Social Cognition, Vol. 38, No. 6, 2020, pp. 489–520

HAS THE EFFECT OF THE AMERICAN FLAG ON POLITICAL ATTITUDES DECLINED OVER TIME? A CASE STUDY OF THE HISTORICAL CONTEXT OF AMERICAN FLAG PRIMING

Travis J. Carter
Roanoke College

Gayathri Pandey
Cornell University

Niall Bolger
Columbia University

Ran R. Hassin
Hebrew University of Jerusalem

Melissa J. Ferguson
Yale University

We report findings from a meta-analysis on all published and unpublished studies from our labs (total N = 9,656) examining the priming effect of the American flag on political attitudes. Our analyses suggest that, consistent with the studies we originally published in 2011 (T. J. Carter et al., 2011b), American flag primes did create politically conservative shifts in attitudes and beliefs during the initial time period when data were collected (even excluding the published studies), but this effect has since declined over time to be roughly zero, though we believe that other interpretations, including false positives, are plausible. We discuss possible interpretations of this decline effect and the importance of considering the historical context in

Supplemental materials are available online at https://doi.org/10.6084/m9.figshare.c.4881351.

The authors declare no conflict of interest with the content of this article. Author contributions: Analyzed data: TJC; Wrote paper: MJF, TJC, RRH, NB. Data collection: MJF, TJC, GP.

Thank you to Felix Thoemmes for his comments on an earlier draft of this article. RRH would like to thank the IIAS for a wonderful semester that allowed for much (critical) thinking; he would also like to thank the Templeton Foundation for their continuous support.

Address correspondence to Travis J. Carter at 221 College Lane, Salem, VA 24153. E-mail: tjcarter@roanoke.edu.

relation to the priming effects of symbols whose meaning is not static over time. We also highlight the value of publicly posting data, emptying file drawers, and conducting direct as well as conceptual replications.

Keywords: replication, priming, flag priming, meta-analysis

How do the symbols of a nation influence a citizen's everyday attitudes and behaviors? Despite an extensive literature in the social sciences on how a nation might influence the citizenry in both subtle and obvious ways (e.g., Billig, 1995; Gellner, 2009), there has only recently been empirical psychological work testing whether and how that might happen. Using the basic logic of decades of theory on cognition (e.g., Schröder & Thagard, 2013), studies have found that subtle exposure to potent national symbols, such as a national flag, unintentionally activates various attitudes, beliefs, and behaviors (Butz, Plant, & Doerr, 2007; Callahan & Ledgerwood, 2016; T. J. Carter, Ferguson, & Hassin, 2011a, 2011b, hereafter CFH; Chan, 2017; Devos & Banaji, 2005; Devos & Ma, 2008, 2012; Ehrlinger et al., 2011; Ferguson & Hassin, 2007; Gangl, Korgler, & Kirchler, 2016; Hassin, Ferguson, Shidlovski, & Gross, 2007; Kalmoe & Gross, 2016; Kemmelmeier & Winter, 2008; Ma & Devos, 2013; Nosek, 2005; Rydell, Hamilton, & Devos, 2010).

In 2011, three of the present authors published two experiments in *Psychological Science* demonstrating one such effect of an American flag prime: a shift toward conservative/Republican attitudes and behaviors (T. J. Carter et al., 2011b). In Study 1, conducted around the 2008 American presidential election between Republican John McCain and Democrat Barack Obama, relative to control participants, participants primed with an American flag while reporting their voting intentions reported more conservative political beliefs, a greater intention to vote for—and a greater likelihood of having actually voted for—the Republican candidate in the election. Primed participants then expressed less satisfaction with Obama's job performance eight months later. In Study 2, conducted in early 2010, those primed with the American flag immediately expressed more conservative political beliefs than participants in the control condition.

Critically, we were not claiming in CFH that, in Study 1, participants' exposure to the flag prime *directly* affected their satisfaction eight months later. We assumed there was a psychologically crucial chain of events in the interim. That is, the presentation of the flag prime influenced participants' voting intentions (and other attitude and beliefs measures) directly. Those changes in voting intentions subsequently influenced their actual vote; indeed, their conscious intentions mediated the effect of the prime on the voting behavior. We assumed that participants' reported behavioral intentions and their actual (voting) behavior then affected their reported satisfaction eight months later. The notion that declared behavioral intentions in the unique setting of an experiment, and then the highly salient and unique behavior of voting, could influence attitudes eight months later is in line with many decades of research on people's tendency to remain consistent with their expressed behavioral intentions and actual behavior (e.g., see review

in Gawronski & Strack, 2012). To imagine how this could occur, compare two people. One was nudged by priming to say she will vote for McCain. The other, identical in every way, was in the control condition, and says she will vote for Obama. Having declared those voting intentions, the likelihood they will actually vote for McCain and Obama, respectively, increases (Greenwald, Carnot, Beach, & Young, 1987). The first, having voted for McCain, is less satisfied with Obama eight months later—after all, she did vote for McCain, so she is naturally more disappointed. Thus, we interpreted this as the rolling consequences of immediate effects, not a priming effect lingering in participants' heads for a year.

THE CURRENT STUDY

Since 2006, we have conducted many experiments testing the influence of American flag primes on expressions of political beliefs and attitudes. In order to understand and extend the findings in CFH (and the exploratory efforts that preceded them), we conducted multiple experiments that would be deemed direct or conceptual replications—some of which were reasonably successful replications, and some of which were clear failures to replicate—none of which we have attempted to publish.

In the present article, we used all of these data to consider whether the most basic claim reported in CFH—that exposure to American flag primes shifts political attitudes rightward—was generally supported. Given that many of those studies did not show this effect, we began with a fairly simple hypothesis: that the effect of flag priming decreased over time. This hypothesis is based on a simple premise supported by decades of theory and research in cognitive and social psychology (e.g., Schröder & Thagard, 2013): exposure to a stimulus will activate whatever idiosyncratic knowledge an individual associates with that stimulus. Culture in general, and the political atmosphere in particular, change over time in myriad ways, meaning that the associated knowledge—and priming effects—change as well. Our initial experiments were run after roughly five years of a Republican presidency—one profoundly shaped by the tragic events of September 11th, 2001 (aka 9/11) and the military and political confrontations that followed. There was discussion then of the American flag being "branded" by Republicans (e.g., see Carney, Jost, Gosling, & Potter, 2008), as well as by 9/11 (Landau et al., 2004; Skitka, 2005). This formed the basis for our original hypothesis that American flag primes would produce conservative shifts in attitudes and beliefs.

The election of Democrat Barack Obama in late 2008 seemed to herald a shift in the American political winds and culture more widely. The election of the first African-American president was seen by many as indicating a potentially "post-racial" era (Steele, 2008), and pundits and politicians expected a new wave of progressivism (Beinart, 2016). There has been considerable debate over how American culture in fact changed over the Obama years, with some pollsters and social scientists reporting significant change in Americans' racial and other progressive attitudes (Beinart, 2016; Devos & Ma, 2012; Dimock, 2017; Ma & Devos, 2013) and others suggesting little permanent change on some dimensions (Goldman & Mutz,

2014; Schmidt & Nosek, 2010). Although it will doubtless require many years for historians and social scientists to fully understand how the country changed over the Obama era, there has been sufficient speculation of cultural and political change that we expected that people's associations with the American flag might have changed over that time. Specifically, we predicted that the "Republicanism" of the American flag diminished over Obama's (Democratic) presidency.

However, there is a clear alternative hypothesis to consider, namely that those initial positive results were flukes, and any apparent change in the priming effect over time is simply a regression toward the true effect size (i.e., zero). We discuss this possibility at length below, and endeavored to apply various analytical approaches to ensure it is considered seriously. As such, we sought to quantify and evaluate *all* of the data we have collected related to American flag primes on political measures using a meta-analytic approach. This article represents an attempt to test the decline hypothesis, and to do so as transparently as possible. *Crucially, in this article we empty our file drawer completely* (see Lane, Luminet, Nave, & Mikolajczak, 2016). In addition, we are making available all of the raw data and the R code used to reproduce the analyses reported below.[1]

METHOD

STUDY SELECTION

We began by assembling data from all of the studies the original authors have conducted testing the influence of American flag primes on the expression of political attitudes along the liberal-conservative continuum. (Based on an exhaustive search of our paper and digital files, we are reasonably confident that we found every study in our labs on this effect.) We then selected *every* experiment we conducted that included 1) a manipulation of the presence/absence of an American flag, and 2) measures conceptually related to political ideology (e.g. beliefs about specific political issues, voting intentions, attitudes towards political candidates, etc.). In addition to the two published studies (CFH Studies 1 and 2), we identified a total of 33 unpublished studies, the earliest of which was conducted in 2006 and the most recent in the fall of 2016.

In addition, because we are reporting the contents of our own file drawer, we do not include the Many Labs study (Klein et al., 2014) in the analyses by default. However, we also conducted the analyses with this study included; its inclusion did not change any of the results or conclusions reported below. The data from this study are included along with the raw data from all of the other studies to make it easy for others to reproduce these analyses. The basic details of each of the studies we considered for inclusion can be found in Table 1.

To be clear, this is not meant to be an exhaustive meta-analysis of *all* experiments that fit these criteria; we did not put out a general call for such experiments

1. We also encourage readers to contact us for the survey materials and priming methods used in the studies reported here.

