (observed and unobserved) estimates and cannot identify the absolute publication probabilities.

Two modeling assumptions should be noted: First, as in Andrews and Kasy (2019), standard errors are assumed to be uncorrelated with the estimated underlying effect parameter. This assumption would be violated, for instance, if researchers increased sample sizes and included more control variables only to the extent that the resulting gains in precision were needed in order to distinguish the expected effect size from zero. If researchers have approximate knowledge of $\hat{\theta}_i$ prior to their study and aim for, say, 80% power, this could generate a similar pattern to that of publication bias. This possibility is discussed further in the "Publication Bias versus Alternative Interpretations" section. Second, we rely on the

[8]This model is similar to a recent maximum likelihood approach (Andrews and Kasy 2019) but differs in that it requires priors for the true effect, effect heterogeneity, and publication bias. Priors tend to improve stability when the sample of studies is small, as even substantively weak priors can rule out implausible parameter values, though it does require researchers to defend how their prior choices reasonably encode prior knowledge.

assumption that the publication probability of significant relative to nonsignificant results is independent of their standard error. In other words, we assume that if a positive significant result is twice as likely to be published as a nonsignificant result, then this is equally true for precise and imprecise studies.[9]

The bias model shares two parameters of interest with the baseline model (mean and variance of the effect distribution), in addition it has two bias parameters for the relative publication probabilities that (may) have distorted the publication record. Each parameter requires a prior that summarizes prior knowledge. With small data samples, these priors are expected to affect the inference.

For the mean of the effect distribution in both the baseline and the bias models, we use a normally distributed zero-centered prior with a standard deviation of 0.5 for the average effect. This reflects a belief that the true average effect is unlikely to exceed 1 in absolute value (the standard methods reported above gave point estimates in the range 0.08–0.45). For the standard deviation of the effect distribution in both models, we use an inverse gamma prior with parameters 2 and 0.75. This distribution ensures that the prior is positive and that the effect heterogeneity is unlikely to be smaller than 0.1 or larger than 2 (the standard methods reported above gave τ estimates in the range 0.1–0.6). Although we believe these priors are reasonable in the present context, they should not be taken to be default priors for all applications of the model.

We estimate three variants of the bias model, where we vary how the relative publication probability parameters, τ_k, are specified and assigned priors:

1. **Neutral prior model—two parameters, neutral priors:** In this version of the model, the publication probabilities for *s*−, *ns*, and *s*+ results vary freely relative to each other, resulting in two independent parameters. The publication probabilities can take any value in the $[0, \infty]$ range, and take the value of 1 when there is no bias. As a neutral prior, we use an exponentiated zero-centered normal prior with a standard deviation of 2. This prior has a median value of 1, and implies that we find it equally likely that the bias parameter is above 2 or below 0.5 (both possibilities have a 36% chance under the prior), or above 10 or below 0.1 (both with about 12% probability).
2. **Single bias parameter model—one parameter, neutral prior:** In this version of the model, *s*− and *ns* results have the same publication probability relative to *s*+ results. Limited information in the study sample for identifying two publication bias parameters makes the case for using a single parameter for both *s*− and *ns* studies. We assign it the neutral prior described above.
3. **Assumed bias model—two parameters "assumed bias" priors:** In this version of the model, all type results are again allowed to have independent relative publications probabilities, resulting in two parameters to be identified (as in the neutral prior model). Rather than a neutral prior, however, they are assigned priors that reflect the belief that publication bias is likely. In this model we use a gamma prior with parameters 1 and 2. This prior assigns only a 13% probability of reverse publication bias (i.e., that insignificant coefficients or coefficients with unexpected sign have a larger publication probability than significant coefficients with the expected sign), and implies that substantial probability bias is possible. It gives a 63% probability that *s*+ results are more than twice as likely to be published than *s*− or *ns* results, and 18% prior probability that they are more than ten times as likely to be published.

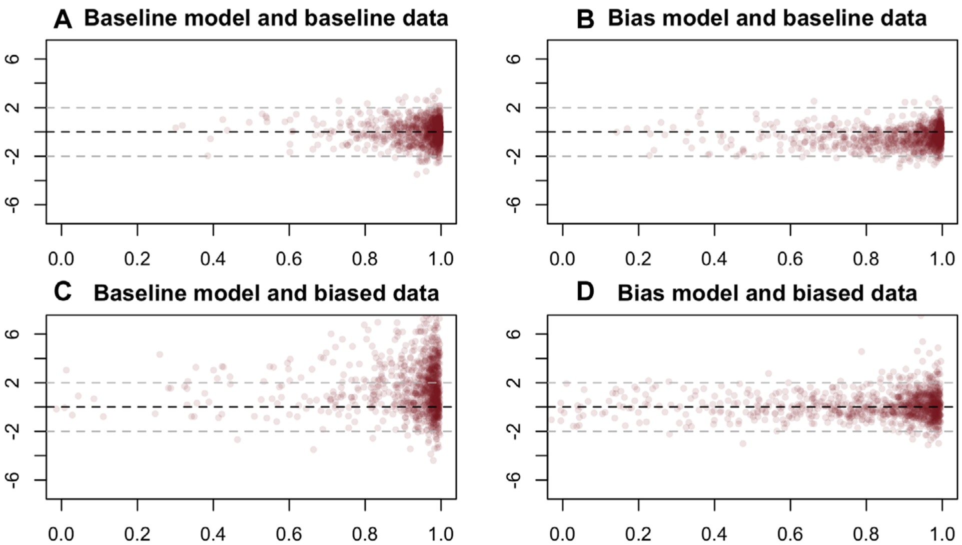
The models were coded in the probabilistic modeling language Stan and estimated using Rstan v 2.19.2.

