Bayesian evaluation of informative hypotheses for synthesizing evidence from diverse statistical models[*]

Lion Behrens   Simon Ellerbrock   Rebecca M. Kuiper[†]

August 15, 2019

Abstract. In light of current credibility and replication crises, the gold standard for evidence is an empirical result which is consistent across multiple studies. Currently, the most prominent approaches to the systematic integration of statistical evidence are meta-analysis and Bayesian updating. These methods combine individual effect estimates or posterior distributions and come with the drastic shortcoming that they can only be applied if homogeneous statistical models sharing a common functional form are underlying all studies of interest. Since social and political researchers typically tackle research problems using diverse statistical methods, their applicability remains limited in these domains. Remarkably, the framework of *informative hypotheses* has received no attention in political science research. As a counterframework to Null Hypothesis Significance Testing, we introduce Bayesian evaluation of informative hypotheses to the field of political science and suggest a procedure to aggregate evidence from any general statistical model. First, substantive expectations are translated into informative hypotheses that can go far beyond NHST. Second, evidence is aggregated over studies by updating prior and posterior model probabilities of informative hypotheses rather than point estimates or posterior distributions of regression coefficients. We provide a series of Monte Carlo simulations and demonstrate the use of our approach on a set of studies investigating determinants of right-wing voting.

Keywords: *Evidence synthesis, Bayesian estimation, informative hypotheses, meta-analysis, replication research.*

[*] Paper prepared for the 9th Annual Conference of the European Political Science Association, Belfast, June 20-22, 2019. We thank Irene Klugkist, Vincent Buskens and Werner Raub for helpful and profound discussions.

[†] Lion Behrens (corresponding author) is Researcher at the Collaborative Research Center 884, University of Mannheim. E-Mail: behrens@uni-mannheim.de. Simon Ellerbrock is Researcher at the Mannheim Center for European Social Research. Rebecca M. Kuiper is Assistant Professor at the Faculty of Social and Behavioural Sciences, Utrecht University. This study was supported by the Netherlands Organization for Scientific Research (NWO; VENI Grant 451-16-019).

Introduction

This manuscript suggests Bayesian evaluation of informative hypotheses as an approach to the formal integration of evidence from diverse statistical models. In science, the gold standard for evidence is an empirical result which is consistent across multiple studies. Throughout wide branches of the behavioral sciences, the replicability of findings has recently been seriously called into question (Open Science Collaboration 2015), leading certain scholars to even diagnose "credibility" and "replicability" crises. Also specifically in the political sciences, doubts about the robustness of empirical evidence are raised in recent years and several scholars have called for "meta-scientific introspection" (Wuttke 2019, Gill 2019). It is important to realize, however, that volatile results across studies should not immediately take political researchers by surprise. Even if a set of studies is researching a jointly underlying "true" effect in a certain population of interest, there is a whole array of reasons why empirical results might fluctuate and seem to lack robustness. The two main sources for this are as striking as they are simple: Most effects are small and most studies are not without issues. Evidently, when repeatedly sampling from small effect sizes, a mixture of positive, negative, and null results may be inconvenient, but naturally expected. Moreover, even medium-sized and more influential relations might not be detected in individual analyses as political scientists are frequently confronted with issues like selection bias, measurement error, lacking variation on key variables, and scarce sample sizes that might all distort correct effect estimates.

To counterbalance these problems, several methodological approaches for the formal integration of evidence that was retrieved from a variety of studies have been developed. First, meta-analysis has been suggested as a framework to estimate pooled effect sizes over all studies of interest, considering individual estimates and intra-study heterogeneity of effects (Cooper et al. 2009, Lipsey & Wilson 2001). Second, in Bayesian updating, evidence is combined by merging various posterior distributions via so-called *power priors* (Chen & Ibrahim 2000). Bayes theorem is repeatedly applied to a set of studies $t = 1, ..., T$, where initial prior specifications remain uninformative and the posterior distribution of Study t is used to define the prior distribution in Study $t + 1$. Repeatedly executing this process arrives at a posterior distribution that comprises information from all included studies. Against their large promises for the systematic integration of evidence across empirical research, such methodologies for evidence synthesis meanwhile take on a fixed place in the political scientist's toolbox (see, for instance, Doucouliagos & Ulubaşoğlu 2008, Costa 2017, Amengay & Stockemer 2018).

The current state-of-the-art methods come with an important shortcoming. Since regression coefficients or effect sizes are directly aggregated into a joint estimate or posterior distribution, integration is restricted to parameter estimates that share a common functional form and can be expressed in comparable units. Aggregating evidence that stems from diverse statistical models, on the other hand, is unfeasible, as coefficients from linear, logit, or count regressions would need to be directly transformed to a common scale. Notably, in the social and political sciences, studies typically employ very diverse statistical models to investigate a joint underlying problem. Thus, the applicability of current approaches for the formal integration of statistical evidence is, at best, limited. Remarkably, while being developed in the medical and psychological sciences for over a decade, there has been no attention to *Bayesian evaluation of informative hypotheses* in political science research. In this framework of hypothesis evaluation, substantive theoretical expectations with regard to one or multiple parameters of interest are confronted with empirical data using a series of (in)equality constraints like

$$
\begin{aligned}
&H_0 : \theta_1 = \theta_2 = \theta_3 \qquad \text{(Equality constrained)},\\
&H_1 : \theta_1 > \theta_2 > \theta_3 \qquad \text{(Inequality constrained)},\\
&H_2 : \{\theta_1 > \theta_2\} < \theta_3 \qquad \text{(Inequality constrained)}.
\end{aligned} \tag{1}
$$

These can represent various degrees of complexity and go far beyond the traditional null hypothesis testing framework.

Exploiting the framework of informative hypotheses, we suggest a Bayesian approach to evidence synthesis that is viable even when diverse statistical models are underlying and current methods for evidence synthesis fail. Our approach circumvents the "common scale" problems of meta-analysis and Bayesian updating by taking a *hypothesis evaluation* rather than an *estimation*-based approach. Researchers formulate parameter expectations across all studies of interest and construct scale- and model-independent evidence measures in every study using Bayes factors and posterior model probabilities. These measures are then aggregated over all analyzed studies which forms an overall quantification for the hypotheses at hand. In this way, researchers are facilitated to combine evidence regarding single or multiple quantities of interest even from a heterogeneous set of studies, and have the choice to do so via classical null, one-sided or far more complex hypotheses that directly represent their theoretical expectations.

The contribution of this manuscript is two-fold. First, we introduce the framework of informative hypotheses to the field of political science and show how researchers can use these to directly evaluate their theoretical expectations of interest against collected data (Section 2 and 3). Second, we outline a Bayesian approach how the evaluation of informative hypotheses can be used to tackle the problem of evidence synthesis from diverse statistical models (Section 4). Section 5 evaluates the proposed approach in a series of Monte Carlo simulations and shows that it renders correct results for various magnitudes of population effects, sample sizes, and hypothesis specifications. In Section 6, we demonstrate the use of our approach on a set of studies investigating determinants of right-wing voting.

Informative Hypotheses

This section lines out the basis of our approach and presents the framework of informative hypotheses. Usually, across the political sciences, our hypothesis tests are restricted to single variables for which we test an effect being different from zero. However, social science theories go beyond the falsification of null effects. Throughout many of our subfields, decades of theory-building have led to the emergence of clear theoretical expectations not only about the presence, but importantly the order of various individual effects.

A topic that received profound attention in recent years are citizens' attitudes towards parties that are located at the extreme right of the political spectrum. The most prominent explanations for spatial variations in electoral support for radical right parties point towards the importance of immigration as a driver of differences across social contexts. This line of research mostly draws on theories of group conflict, group threat, and social identity (Bélanger & Pinard 1991, Blumer n.d., Quillian 1995, Tajfel & Turner 1986), and intergroup contact theory (e.g. Allport et al. 1954, Pettigrew 1998). According to Pettigrew & Tropp (2008), immigration increases the opportunity for contact between natives and immigrants, which can under favorable conditions lead to a decrease in prejudices against immigrants, and by association to a decreased likelihood of voting for radical right parties. In contrast, proponents of *Realistic Group Threat Theory* (Jackson 1993) and *Ethnic Competition* (Bélanger & Pinard 1991) argue that increasing numbers of immigrants moving to a social context would result in greater perceived threat and thus lead to negative attitudes towards immigration. Figure 1 shows how we arrive at opposing hypotheses for the effect of an immigrant influx when incorporated in statistical analyses.

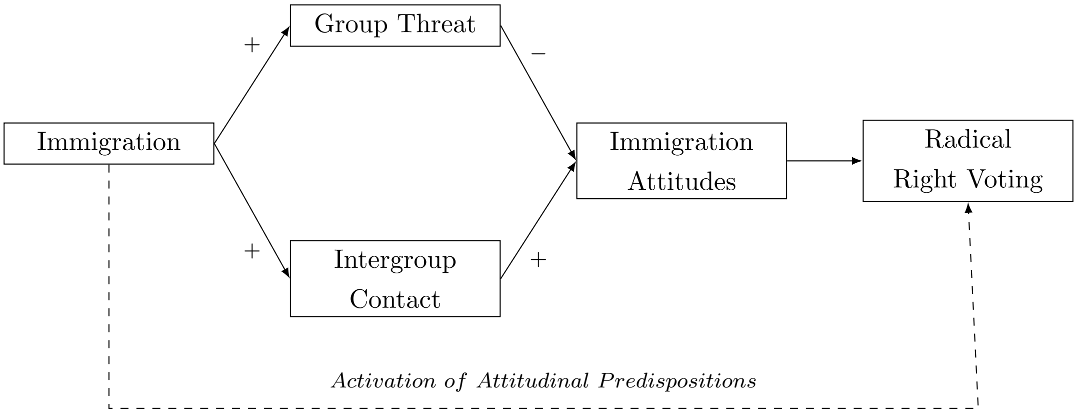


Figure 1 Two opposing hypotheses for the effect of immigration on radical right voting.