TABLE 1. Attributes of Studies Considered for Inclusion in the Meta-Analysis

Study	Date	*N*	*n*	Manipulation	Format	Primary Measure	Design changes
	Jun 20, 2006	111	0	VGQ	Paper/pencil	Political Beliefs	
	Nov 29, 2006	52	0	Subliminal	Computer (Lab)	Political Beliefs	
	Feb 4, 2007	130	0	VGQ	Paper/pencil	Approval	
	Jan 18, 2008	86	0	Survey Corner	Paper/pencil	Voting	
	Feb 24, 2008	49	0	Survey Corner	Paper/pencil	Voting	
CFH Study 1	Oct 11, 2008	183	8	Survey Corner	Online	Voting	
	Oct 16, 2008	126	0	Survey Corner	Paper/pencil	Voting	
	Jan 31, 2009	80	0	Survey Corner	Online	Political Beliefs	
	Mar 13, 2009	95	0	Subliminal	Computer (Lab)	Political Beliefs	CL
CFH Study 2	Apr 1, 2009	70	0	VGQ	Paper/pencil	Political Beliefs	
	Dec 7, 2009	46	0	Survey Corner	Paper/pencil	Approval	
	Feb 28, 2010	317	0	Survey Corner	Online	Political Beliefs	
	Aug 24, 2010	123	0	VGQ	Computer (Lab)	Political Beliefs	DR
	Jul 25, 2012	150	5	Survey Corner	Online	Voting	CL
	Jul 31, 2012	176	3	Candidate Photo	Online	Speech	CL, DR, CB
	Oct 15, 2012	511	4	Candidate Photo	Online	Voting	CL, CB
*	Oct 27, 2012	415	4	Survey Corner	Online	Voting	
	Nov 3, 2012	121	3	Candidate Photo	Online	Voting	CB
	Nov 4, 2012	91	1	Survey Corner	Online	Voting	
	Feb 5, 2013	200	1	Survey Corner	Online	Approval	
	Feb 19, 2013	200	1	Survey Corner	Online	Approval	
	Mar 6, 2013	580	1	Survey Corner	Online	Approval	
	Mar 9, 2013	499	2	Candidate Photo	Online	Approval	CB
	Mar 16, 2013	652	0	Survey Corner	Online	Warmth	
	Mar 26, 2013	299	0	Survey Corner	Online	Warmth	
	Apr 23, 2013	451	0	Candidate Photo	Online	Speech	CB
	Sep 13, 2013	302	0	Candidate Photo	Online	Speech	CB
Many Labs*	Sep 15, 2013	4896	0	VGQ	Online	Political Beliefs	CL
	Sep 26, 2013	601	5	Candidate Photo	Online	Speech	CL, CB
	Oct 13, 2013	408	2	Survey Corner	Online	Voting	CL
	Oct 23, 2013	249	1	Survey Corner	Online	Voting	CL
	Dec 20, 2013	801	0	Survey Corner	Online	Voting	
	Feb 10, 2014	794	0	Survey Corner	Online	Political Beliefs	
	Apr 13, 2014	111	0	Survey Corner	Computer (Lab)	Political Beliefs	
	Nov 5, 2016	992	17	Survey Corner	Online	Political Beliefs	
*	Nov 11, 2016	76	7	Survey Corner	Computer (Lab)	Political Beliefs	

Note. The date listed is the estimated midpoint of the data collection period for each study. *N* = total number of participants after exclusions; *n* = number of participants excluded for inattention; VGQ = Visual Geography Quiz; CL = Collapsed across other factors (e.g., order, speaker); DR = Dropped priming condition (e.g., UK flag priming); CB = Combined American flag priming conditions (e.g., lapel pin and photo background American flag priming conditions).

**Indicates a study that was excluded from the main analyses; including them does not impact the results.*

to other researchers. Rather, it is meant to shed light on our own data collection efforts—and to empty our file drawer—by making the results of unpublished studies public. That is, the results of this analysis represent the totality of our own empirical investigations on this particular topic, a step we believe is useful for the field as a whole.

GENERAL ANALYTICAL APPROACH

As is the case in any meta-analysis, we had to make a variety of analytical decisions at every stage of the process. In every case, we attempted to choose the approach that would produce the most accurate estimate of the effect—without considering its impact on the outcome, and always erring on the side of caution, recognizing that we might not detect the intrusive influence of our own desires or expectations on the process (Wilson & Brekke, 1994). By default, we assumed every piece of data should be included, with exceptions made only when justified by standard practices in the field (e.g. excluding participants who failed an attention check).

In the sections below, we provide a general rationale for how we made decisions related to manipulated independent variables, moderators, dependent measures, participant exclusions, and study exclusions. A more detailed rationale for the specific decisions made for each study can be found in the supplemental materials. Ultimately, we believe that the specific decisions we made are reasonable and represent the most accurate test of the hypotheses being considered. Nevertheless, we acknowledge that other reasonable people could disagree. Although our decisions rarely altered the outcome (as reported below), we have made the raw data publicly available so that those same reasonable people may test an alternative set of decisions for themselves.

Independent Variables. Because this analysis is intended to track the magnitude of the effect of an American flag prime on political attitudes and beliefs over time, it is important to use the same baseline of comparison (i.e., a proper control condition) across all studies: either no prime at all or a purely neutral image in place of the flag. Similarly, to ensure an apples-to-apples comparison, for each study in the meta-analysis, we must have a single mean value representing the American prime condition and a mean value representing the control condition. However, some of the studies we conducted—particularly after the publication of CFH—involved additional complexity that needed to be simplified in order to achieve the crucial comparison of an American flag condition to a control condition.

These decisions were typically straightforward. Whenever a factor was simply added to the design (e.g. whether the flag prime was presented alongside a Republican or a Democratic politician), we simply collapsed across that factor to create a two-cell design. Whenever there were multiple conditions involving an American flag (e.g., a flag lapel pin vs. a free-standing flag in the background of the image), we combined them into a single American flag condition. There were a few rare cases when the decision was not quite as simple—typically when there was a

condition that clearly did not involve an American flag prime, but was not obviously a neutral control (e.g., United Kingdom flag priming condition). A condition was dropped from the analyses only when it seemed that including participants in that condition would be more likely to produce a misleading than an illuminating effect size estimate. (This applied to only two studies.) All such decisions are described in the supplemental materials and summarized in Table 1.

We conducted the analyses only after making these particular decisions. (Readers are free to conduct analyses with an alternative set of decisions, of course.) However, we did make at least some attempt to discern whether these decisions mattered, after the fact. Each study was classified by whether or not any priming conditions were combined, whether or not any priming conditions were dropped from the analyses, and whether or not there were any other non-priming factors that were collapsed across. Although it is not definitive evidence that these decisions were not problematic, we note that none of these classification variables moderated the findings reported below (all ps > .18).

Dependent Measures. As described above, the focus of the present analysis is the influence of American flag primes on the conceptual dependent variable examined in CFH: endorsement of political conservatism (broadly construed). All of the studies included in this analysis had at least one measure that hewed closely to that conceptual variable. Some of the studies also included measures expressly intended for projects unrelated to political ideology (e.g., racial attitudes). Any measure that was not explicitly related to political ideology was not considered for the analysis, but all such measures are described in the supplemental materials and included in the available raw data.

Following standard procedures in the field, when a given measure was a composite of multiple responses (e.g., a questionnaire with multiple items assessing political beliefs about various issues; warmth ratings towards multiple political figures), these composites contained *all* relevant responses. All measures were scored such that higher numbers indicated greater endorsement of conservatives/Republicans and lower numbers indicated greater support for liberals/Democrats, either through reverse scoring or by creating difference scores (following the procedures used in CFH for structurally similar measures). The precise aggregation procedures for each measure are described in the supplemental materials and can be seen explicitly in the R code used to process each study's raw data into its final form.

One particularly important issue is how to treat multiple relevant dependent measures from the same study. Considering them as separate effects would give studies with multiple measures far too much weight, especially given that multiple responses from the same participants will almost certainly be positively correlated. The standard approaches to dealing with such non-independent effect sizes (see Borenstein, Hedges, Higgins, & Rothstein, 2009; Cheung, 2014) either involve using a single effect size to represent each study or accounting for the non-independence statistically. Because both approaches have strengths and

weaknesses, we opted to use a version of each in order to maximize the robustness of the results. Converging evidence would therefore be better than either approach alone.

For the first approach—a single effect size to represent each study—one option is aggregation: simply create an equally weighted average of every measure in a given study. This seemed ill suited for the present study for several reasons. In many cases, a study included one measure that very directly measured endorsement of conservative/Republican policies over liberal/Democratic policies (e.g., voting intentions) and one that was less direct or in a wholly different modality (e.g., a measure of implicit attitudes toward particular political figures). Although both are certainly relevant to the conceptual variable, the less direct measures may dilute the aggregate's representation of the conceptual variable. In addition, because it is generally assumed that the direct effects of a prime can fade over the subsequent couple of minutes (e.g., E. T. Higgins, 1996; Wilson & Capitman, 1982; Wyer & Srull, 1986), a measure presented simultaneously with the prime would likely show a stronger effect than a measure presented several minutes later. An aggregate of both measures would thus be a weaker test than the first measure alone. Thus, when a study contained more than one relevant dependent measure, we designated one *primary measure* to represent that study in the analysis. This decision was based firstly on the degree to which it reflected endorsement of conservative/Republican policies relative to liberal/Democratic policies, secondarily on its temporal proximity to the prime, and finally examining the number of missing responses from each measure (preferring more data to less). To be clear: Which measure was chosen as primary was made prior to conducting the analyses reported below, and the magnitude or direction of a measure's effect size was *not* taken into account.

The major downside to using only one measure from each study is that it necessarily involves a loss of valid information. The second approach is meant to overcome this particular weakness by making use of all available data and accounting for sources of non-independence statistically. Specifically, given that we had access to all of the raw data, we used linear mixed-effects models to analyze participants' individual responses. By using nested random factors for participant and study, we can account for the non-independence of different responses from the same participants as well as different participants in the same study. Although it does mean using responses to measures that are further from the conceptual variable (and temporally further from exposure to the prime), this approach has the benefit of making full use of the available data. Thus, the weaknesses of one approach are overcome by the strengths of the other. As the results show, these two approaches yielded nearly identical results.

Measured Moderators. As part of our exploration of the nature of the priming effect and its boundary conditions, a number of individual differences measures were identified as potential moderating variables. However, in the interest of simplicity and remaining conservative in our analytical approach (i.e., to avoid a fishing expedition), we did not account for any of these variables in the analyses. They

are described in the relevant section for each study in the supplemental materials, and they are included in each study's full data file.