The estimation of each model results in a random sample of parameter values from that model's posterior distribution. The posterior distribution summarizes the parameter uncertainty that remains after the prior distribution has been updated in light of the data, and a comparison of the prior and posterior distributions for each parameter shows us the extent to which the data were informative for the different parameters.

To test the model, we first draw 1,000 samples from the prior distribution of each parameter. For each set of parameter draws, we generate 31 synthetic observations with the same standard errors as in our actual study sample, resulting in 1,000 data samples generated by a wide range of parameter values. Because we know the true parameter values generating each simulated data set, we can compare estimates produced by our inference model to the true parameter value. Following Betancourt (2020), each estimation is summarized using two scores: the posterior *z*-score, which measures how our point estimate

[9] If nonsignificant studies are published—but only if they are precise—the model would view these nonsignificant studies as evidence of a weaker publication bias. Because all the imprecise studies would be significant, the model would infer that these just randomly happened to estimate larger effects θ_i.

FIGURE 3 Testing Model Performance on Simulated Data



Note: Posterior *z*-score and posterior contraction for the average effect parameter; 1,000 synthetic data sets were generated using random draws from the parameter prior distributions in both the baseline (no publication bias) and extended (with publication bias) models. Both models were used to estimate the average effect in the effect distribution, and the posterior *z*-score and posterior contraction were calculated for each estimation.

varies around the true parameter value,[10] and the posterior contraction, which expresses how precisely the parameter is identified relative to the prior uncertainty.[11]

We use this setup also to compare the risks of using a misspecified model: by drawing 1,000 simulated data sets from the baseline model without publication bias as well as 1,000 simulated data sets from the assumed bias model. By using both models to draw inference from each synthetic data sample, we can compare the consequences of ignoring publication bias when it is present, to the consequences of allowing for it when it is not.

This gives us the four scatter plots for the estimated average effect in the effect distribution, shown in Figure 3. Panels A and D show the results when the inference model is correctly specified relative to the data-generating process. In both cases, the posterior *z*-score is scattered within a small range symmetrically around 0, and the posterior contraction is typically high, signifying that the parameter is well identified by the data. Panels B and C show the results of the two types of misspecification. Allowing for publication bias when it is not present (panel B) leads to a slight underestimation of the effect mean parameter on average, but the posterior *z*-scores remain concentrated within a reasonable [−2, 2] range. Analyzing data generated with publication bias using a model that does not allow for it (panel C), however, causes the *z*-scores to spread out to higher values, indicating a combination of positive bias and misleading precision.

These tests indicate that the model allowing for publication bias yields substantially more precise inference: When the data were from the nonbiased process, the true parameter was outside the 95% credibility intervals 4.1% of the time when using the no-bias model and 4.7% when using the bias model. When the data were from the biased process, however, the true parameter was outside the 95% credibility intervals 39% of the time when using the no-bias model, and 6.1% when using the extended model.

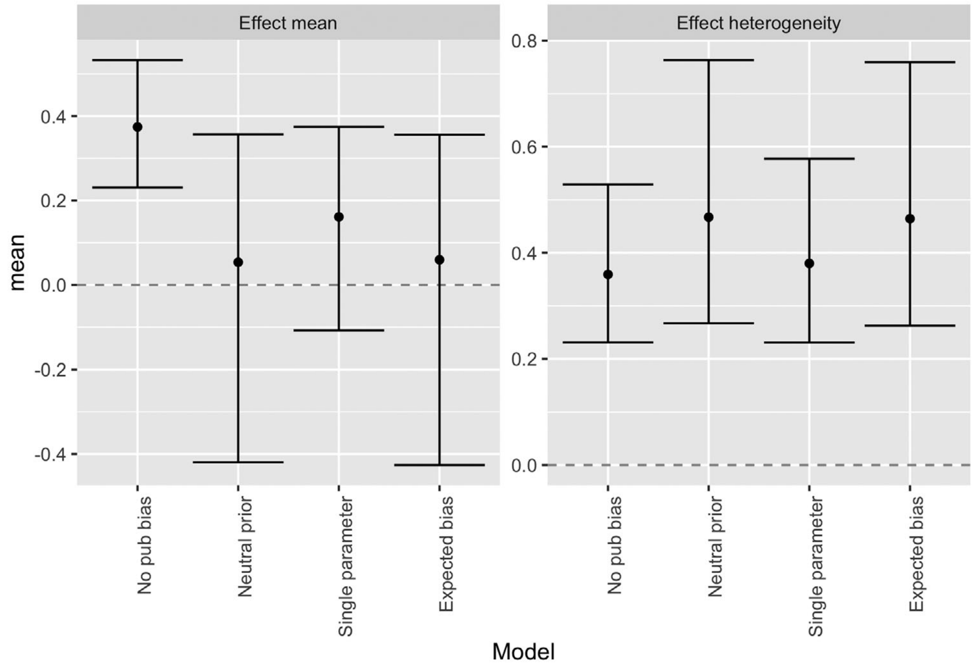
Having tested how the two main models perform on synthetic data, we estimate the baseline model and the three variants of the bias model on our actual data sample. All chains converged and diagnostics were fine (see details in SI, p. 8). Results for the key parameters across all models are shown in Figure 4. Results for all models and parameters are available in Table A2 (SI, p. 4).