Let $\theta = [\theta_1, \theta_2, ..., \theta_K]^T$ be the parameters of a GLM, GLMM or SEM. Informative hypotheses H_m provide a framework to specify any number of expectations using equality (=), inequality (>, <) or the absence (,) of constraints regarding any number of parameters in θ. (In)equality constrained hypotheses take on the form of

$$H_m : S\theta = 0, R\theta > 0 \tag{2}$$

for $m = 1, ..., M$. S is a $q_s \times K$ matrix and R is a $q_r \times K$ matrix representing the equality and inequality constraints of Hypothesis H_m. The number of rows in S and R equals the number of equality (S) and inequality (R) constraints. The number of columns equal the length of θ. Often, researchers are interested in a single quantity of interest. Hence, in its most simple form, an informative hypothesis regarding one predictor of interest θ_1 can be formulated as $H_1 : \theta_1 > 0$. Additionally, complementing hypotheses $H_2 : \theta_1 < 0$ and $H_0 : \theta_1 = 0$ can be introduced. Jointly, these form the hypothesis set

$$H_M^1 = \begin{array}{l} H_0 : \theta_1 = 0 \\ H_1 : \theta_1 > 0 \\ H_2 : \theta_1 < 0 \end{array} . \tag{3}$$

For each hypothesis H_m in H_M^1, it holds that $\theta = [\theta_1]^T$ and $K = 1$. In order to express H_0 using (in)equality constraints,

$$S\theta = [1][\theta] = 0, R\theta = [0][\theta] > 0.$$

The equality constraint is specified in S, reading as $1 \times \theta_1 = 0$. Since H_0 does not include an inequality constraint, R is left empty. Hence, for H_0, it holds that $q_s = 1, q_r = 1$. Similarly, H_1 is defined as

$$S\theta = [0][\theta] = 0, R\theta = [1][\theta] > 0$$

and H_2 as

$$S\theta = [0][\theta] = 0, R\theta = [-1][\theta] > 0.$$

As we will outline in Section 3, Bayesian evaluation of informative hypotheses evaluates all specified hypotheses jointly by comparing the extent to which the data at hand is in line with each expectation. Although for many researchers, substantial expectations are limited to one quantity of interest, the possibilities of formulating informative hypotheses go far beyond this framework. For instance, researchers might be interested in multiple parameters $\theta = [\theta_1, \theta_2, \theta_3]^T$. One goal might be to investigate how these parameters *jointly* affect the outcome. In this multiple parameter case, the substantive hypotheses might be formulated as $H_1 : \{\theta_1, \theta_2, \theta_3\} > 0$ and $H_2 : \{\theta_1, \theta_2, \theta_3\} < 0$ to form the set

$$H_M^2 = \begin{array}{l} H_3 : \{\theta_1, \theta_2, \theta_3\} > 0 \\ H_4 : \{\theta_1, \theta_2, \theta_3\} < 0 \end{array}. \tag{4}$$

Here, H_3 is defined as

$$S = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0, \qquad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} > 0 \tag{5}$$

and H_4 is defined as

$$S = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0, \qquad R = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} > 0 \tag{6}$$

respectively. Taken even further, theoretical considerations often imply different estimates to be of different strengths. For example, a certain parameter θ_1 can be expected to affect an outcome stronger than θ_2 and θ_3, while the literature might be inconclusive about how the latter two are related. Here, a researcher might want to evaluate a set of hypotheses like

$$H_M^3 = \begin{array}{l} H_5 : \theta_1 > \theta_2 > \theta_3 > 0 \\ H_6 : \theta_1 > \{\theta_2 < \theta_3\} > 0 \end{array}. \tag{7}$$

Like H_3, H_5 still states that all coefficients in θ are positive, but additionally formulates order constraints in a way that $\theta_1 > \theta_2$, $\theta_1 > \theta_3$ and $\theta_2 > \theta_3$ while $\{\theta_1, \theta_2, \theta_3\} > 0$. Hence, H_5 extends the equality (S) and inequality (R) constraints of H_3 and is defined as

$$S = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 0, \qquad R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} > 0. \tag{8}$$

For instance, the fourth row in R reads as $1 \times \theta_1 + (-1) \times \theta_2 + 0 \times \theta_3 > 0$. The strength of informative hypotheses is that any kind of theoretical expectations towards the relation among parameters can be directly expressed and tested. Importantly, unstandardized regression coefficients of θ do not only consider the strength of relationship between y and x_1 versus x_2, but also inherit the scale of the respective predictor variables. Hence, when evaluating order constraints among predictor variables like in H_m^3, hypotheses consider standardized regression coefficients.

Bayesian Evaluation of Informative Hypotheses

So far we have introduced the concept of informative hypotheses and have shown how these can be constructed applying a set of equality and inequality constraints (S, R) to θ. As a difference to Null Hypothesis Significance Testing, it became apparent that these directly examine theoretical expectations towards the parameters of interest. Following, we show how such are evaluated. Bayesian evaluation of informative hypotheses uses the Bayes factor to quantify evidence for a set of competing expectations rather than the p-value. This section fulfills two goals. First, we build intuition on the Bayes factor in general. Second, we review a simple approach to construct computationally tractable Bayes factors for informative hypotheses given any general statistical model.

THE BAYES FACTOR

Assume we observe data y, X, where X is treated as fixed. Conceptually, from Bayes Theorem we obtain that the posterior probability of Hypothesis H_m given y is

$$P(H_m|y) = \frac{P(y|H_m)P(H_m)}{P(y)} \tag{9}$$

where the key term $P(y|H_m)$ expresses the probability that observed data y are generated under a specified hypothesis H_m. The key to Bayesian model comparison is to directly compare this quantity for different hypotheses H_m and $H_{m'}$. From Equation 9, it follows that the posterior odds of H_m against $H_{m'}$ are constructed by

$$\frac{P(H_m|y)}{P(H_{m'}|y)} = \frac{P(y|H_m)}{P(y|H_{m'})} \frac{P(H_m)}{P(H_{m'})}. \tag{10}$$

Table 1 Degree of evidence in favor of H_m when evaluated against a classical null hypothesis.

BF_{m0}	Evidence in favor of H_m
1 to 3	anecdotal
3 to 20	positive
20 to 150	strong
>150	very strong

In words, posterior odds = Bayes factor × prior odds. Hence, we can define the Bayes factor as

$$\begin{aligned} BF_{mm'} &= \frac{P(y|H_m)}{P(y|H_{m'})} \\ &= \frac{P(H_m|y)}{P(H_{m'}|y)} \frac{P(H'_m)}{P(H_m)} \\ &= \frac{P(H_m|y)/P(H_{m'}|y)}{P(H_m)/P(H_{m'})}. \end{aligned} \tag{11}$$

At this stage, it is important to note from Equation 11 that much like traditional model fit criteria like the AIC or BIC, the Bayes factor finds a compromise between two hypotheses' posterior and prior odds. Large posterior odds reward hypotheses that fit well given the data at hand as the Bayes factor grows with $P(H_m|y)$. Large prior odds penalize simplistic hypotheses that were very likely to find support from the data *a priori* and reward specific hypotheses that are harder to falsify as the Bayes factor diminishes with $P(H_m)$.

Once we parameterize H_m and $H_{m'}$ using $\theta = [\theta_1, \theta_2, ..., \theta_K]^T$ in empirical analyses as outlined in Section 2, we can rewrite Equation 11 to

$$\begin{aligned} BF_{mm'} &= \frac{m(y|H_m)}{m(y|H_{m'})} \\ &= \frac{\int f(y|\theta)p(\theta \in H_m)d\theta}{\int f(y|\theta)p(\theta \in H_{m'})d\theta}, \end{aligned} \tag{12}$$

and denote the Bayes factor $BF_{mm'}$ as the ratio of the two hypotheses' marginal likelihoods. The marginal likelihood $m(y|H_m)$ is defined as the likelihood of the data at hand $f(y|\theta)$ under all values in agreement with H_m given a certain prior distribution $p(\theta \in H_m)$. Hence, $BF_{mm'}$ comparatively quantifies the relative plausibility of two competing hypotheses given data at hand. This is fundamentally different from Null Hypothesis Significance Testing, which is restricted to the quantification of evidence *against* single point expectations. If the Bayes factor $BF_{mm'}$ is 10, this means that after observing the data, support for H_m is 10 times more likely than for $H_{m'}$ and vice versa. When $H_{m'}$ resembles a classical null hypothesis, Jeffreys (1961) and Kass & Raftery (1995) suggested guidelines to the interpretation of Bayes factors as given in Table 1. Note that an important reference value for the interpretation of the Bayes factor is the number 1. If $BF_{mm'} > 1$, H_m obtains more support from the data than $H_{m'}$. If $BF_{mm'} < 1$, $H_{m'}$ obtains more support from the data than H_m.