Study Exclusions. From the 33 unpublished studies identified as being relevant, only two stood out as sufficiently problematic to warrant removal from the main analyses. The two studies (dated October 27, 2012, and November 5, 2016) were conducted in multiple sessions (similar to CFH Study 1), with baseline measures assessed in the first session and the priming manipulation administered in a second, separate online session. In both cases, random assignment to priming condition in the second session failed to eliminate differences between conditions on important measures taken during the first session, such as party affiliation and political ideology. With baseline differences between conditions so closely related to our dependent measures, we believed that the outcomes from these studies—even controlling for the baseline differences—were more likely to obscure than illuminate the effect of an American flag prime on political conservatism, thus warranting their exclusion from the main analysis.[2] It is worth noting that when these studies were included in the meta-analyses, both were identified as outliers by examining the studentized residuals (*ps* < .041). Further, including these studies does not alter the conclusions of any of the analyses reported below. As mentioned above, the raw data are available as part of the supplemental materials.

In addition, because we are reporting the contents of our own file drawer, we do not include the Many Labs study (Klein et al., 2014) in the analyses by default. However, it is important to note that we also conducted the analyses with this study included; its inclusion did not change any of the results or conclusions reported below. The data from this study are included along with the raw data from our own studies to make it easy for others to reproduce these analyses.

Participant Exclusions. By default, all valid responses from every participant were included in every analysis. Put differently, we did not use missing data from a single measure to exclude a participant's data entirely (though obviously if a participant did not respond to a given dependent measure, they would not be part of any analysis of that dependent measure). Following standard practices in the field, there are two potential reasons for excluding a participant's data entirely: lack of attention to the task and suspicion of the prime. In all online studies, which includes most of the studies reported here, it is presumed that some participants are simply attempting to complete the experiment as quickly as possible in order to get paid, and thus are not providing accurate or thoughtful answers to the questions (see Oppenheimer, Meyvis, & Davidenko, 2009). To ensure that participants were putting in some minimal amount of effort—actually reading the questions and giving somewhat meaningful responses—some studies included a very simple attention check question involving easy arithmetic problems (e.g., "What is 6 + 7?"). As is often done in such experiments, participants who failed to give

2. Although there were some between-condition differences on party affiliation/political ideology in other studies, because those measures were assessed after exposure to the priming manipulation, it was impossible to distinguish between legitimate baseline differences and priming effects.

the correct response to this question were excluded for inattention. For the published studies, we followed the same criteria as the original publication when it came to excluding inattentive participants. The number of inattentive participants excluded from each study can be found in Table 1, and the specific criteria used for all unpublished studies can be found in the supplemental materials.

Priming studies in particular are vulnerable to participant suspicion (e.g., Kleiman, Sher, Elster, & Mayo, 2015); if a participant notices the prime and becomes suspicious of the researchers' intentions, she may change her behavior as a result. Although the online format limits the effectiveness of suspicion probes, some studies included general free response questions designed to identify participants who noticed the prime and found it odd. However, we decided to retain participants who expressed suspicion in the analyses for two reasons. First, based on an informal examination of participants' responses to the suspicion probe questions across all of the studies in the analysis, the few participants who mentioned noticing the flag prime rarely expressed concern that it might have influenced their responses, indicating that deliberate shifts in responses due to suspicion were exceedingly rare. Second, unlike attention checks, which are designed to have a clear criterion for passing or failing, suspicion checks are more subjective—particularly when they involve written responses to open-ended questions that differ from study to study—making it difficult to identify a clear standard for exclusion that could be applied consistently across every study. Third, assuming that reactance effects would be stronger than demand effects, including suspicious participants should only serve to make any observed effect of the flag prime weaker, making their inclusion the more conservative approach. Although we doubt that suspicion had any effect on the results, particularly given the low rate of suspicion across studies, which remained fairly constant over time, it seems likely that any effect it might have had would favor the null hypothesis.

Thus, the analyses reported below exclude inattentive participants. (The results do not change when all participants are included; see supplemental materials.)

ANALYSES

After the study and participant exclusions described above, the main analyses reported below included a total of 33 studies (N = 9,656). All analyses were conducted using the statistical software program R (version 3.4.0, R Core Team, 2017). As mentioned above, we used two different approaches to analyze the data in aggregate: meta-analyses and linear mixed-effects models. To reiterate, we conducted a meta-analysis of each study's primary measure, which allows a focused analysis on the measures we identified as being the best representation of the conceptual variable. However, because the other dependent measures certainly are relevant to the question at hand, we also analyzed participants' individual responses to every measure using linear mixed-effects models. The specifics of the meta-analyses and the linear mixed-effects models are described in detail below.

Meta-Analysis. The meta-analyses reported below analyzed the effect size of each study's primary measure. For each study, we calculated the effect size as Hedges *g*: the standardized mean difference (SMD) between the flag and control conditions (Lipsey & Wilson, 2001), with the Hedges (1981) correction for positive bias. Specifically, the following formulas were used (with subscripts *f* and *c* referring to the Flag and Control conditions, respectively):

$$g = \frac{\bar{X}_{\mathrm{f}} - \bar{X}_{\mathrm{c}}}{S_{\mathrm{pooled}}} \times \left(1 - \frac{3}{4(n_{\mathrm{f}} + n_{\mathrm{c}}) - 9}\right)$$

$$S_{\mathrm{pooled}} = \sqrt{\frac{(n_{\mathrm{f}} - 1)S_{\mathrm{f}}^2 + (n_{\mathrm{c}} - 1)S_{\mathrm{c}}^2}{n_{\mathrm{f}} + n_{\mathrm{c}} - 2}}$$

Because meta-analyses are typically conducted in order to make broader inferences, fixed-effects models, which limit the inferences to the particular pool of studies under consideration, are rarely preferred over random-effects models (see Hunter & Schmidt, 2000). Given that our stated intention is to analyze the results of our own data-collection efforts, a fixed-effects approach might actually be appropriate in this case. However, given the heterogeneity in our methods, coupled with our assumption that there is considerable heterogeneity in the effect size over time, it seemed best not to assume that the true effect size is the same across all studies. Thus, we used a random-effects model to estimate the overall effect size. As is typical, the analysis weights each study's observed effect size by the precision of the measure (the inverse of the sampling variance). Thus, more reliable studies—typically those with the largest sample sizes—are weighted more heavily in the effect size estimation.

The meta-analyses were conducted in R using the metafor package (version 1.9–9; Viechtbauer, 2010). We used restricted maximum likelihood (REML) to estimate residual heterogeneity (τ^2), which is the default estimator in metafor, and which is appropriate here as it is an approximately unbiased estimator of the SMD (Viechtbauer, 2005). To account for any uncertainty in the estimated heterogeneity, we also applied the Knapp and Hartung correction (Knapp & Hartung, 2003; see Viechtbauer, López-López, Sánchez-Meca, & Marín-Martinez, 2015), which bases its estimates on a *t* distribution with $k - j$ degrees of freedom (where *k* is the number of studies and *j* is the number of coefficients in the model, including the intercept), rather than a normal distribution. For the mixed-effects analyses, the omnibus test of the moderators is based on an *F* distribution with *i* and $k - j$ degrees of freedom (where *i* is the number of coefficients in the model).[3]

3. This was intended to be the more conservative approach. Indeed, when estimates are made without the Knapp-Hartung correction (i.e., assuming a normal distribution), the results are actually somewhat stronger.

For each of the random and mixed-effects meta-analyses reported below, we followed the standard practice of testing whether any studies in each analysis would be considered outliers or highly influential (Viechtbauer, 2010; Viechtbauer & Cheung, 2010). In the interest of completeness, we report the results of those tests as well as the results of versions of that same analysis with outlier and highly influential studies excluded. However, in the interest of remaining as conservative as possible, because there are no ironclad exclusion criteria for outliers or influence, we believe that the full version of each analysis should be considered the most valid one.

Linear Mixed-Effects Models. We tested for the effect of the prime on participants' individual responses using linear mixed-effects models with the lme4 package (version 1.1–13; Bates, Mächler, Bolker, & Walker, 2015). We used the lmerTest package (version 2.0–33; Kuznetsova, Brockhoff, & Christensen, 2018) to calculate *p* values for significance tests (using Satterthwaite's approximations for the degrees of freedom) and the lsmeans package (version 2.26–33; Lenth, 2016) to estimate cell means and to perform any post hoc or pairwise comparisons.

The fixed effect of prime condition was treated in the analysis as a factor using deviation contrasts (Control = –0.5, Flag = +0.5), and participants' individual responses were converted to *z*-scores (separately for each measure within each study) to ensure that all responses were on the same scale. For the random effects structure, we started with a maximal model (Barr, Levy, Scheepers, & Tily, 2013) and then followed procedures described by Bates and colleagues (Bates, Kliegl, Vasishth, & Baayen, 2018) whereby random-effects parameters were removed until the model was identifiable and reliably converged with valid random-effects parameters (see supplemental materials). The resulting random-effects structure, used for all models, included by-study random intercepts and slopes, and random intercepts for participants (nested within study), without allowing for correlated random intercepts and slopes. For instance, the model testing for the overall effect of the prime was as follows:

$$Y_{ijk} = \beta_0 + \beta_1 \,\mathrm{Prime}_{jk} + \mathrm{Study}_{0k} + \mathrm{Study}_{1k} \times \mathrm{Prime}_{jk} + \mathrm{Participant}_{0jk} + \varepsilon_{ijk}$$

$$\left(\mathrm{Study}_{0k} \;\; \mathrm{Study}_{1k}\right) \sim N\left(0, \begin{bmatrix} \tau_{00}^2 & 0 \\ 0 & \tau_{11}^2 \end{bmatrix}\right)$$

$$\mathrm{Participant}_{0jk} \sim N(0, \omega^2)$$

$$\varepsilon_{ijk} \sim N(0, \sigma^2)$$

where Y_{ijk} is the i^{th} response of the j^{th} participant in the k^{th} study, β_0 is the fixed-effect intercept, β_1 is the slope of the fixed-effect of priming condition (Prime_{jk}), Study_{0k} and S_{1k} are the random intercept and slope for Study (with variances τ_{00}^2 and τ_{11}^2), respectively, P_{0jk} is the random intercept for Participant (with variance

ω^2), and ε_{ijk} is the residual error (with variance σ^2). Models testing study-level moderators used the same random effects structure and added fixed effects for the main effect of that moderator ($\beta_2 \times \text{Mod}_k$) and its interaction with priming condition ($\beta_3 \times \text{Mod}_k \times \text{Prime}_{jk}$). It should be noted that because responses for each measure within a given study were converted to *z*-scores (i.e., each individual measure had $M = 0$ and $SD = 1$), it was not strictly necessary to include by-study random intercepts. They were nonetheless included in order to account for random variability of the priming effect (i.e., random slopes for priming condition) within each study. (This fact may also explain why models that allowed for correlated random slopes and intercepts often failed to converge.)