Beginning with the estimates for the effect mean (left panel), the first thing to note is that the baseline model without publication bias gives almost identical results to

[10] The posterior *z*-score is the difference between the posterior average (used as a point estimate) and the true parameter value, scaled by the standard deviation of the posterior samples.

[11] The posterior contraction is found by dividing the standard deviation of the posterior samples of a parameter by the standard deviation of the prior samples of a parameter, and subtracting the answer from 1.

FIGURE 4 Bias Models—Results



Note: The figure shows the posterior mean and 95% credibility intervals for four Bayesian models: (1) a baseline model without publication bias, (2) an extended model with two bias parameters (for results type *s*− and *ns* relative to *s*+) assigned symmetric neutral priors, (3) an extended model with one bias parameter (for *s*− and *ns* combined, relative to *s*+) assigned a symmetric neutral prior, and (4) the extended model with two bias parameters—but with a prior that expects bias favoring type *s*+ results.

a standard RE model estimated using the REML algorithm. This confirms that the core of the Bayesian model is a reasonable analogue to standard tools.

Second, all models allowing for publication bias indicate substantially smaller effect means and have similar upper bounds on their 95% credibility intervals (∼0.3) that exclude six of the eight point estimates from the standard REs models shown in Table 2.

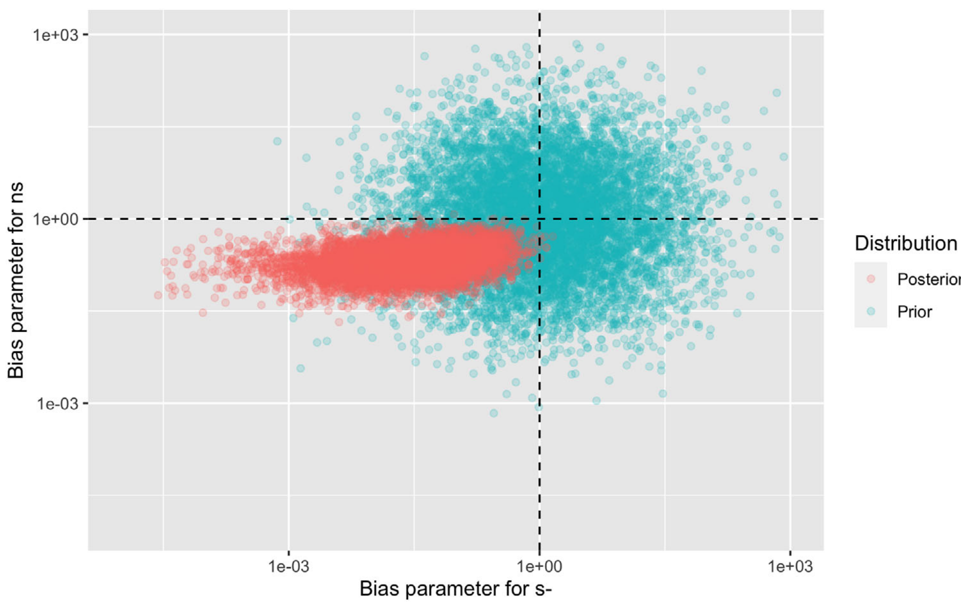
Third, the greatest difference between the bias models concerns the lower bound of the credibility intervals. Our data sample contains no statistically significant studies with the "wrong" (i.e., negative) sign, even though we would expect such studies to show up, given the large effect variation we estimate and the large standard errors we observe. When results with negative sign are given their own bias parameter, the model consequently estimates a very high publication bias for this type of results. And this, in turn, means that there *could* be unpublished studies that would pull the true effect mean substantially down—hence the lower value of the lower bound in the two-parameter models.

Turning to effect heterogeneity across studies (right panel in Figure 4), the point estimates are more similar across models. Not surprisingly, however, the precision of these estimates is strongly reduced when the publication record is potentially distorted by bias.

In sum, the results show that the data from our meta-analytic sample are strongly consistent with what we would see if there were extensive publication bias. This is particularly evident when we compare samples from the prior and posterior distributions of the neutral bias model (Figure 5). Although the prior on a logarithmic scale gives equal probability to bias in both directions, the posterior draws are strongly concentrated in the quadrant where both bias parameters are below 1, with particularly strong bias inferred for *s*− studies.

One of our key inferential questions concerns the true (latent) distribution of causal effects. Although this distribution is characterized by the mean and variance parameters inferred above, these parameters are both imprecisely inferred and correlated in the posterior distribution (see Figure A5, SI. p. 13). To average across the (correlated) uncertainty in these parameters, we assess the posterior predictive distribution: For each parameter draw from the posterior distribution, we generate 100 REs from the estimated effect distribution. Pooling these, our model indicates a 50% probability that an effect will be in the [−0.3, 0.35] range, and a 95% probability that

FIGURE 5 Comparison of Samples from Prior and Posterior Distributions



Note: Scatterplot of paired samples from the prior and posterior distribution of the neutral bias model. Values below the dashed lines indicate publication bias favoring s+ studies; values exceeding the dashed lines indicate reversed publication bias.

it will be in the [−1.1, 1] range. Thus, although local immigration does not appear to be important in the average case, the average case conceals a large amount of effect heterogeneity—to the extent that estimates at the upper and lower bounds should be considered as politically important effects. These findings also imply that single study estimates have limited external validity given the large effect variation indicated by the literature as a whole.