Construction of Bayes factors depends on the construction of the integrals $\int_\theta f(y|\theta)p(\theta)d\theta$. For some elementary cases, this integral might be evaluated analytically. Once the dimensionality in θ increases, integrals need to be approximated by numerical methods. As a curse of dimension, these can be extremely computationally heavy and numerically unstable. Note however that we are interested in the calculation of Bayes factors specifically for *informative hypotheses*. The last decade has whitnessed several studies dealing with the evaluation of informative hypotheses under the use of the Bayes factor. This approach was pioneered by Klugkist et al. (2005) in the context of ANOVA models and extended to contingency tables (Klugkist et al. 2010) and multivariate normal linear models (Mulder et al. 2010). Recently, Gu et al. (2017) have proposed a computationally tractable Bayes factor for evaluating informative hypotheses based on any general statistical model.

The key to this approach is the introduction of an unconstrained hypothesis

$$H_u : \theta_1, \theta_2, ..., \theta_k, \tag{13}$$

which does not empose any constraints and thus covers the parameter space in θ as a whole and to evaluate informative hpyotheses H_m against the unconstrained alternative H_u. Take Equation 12. Using the denotion of Chib (1995), we can rewrite the ratio of marginal likelihoods as

$$BF_{mu} = \frac{f(y|\theta)p(\theta_m)}{p(\theta_m|y)} \bigg/ \frac{f(y|\theta)p(\theta_u)}{p(\theta_u|y)} \tag{14}$$

where $f(y|\theta)$ is the sampling density, $p(\theta_m)$ is the prior density under H_m, and $p(\theta_m|y)$ is the posterior denstiy under H_m. Notably, all (in)equality constrained hypotheses as presented in Section 2 are nested under the unconstrained hypothesis H_u. Hence, we define $p(\theta_u)$ as an encompassing prior for θ given no constraints. As a result, the prior density of θ under any constrained hypothesis is proportional to $p(\theta_u)$ for the parameter space in θ that is covered by H_m and given by

$$p(\theta_m) = \frac{1}{c_m} p(\theta_u) \text{ if } \theta \in H_m, \tag{15}$$

where c_m is a normalizing constant for the prior density

$$c_m = \int_{\theta \in H_m} p(\theta_u) d\theta. \tag{16}$$

Similarly, the posterior density of θ under any constrained hypothesis is proportional to the posterior $p(\theta_u|y)$ under the unconstrained hypothesis H_u for the parameter space in θ that is covered by H_m and given by

$$p(\theta_m|y) = \frac{1}{f_m} p(\theta_u|y) \text{ if } \theta \in H_m, \tag{17}$$

where f_m is a normalizing constant for the posterior density

$$f_m = \int_{\theta \in H_m} p(\theta_u|y) d\theta. \tag{18}$$

When exploiting our denotion of the Bayes factor from Equation 14 and substituting (15) and (17), we see that

$$\begin{aligned} BF_{mu} &= \frac{f(y|\theta)\frac{1}{c_m}p(\theta_u)}{\frac{1}{f_m}p(\theta_u|y)} \Bigg/ \frac{f(y|\theta)p(\theta_u)}{p(\theta_u|y)} \\ &= \frac{\frac{1}{c_m}p(\theta_u)}{\frac{1}{f_m}p(\theta_u|y)} \Bigg/ \frac{p(\theta_u)}{p(\theta_u|y)} \\ &= \frac{\frac{1}{c_m}}{\frac{1}{f_m}} = \frac{f_m}{c_m}, \end{aligned} \tag{19}$$

which reduced Equation 12 to a simple fraction.[1] For equality constrained hypotheses, Equation 12 is equal to the known Savage–Dickey density ratio (Dickey 1971)

$$BF_{mu} = \frac{p(\theta \in H_m|y)}{p(\theta \in H_m)}. \tag{20}$$

The numerator f_m quantifies the fit of hypothesis H_m and is the percentage of the posterior probability mass of these parameters that are in agreement with H_m. The denominator c_m quantifies the complexity of hypothesis H_m and is the percentage of the prior probability mass that is in agreement with H_m. We see how this derivation of the Bayes factor for (in)equality constrained hypotheses is exactly in line with the intuition that we built in Equation 11. f_m is a measure of fit and rewards hypotheses that fit the data well, hence that obtain more probability mass in the region specified by H_m. c_m is a measure of how likely it was to find this hypothesis *a priori*, with vague hypotheses covering large parts of θ being penalized and specific hypotheses being rewarded. The Bayes factor comparing two informative hypotheses directly is then given by

$$BF_{mm'} = \frac{f_m/c_m}{f_{m'}/c_{m'}}. \tag{21}$$

Take the inequality constrained hypotheses $H_7 : \theta_2 > \theta_1$ and $H_8 : \theta_2 < \theta_1$. Figure 2 considers two conceptual examples. Across both subfigures, the prior distribution for each H_m is constructed such that $\theta_2 > \theta_1$ is not favored over $\theta_2 < \theta_1$, as exactly 50% of its probability mass is in agreement with each expectation. In the first subfigure, all posterior mass lies in the region agreeing with H_m after incorporating the information from the data. Hence, we can directly read off that here $BF_{7u} = 1/0.5 = 2$. Vice versa, $BF_{8u} = 0/0.5 = 0$, and BF_{78} tends to infinity. In the first subfigure, $H_7 : \theta_2 > \theta_1$ would hence find highly superior support over $H_8 : \theta_2 < \theta_1$. In the second subfigure, parameter estimates did not favor any hypotheses, with BF_{78} reducing to 1.

[1] Note that given a prior $p(\theta_u)$, the Bayes factor for informative hypotheses obtains an upper bound of $1/c_m$, given by $BF_{mu} = f_m/c_m \leq 1/c_m$. This is different than the range of Bayes factor values constructed against null hypotheses as reviewed in Table 1.

Note how the derivation of a Bayes factor for informative hypotheses reduces hypothesis evaluation of any complexity to the following process for applied researchers.

1. *Specify a hypothesis H_m based on direct theoretical considerations.*

2. *Construct a low-informative prior distribution for H_m. Evaluate the proportion of this prior distribution in agreement with H_m.*

3. *Provide the parameter estimates $\theta = [\theta_1, \theta_2, ..., \theta_K]^T$ of a GLM, GLMM or SEM and their covariance matrix Σ_θ.*

4. *Construct the resulting posterior. Evaluate the proportion of this posterior distribution in agreement with H_m.*

Bayesian evaluation of informative hypotheses for any general statistical models faces a key challenge in that results are sensitive to the specification of the prior distributions for the model parameters θ. Prior distributions that are extremely vague will lead the Bayes factor to prioritize the most simple hypothesis regardless of the data. This makes the influence of the prior in Bayesian hypothesis testing systematically different from its influence on Bayesian parameter estimation and has been prominently postulated as Lindley's paradox. In order to increase the robustness of Bayesian hypothesis testing to the prior variance, an array of procedures has been proposed, including the local Bayes factor (Smith & Spiegelhalter 1980), the intrinsic Bayes factor (O. Berger & Pericchi 1996), the fractional Bayes factor (O'Hagan 1995), and the partial Bayes factor (O'Hagan 1995). See Gill (2014) for an excellent summary.

In the framework of informative hypotheses, to avoid prior specifications that are too vague, Gu et al. (2017) suggest to specify the prior variance considering the fractional Bayes factor approach of O'Hagan (1995). In this approach the prior is automatically generated using a fraction of the likelihood. The resulting fractional prior is specified using a noninformative prior and a proportion of the likelihood

$$p(\theta|y^b) \propto f(y|\theta)^b p(\theta) \tag{22}$$

where $p(\theta)$ is the noninformative prior distribution and $f(y|\theta)^b$ is the fraction of the likelihood determined by b. From Equation 22, the prior variance is derived. The resulting prior mean of θ is adjusted around the focal point of interest θ^* in H_m such that Equation 22 turns into

$$p(\theta|y^b) \approx N(\theta^*, \hat{\Sigma}_\theta/b) \tag{23}$$

such that the prior distribution for θ under both hypotheses are essentially equivalent and no hypotheses is favored by the prior (like in Figure 2).