THE EFFECT OF TIME

Our primary hypothesis was testing for a decline of the flag-priming effect over time. In order to test for a decline effect—that is, to look at the passage of time as a potential moderator as part of a meta-analysis—it was necessary to use a single date to represent each study. Because most studies were carried out over a period of several weeks, this was, by necessity, an approximation in many cases. To arrive at a single representative date, we attempted to determine the start and end dates of data collection for each study, and we used the midpoint of those two dates. Whenever possible, we determined a study's start and end dates via date and time stamps associated with participants' responses. When that information was not available, such as for paper and pencil studies, we estimated the dates using all available information (e.g., creation/modification dates of the digital versions of the study materials and data files). Although there are undoubtedly errors in these estimates, our digital forensics were generally fairly successful, allowing us to narrow the start and end dates to within 1–2 weeks in all but a few cases. Because the broader timeframe of our investigation is measured in years, it seems unlikely that these minor discrepancies would substantially reduce the accuracy of our conclusions.[4]

Of great concern—given that our primary hypothesis concerns the effect of an exogenous variable like Time acting as a moderator in a meta-analysis—is that any observed decline effect is spurious (i.e., due to a confound). In other words, there is some explanation for why the earlier studies tended to show positive effects and later studies did not, other than a decline in the actual effectiveness of the prime over time. We discuss several such possibilities below.

4. In the context of a linear mixed-effects analysis, it would be possible to use participants' individual participation date as the moderating variable. However, because we did not have participant-level date information for every study, that approach may be somewhat problematic. Doing so would require using the date associated with the overall study, effectively making Time a study-level variable for some studies and a participant-level variable for others. Given that most studies were completed within a few months, it seemed the more conservative approach to simply consider Time a study-level variable. (Indeed, the results do in fact grow stronger when Time is treated as a participant-level variable.)

Study Properties. One possibility is that there were changes in the study methodology that led the initially observed effects to grow smaller over time. It is worth noting that this explanation logically implies that the priming effect is indeed real (rather than spurious), even if it is highly contextual. We address this by considering properties of the studies that could act as potential moderators instead of time: priming manipulation, study format, and the type of dependent measure. (Information about the study properties is summarized in Table 1.)

Priming manipulation. This set of studies used slight variations on one of four basic approaches to the priming manipulation. The most common method (k = 22 studies) was manipulating the presence or absence of a small American flag in the top left corner of a survey ("Survey Corner"). The next most common (k = 7 studies) was presenting photographs of American politicians—typically presidential candidates—with an American flag digitally added/removed from the photograph ("Candidate Photo"). This approach was actually two different manipulations being tested as boundary conditions: the flag prime appeared either as a pin on the lapel of the politician or in the background of the photograph. Because these two manipulations always appeared as different conditions in the same study and were always combined into a single flag prime condition, they are not distinguishable here. The next most common method (k = 4 studies) involved a bogus task involving photographs of places that included an American flag or had the flag digitally removed ("Visual Geography Quiz"[5]). The least common method (k = 2 studies) involved a bogus computerized categorization task that presented the American flag or a control stimulus subliminally. The variations of the different techniques used by each study are described in the supplemental materials.

<D>*Study format.* The studies also varied considerably in how participants provided their responses. The majority of studies (k = 24 studies) were conducted online, typically (though not always) recruiting participants from Amazon's Mechanical Turk. The next most common format (k = 7 studies) was a simple paper and pencil survey. Finally, some studies (k = 5 studies) were conducted in a controlled laboratory environment, collecting responses via a laboratory computer. It was not entirely clear how best to create distinct categories, made even more difficult because we lack complete information about precisely how the paper and pencil studies were administered. That is, some paper and pencil studies were conducted in public spaces around a college campus; some were given to participants as an interim task during an unrelated study in the laboratory; and some studies used a combination of the two approaches with no way to distinguish which participants received which administration. Ultimately, we settled on two different categorization schemes; studies were categorized as being conducted online vs. in person (paper/pencil or in the lab), and as being computer-based vs. paper and pencil–based. Both of these possibilities are tested below. The online

5. In some studies, the bogus task was a "visual geography quiz," which involved guessing in which city the photo had been taken. In other studies (e.g., CFH Study 2), the bogus task involved estimating the time of day the photo had been taken, and it was presented to participants as a "daylight estimation quiz."

vs. in-person categorization also neatly divides studies by the population being sampled (Mechanical Turk worker vs. college student), allowing it to serve a dual purpose.

One caveat is important to keep in mind, however: the format of the study is somewhat confounded with the date the study was conducted, with paper and pencil surveys being used early on, then phased out in favor of online studies, primarily in order to recruit considerably larger samples.

Dependent measures. There are many ways to measure political beliefs and attitudes, and although they should all be conceptually related, it is possible that some measures will be more closely associated with the construct(s) being primed, and therefore more strongly affected by the priming manipulation. We attempted to create a categorization scheme that acknowledged conceptual differences between measures without being overly cumbersome and reducing the power to detect these differences. We settled on five categories: 1) expressions of the intention to vote for a Republican candidate over a Democratic candidate (Voting); 2) expressions of (dis)agreement with statements about specific political issues (Political Beliefs); 3) expressions of (dis)approval with the job performance of a politician or political party (Approval); 4) ratings of warmth or positivity/negativity toward politicians and/or political parties (Warmth); and 5) expressions of support for a specific policy after reading the transcript of a speech delivered by a Republican or Democratic politician advocating that policy (Speech).

As was the case for study format, the type of dependent measure used was also somewhat confounded with time. The speech evaluation measures, for instance, were not used until roughly 2012, whereas the political beliefs and voting measures were mostly used very early on. We address this issue in the Results.

False Positives. As discussed above, it is possible that, rather than a real priming effect being moderated by some other factor(s), any positive results (including CHF, 2011) are spurious, attributable to issues with methodology. We address several specific versions of this hypothesis in the following sections.

p-hacking. One version of the false-positive hypothesis is that the apparent decline is due to various forms of *p*-hacking in the data analysis and reporting. That is, the main researcher degrees of freedom that can be exploited (see Wicherts et al., 2016) include choosing advantageous exclusion criteria, not reporting additional dependent variables or experimental conditions, and selective reporting of significant analyses. Thus, a version of this concern would be that we chose the analytical approach that yielded positive results in the early studies, but later results were constrained by those choices and were thus more likely to produce a (true) null effect. This particular concern seems not to apply in this case, because we are applying the same conservative analytical approach to every study. (And in any case, the results hold regardless of the specific set of decisions we made.) As described above and in the supplemental materials, we attempted to make analytical decisions—including participant and study exclusion criteria, the aggregation

or exclusion of conditions, and the selection of the primary measure—that were reasonable, consistent, and conservative, and all such decisions were made a priori. Most importantly, we are being completely transparent with these decisions, and the results of alternative analyses indicate that the particular set of decisions we made do not affect the results. By making available the raw data and the code used to produce the results reported below, we are trying to make it as easy as possible for other researchers to conduct alternative analyses using different assumptions or alternative criteria, thereby directly evaluating the impact of our decisions on the outcome.

Small-sample bias. Relatedly, we must also contend with the possibility that the observed decline effect is due to the fact that the early studies generally used much smaller samples than later studies. That is, larger studies tend to more accurately estimate the true effect size of a given phenomenon, and the smaller early studies were simply less precise than later studies. (It is worth reiterating that, in the meta-analysis, each study's effect size is weighted by its precision, a function of the study's sample size, ensuring that smaller studies do not exert an undue influence on the outcomes. In the linear mixed-effects models, the individual participant is the unit of analysis, which applies equally across time.) However, this possibility is not sufficient on its own; there still must be some reason why the early studies tended to show positive results, and later studies did not.

In a typical meta-analysis, the concern about small studies relates to publication bias: Null or negative results are systematically censored from the set of studies being considered, thus inflating the estimated effect size. In the present analysis, because the set of studies under consideration is entirely those with which we were directly involved, we can be confident that we are not systematically excluding null or negative results. It is also worth noting that, based on our usual lab record-keeping practices (i.e., to keep digital records of the full materials from every study), we have no reason to believe that our search would have been biased against null results. Further, although we are confident that publication bias is not an issue in this case, in the interest of being both thorough and prudent, we endeavored to use all of the means at our disposal to test for this and other potential sources of bias. A lengthy discussion of our approach to this issue, along with the results of several empirical tests for publication bias, can be found in the supplementary materials. In short, we believe the results of these tests support the conclusion that the results we report below are not compromised by publication bias.