Publication Bias versus Alternative Interpretations

We find publication bias to be the most plausible explanation for the strong negative correlation between precision and effect sizes in our meta-analytic sample. We find it plausible that research ideas are more likely to be pursued vigorously if they show promise in the sense of finding statistically significant results in a direction consistent with theory and earlier estimates. Such estimates are also likely to face less critical scrutiny from editors and referees than a paper that "finds nothing," or has a significant result that implies the opposite of what researchers believe is true. Furthermore, publication bias can explain strong discontinuities in the distribution of published significance values: *p*-Values that are "barely significant" are substantially more common than those "barely non-significant" (Gerber and Malhotra 2008). Publication bias also provides the simplest explanation for the fact that most published research findings are statistically significant (Gelman, Skardhamar, and Aaltonen 2020). Finally, Kaufmann and Goodwin's (2018) meta-analysis does not find clear signs of reporting bias as captured by the systematic relationship between effect size and precision. The authors suggest that the reason might be that they include studies where immigration is merely a control variable in the analysis, that is, that there is no reporting bias when there are no stakes attached to finding a statistically significant coefficient on immigration.

Having made this case, however, we accept that the relationship between studies' precision and their effect size may have other explanations. The most credible alternative hypothesis, in our view, is that precision is costly and that researchers incur the cost of increased precision only when this is required to distinguish effects from zero. This hypothesis assumes, in addition to precision being costly and researchers caring about attaining statistical significance rather than overall precision, that researchers can accurately predict the size of the effect they plan to study. We find these assumptions less reasonable than those supporting the publication bias interpretation. Although the costs of randomized trials clearly scale with the number of participants, when researchers use administrative data, it is less clear that extending data periods or analyzing aggregate data from a larger set of municipalities adds substantial cost. If a researcher believes to have found a context where the effect is considerably larger than is typically the case, the precision of the estimate establishing this result would be seen as highly valuable and worth the extra cost. Finally, we doubt that researchers on average are accurate in predicting the effects they estimate. Requiring researchers to preregister hypotheses and analyses strongly reduces the probability of significant results (Kaplan and Irvin 2015), and preregistered replications tend to find smaller effects and fewer significant results than we should expect had the original results been random draws from the effect distribution (Camerer et al. 2018).

To further assess the plausibility of this hypothesis, we suggest a simple Bayesian model (see SI, p. 10). In terms of intuition, we assume that all researchers draw an effect θ_i and have some awareness of its level prior to running their statistical study. Next, they choose a standard error that achieves some "optimal" statistical power, and because everyone agrees on the optimal power level, their standard errors differ only to the extent that they expect their specific effects θ_i to differ. Defining statistical power as the probability that an estimate $\hat{\theta}_i$ will be of type $s+$, we can now use the observed share of type $s+$ studies to infer the optimal power level. With our inferred optimal power in hand, we can ask for each individual study: "What true effect would a researcher have to assume for their study to have optimal power with this observed standard error?"

Estimated on our study sample, the model implies that researchers in this field have aimed for a power of 0.8 (95% CI = [0.7, 0.9]) when this parameter is given a flat prior on the 0–1 range. The estimated power level is used to infer the ex ante effect size expected by the authors of the different studies under the optimal power assumption, and the actual estimates reported are compared to these (inferred) expectations to see how precise the expectations were. Although we assign the precision parameter a broad prior in our model (with only 13% probability of being below 1 in absolute value), the estimated prediction error is estimated at close to zero (0.02, 95% CI = [0.00, 0.05]). Put differently, this model implies that all studies—including the most imprecise—are highly powered relative to ex ante predictions, and that researchers in this field have highly accurate and well-calibrated expectations regarding the effects they are studying. In other words, according to this view, research essentially always confirms researchers' exact a priori beliefs.

P-Hacking and Forking Paths

The models discussed so far share the assumption that a study's reported standard error accurately describes how an effect estimator $\hat{\theta}_i$ is distributed around some true effect θ_i, whereas the scale parameter σ describes how the true effect sizes θ_i are distributed around the average effect μ. In reality, however, there is always more than one estimate available to a researcher, in the sense that the final estimate is shaped by choices made throughout the research process. Researchers have to make judgment calls and different researchers estimating the same effect on the same data may consequently reach quite different results (Orben and Przybylski 2019). Although somewhat speculative, we can attempt to briefly assess how such issues would influence our results and our interpretation of them.