The R package "Bain" (Gu et al. 2017) implements this constraint and reduces researchers' tasks

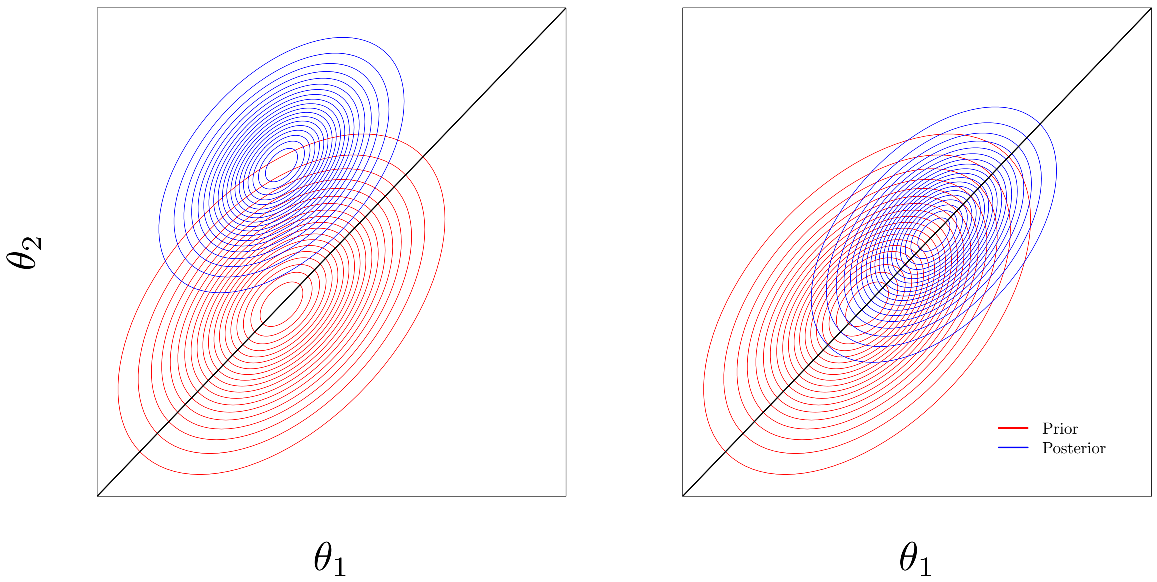


Figure 2 Bayesian evaluation of informative hypotheses as a trade-off between fit and complexity.

to (1.) and (3.). As input, researchers merely need to provide the informative hypotheses under evaluation using the sets of equality and inequality constraints S, R, the vector of parameter estimates θ over which these should be evaluated and their covariance matrix Σ_θ. This makes Bayesian evaluation of informative hypotheses tractable for researchers of varying methodical backgrounds. Instructions and an example of how to use Bain is provided in the supplementary material of this article and can be found in Hoijtink et al. (2019).

A Bayesian Approach to Evidence Synthesis

So far all we have done is that we introduced the concept of informative hypotheses and reviewed how such can be conveniently evaluated based on any general statistical model under the use of the Bayes factor. We argue that this provides a suitable framework for political science researchers to directly evaluate their theoretical expectations against data at hand. Now, against the background of replicability concerns across the behavioral sciences and in political science, we would like to execute one further step and suggest what this approach means for evidence synthesis over studies.

Assume a researcher is confronted a set of studies that evaluates the effect of immigration influx on radical right voting as portrayed in Figure 1. In every individual study, these concepts might be operationalized using different measurements, variables and statistical models yielding functionally incompatible parameter estimates of θ. However, it is important to note that per study, study-specific hypothesis sets can be constructed such that they translate the underlying theoretical expectations into sets that draw on the specific estimates used in each particular study.[2] In a study employing a correlational design, this construct might have been measured through a single real-valued variable exerting one linear effect. This will lead to a hypothesis set similar to the form of H_M^1 in Equation 3

[2] This implies that two hypothesis sets that represent the same theoretical expectation might differ over two studies.

(Kuiper et al. 2013). A second study might have investigated the same construct using an experimental approach and various treatment and control groups. In this setting, the same underlying theoretical expectation might be operationalized much more like the form of H_M^3 in Equation 4. We argue that for evidence synthesis taking a Bayesian hypotheses evaluation approach, this does not constitute a problem at all. In the following, we outline why.

Given that in a set of $1, .., T$ studies Bayes factors BF_{mu} are constructed for all substantive hypotheses under consideration, the cornerstone to our aggregation method is to compute posterior model probabilities (PMPs) for each hypothesis within each study. These provide us with the key for evidence synthesis. In general, PMPs are computed using

$$PMP_m = \frac{\pi_m BF_{mu}}{\sum_m \pi_m BF_{mu}}, \tag{24}$$

where π_m constitutes an hypothesis' prior model probability (PrMP)[3] and BF_{mu} denote the Bayes factors for each hypothesis under consideration. Posterior model probabilities quantify the relative evidence for each hypothesis H_m on a range from 0-1 and jointly sum up to 1. Thus, they quantify the degree to which each hypothesis is supported by the data on a fixed scale. The degree of certainty that a certain hypothesis is the best out of the specified set is PMP_m. Additionally, PMPs can be interpreted as Bayesian error probabilities given the data and underlying hypotheses. The error probability if a hypothesis H_m is selected out of its set is $1 - PMP_m$.

It is important to realize what the described process means for quantitative evidence synthesis. Starting with parameter estimates from a specific model and given that substantive interests regarding the topic under study are translated into informative hypotheses, posterior model probabilities can comparatively quantify evidence over studies on a fixed scale between 0 and 1 *even if these apply different research designs, statistical models and measurements.* It is this asset with which we overcome current approaches to the aggregation of statistical evidence.

For a set of T studies, evidence can be combined by setting the PrMP $\pi^0_{t,m}$ within Study t equal to the PMP $\pi^1_{t,m}$ from Study $t-1$. For the very first study, each hypothesis receives an equal PrMP of $\pi^0_{t=1,m} = 1/M$, where M is the number of all evaluated hypotheses (including the unconstrained H_u). Given that $\pi^0_{t,m}$ represent prior probabilities, $\pi^1_{t,m}$ denote posterior probabilities for Hypothesis H_m in Study t and $\pi^0_{t,m} = \pi^1_{t-1,m}$for $t = 2, ..., T$, all extracted Bayes factors BF^t_{mu} can be combined into an overall evidence measure $\pi^1_{T,m}$ for every hypothesis using

$$\pi^1_{T,m} = \frac{\pi^0_{t,m} BF^t_{mu}}{\sum_{t,m} \pi^0_{t,m} BF^t_{mu}} \tag{25}$$

for $t = 1, \ldots, T$. Thus, PMPs are first constructed in Study 1 and subsequently used as respective PrMPs for Study 2, whose PMPs provide the starting point for Study 3 until all studies are included. Reusing $H_7 : \theta_2 > \theta_1$ and $H_8 : \theta_2 < \theta_1$ that we introduced in the last section and examined in Figure 2 for one study, Figure 3 symbolizes the concept of evidence synthesis when combining evidence from 25 studies for three different scenarios of Bayes factor combinations. Figure 4 summarizes the whole procedure that we have built in this article so far. Since the emerging result is independent of the

[3] Note that this prior model probability π_m for Hypothesis H_m is not to be confused with the prior distribution $p(\theta)$ over model parameters θ from the last section.

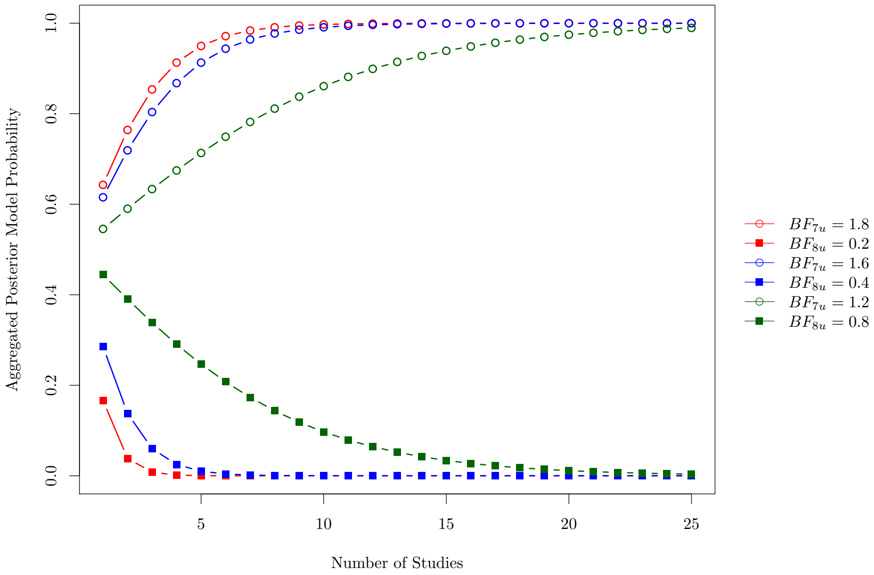


Figure 3 Bayesian evidence sythesis of informative hypotheses.