Selective stopping. There is, however, another possible explanation for why the early studies tended to show positive results. The concern is a form of *p*-hacking during data collection known as selective stopping: frequently checking the results during data collection and selectively continuing or ceasing data collection based on whether the results are significant. Thus, the issue would be that selective stopping was more likely to be applied in the early studies than the later studies. Although we did not engage in the more extreme forms of selective stopping,

we did not employ a unilateral stopping rule for all of the studies reported here—particularly for those conducted before the importance of such rules was made apparent (i.e., around the time of the publication of Simmons, Nelson, & Simonsohn, 2011). To ensure that even a minor amount of selective stopping could not account for the present results, we conducted a series of simulations. The results of these simulations, which are fully reported in the supplemental materials, clearly demonstrate that even if it were a concern, selective stopping alone is not sufficient to produce a false positive. Only in combination with a high degree of publication bias did it lead to faulty conclusions. As stated above, it seems very unlikely that even a small amount of publication bias is present, thus minimizing concerns that selective stopping can explain the observed results.

RESULTS

For each hypothesis test we report below, we tested the priming effect using the two different analytical approaches described above: meta-analysis of study-level effect sizes using each study's primary measure and linear mixed-effects models examining individual participants' responses. If both approaches produce the same results, we can be confident that the conclusions do not depend on the strengths or weaknesses in either approach.

Before we turn to the main analyses of interest—the effect of flag priming over time—we present an assessment of the overall evidence for a priming effect. We later turn to the potential moderating role played by properties of the studies themselves.

OVERALL PRIMING EFFECT

We first wished to characterize the overall priming effect. In each of the analyses reported below, we used the study and participant exclusion criteria defined above, analyzing the two published studies (CFH Studies 1 and 2) and 31 unpublished studies (excluding only the studies specified above), for a total of $k = 33$ studies ($N = 9{,}656$ participants).

Random-Effects Meta-Analysis. As can be seen in Table 2, the basic random-effects meta-analysis of each study's primary dependent measure estimated an overall priming effect that was not significantly different from zero ($\bar{g} = 0.038$, $p = .145$). Examining the studentized residuals revealed only one study (dated March 6, 2013) to be an outlier ($z = -2.19$, $p = .029$). With this outlier removed from the analysis, the estimated effect size grows only slightly ($\bar{g} = .046$), though it does become marginally significant ($p = .067$). Following the procedures described by Viechtbauer and Cheung (2010), no studies were identified as being highly influential. Although we present the full description of our efforts to assess publication bias in the supplemental materials, as is customary when presenting the results of a meta-analysis, we present a funnel plot (Figure 1), and include the results after applying the trim and fill procedure (Duval & Tweedie, 2000) in Table 2. (It is worth noting

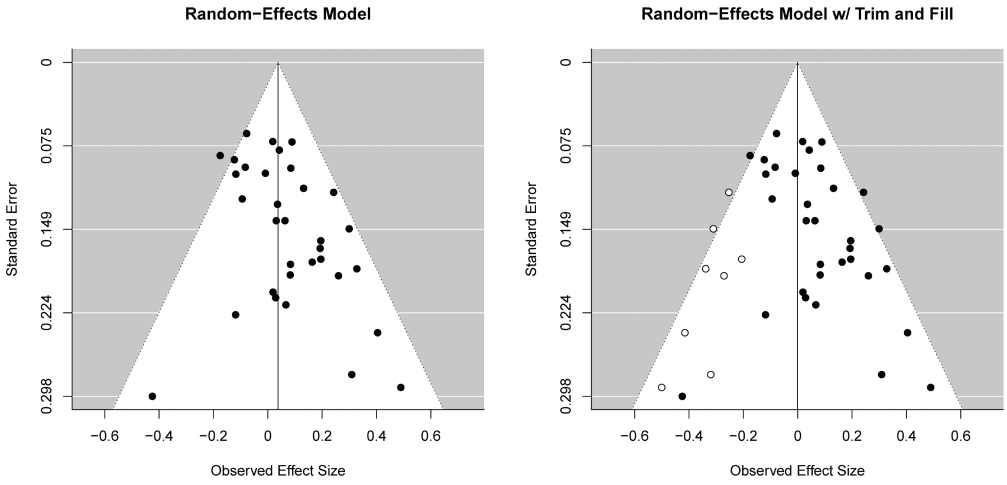

FIGURE 1. Funnel plot of the observed effect sizes (solid circles) of the primary measures in the main meta-analysis (left panel), and including filled-in studies (empty circles) resulting from the trim and fill procedure (right panel).

that, despite its ubiquity, there are serious concerns about the validity of inferences made on the trim and fill procedure, as discussed in the supplemental materials; see also E. C. Carter, Schönbrodt, Gervais, & Hilgard, 2019; Terrin, Schmid, Lau, & Olkin, 2003.)

Linear Mixed-Effects Analysis. The detailed results of the linear mixed-effects analysis of the overall effect of the prime on individual participant responses can be found in Table 3. Consistent with the meta-analysis, this analysis found no overall priming effect ($b = 0.041$, $p = .132$).

MAIN ANALYSES: TESTING FOR A DECLINE EFFECT

As described above, our main analysis is to test the a priori hypothesis that the effect of the flag prime on expressions of political conservatism changed over time.[6] Using each study's representative date, we calculated the amount of time (in

6. Although the basic random-effects model did not show evidence of a large amount of residual heterogeneity, as indicated by the non-significant Cochran's *Q* statistic and the relatively low I^2 statistic (an estimate of the amount of observed heterogeneity not due to random error), we nonetheless proceeded with the planned analysis for several reasons. First, Cochran's *Q* suffers from low power (e.g., Hardy & Thompson, 1998). Second, there is also quite a bit of uncertainty about the I^2 statistic as well (Huedo-Medina, Sánchez-Meca, Marín-Martinez, & Botella, 2006). Indeed, although the point estimate for the I^2 statistic was roughly in the range for very mild heterogeneity (< 25%; (J. P. T. Higgins & Thompson, 2002), the wide confidence intervals, which included values that would be categorized as moderate or large, indicate that these results cannot confidently be interpreted as lacking residual heterogeneity (Ioannidis, 2008; Ioannidis, Patsopoulos, & Evangelou, 2007). Third, particularly when there is uncertainty in the estimates of heterogeneity, prior beliefs can serve to guide interpretation of the statistical tests (Pereira, Patsopoulos, Salanti, & Ioannidis, 2010).

TABLE 2. Random-Effects Meta-Analysis of Primary Measures

			Effect size estimate			Residual heterogeneity			
	k	*N*	$\bar{g}$	*t*	*p*	$\hat{\tau}^2$	Q_e	*p*	I^2
Primary measures	33	9,656	0.038 [–0.014, 0.089]	1.50	.145	0.005 [0.000, 0.022]	41.78	.115	25.32% [0.00, 60.78]
Excluding outliers	32	9,076	0.046 [–0.003, 0.095]	1.90	.067	0.003 [0.000, 0.019]	35.65	.259	15.29% [0.00, 56.16]
Trim and fill	41	-	–0.001 [–0.053, 0.051]	–0.04	.967	0.009 [0.002, 0.045]	66.19	.006	34.90% [11.57, 73.66]

Note. The Trim and fill model is the model based on the estimated 8 (*SE* = 3.79) missing studies being included with the rest of the data. For the effect size estimate (), the estimated amount of residual heterogeneity (), and the percentage of variability due to heterogeneity (I^2), 95% confidence intervals are reported in brackets.

years) that had elapsed since the earliest date in the analysis (i.e., the zero point) and treated it as a continuous predictor.

Mixed-Effects Meta-Analysis. We fitted a mixed-effects meta-regression model to test the hypothesized decline in the effect size over time. As can be seen in Table 4, the analysis produced a significant intercept (b = 0.203, p = .006), as well as a significant effect for time (b = −0.027, p = .014). This can be interpreted as a relatively small but positive flag priming effect when data collection began, which has declined by 0.027 in each year since (see Figure 2). The same results obtain when excluding the one study identified as an outlier (dated March 6, 2013, z = −2.46, p = .014) or when excluding the one study identified as highly influential (dated November 5, 2016).

Although none of the other study-attribute variables tested showed potential as moderators (as described below), as a check on robustness, we tested whether this decline effect would remain when controlling for other study variables in a separate meta-regression controlling for each moderator. In only one case, when controlling for study format (online vs. in-person studies), did the effect of time drop to marginal significance (p = .099). Importantly, the effect of the study format moderator was itself not significant (p = .694). The fact that study format offers virtually no predictive power in a model with time strongly indicates that the marginally significant result for study format observed above is simply due to it being confounded with time, and the slight drop in the significance of time as a moderator is therefore not surprising.

There are two points that we find important to highlight here. First, as Figure 2 clearly shows, study size is confounded with time. That is, sample sizes were smaller in earlier than later studies. It is critical therefore to note that the meta-analytic approach we adopted addresses this directly by taking sample size into account (via its weighting procedure). However, one could be concerned that the larger studies conducted later are nonetheless "tipping the scales" to produce a

TABLE 3. Overall Priming Effect: Linear Mixed-Effects Model

Effect	*SD*		*b*		*t*	*df*	*p*
Random effects							
Study							
Intercept	0.000	[0.000, 0.020]					
Prime	0.083	[0.000, 0.143]					
Participant (in Study)							
Intercept	0.820	[0.803, 0.836]					
Residual	0.569	[0.559, 0.580]					
Fixed effects							
Intercept			–0.000	[–0.020, 0.019]	–0.05	9175.95	.963
Prime			0.041	[–0.011, 0.098]	1.56	23.94	.132

Note. The results of the linear mixed-effects models predicting participants' individual responses on all dependent measures (N = 9,656). Numbers in brackets are 95% confidence intervals.

decline despite no initial effect. Importantly, when examining just the early studies (i.e., prior to 2011, when a gap in data collection occurred), which ensures that the analysis is unaffected by the larger studies conducted later, there is indeed a significant overall effect ($\bar{g}$ = 0.142, p = .011). This indicates that there was indeed an initial positive effect from which to decline.