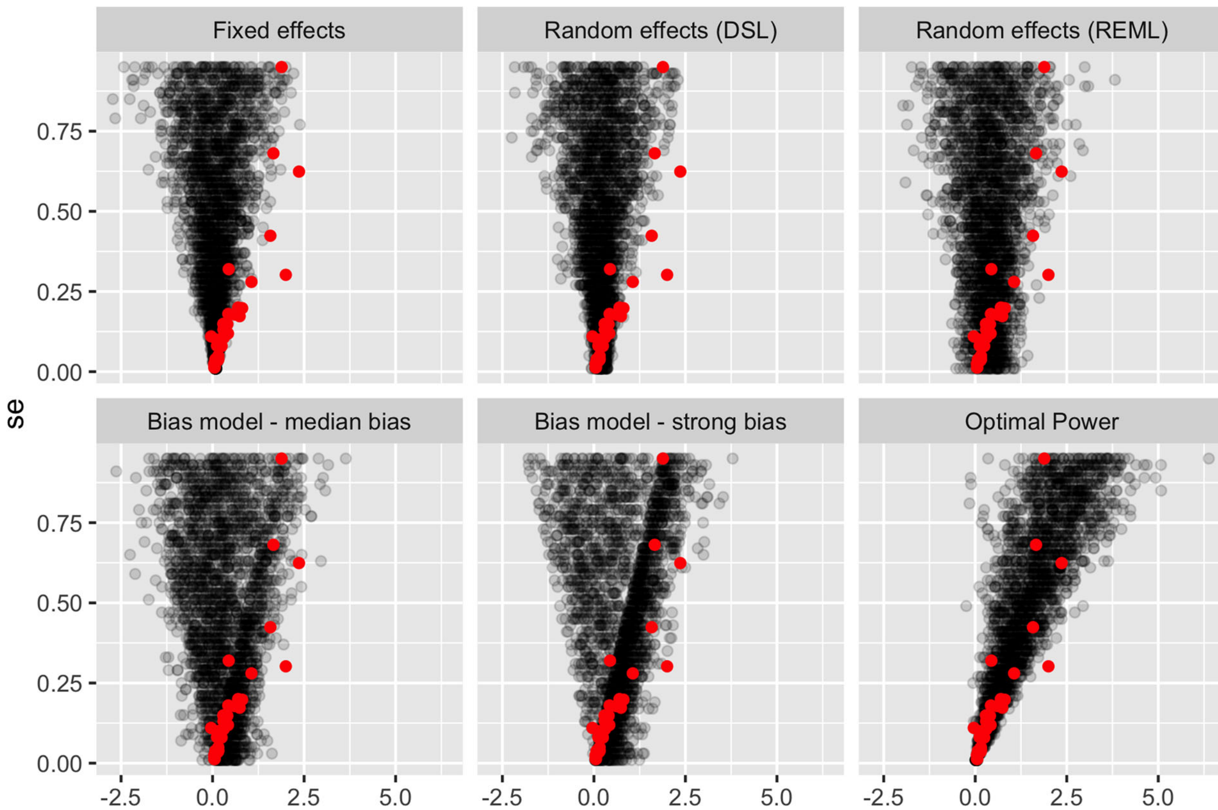
An estimator can be thought of as a lottery ticket that may or may not give a type $s+$ result when applied to a given data set. When there are multiple choices that can be made, resulting in different (all defensible) estimates, the p-hacker simply tries out one lottery ticket after another until they find a winner. With "normal" publication bias, the researcher whose study estimates a θ_i with a low value will typically get a nonsignificant result and shelve the study. A p-hacker will find some specification that makes even this study significant and then publish. Because such studies will tend to be only marginally significant, they will increase the share of published estimates with p-values right below the threshold of significance. This influences the distribution of observed p-values, and a p-curve analysis will interpret this as evidence of a lower evidential value of the literature as a whole.

This problem is exacerbated when we allow for research errors and limited knowledge, which implies that the choice of estimator is made from a larger set that also includes biased and flawed specifications that are not identified as such by researchers or peers.[12] This increases the dispersion of results, giving the appearance of increased effect heterogeneity.

Importantly, we do not believe that extreme and deliberate p-hacking is widespread. However, similar (albeit perhaps dampened) problems follow from the more widespread practice of working with the data while refining the empirical strategy—likened by Gelman and Loken (2014) to walking through a "garden of forking paths." The heterogeneity in model specifications that we observe in the literature (see the discussion in the

[12]As an example, consider early analyses that found spurious results by using two-way FE specifications with incorrectly specified standard errors (Bertrand, Duflo, and Mullainathan 2004).

FIGURE 6 Expected and Actual Pattern of Coefficients and Standard Errors by Model



Note: Each panel uses estimates from a specific meta-analytic model to generate 50 simulated "published studies" for standard errors ranging from 0.01 to 0.96 in increments of 0.02. The actual studies are superimposed. Two plots are included from the bias model—one using the median estimated bias, the other using the boundary value from the 95% credibility interval.

"Overview of the Studies" section) shows a lack of agreement on what is the correct specification. The ip-value expresses the probability of obtaining an estimate as extreme as the current one if the true effect is zero and we used our current estimator on multiple, independent random samples from the same population. If, however, researchers refine, adjust, and shape their analysis to the specific samples drawn, then the estimator used on the repeated samples will vary, and the true probability of getting significant results on the repeated samples may be well above the stated *p*-value.

Conclusion

Reviews of the literature on the growth of anti-immigration parties, such as Golder (2016) and Arzheimer (2018), typically discuss the role of immigration in explaining the rise of these parties, but they reach conflicting conclusions. Arzheimer (2018) concludes that immigration is important, whereas Golder (2016) argues that causal evidence is weak. Our article supplements conventional reviews of the literature by collecting and analyzing estimates of the relationship between local immigration and the vote shares of anti-immigration parties from studies that employ estimation strategies explicitly addressing the issues related to causal inference.

Most published studies find a positive relationship between immigration and anti-immigration parties' vote shares. In our data, conventional meta-analytic RE models suggest that a 1 percentage point increase in immigration increases the anti-immigration vote share by 0.4 percentage points. The magnitude implies political importance in areas where immigration is increasing rapidly. However, the available studies show a clear pattern of effect estimates declining with increased power and precision, which we believe reflects (potentially strong) reporting bias in the literature. To account for this, we introduce a Bayesian meta-analytic model that jointly accounts for effect heterogeneity and reporting bias. Using this model, we find that the average effect of immigration is likely substantially below the estimates from standard RE models, with a point estimate close to zero. Although there is important uncertainty in this estimate, and evidence of substantial effect variation, these results suggest that immigration will only be of political importance in certain contexts.

As a technique, systematic meta-analyses are used to identify a comprehensive set of studies that address the same research question in order to pool and summarize the evidence base as a whole. In our case, different meta-analytic techniques produce widely differing results. This may appear to replace one problem with another: Instead of cherry-picking individual studies to support your prior beliefs, you can now cherry-pick meta-analytic models to do the same. We argue, however, that most of these models are based on assumptions that are implausible given the data at hand.