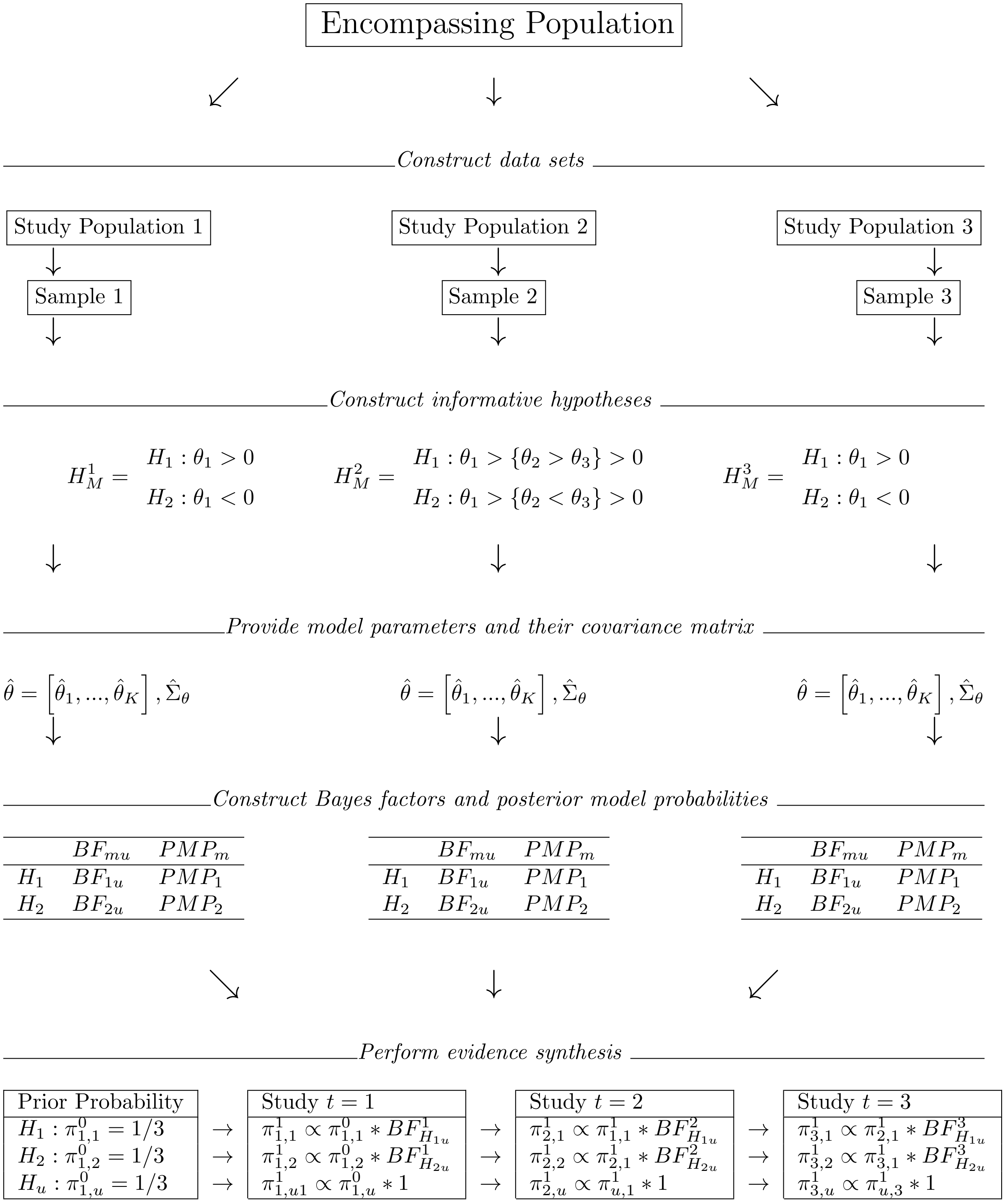
order in which the T studies are arranged, this procedure equips researchers with a framework to aggregate evidence for informative hypotheses of any kind, which is new from our article. Evidence synthesis using Bayesian evaluation of informative hypotheses thus overcomes the issue of combining effect estimates in meta-analysis and Bayesian updating by translating these into Bayes factors quantifying support for informative hypotheses, and aggregating Bayes factors over studies using prior and posterior model probabilities.

Summing up, the approach presented here constitutes a general method for hypothesis evaluation that we present as a counterframework to Null Hypothesis Significance Testing. Turning to this approach, theoretical considerations can be directly confronted with data rather than circumventing testing by the introduction of null hypotheses of no substantive interest. In light of current credibility and replicability crises, this approach allows for evidence synthesis from any general statistical model. For each study, this approach draws on the estimated parameters and their covariance matrix, and constructs Bayes factors using low-informative default priors. Lastly, Bayes factors are combined over studies using the framework of prior and posterior model probabilities. In the next section, we will evaluate this method's performance in a large range of scenarios. Afterwards, we will demonstrate it on our motivating example of right-wing voting.

Monte Carlo Study

EXPERIMENTAL SET UP

When investigating the behavior of the presented approach over a set of studies, we cannot contrast its performance to conventional methods for evidence synthesis since we focus on cases where these are not applicable. In order to evaluate our method, we need to show that after aggregating evidence



Encompassing Population

Construct data sets

Study Population 1 → Sample 1

Study Population 2 → Sample 2

Study Population 3 → Sample 3

Construct informative hypotheses

$H_M^1 = \begin{array}{l} H_1 : \theta_1 > 0 \\ H_2 : \theta_1 < 0 \end{array}$

$H_M^2 = \begin{array}{l} H_1 : \theta_1 > \{\theta_2 > \theta_3\} > 0 \\ H_2 : \theta_1 > \{\theta_2 < \theta_3\} > 0 \end{array}$

$H_M^3 = \begin{array}{l} H_1 : \theta_1 > 0 \\ H_2 : \theta_1 < 0 \end{array}$

Provide model parameters and their covariance matrix

$\hat{\theta} = \left[\hat{\theta}_1, ..., \hat{\theta}_K\right], \hat{\Sigma}_\theta$

$\hat{\theta} = \left[\hat{\theta}_1, ..., \hat{\theta}_K\right], \hat{\Sigma}_\theta$

$\hat{\theta} = \left[\hat{\theta}_1, ..., \hat{\theta}_K\right], \hat{\Sigma}_\theta$

Construct Bayes factors and posterior model probabilities

	BF_{mu}	PMP_m
H_1	BF_{1u}	PMP_1
H_2	BF_{2u}	PMP_2

	BF_{mu}	PMP_m
H_1	BF_{1u}	PMP_1
H_2	BF_{2u}	PMP_2

	BF_{mu}	PMP_m
H_1	BF_{1u}	PMP_1
H_2	BF_{2u}	PMP_2

Perform evidence synthesis

Prior Probability		Study $t = 1$		Study $t = 2$		Study $t = 3$
$H_1 : \pi^0_{1,1} = 1/3$	→	$\pi^1_{1,1} \propto \pi^0_{1,1} * BF^1_{H_{1u}}$	→	$\pi^1_{2,1} \propto \pi^1_{1,1} * BF^2_{H_{1u}}$	→	$\pi^1_{3,1} \propto \pi^1_{2,1} * BF^3_{H_{1u}}$
$H_2 : \pi^0_{1,2} = 1/3$	→	$\pi^1_{1,2} \propto \pi^0_{1,2} * BF^1_{H_{2u}}$	→	$\pi^1_{2,2} \propto \pi^1_{2,1} * BF^2_{H_{2u}}$	→	$\pi^1_{3,2} \propto \pi^1_{3,1} * BF^3_{H_{2u}}$
$H_u : \pi^0_{1,u} = 1/3$	→	$\pi^1_{1,u1} \propto \pi^0_{1,u} * 1$	→	$\pi^1_{2,u} \propto \pi^1_{u,1} * 1$	→	$\pi^1_{3,u} \propto \pi^1_{u,3} * 1$

Figure 4 Evidence synthesis using Bayesian evaluation of informative hypotheses.

from a set of studies that employs diverse statistical models, aggregated evidence is consistently rendered for those hypothesis that are *true*, or, if no true hypothesis is specified, for those that provide the closest representation of the underlying data generating process. If all substantive hypotheses approximate it poorly, then our approach performs well if these are rejected and most support is rendered for the unconstraind hypothesis H_u. We investigate this via three separate Monte Carlo experiments. Within each experiment, we combine evidence from four artificial studies. The encompassing population that these draw sample data from is characterized by three latent parameters $\theta = [\theta_1, \theta_2, \theta_3]$ that we aim to make inferences on and their true relation is forced to be $\theta_1 > \theta_2 > \theta_3 > 0$. When following our approach, applicants necessarily find themselves in one of three scenarios.

- They might investigate a set of hypotheses H_m that covers the whole possible parameter space and by definition include the true hypothesis that is to be detected.
- They might investigate a more complex hypothesis set that does not cover the whole possible parameter space, but *does* include the true hypothesis.
- They might specify a set that does not cover the whole parameter space which *does not* include the true hypothesis.

Each respective experiment investigates our method's performance in one of these scenarios. Table 2 provides a summary of the scenarios that we evaluate.

HYPOTHESIS SETS

In general, we acknowledge that there are two fundamentally distinct types of informative hypotheses, namely equality and inequality constrained formulations. Hence, across all three experiments, we separetely examine our experimental conditions for hypothesis sets that (i) consist only of inequality constraints and (ii) include equality constraints. In Experiment 1, we investigate our method's performance to correctly identify the true hypothesis given that the specified hypothesis set H_m covers the whole parameter space. Here, H_m focuses on one parameter of interest. To maximize the difficulty for our method to perform well, we focus on the smallest effect θ_3 and most evidence is expected to be rendered for H_1. The first set H_m^i consists of two directional hypotheses that differentiate between θ_3 being positive or negative. The second set H_m^e includes an equality constrained hypothesis that incorporates $H_0 = \theta_3 = 0$.

In Experiment 2, we evaluate performance for a set that merely investigates a subset of θ and does entail a true substantive hypothesis. We focus on a multiple parameter case and explicitly add an unconstrained hypothesis H_u. If our method performs well, most evidence will be rendered for H_3. As in Experiment 1, we investigate a separate hypothesis set where an equality constrained hypothesis is added.