Crucially, if one is concerned that the CFH studies were a fluke, then including them in the meta-analysis (and especially in the previous analysis) may bias the results. (Note, of course, that the logic of meta-analyses suggests that all studies should be included, including flukes.) However, to be conservative, we examined the decline effect without the two studies from CFH (i.e., just the unpublished data), and indeed, it is significant even without them ($b_{intercept}$ = 0.164, p = .032; b_{time} = –0.022, p = .049; see Table 4).

Linear Mixed-Effects Analysis. As mentioned above, Time was treated as a study-level variable, using the same basic approach as the other moderator tests: Prime, Time, and the Prime × Time interaction as fixed effects, with by-study random intercepts and slopes, and random intercepts for participants (nested within study). The results of this analysis correspond closely to the results of the meta-analysis (see Table 5). Indeed, the parameter estimates for the main effect for Prime (b = 0.194, p = .008), as well the Prime × Time interaction (b = –0.025, p = .023), are very similar to the intercept and Time parameters from the mixed-effects meta-analysis above. In models controlling for the other study-level moderators tested below (whether any priming conditions were combined or dropped; whether any independent variables were included; type of manipulation; online vs. in person; paper/pencil vs. computer; type of dependent measure), the main effect for Prime remained significant in all cases (all ps < .019) though the Prime × Time interaction dropped to non-significance when controlling for study format (online vs. in

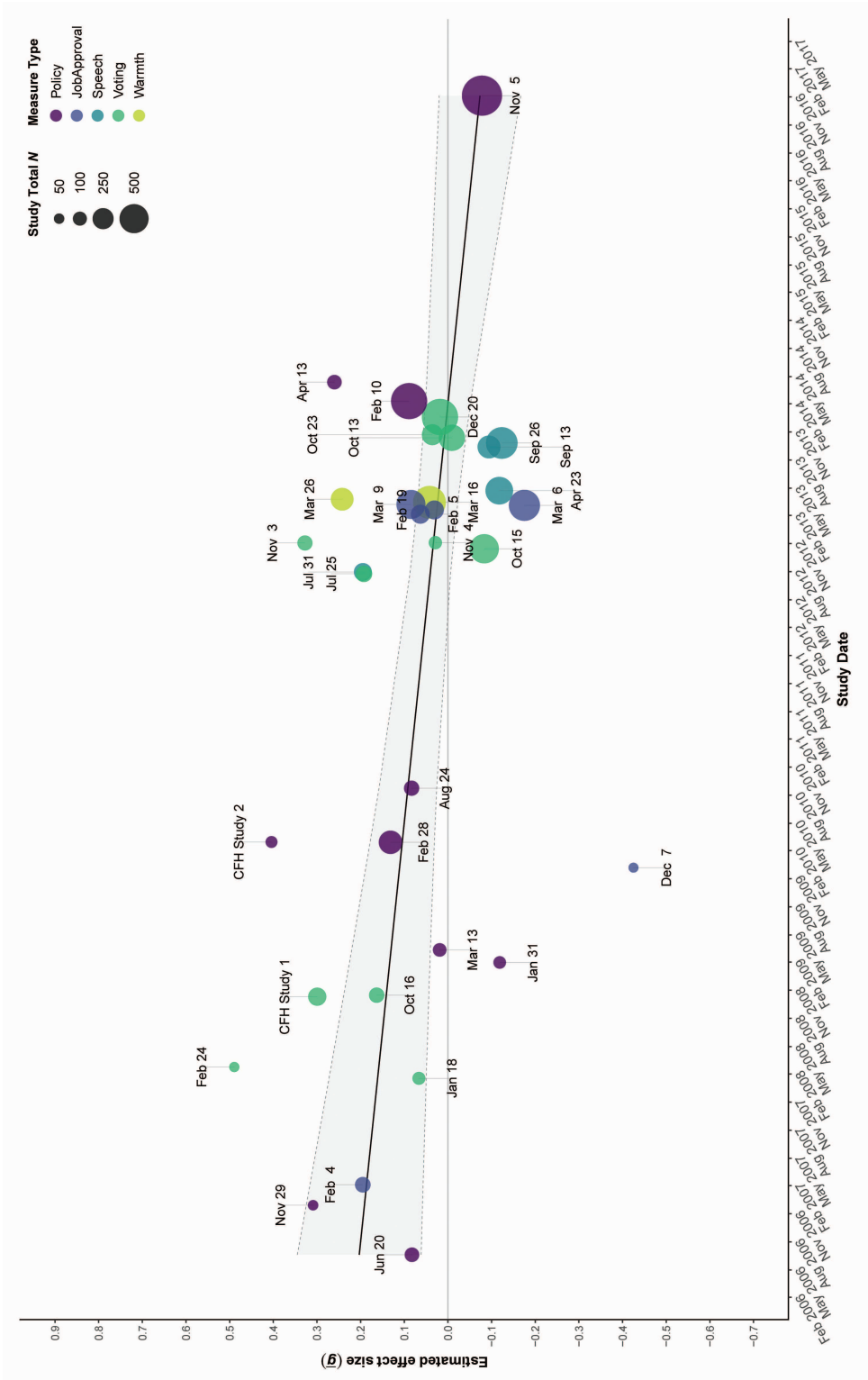

FIGURE 2. Observed effect sizes of the primary measures over time by measure type. The size of each circle indicates the study's total sample size. The solid black line depicts the predicted values generated by the mixed-effects meta-regression, with the grey shaded region depicting the 95% confidence region about that prediction.

TABLE 4. Time as a Moderator: Mixed-Effects Meta-Regressions

	Parameter estimates				Moderators		Residual heterogeneity			
	Parameter	***b***	***t***	***p***	***F*** (Q_m)	***p***	$\hat{\tau}^2$	Q_e	***p***	I^2
Primary measures	Intercept	0.203	2.92	.006	6.85	.014	0.002	33.97	.326	11.91%
k = 33		[0.061, 0.345]					[0.000, 0.017]			[0.00, 53.63]
N = 9,656	Time	–0.027	–2.62	.014						
		[–0.047, –0.006]								
Primary measures	Intercept	0.213	3.31	.002	8.28	.007	0.000	27.94	.574	0.02%
(Excl. outliers)		[0.082, 0.345]					[0.000, 0.014]			[0.00, 47.64]
k = 32	Time	–0.026	–2.88	.007						
N = 9,076		[–0.045, –0.008]								
Primary measures	Intercept	0.201	2.56	.016	4.47	.043	0.003	33.97	.282	15.73%
(Excl. influential)		[0.041, 0.360]					[0.000, 0.019]			[0.00, 54.77]
k = 32	Time	–0.026	–2.11	.043						
N = 8,664		[–0.051, –0.001]								
Primary measures	Intercept	0.164	2.56	.016	4.24	.049	0.002	30.97	.367	11.22%
(Excl. CFH)		[0.015, 0.313]					[0.000, 0.016]			[0.00, 52.63]
k = 31	Time	–0.022	–2.06	.049						
N = 9,403		[–0.043, –0.000]								

Note. The intercept parameter is the estimated effect size at the beginning of data collection. The parameter for Time estimates the decline from that initial effect size in each subsequent year. Numbers in brackets are the upper and lower limits of 95% CIs.

TABLE 5. Time as a Moderator: Linear Mixed-Effects Model

Model and effects	*SD*		*b*		*t*	*df*	*p*
Random effects							
Study							
Intercept	0.000	[0.000, 0.020]					
Prime	0.064	[0.000, 0.120]					
Participant (in Study)							
Intercept	0.820	[0.803, 0.836]					
Residual	0.569	[0.559, 0.580]					
Fixed effects							
Intercept			0.002	[–0.059, 0.062]	0.06	8250.35	.950
Prime			0.194	[0.057, 0.332]	2.75	51.40	.008
Time			–0.000	[–0.009, 0.008]	–0.08	7920.20	.938
Prime × Time			–0.025	[–0.046, –0.005]	–2.39	29.76	.023

Note. Results of linear mixed-effects model testing decline effect (N = 9,656). Numbers in brackets are 95% confidence intervals.

person; $p = .152$), and to marginal significance in two other cases (ps < .075). However, because none of the moderators' main effects showed any evidence of having predictive power (all ps > .83), nor did they interact with Prime (all ps > .21), we believe these models should be interpreted cautiously, and do not detract from the conclusion that the priming effect declined over time.

Considering the consistency of the results from both approaches, the decline result seems to be reasonably robust, though we discuss other interpretations below.

MODERATOR ANALYSES: STUDY ATTRIBUTES

Next, we report the planned tests of whether properties of the studies themselves serve as potential moderators, as described above.

Mixed-Effects Meta-Analysis. As can be seen in Table 6, the overall moderator tests (Q_m) for type of priming manipulation, study format (in person vs. online, and paper and pencil vs. computer), and type of dependent measure were nonsignificant (all ps > .065), although some of the individual parameter estimates were significant or marginally significant. When the analyses were conducted without studies identified as outliers or highly influential, some of the individual model parameters became significant, though not the overall test of moderators. As mentioned above, we consider the first analysis (including outlying or influential studies) to be the more valid result.