To see this in our case, in Figure 6 we superimpose the actual estimates from our sample on a plot displaying the way precise and imprecise estimates would be distributed under the assumptions integral to the different approaches discussed above.

The plots make a simple point: The strong correlation between estimates and standard errors in our study sample is extremely unlikely unless you believe in either publication bias or the optimal power model. Starting with the FE model, the most precise studies fit neatly into the sharp "tip" indicating the true (uniform) effect parameter, but implying that almost every study with a standard error above 0.15 is an extreme outlier. The RE models make the imprecise estimates more plausible by broadening the funnel, but this implies a trade-off as it makes the model less able to explain why all the precise studies seem to converge on similar values. The REML and the DSL (i.e. DerSimonian-Laird) RE models choose different points on this trade-off curve. Turning to the Bayesian bias model, this predicts that studies passing through the publication filter form a dense band within which most of the published estimates fall. The band is more striking, the stronger the bias assumed. The optimal power model, on the other hand, explains the same pattern as the outcome of well-informed researchers whose goal when choosing precision is to have an 80% chance of getting a statistically significant result. Ongoing changes in research practices, such as credibly preregistered expectations of effect sizes prior to data collection, should allow for additional comparisons of the optimal power and reporting bias models in the future.

One concern with using Bayesian models that adjust for reporting bias is that the meta-analytic results will be less precise than those estimated using standard models. This, however, is a feature, not a bug: The Bayesian model explicitly recognizes the possibility that reporting bias is skewing the published record, and uncertainty with regard to the strength of this bias *should* make us less certain as to what the true effects and effect variability are. Reporting bias implies that the published record is less informative as a whole, and ignoring it simply ensures that we are most likely wrong even when the results look precise.

We would like to highlight two important implications for research in the field. The first concerns the extent to which results generalize across contexts. Both the standard and the bias models indicate strong effect heterogeneity, which implies that results from one country are unlikely to generalize to other countries. It is important to realize that this is not necessarily a consequence of the strong emphasis on internal validity that characterizes the studies in our sample (see Aronow and Samii 2016). Across-study heterogeneity is widespread (Vivalt 2020), and an improved understanding of what it reflects and what it implies for external validity is required. For substantive research, it appears to be insufficient to study the role of objective factors without understanding when the issue is politicized, for instance, through the media (Hopkins 2010) or by political elites (Grande, Schwarzbözl, and Fatke 2019).

Second, the field appears to have been too preoccupied with voters' *local* exposure to immigration as a source for growth of anti-immigration parties. Clearly, such exposure can be important, but its emphasis in the literature seems exaggerated. This does not imply that the immigration issue is not an important driver of anti-immigration voting; our studies rely on within-country variation, which means that they are not informative about the impact of the national level of immigration. The studies estimate relative and not absolute effects of immigration, which is an important distinction. Immigration at the national level might be more important for voters than local immigration, and the immigration policy is decided at the national level. Estimating absolute effects is difficult because exogenous variation is harder to find, but is an important task for future research.

References

Akkerman, Agnes, Cas Mudde, and Andrej Zaslove. 2014. "How Populist Are the People? Measuring Populist Attitudes in Voters." *Comparative Political Studies* 47(9): 1324–53.

Alba, Richard, and Nancy Foner. 2017. "Immigration and the Geography of Polarization. " *City & Community* 16(3): 239–43.

Amengay, Abdelkarim, and Daniel Stockemer. 2018. "The Radical Right in Western Europe: A Meta-Analysis of Structural Factors." *Political Studies Review* 17(1): 30–40.

Andrews, Isaiah, and Maximilian Kasy. 2019. "Identification of and Correction for Publication Bias." *American Economic Review* 109(8): 2766–94.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton, NJ: Princeton University Press.

Aronow, Peter M., and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* 60(1): 250–67.

Arzheimer, Kai. 2009. "Contextual Factors and the Extreme Right Vote in Western Europe, 1980–2002." *American Journal of Political Science* 53(2): 259–75.

Arzheimer, Kai. 2018. "Explaining Electoral Support for the Radical Right." In *The Oxford Handbook of the Radical Right*, ed. Jens Rydgren. Oxford: Oxford University Press, 143–65.

Barone, Guglielmo, Alessio D'Ignazio, Guido de Blasio, and Paolo Naticchioni. 2016. "Mr. Rossi, Mr. Hu and Politics. The Role of Immigration in Shaping Natives' Voting Behavior." *Journal of Public Economics* 136: 1–13.

Becker, Sascha, and Thiemo Fetzer. 2017. "Does Migration Cause Extreme Voting?" Competitive Advantage in the Global Economy (CAGE) Working Paper No. 306: https://warwick.ac.uk/fac/soc/economics/research/centres/cage/manage/publications/306-2016_becker_fetzer.pdf (accessed November 25, 2020).

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. "How Much Should We Trust Differences-In-Differences Estimates?." *Quarterly Journal of Economics* 119(1): 249–75.

Betancourt, Michael. 2020. "Towards A Principled Bayesian Workflow." Manuscript, April 2020: https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html (accessed November 25, 2020).

Borenstein, Michael, Larry Hedges, Julian Higgins, and Hannah Rothstein. 2009. *Introduction to Meta-Analysis*. New York: John Wiley & Sons.

Borjas, George J. 2003. "The Labor Demand Curve *Is* Downward Sloping: Reexamining the Impact of Immigration on the Labor Market." *Quarterly Journal of Economics* 118(4): 1335–74.