In Experiment 3, we evaluate the special case of a set that merely investigates a subset of θ and does not entail a true substantive hypothesis. Thus, we aim to reject both substantive expectations that are bad representations of the underlying population structure and expect most support to be rendered for the unconstrained hypothesis H_u.

Encompassing population	$\theta_1 > \theta_2 > \theta_3 > 0$		
	Experiment 1	Experiment 2	Experiment 3
H_m^i	$H_1:\ \theta_3 > 0$ $H_2:\ \theta_3 < 0$	$H_3:\ \theta_1 > \theta_2 > \theta_3$ $H_4:\ \{\theta_1 < \theta_2\} > \theta_3$	$H_5:\ \theta_2 > \{\theta_1, \theta_3\}$ $H_6:\ \theta_1 < \theta_2 < \theta_3$
H_m^e	$H_1:\ \theta_3 > 0$ $H_0:\ \theta_3 = 0$ $H_2:\ \theta_3 < 0$	$H_3:\ \theta_1 > \theta_2 > \theta_3$ $H_4:\ \theta_1 = \theta_2 = \theta_3$	$H_5:\ \theta_2 > \{\theta_1, \theta_3\}$ $H_6:\ \theta_1 = \theta_2 = \theta_3$
Varying factors in encompassing population	$R^2 = 0.25$, 0.15, 0.02; θ-ratios $= 4:2:1$ and $1.77:1.33:1$		
Varying factors per study	Sample sizes from n=50 until n=500 by n=25		

Table 2 Experimental set up and conditions.

STATISTICAL MODELS

We combine evidence from four GLMs mimicking four different studies that sample from a jointly underlying encompassing population. GLMs can be written as:

$$f(\hat{y}) = X\beta, \tag{26}$$

where $\hat{y} = E(y)$ denotes the expectation of the dependent variable y and $X\beta$ is the linear predictor with independent variables $X = [x_1, x_2, ..., x_K]$ and model parameters $\beta = [\beta_1, \beta_2, ..., \beta_K]$ for $k = 1, .., K$. $\beta = [\beta_1, \beta_2, \beta_3]$ are our estimates of $\theta = [\theta_1, \theta_2, \theta_3]$. In our simulation study, two GLMs set $f(\hat{y}) = y$ and estimate linear regressions. One GLM sets $f(\hat{y}) = log(\frac{\hat{y}}{1-\hat{y}})$ and builds a logistic regression model. The final model estimates probit regression and sets $f(\hat{y}) = \Phi^{-1}\hat{y}$.

EXPERIMENTAL CONDITIONS

In each experiment, we use a full factorial 3 x 2 x 19 design. The following factors vary:

- The overall level R^2 by which the latent parameters θ are related to the outcome variable in the encompassing population. Following Cohen (1992), we vary between small ($R^2 = 0.02$), medium ($R^2 = 0.15$) and large ($R^2 = 0.25$) levels of association. Given that a certain association level is chosen, the individual shares on the R^2 are distributed equally over all three latent parameters.

- The ratios between the latent factors in the encompassing population. When distributing the overall R^2 over the three parameters, we do so in two different ways. Once, the true parameters take on ratios of 2:1 implying that $\theta_1 = 2 \times \theta_2 = 4 \times \theta_3$. In a second set of scenarios, we set the ratios to 1.33:1 resulting in $\theta_1 = 1.33 \times \theta_2 = 1.77 \times \theta_3$. This is important for informative hypotheses regarding different strengths of latent parameters (for instance $H : \theta_1 > \theta_2$), as the power to yield correct inferences might depend on the true strengths being substantively or moderately different.

- The sample size which each individual study takes on. This is varied between n=50 and n=500 in increments of n=25. For reasons of simplicity, we let every individual study take on the same sample size.

For each of these 3 x 2 x 19 = 114 experimental conditions, 4 x 100 datasets are generated (100 datasets for each individual study). Following, Bayes factors are constructed using the R package Bain (Gu et al. 2017) and evidence is synthesized over all four studies. Thus, within each experimental condition, 100 aggregated results emerge.[4]

REPORTED QUANTITIES

To summarize our Monte Carlo experiments, we focus on two quantities of interest. The first quantity is the *true hypothesis rate* (THR) which has been used as the equivalent to power in frequentist decision-theory to evaluate the performance of Bayesian hypothesis evaluation (Kuiper et al. 2015). This quantity summarizes the share of times most support is rendered for the correct hypothesis. If the evaluated set does include a correct hypothesis, most evidence is expected to be rendered for H_u. Equivalently to guidelines for statistical power in null hypothesis significance testing, we consider THRs of at leat 0.8 desirable.

Second, we evaluate *how extremely* the support for the preferred hypothesis differs from the support for the competing expectations. This is important because when H_1 is true, a set of aggregated posterior model probabilities of, for instance, $(PMP_{H_1} = 0.55, PMP_{H_2} = 0.05, PMP_{H_u} = 0.40)$ would contribute to a high THR, while in applied settings lets researchers inconclusive about the underlying population structure. Here, we report the final PMP rendered for the true hypothesis.

Preliminary Monte Carlo Results

INEQUALITY CONSTRAINED HYPOTHESES

We discuss the results of our experiments separately for the cases of hypothesis sets H^i_m that include only inequality constrained hypotheses and H^e_m which incorporate equalities as well. Figure 5 displays the results when using the former. On the y-axis, we present (i) the true hypothesis rate, that is the share of times the PMP was correctly rendered highest for the hypothesis that we expect to outperform its counterparts (H_1 in Experiment 1, H_3 in Experiment 2, H_u in Experiment 3) and (ii) the average PMP that was rendered for this hypothesis over the 100 simulation runs. The x-axis displays the sample size that each underlying model was based on. Additionally, this relation is qualified for the three R^2 with which all three variables are jointly related to their dependent outcome in the encompassing population. In the main text, we limit this discussion to cases where the true ratios between the population coefficients are set to 2:1. Results for ratios 1.33:1 are reported in the supplemental material of this article and lead us to similar conclusions in every evaluated scenario.

[4] These are preliminary simulations. The final version of this document will report our results based on a significantly higher number of iterations.

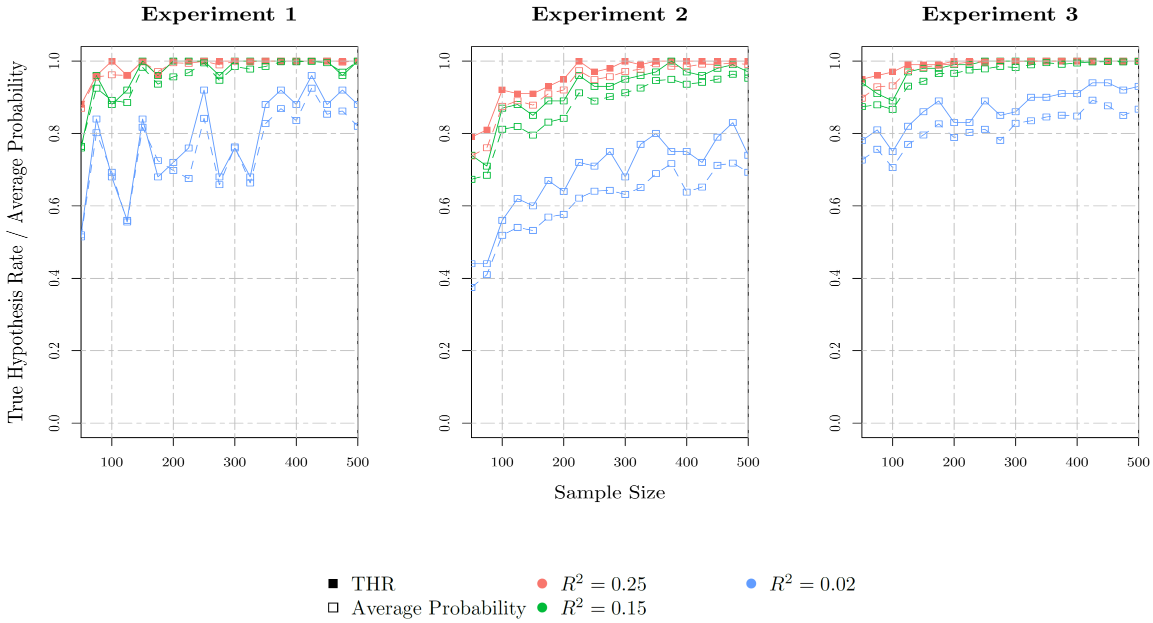


Figure 5 True hypothesis rates and average probability, inequality constrained hypotheses.

For large and moderate effect sizes, that is when the true overall association with which all three variables are jointly related to their dependent outcome in the encompassing population is set to $R^2 = 0.25$ or $R^2 = 0.15$, THRs consistently exceed the conventional power threshold of 0.8 already for the smallest investigated sample sizes and render THRs of at least 0.9 when $n = 200$. Given that effect sizes in the underlying population are small ($R^2 = 0.02$), more sample size is needed to correctly infer the underlying parameter relations, with consistent results emerging for sample sizes of at least $n = 400$. This is behavior that is well in line with our expectations and should not take readers by a surprise as one naturally needs large sample sizes to detect small effects.