TABLE 6. Study Attributes as Moderators: Mixed-Effects Meta-Regressions

	Parameter estimates				Moderators		Residual heterogeneity			
	Parameter	***b***	***t***	***p***	***F*** (Q_m)	***p***		Q_e	***p***	I^2
Manipulation	Intercept	0.079	1.53	.137	1.13	.353	0.004	37.46	.135	23.19%
k = 33	(Corner of survey)	[–0.027, 0.185]					[0.000, 0.024]			[0.00, 62.82]
N = 9,656	Candidate Photo	–0.199	–1.60	.121						
		[–0.453, 0.056]								
	Subliminal	0.089	0.34	.738						
		[–0.450, 0.628]								
	VGQ	0.180	1.02	.317						
		[–0.182, 0.542]								
In person vs. Online	Intercept	0.083	2.39	.023	3.66	.065	0.003	37.39	.199	19.57%
k = 33	(In person)	[0.012, 0.155]					[0.000, 0.020]			[0.00, 57.81]
N = 9,656	Online	–0.134	–1.91	.065						
		[–0.276, 0.009]								
Paper vs. Computer	Intercept	0.088	2.00	.054	2.03	.164	0.004	39.16	.149	22.15%
k = 33	(Paper)	[–0.002, 0.177]					[0.000, 0.022]			[0.00, 60.51]
N = 9,656	Computer	–0.125	–1.43	.164						
		[–0.305, 0.054]								
Dependent measure	Intercept	0.034	1.28	.210	1.42	.254	0.005	34.73	.178	23.52%
(All)	(Political Beliefs)	[–0.021, 0.089]					[0.000, 0.022]			[0.00, 58.10]
k = 33	Approval	–0.098	–0.93	.358						
N = 9,656		[–0.312, 0.116]								
	Speech	–0.208	–1.86	.073						
		[–0.437, 0.021]								
	Voting	0.079	0.90	.375						
		[–0.101, 0.259]								
	Warmth	0.166	1.23	.229						
		[–0.110, 0.442]								
Dependent measure	Intercept	0.048	1.65	.109	0.42	.521	0.006	41.43	.100	27.16%
(Policy vs. others)	(Political Beliefs)	[–0.011, 0.107]					[0.000, 0.024]			[0.00, 61.20]
k = 33	Others	–0.038	–0.65	.521						
N = 9,656		[–0.155, 0.080]								

Note. The intercept parameter for each model represents the estimated effect size at the baseline level of the moderator, which is indicated in parentheses. 95% confidence intervals for model parameters and estimates are reported in brackets.

Linear Mixed-Effects Analysis. For each of the moderators described above, we conducted a linear mixed-effects analysis predicting participants' individual responses using Prime condition, the moderating variable (with one level serving as the reference level), and their interaction as fixed effects, with random intercepts for Study and for Participant nested within Study. The results of these analyses (see Table 7) largely adhere to the corresponding meta-analyses. Although some of the individual parameter estimates were significant, none of the overall tests of the moderator were significant, indicating that those results should be interpreted with caution.

Overall, there does not seem to be strong evidence that variability in how the study was conducted made a substantial difference in the magnitude of the priming effect.

GENERAL DISCUSSION

The present study was designed to attempt to clarify the nature of the earlier published findings (CFH) showing that American flag primes produced conservative shifts in political beliefs and values. We hypothesized that exposure to an American flag produced conservative shifts for several years after we began collecting data in 2006, but this effect then declined to be indistinguishable from zero in the following years.

There is, of course, an obvious alternative hypothesis: that the initially observed effect was spurious, and any apparent decline was simply coincidental with methodological changes or artifacts. We endeavored to make sure this alternative was fairly considered in our analyses, and, based on these analyses, we conclude that it is not the best explanation for our data. We do maintain, though, that it remains a plausible explanation of the results, and that reasonable people might prefer this explanation. To be sure, the evidence in favor of a real effect that has since declined is probabilistic and incomplete. Although we think it unlikely, we cannot rule out the possibility that despite our practice of documenting and preserving every study we conduct and our best efforts to be exhaustive in our search, an early study might not have been properly documented and failed to show up in our search. If we had had the foresight to realize we were studying a dynamic phenomenon, our approach would have been more systematic (not leaving large gaps between periods of data collection, more consistent use of measures and priming methodologies, etc.), which would have allowed a more definitive test. Unfortunately, it is also not possible for us to follow the prescribed approach: conduct a direct replication circa 2008.

Still, we believe that the evidence we have gathered here broadly supports the conclusion that the effect was real and then declined over time. Why would this be the case? To be clear, we are not contending that it is *time* per se that caused the decline; time is merely an index of change in some other psychological variable(s). The objective of the present article is to document the decline effect while we continue working to understand the underlying mechanism. However, we can offer some speculation about plausible mechanisms.

TABLE 7. Study Attributes as Moderators: Linear Mixed-Effects Models

Moderator and effects	*b*		*t*	*F*	*df*	*p*
Manipulation						
Parameter estimates						
Intercept	0.001	[–0.041, 0.044]	0.05		17132.01	.958
Prime (Corner of survey)	0.058	[–0.042, 0.155]	1.14		92.07	.257
Candidate Photo	0.005	[–0.097, 0.106]	0.09		16917.43	.926
Subliminal	–0.000	[–0.244, 0.244]	–0.00		17248.37	1.000
VGQ	–0.001	[–0.094, 0.092]	–0.02		17111.98	.983
Prime × Candidate Photo	–0.118	[–0.363, 0.127]	–0.91		51.55	.365
Prime × Subliminal	0.141	[–0.382, 0.671]	0.51		189.75	.607
Prime × VGQ	0.063	[–0.214, 0.339]	0.43		31.53	.667
Overall tests						
Manipulation				0.01	3, 16126.9	.998
Prime × Manipulation				0.47	3, 31.6	.704
In person vs. Online						
Parameter estimates						
Intercept	0.001	[–0.031, 0.033]	0.05		9914.97	.960
Prime (In person)	0.085	[0.015, 0.154]	2.39		104.76	.019
Online	–0.002	[–0.066, 0.061]	–0.07		9914.97	.945
Prime × Online	–0.132	[–0.271, 0.007]	–1.86		104.76	.066
Paper vs. Computer						
Parameter estimates						
Intercept	–0.001	[–0.040, 0.039]	–0.03		9541.08	.980
Prime (Paper)	0.087	[0.001, 0.172]	1.98		118.36	.050
Computer	0.001	[–0.078, 0.079]	0.01		9541.08	.989
Prime × Computer	–0.116	[–0.287, 0.055]	–1.33		118.36	.187
Dependent measure						
Parameter estimates						
Intercept	0.001	[–0.024, 0.025]	0.06		15383.94	.954
Prime (Political Beliefs)	0.036	[–0.023, 0.097]	1.22		43.54	.230
IAT	–0.019	[–0.172, 0.133]	–0.25		6251.98	.802
Approval	0.002	[–0.054, 0.058]	0.06		8669.08	.950
Speech	0.033	[–0.062, 0.127]	0.68		15077.74	.499
Voting	–0.012	[–0.067, 0.044]	–0.41		10574.03	.681
Warmth	0.000	[–0.054, 0.054]	0.00		9003.35	.997
Prime × IAT	0.078	[–0.227, 0.385]	0.50		5567.70	.619
Prime × Approval	0.027	[–0.091, 0.142]	0.46		694.31	.646
Prime × Speech	–0.255	[–0.477, –0.033]	–2.23		50.80	.030
Prime × Voting	0.048	[–0.071, 0.165]	0.79		334.55	.429
Prime × Warmth	0.052	[–0.063, 0.164]	0.90		521.91	.370
Overall tests						
Dependent measure				0.13	5, 9638.2	.987
Prime × Dependent measure				1.09	5, 253.0	.369

Note. Results of the fixed-effects terms from the linear mixed-effects analysis testing properties of the study as moderating variables (see supplemental materials for details on the random effects). The number of studies and participants was the same for all models ($k = 33$; $N = 9{,}656$). The parameter estimate for the main effect of Prime is equivalent to the priming effect at the reference level of the moderating variable (given in parentheses). For moderators with more than two levels, an overall main effect of the moderator and its interaction with prime are also provided. Numbers in brackets are 95% confidence intervals.

Although the rationale we developed above is possible (i.e., a decline effect due to changes in the meaning of the stimulus), in the end what we document here is a change over time that correlates with duration of a Democratic administration. Like every correlation, there may be other explanations.

As we argued above, the purported historical changes in social and political culture that occurred during this decade of research may have changed the meaning of the flag, and thus the knowledge associated with it. Although we can offer no direct evidence in support of this idea (which is beyond the scope of the present article), the logic is fairly straightforward, and supported by theory (e.g., Schröder & Thagard, 2013). Because exposure to a stimulus activates the information associated with it, if those associations change, so will any priming effects produced by that stimulus. The election of Democrat Barack Obama in late 2008 was widely seen as a historical change to political culture in America. This leads to an obvious question, which we attempted to address in CFH Study 2: Could American flag-priming effects be a reflection of the political views of the party currently occupying the White House, rather than being inherently tied to the Republican party and political conservatism? The current analysis suggests that the "Republicanism" of the American flag did indeed diminish over Obama's presidency. This interpretation is supported by other findings showing that implicit associations with U.S. symbols have changed over this approximate time period (Devos & Ma, 2012; Ma & Devos, 2013).

Thus, given the amount of data we collected initially and the robustness of the effects of flag priming in other countries (Gangl et al., 2016; Hassin et al., 2007, 2009; Kardosh, Carter, Ferguson, & Hassin, 2017; Sibley, Hoverd, & Duckitt, 2011), as well as flag priming effects in the United States on other dependent variables (Callahan & Ledgerwood, 2016; Chan, 2017; Kalmoe & Gross, 2016), the phenomenon of flag priming seems to be real, though inconsistent and something of a moving target. Indeed, the outcome of the 2016 American presidential election—a shift in power toward the Republican Party after eight years of a Democratic presidential administration, coupled with near-unprecedented backlash from progressives and Democrats—may indicate that the target has moved yet again.

To be sure, the fact that the effect was indistinguishable from zero in the latest studies in our analysis is not necessarily an indication that the flag has taken on a neutral/moderate association. It is also possible that some unmeasured moderator is creating a polarizing effect such that the prime pushes people in opposite directions, which are cancelling each other out. We note that other researchers have identified moderators for the effect of the American flag on Republicanism. Kalmoe and Gross (2016) conducted three experiments surrounding the 2012 and 2016 American presidential elections. Across all three experiments, they found evidence of a modest advantage for Republican candidates among flag-primed participants high in symbolic patriotism, racial prejudice, and Republican affiliation. Another recent paper (Chan, 2017) reports a polarizing effect: When primed with an American flag, compared with a control condition, Democrats more strongly endorsed Democratic values, and Republicans more strongly endorsed Republican values. Both of these papers indicate idiosyncrasies in how American

flag primes influenced participants' attitudes and beliefs. Learning more about those idiosyncrasies is crucial to reconciling the findings in America from our own lab as well as these completely independent labs, and with the different findings that have emerged in other countries. It does seem plausible that, with a complete model of how priming effects work in their proper historical and cultural context—with good measures of the relevant moderators—it would be possible to detect the effects of American flag primes more accurately. That work is ongoing.