Brunner, Beatrice, and Andreas Kuhn. 2018. "Immigration, Cultural Distance and Natives' Attitudes Towards Immigrants: Evidence from Swiss Voting Results." *Kyklos* 71(1): 28–58.

Camerer, Colin F., et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science Between 2010 and 2015." *Nature Human Behaviour* 2(9): 637–44.

Card, David, Christian Dustmann, and Ian Preston. 2012. "Immigration, Wages, and Compositional Amenities." *Journal of the European Economic Association* 10(1): 78–119.

Caselli, Mauro, Andrea Fracasso, and Silvio Traverso. 2018. "Globalization and Electoral Outcomes: Evidence from Italy." Working Paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3194705 (accessed November 25, 2020).

Cavaille, Charlotte, and Jeremy Ferwerda. 2018. "How Distributional Conflict over In-Kind Benefits Generates Support for Anti-Immigrant Parties." Working Paper: https://charlottecavaille.files.wordpress.com/2018/10/cc-paper.pdf (accessed Novembet 25, 2020).

Chasapopoulos, Panagiotis, Arjen van Witteloostuijn, and Christophe Boone. 2018. "Immigration and Electoral Support for the Radical Right: Evidence from Dutch Municipalities." In *Migration Policy in Crisis*, eds. Ibrahim Sirkeci, Emilia Lana de Freitas Castro, and Ülkü Sezgi Sözen. London, UK: Transnational Press, 109–32.

Chletsos, Michael, and Stelios Roupakias. 2018. "Immigration and Far-Right Voting: Evidence from Greece." Working Paper: https://mpra.ub.uni-muenchen.de/88545/1/MPRA_paper_88545.pdf (accessed November 25, 2020).

Colantone, Italo, and Piero Stanig. 2018. "The Trade Origins of Economic Nationalism: Import Competition and Voting Behavior in Western Europe." *American Journal of Political Science* 62(4): 936–53.

Dal Bó, Ernesto, Frederico Finan, Olle Folke, Torsten Persson, and Johanna Rickne. 2018. "Economic Losers and Political Winners: Sweden's Radical Right." Working Paper: http://perseus.iies.su.se/~tpers/papers/Draft180902.pdf (accessed November 25, 2020).

Dinas, Elias, and Joost van Spanje. 2011. "Crime Story: The Role of Crime and Immigration in the Anti-Immigration Vote." *Electoral Studies* 30(4): 658–71.

Dinas, Elias, Konstantinos Matakos, Dimitrios Xefteris, and Dominik Hangartner. 2019. "Waking Up the Golden Dawn: Does Exposure to the Refugee Crisis Increase Support for Extreme-Right Parties?" *Political Analysis* 27(2): 244–54.

Dustmann, Christian, Kristine Vasiljeva, and Anna Piil Damm. 2019. "Refugee Migration and Electoral Outcomes." *Review of Economic Studies* 86(5): 2035–91.

Dustmann, Christian, Tommaso Frattini, and Ian P. Preston. 2012. "The Effect of Immigration Along the Distribution of Wages." *Review of Economic Studies* 80(1): 145–73.

Dustmann, Christian, Uta Schönberg, and Jan Stuhler. 2016. "The Impact of Immigration: Why Do Studies Reach Such Different Results?." *Journal of Economic Perspectives* 30(4): 31–56.

Edo, Anthony, Yvonne Giesing, Jonathan Öztunc, and Panu Poutvaara. 2018. "Immigration and Electoral Support for the Far-Left and the Far-Right." Working Paper: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3167728 (accessed November 25, 2020).

Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102(6): 460–5.

Gelman, Andrew, Torbjørn Skardhamar, and Mikko Aaltonen. 2020. "Type M Error Might Explain Weisburd's Paradox." *Journal of Quantitative Criminology* 36(2): 295–304.

Gerber, Alan, and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3(3): 313–26.

Gerdes, Christer, and Eskil Wadensjö. 2008. "The Impact of Immigration on Election Outcomes in Danish Municipalities." IZA DP No. 3586: https://www.iza.org/en/publications/dp/3586/the-impact-of-immigration-on-election-outcomes-in-danish-municipalities (accessed November 25, 2020).

Golder, Matt. 2016. "Far Right Parties in Europe." *Annual Review of Political Science* 19: 477–97.

Grande, Edgar, Tobias Schwarzbözl, and Matthias Fatke. 2019. "Politicizing Immigration in Western Europe." *Journal of European Public Policy* 26(10): 1444–63.

Guolo, Annamaria, and Cristiano Varin. 2017. "Random-Effects Meta-Analysis: The Number of Studies Matters." *Statistical Methods in Medical Research* 26(3): 1500–18.

Hainmueller, Jens, and Daniel J. Hopkins. 2014. "Public Attitudes Toward Immigration." *Annual Review of Political Science* 17: 225–49.

Halla, Martin, Alexander F. Wagner, and Josef Zweimüller. 2017. "Immigration and Voting for the Far Right." *Journal of the European Economic Association* 15(6): 1341–85.

Harmon, Nikolaj A. 2018. "Immigration, Ethnic Diversity, and Political Outcomes: Evidence from Denmark." *Scandinavian Journal of Economics* 120(4): 1043–74.

Hopkins, Daniel. 2010. "Politicized Places: Explaining Where and When Immigrants Provoke Local Opposition." *American Political Science Review* 104(1): 40–60.

Kaplan, Robert M., and Veronica L. Irvin. 2015. "Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased Over Time." *PloS One* 10(8): 1–12.