Moreover, across all experimental conditions, not only is most evidence rendered for the hypothesis that is expected to perform best, but the average final PMP that is observed after aggregating evidence from the four studies is consistently showing clear support (meaning high PMPs) for these expectations as well. Aggregated PMPs are consistently high for and take on low values for their counterparts. This reaffirms us that evidence synthesis using Bayesian evaluation of inequality constrained hypotheses correctly identifies true parameter relations for hypothesis sets of different forms and a wide range of scenarios that researchers might face.

EQUALITY CONSTRAINED HYPOTHESES

Figure 6 presents the performance ouf our three investigated hypothesis sets including equality constraints (H_m^e) over our preliminary simulations. Results differ from those achieved for inequality constraints over all three evaluated hypothesis sets. Specifically, we observe power issues under the inclusion of equality constraints. Only in a minority of simulation scenarios, true hypotheses are correctly identified after aggregating results over all four models. While for instance in Experiment

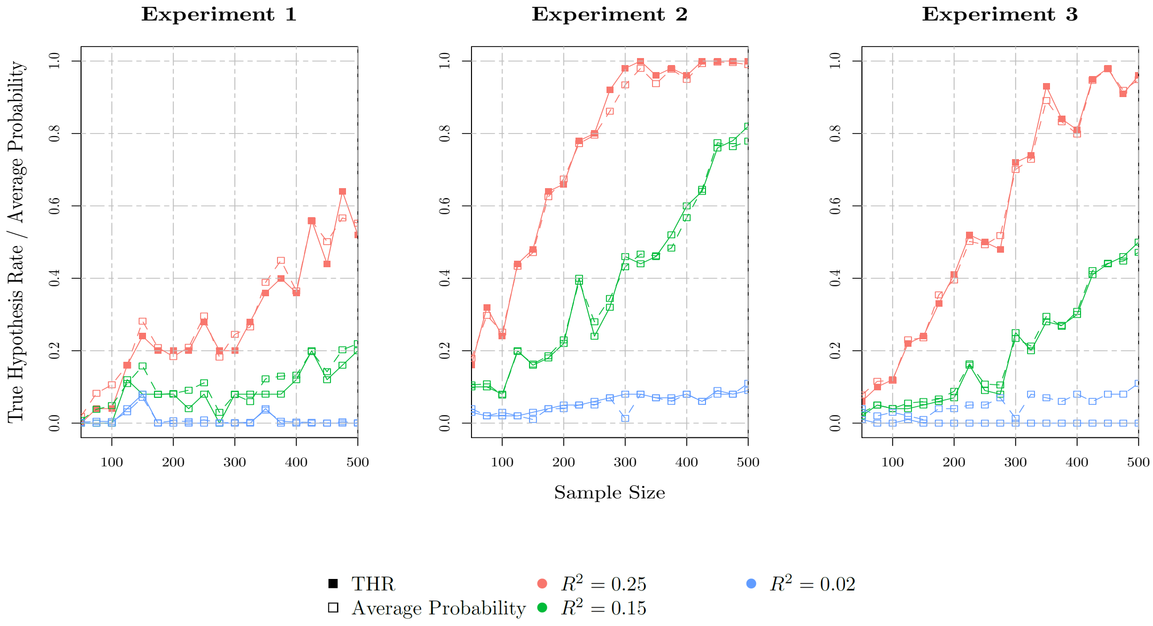


Figure 6 True hypothesis rates and average probability, including equality constrained hypotheses.

3, Hypothesis 5 is correctly identified in over 80% of the cases for medium sample sizes ($n > 350$) when population effects are large ($R^2 = 0.25$), for most evaluated scenarios, true hypothesis rates stay below desirable thresholds. From further inspecting our simulation output, we note that in these experimental conditions, the equality constrained hypotheses H_0, H_4 and H_6 receive most support from Bayesian evidence synthesis although not being in line with the underlying population structure. Thus, when evaluating equality constraints, researchers might very well find convincing support for wrong null hypotheses.

Summing up, given that inequality constrained hypotheses are specified, researchers are well advised to directly evaluate their theoretical expectations using Bayesian evaluation of informative hypotheses, and can confidently use this approach for the formal integration of evidence from diverse statistical models. On the other hand, the presented approach seems to underperform when traditional null hypotheses are evaluated.

Plan to Proceed: Application to Right-Wing Voting

So far we have (i) reviewed Bayesian evaluation of informative hypotheses, (ii) introduced our approach to evidence synthesis and (iii) evaluated its performance in Monte Carlo simulations. In the remainder of this manuscript, we plan to demonstrate our approach on a set of studies from the motivating example presented in Section 2. Electoral support for radical right parties is one of the most researched topics in the field of political behavior nowadays (Arzheimer 2018). Over 800 studies have been published in recent decades on radical right parties in Europe alone[5]. Given the unusually large body

[5] Kai Arzheimer has collected 857 titles in his *Extreme Right Bibliography.* The list can be accessed via https://www.kai-arzheimer.com/extreme-right-western-europe-bibliography

of literature, this research area provides and ideal opportunity to apply our new method of evidence synthesis. To showcase the use of our method on a substantive problem, the future plan for our paper is to

- Scan the literature on right-wing voting in Europe
- Identify a set of studies that tackle a jointly underlying research problem
- Apply our method of evidence synthesis to this set of studies

One reason for the vast interest in the topic is the immense regional differences in electoral support for radical right parties both between and within countries, which frequently serve as empirical puzzles driving the research agenda (Arzheimer 2009, Lubbers et al. 2002, Rydgren & Ruth 2013, Stecker & Debus 2019, Stockemer 2018). While Portugal, Spain and Malta have never had a relevant radical right party, the *Front National* in France, the *Freiheitliche Partei Östereichs*, and the *Schweizerische Volkspartei* (SVP) in Switzerland have been more or less consistently successful at the ballot boxes (Arzheimer 2018). The within-country variations are even more pronounced. The vote share of the SVP ranged between 0 and 90 percent across Swiss municipalities (Stockemer 2018) and support for the *Alternative für Deutschland* ranged between 4.9 and 35.5 percent across German electoral districts in the 2017 federal election[6].

Since radical right parties regularly centre their election campaigns around the topic of immigration, it seems natural for researchers to look at actual immigration as an explanation for electoral support for these parties. Two opposing theoretical arguments are pivotal to explaining anti-immigrant sentiment and by association radical right voting. *Realistic Group Threat Theory* proposes that immigration can lead to a conflict over the distribution of resources between natives and immigrants. This competition creates feelings of threat and in turn makes citizens prone to support radical right parties. According to Pettigrew & Tropp (2008), immigration increases the opportunity for contact between natives and immigrants, which leads to a decrease in prejudices against immigrants. With anti-immigrant attitudes being one of the major determinants of radical right voting (Arzheimer 2018), in consequence, immigration would result in less support for radical right parties. The debate on which of these mechanisms is more prevalent is still ongoing. Some scholars find that higher immigrant inflow leads to an increase in voting for radical right parties (e.g. Halla et al. 2017, Lubbers et al. 2002, Rydgren & Ruth 2011), while others find the reverse effects (e.g. Teney 2012). On a macro level, besides immigration and ethnic diversity, the concepts most frequently studied in previous research include regional deprivation measured as unemployment on the contextual level (Coffé et al. 2007, Smith 2010) and levels of insecurity in the form of crime rates (Arzheimer 2009, Golder 2003) in a given social context.

Previous attempts at combining evidence from various studies had clear limitations. Amengay & Stockemer (2018) investigate the findings of 48 selected peer-reviewed articles. They could only include studies which are based on similar dependent variables, i.e. vote share, and models with the same functional form. However, research in the field is regularly based on a variety of model specifications. Analytical tools in this field include linear regression models using aggregated electoral data (Rydgren

[6] Source: Bundeswahlleiter, retrievable at https://www.bundeswahlleiter.de/bundestagswahlen/2017/ergebnisse.

& Ruth 2011, Stockemer 2018), binary logistic regressions with individual vote choice as dependent variable (Arzheimer 2009), multinomial regressions models (Zhirkov 2014), weighted spatial models (Teney 2012), and tobit models (Golder 2003). Furthermore, Amengay & Stockemer (2018) simply study existing statistically significant relationships between variables instead of combining evidence for substantial hypotheses.

We intend to overcome these limitations by applying Bayesian evidence synthesis. For instance, we might compare a) extant evidence for group threat and contact hypotheses and b) the contextual factors driving regional variations in electoral support for radical right parties. We can evaluate the relative importance of immigration, regional deprivation and insecurity for the explanation of radical right voting.

Conclusion

The systematic integration of evidence from a variety of studies plays a crucial role throughout all the sciences. In the natural and medical sciences, quantitative evidence synthesis constitutes *the* gold standard for pushing the frontiers of knowledge forward. There is little doubt that also our disciplines of the social and political sciences can considerably profit from such endeavors. However, merely copying methodical approaches that were aimed at different fields and backgrounds comes with major shortcomings. Rather, if we as a discipline face specific challenges like the synthesis of evidence from diverse studies, we need to rely on specific methodical toolsets.

This article worked towards this end and presents a method that can synthesize evidence even for models and research design as diverse as applied in our fields. By aggregating evidence for informative hypotheses rather than estimated effect sizes, our method works around limitations that are inherent to currently used methods. In our simulations, we have shown that if hypotheses are carefully specified, the presented approach consistently provides confident results that researchers can rely on. This expands the possibilities that researchers haven when performing quantitative evidence synthesis.