Kaufmann, Eric. 2018. *Whiteshift: Populism, Immigration and the Future of White Majorities.* London: Penguin UK.

Kaufmann, Eric, and Matthew J. Goodwin. 2018. "The Diversity Wave: A Meta-Analysis of the Native-Born White Response to Ethnic Diversity." *Social Science Research* 76: 120–31.

Mehic, Adrian. 2019. "Immigration and Right-Wing Populism: Evidence from a Natural Experiment." Working Paper 2019:5, Department of Economics, Lund University: https://project.nek.lu.se/publications/workpap/papers/wp19_5.pdf (accessed November 25, 2020).

Mendez, Ildefonso, and Isabel M. Cutillas. 2014. "Has Immigration Affected Spanish Presidential Elections Results?." *Journal of Population Economics* 27(1): 135–71.

Mudde, Cas. 2013. "Three Decades of Populist Radical Right Parties in Western Europe: So What?." *European Journal of Political Research* 52(1): 1–19.

Newman, Benjamin J., and Yamil Velez. 2014. "Group Size versus Change? Assessing Americans Perception of Local Immigration." *Political Research Quarterly* 67(2): 293–303.

Norris, Pippa. 2005. *Radical Right: Voters and Parties in the Electoral Market.* Cambridge: Cambridge University Press.

Orben, Amy, and Andrew K. Przybylski. 2019. "The Association Between Adolescent Well-Being and Digital Technology Use." *Nature Human Behaviour* 3: 173–82.

Otto, Alkis Henri, and Max Friedrich Steinhardt. 2014. "Immigration and Election Outcomes—Evidence from City Districts in Hamburg." *Regional Science and Urban Economics* 45: 67–79.

Paluck, Elizabeth Levy, Seth A. Green, and Donald P. Green. 2019. "The Contact Hypothesis Re-Evaluated." *Behavioural Public Policy* 3(2): 129–58.

Piketty, Thomas. 2018. "Brahmin Left vs Merchant Right: Rising Inequality and the Changing Structure of Political Conflict." Working Paper, Paris School of Economics: http://piketty.pse.ens.fr/files/Piketty2018.pdf (accessed November 25, 2020).

Quillian, Lincoln. 1995. "Prejudice as a Response to Perceived Group Threat: Population Composition and Anti-Immigrant and Racial Prejudice in Europe." *American Sociological Review* 60(4): 586–611.

Röver, Christian. 2020. "Bayesian Random-Effects Meta-Analysis Using the Bayesmeta R package." *Journal of Statistical Software* 93(6): 1–51.

Schaub, Max, Johanna Gereke, and Delia Baldassarri. 2019. "Strangers in Hostile Lands: Exposure to Refugees and Right-wing Support." Working Paper: https://static1.squarespace.com/static/50e64c35e4b02b36141d5175/t/5d6941e5e543ab0001df9a3b/1567179243467/Schaub%2C+Gereke%2C+Baldassarri+2019+Refugees+WP.pdf (accessed November 25, 2020).

Sekeris, Petros, and Chrysovalantis Vasilakis. 2016. "The Mediterranean Refugees Crisis and Extreme Right Parties: Evidence from Greece." Working Paper: https://mpra.ub.uni-muenchen.de/72222/1/MPRA_paper_72222.pdf (accessed November 25, 2020).

Sides, John, and Jack Citrin. 2007. "European Opinion about Immigration: The Role of Identities, Interests and Information." *British Journal of Political Science* 37(3): 477–504.

Sørensen, Rune Jørgen. 2016. "After the Immigration Shock: The Causal Effect of Immigration on Electoral Preferences." *Electoral Studies* 44(December): 1–14.

Stanley, Tom D., and Hristos Doucouliagos. 2014. "Meta-Regression Approximations to Reduce Publication Selection Bias." *Research Synthesis Methods* 5(1): 60–78.

Stanley, Tom D., Hristos Doucouliagos, and John Ioannidis. 2017. "Finding the Power to Reduce Publication Bias." *Statistics in Medicine* 36(10): 1580–98.

Steinmayr, Andreas. 2016. "Exposure to Refugees and Voting for the Far-Right: (Unexpected) Results from Austria." IZA DP No. 9790: http://ftp.iza.org/dp9790.pdf (accessed November 25, 2020).

Tomberg, Lukas, Karen Smith Stegen, and Colin Vance. 2019. ""The Mother of All Political Problems"? On Asylum Seekers and Elections in Germany." Working Paper: http://hdl.handle.net/10419/203615 (accessed November 25, 2020).

Vasilakis, Chrysovalantis. 2018. "Massive Migration and Elections: Evidence from the Refugee Crisis in Greece." *International Migration* 56(3): 28–43.

Veroniki, Areti, Dan Jackson, Wolfgang Viechtbauer, Ralf Bender, Jack Bowden, Guido Knapp, Oliver Kuss, Julian Higgins, Dean Langan, and Georgia Salanti. 2016. "Methods to Estimate the Between-Study Variance and Its Uncertainty in Meta-Analysis." *Research Synthesis Methods* 7(1): 55–79.

Viechtbauer, Wolfgang. 2010. "Conducting Meta-Analyses in R with the Metafor Package." *Journal of Statistical Software* 36(3): 1–48.

Vivalt, Eva. 2020. "How Much Can We Generalize from Impact Evaluations?" *Journal of the European Economic Association* 18(6): 3045–89.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix A: Selection of main estimates
Appendix B: Additional results
Appendix C: WAAP and FAT-PET-PEESE adjustments for publication bias
Appendix D: Details on the Bayesian selection correction model