Furthermore, this method can also be viewed as a contribution to a wider topic that is inherent in many of our discipline's current discussions, namely the question of how we are evaluating hypotheses in the first place and the ever increasing criticism against the classical framework of null hypothesis significance testing. By revising the informative hypothesis literature, we demonstrate how researchers can leave the ground of testing null hypotheses and directly evaluate their theoretical expectations, may these be investigated in a single or multiple studies under consideration.

However, evidence synthesis via Bayesian evaluation of informative hypothesis is no panacea. As we have shown, performance was rather limited when evaluating equality constrained hypotheses for small and moderate sample sizes. Thus, researchers need to be careful when specifying the hypotheses that are researched. We believe that this article marks an important contribution to the field of political science and hope that it can contribute to the issues our discipline faces today: An increasing call for replication studies and aggregated evidence, but an insufficient methodical toolkit to actually perform this endeavor.

References

Allport, G. W., Clark, K. & Pettigrew, T. (1954), 'The nature of prejudice'.

Amengay, A. & Stockemer, D. (2018), 'The radical right in western europe: A meta-analysis of structural factors', *Political Studies Review* p. online first.

Arzheimer, K. (2009), 'Contextual factors and the extreme right vote in western europe, 1980-2002', *American Journal of Political Science* **53**(2), 259–275.

Arzheimer, K. (2018), Explaining electoral support for the radical right, *in* J. Rydgren, ed., 'The Oxford Handbook of the Radical Right', Oxford University Press.

Bélanger, S. & Pinard, M. (1991), 'Ethnic movements and the competition model. some missing links', *American Sociological Review* **56**, 446–457.

Blumer, H. (n.d.), 'Race Prejudice as a Sense of Group Position, volume = 1, year = 1958', *Pacific Sociological Review* pp. 1–7.

Chen, M.-H. & Ibrahim, J. G. (2000), 'Power prior distributions for regression models', *Statistical Science* **15**(1), 46–60.

Chib, S. (1995), 'Marginal Likelihood from the Gibbs Output', *Journal of the American Statistical Association* **90**(432), 1313–1321.

Coffé, H., Heyndels, B. & Vermeir, J. (2007), 'Fertile Grounds for Extreme Right-Wing Parties: Explaining the Vlaams Blok's Electoral Success', *Electoral Studies* **26**(1), 142–155.

Cohen, J. (1992), 'A Power Primer', *Psychological Bulletin* **112**(1), 155–159.

Cooper, H., Hedges, L. & Valentine, J. (2009), *The handbook of research synthesis and meta-analysis*, Russel Sage Foundation, New York.

Costa, M. (2017), 'How responsive are political elites? a meta-analysis of experiments on public officials', *Journal of Experimental Political Science* **4**(3), 241–254.

Doucouliagos, H. & Ulubaşoğlu, M. A. (2008), 'Democracy and economic growth: A meta-analysis', *American Journal of Political Science* **52**(1), 61–83.
URL: *http://www.jstor.org/stable/25193797*

Gill, J. (2014), *Bayesian Methods: A Social and Behavioral Science Approach (3rd edition)*, Chapman and Hall/CRC, Boca Raton.

Gill, J. (2019), 'Adventures in Replication: An Introduction to the Forum', *Political Analysis* **27**(1), 98–100.

Golder, M. (2003), 'Explaining variation in the success of extreme right parties in western europe', *Comparative Political Studies* **36**(4), 432–466.

Gu, X., Mulder, J. & Hoijtink, H. (2017), 'Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses', *British Journal of Mathematical and Statistical Psychology* (024), 1–43.

Halla, M., Wagner, A. F. & Zweimüller, J. (2017), 'Immigration and voting for the far right', *Journal of the European Economic Association* **15**(6), 1341–1385.

Hoijtink, H., Mulder, J., van Lissa, C. & Gu, X. (2019), 'A tutorial on testing hypotheses using the Bayes factor.', *Psychological Methods* pp. No Pagination Specified–No Pagination Specified.

Jackson, J. W. (1993), 'Realistic group conflict theory: A review and evaluation of the theoretical and empirical literature.', *The Psychological Record* .

Jeffreys, H. (1961), *Theory of Probability (3rd edition)*, Oxford University Press, Oxford.

Kass, R. E. & Raftery, A. E. (1995), 'Bayes Factors', *Journal of the American Statistical Association* **90**(430), 773–795.
URL: *http://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf*

Klugkist, I., Laudy, O. & Hoijtink, H. (2005), 'Inequality constrained analysis of variance: a Bayesian approach.', *Psychological methods* **10**(4), 477–493.

Klugkist, I., Laudy, O. & Hoijtink, H. (2010), 'Bayesian evaluation of inequality and equality constrained hypotheses for contingency tables.'.

Kuiper, R. M., Buskens, V., Raub, W. & Hoijtink, H. (2013), 'Combining Statistical Evidence From Several Studies', *Sociological Methods & Research* **42**(1), 60–81.
URL: *http://journals.sagepub.com/doi/10.1177/0049124112464867*

Kuiper, R. M., Nederhoff, T. & Klugkist, I. (2015), 'Properties of hypothesis testing techniques and (Bayesian) model selection for exploration-based and theory-based (order-restricted) hypotheses', *British Journal of Mathematical and Statistical Psychology* **68**(2), 220–245.

Lipsey, M. W. & Wilson, D. B. (2001), *Practical Meta Analysis*, Vol. 49, SAGE Publications, Thousand Oaks, California.

Lubbers, M., Gijsberts, M. & Scheepers, P. (2002), 'Extreme right-wing voting in western europe', *European Journal of Political Research* **41**(3), 345–378.

Mulder, J., Hoijtink, H. & Klugkist, I. (2010), 'Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors', *Journal of Statistical Planning and Inference* **140**(4), 887–906.
URL: *http://dx.doi.org/10.1016/j.jspi.2009.09.022*

O. Berger, J. & Pericchi, L. (1996), 'The Intrinsic Bayes Factor for Model Selection and Prediction', *Journal of American Statistic Association* **91**, 109–122.

O'Hagan, A. (1995), 'Fractional Bayes Factors for Model Comparison', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1), 99–138.
URL: *http://www.jstor.org/stable/2346088*

Open Science Collaboration (2015), 'Estimating the reproducibility of psychological science', *Science* **349**(6251), 4716–4716.
URL: *http://www.sciencemag.org/cgi/doi/10.1126/science.aac4716*

Pettigrew, T. F. (1998), 'Reactions toward the new minorities of western europe', *Annual Review of Sociology* **24**, 77–103.

Pettigrew, T. F. & Tropp, L. R. (2008), 'How does intergroup contact reduce prejudice? meta-analytic tests of three mediators', *European Journal of Social Psychology* **38**(6), 922–934.

Quillian, L. (1995), 'Prejudice as a response to perceived group threat. population composition and anti-immigrant and racial prejudice in europe', *American Sociological Review* **60**, 586–611.

Rydgren, J. & Ruth, P. (2011), 'Voting for the radical right in swedish municipalities: social marginality and ethnic competition?', *Scandinavian Political Studies* **34**(3), 202–225.

Rydgren, J. & Ruth, P. (2013), ‘Contextual explanations of radical right-wing support in Sweden: socioeconomic marginalization, group threat, and the halo effect’, *Ethnic and Racial Studies* **36**(4), 711–728.

Smith, A. F. M. & Spiegelhalter, D. J. (1980), ‘Bayes Factors and Choice Criteria for Linear Models’, *Journal of the Royal Statistical Society. Series B (Methodological)* **42**(2), 213–220.
URL: *http://www.jstor.org/stable/2984964*

Smith, J. M. (2010), ‘Does crime pay? issue ownership, political opportunity, and the populist right in western europe’, *Comparative Political Studies* **43**(11), 1471–1498.

Stecker, C. & Debus, M. (2019), ‘Refugees Welcome? Zum Einfluss der Flüchtlingsunterbringung auf den Wahlerfolg der AfD bei der Bundestagswahl 2017 in Bayern’, *Politische Vierteljahresschrift* pp. 1–25.

Stockemer, D. (2018), ‘The rising tide: Local structural determinants of the radical right-wing vote in switzerland’, *Comparative European Politics* **16**(4), 602–619.

Tajfel, H. & Turner, J. C. (1986), The social identity theory of intergroup behaviour, *in* S. Worchel & W. G. Austin, eds, ‘Psychology of Intergroup Relations’, Nelson-Hall Publishers, Chicago, pp. 7–24.

Teney, C. (2012), ‘Space matters. the group threat hypothesis revisited with geographically weighted regression. the case of the npd 2009 electoral success’, *Zeitschrift für Soziologie* **41**(3).

Wuttke, A. (2019), ‘Why Too Many Political Science Findings Cannot Be Trusted and What We Can Do About It: A Review of Meta-Scientific Research and a Call for Academic Reform’, *Politische Vierteljahresschrift* **60**(1), 1–19.

Zhirkov, K. (2014), ‘Nativist but not alienated. a comparative perspective on the radical right vote in western europe’, *Party Politics* **20**(2), 286–296.