The American flag is surely one of the most culturally, historically, and socially rich symbols, and one toward which people feel strongly. There are laws governing its handling, and people risk their lives to preserve its honor. Understanding the effects of such a psychologically powerful symbol on the citizenry remains a central question in our ongoing research, and we remain open to whatever the evidence indicates about whether, when, and how this symbol affects people's attitudes and behavior.

On the Authors' Decisions. Readers may wonder about our decision to selectively submit for publication (for the CFH paper) only two of the studies we had conducted. On the one hand, we note that norms surrounding best practices were considerably different then. On the other hand, we did not believe at the time that we were ignoring a file drawer problem. We had both successful as well as unsuccessful studies at the time we submitted the two experiments that became CFH, with the belief that these other studies would form the basis for new lines of research exploring mediators and moderators of our effects. Although we are undoubtedly guilty of the same sort of wishful thinking that plagues many researchers—interpreting successful studies as obviously valid while scrutinizing the faults of null results—a continuously cumulative meta-analysis (CCMA; see Braver, Thoemmes, & Rosenthal, 2014) on all studies up through the publication of CFH shows a consistently positive effect (see supplemental materials). Even if we omit CFH Study 2, the CCMA estimates a positive overall effect for studies conducted up until February 2013 ($p = .013$). Thus, regardless of our motivations, even according to contemporary best practices, the balance of evidence favored the existence of the priming effect when we submitted the manuscript. We fully acknowledge that our reasons for including only these two studies were likely buttressed by our motivated cognition, and that we would not make the same decision today.

CONCLUSION

Overall, we believe that the best interpretation of our results is that the effects of flag priming documented in CFH have declined over time, though we maintain that other interpretations, including the possibility that the effects are spurious, are plausible. Regardless of one's views of the current data, we believe that any priming effects that involve social, cultural, or political knowledge must be assessed within a specific historical context. Others have recently argued for the importance of moderators of priming effects (Cesario, 2014; McGuire, 2013), though it is clearly important to do more than speculate about hidden or unknown moderators (cf.

Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). We view the current work as an illustrative case study about the critical importance of the historical context (time) as a moderator for symbols that are inextricably linked with a dynamic culture, and we believe that the present data offer evidence that this is more than speculation—even if the insights are incomplete. In addition, we view the emptying of one's own file drawer as an increasingly important step toward fully understanding the range and robustness of an effect. We welcome and strongly support the revolution in best practices in the contemporary literature on open science—publicly sharing data, materials, analyses, and unpublished papers—and believe that this will lead to more robust findings going forward. Of course, the same logic applies retroactively; the future of our science also requires efforts like this one, shining disinfecting sunlight on the pre-revolutionary past.

REFERENCES

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2018). *Parsimonious mixed models*. https://arxiv.org/abs/1506.04967v2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beinart, P. (2016, February). Why America is moving left. *The Atlantic*. https://www.theatlantic.com/magazine/archive/2016/01/why-america-is-moving-left/419112/

Billig, M. (1995). *Banal nationalism*. Thousand Oaks, CA: Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. https://doi.org/10.1002/9780470743386

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. https://doi.org/10.1177/1745691614529796

Butz, D. A., Plant, E. A., & Doerr, C. E. (2007). Liberty and justice for all? Implications of exposure to the U.S. flag for intergroup relations. *Personality and Social Psychology Bulletin*, *33*(3), 396–408. https://doi.org/10.1177/0146167206296299

Callahan, S. P., & Ledgerwood, A. (2016). On the psychological function of flags and logos: Group identity symbols increase perceived entitativity. *Journal of Personality and Social Psychology*, *110*(4), 528–550. https://doi.org/10.1037/pspi0000047

Carney, D. R., Jost, J. T., Gosling, S. D., & Potter, J. (2008). The secret lives of liberals and conservatives: Personality profiles, interaction styles, and the things they leave behind. *Political Psychology*, *29*(6), 807–840. https://doi.org/10.1111/j.1467-9221.2008.00668.x

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science*, *2*(2), 115–144. https://doi.org/10.1177/2515245919847196

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011a). Implicit nationalism as system justification: The case of the United States of America. *Social Cognition*, *29*(3), 341–359.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011b). A single exposure to the American flag shifts support toward Republicanism up to 8 months later. *Psychological Science*, *22*(8), 1011–1018. https://doi.org/10.1177/0956797611414726

Cesario, J. (2014). Priming, replication, and the hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Chan, E. Y. (2017). Exposure to the American flag polarizes democratic-republican ide-

ologies. *British Journal of Social Psychology*. https://doi.org/10.1111/bjso.12197

Cheung, M. W. L. (2014). Modeling dependent effect sizes with three-level meta-analyses: A structural equation modeling approach. *Psychological Methods*, *19*(2), 211–229. https://doi.org/10.1037/a0032968

Devos, T., & Banaji, M. R. (2005). American = White? *Journal of Personality and Social Psychology*, *88*(3), 447–466. https://doi.org/10.1037/0022-3514.88.3.447

Devos, T., & Ma, D. S. (2008). Is Kate Winslet more American than Lucy Liu? The impact of construal processes on the implicit ascription of a national identity. *British Journal of Social Psychology*, *47*(2), 191–215. https://doi.org/10.1348/014466607X224521

Devos, T., & Ma, D. S. (2012). How "American" is Barack Obama? The role of national identity in a historic bid for the White House. *Journal of Applied Social Psychology*, *43*(1), 214–226. https://doi.org/10.1111/jasp.12069

Dimock, M. (2017, January 10). How America changed during Barack Obama's presidency. *Pew Research Center*. http://www.pewresearch.org/2017/01/10/how-america-changed-during-barack-obamas-presidency/

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Ehrlinger, J., Plant, E. A., Eibach, R. P., Columb, C. J., Goplen, J. L., Kunstman, J. W., & Butz, D. A. (2011). How exposure to the confederate flag affects willingness to vote for Barack Obama. *Political Psychology*, *32*(1), 131–146. https://doi.org/10.1111/j.1467-9221.2010.00797.x

Ferguson, M. J., & Hassin, R. R. (2007). On the automatic association between America and aggression for news watchers. *Personality and Social Psychology Bulletin*, *33*(12), 1632–1647. https://doi.org/10.1177/0146167207307493

Gangl, K., Torgler, B., & Kirchler, E. (2016). Patriotism's impact on cooperation with the state: An experimental study on tax compliance. *Political Psychology*, *37*(6), 867–881. https://doi.org/10.1111/pops.12294

Gawronski, B., & Strack, F. (Eds.). (2012). *Cognitive consistency: A fundamental principle in social cognition.* New York: Guilford. http://www.worldcat.org/title/cognitive-consistency-a-fundamental-principle-in-social-cognition/oclc/757931779

Gellner, E. (2009). *Nations and nationalism*. Ithaca, NY: Cornell University Press.

Goldman, S. K., & Mutz, D. C. (2014). *The Obama effect: How the 2008 campaign changed white racial attitudes*. New York: Russell Sage Foundation.

Greenwald, A. G., Carnot, C., Beach, R., & Young, B. (1987). Increasing voting behavior by asking people if they expect to vote. *Journal of Applied Psychology*, *72*(2), 315–318.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*(8), 841–856. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<841::AID-SIM781>3.0.CO;2-D

Hassin, R. R., Ferguson, M. J., Kardosh, R., Porter, S. C., Carter, T. J., & Dudareva, V. (2009). Précis of implicit nationalism. *Annals of the New York Academy of Sciences*, *1167*, 135–145. https://doi.org/10.1111/j.1749-6632.2009.04734.x

Hassin, R. R., Ferguson, M. J., Shidlovski, D., & Gross, T. (2007). Subliminal exposure to national flags affects political thought and behavior. *Proceedings of the National Academy of Sciences*, *104*(50), 19757–19761. https://doi.org/10.1073/pnas.0704679104

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Higgins, E. T. (1996). Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. W. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp. 133–168). New York: Guilford.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. https://doi.org/10.1002/sim.1186

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods*, *11*(2), 193–206. https://doi.org/10.1037/1082-989X.11.2.193

Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. Random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*(4), 275–292. https://doi.org/10.1111/1468-2389.00156

Ioannidis, J. P. A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice*, *14*(5), 951–957. https://doi.org/10.1111/j.1365-2753.2008.00986.x

Ioannidis, J. P. A., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *British Medical Journal*, *335*(7626), 914–916. https://doi.org/10.1136/bmj.39343.408449.80

Kalmoe, N. P., & Gross, K. (2016). Cueing patriotism, prejudice, and partisanship in the age of Obama: Experimental tests of U.S. flag imagery effects in presidential elections. *Political Psychology*, *37*(6), 883–899. https://doi.org/10.1111/pops.12305

Kardosh, R., Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2017). *Subliminal exposure to adversary's flag increases bi-national conflict: The Israeli-Palestinian case* [manuscript submitted for publication].

Kemmelmeier, M., & Winter, D. G. (2008). Sowing patriotism, but reaping nationalism? Consequences of exposure to the American flag. *Political Psychology*, *29*(6), 859–879. https://doi.org/10.1111/j.1467-9221.2008.00670.x

Kleiman, T., Sher, N., Elster, A., & Mayo, R. (2015). Accessibility is a matter of trust: Dispositional and contextual distrust blocks accessibility effects. *Cognition*, *142*, 333–344. https://doi.org/10.1016/j.cognition.2015.06.001

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Štěpán Bahník, Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Knapp, G., & Hartung, J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, *22*(17), 2693–2710. https://doi.org/10.1002/sim.1482

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2018). *lmerTest: Tests in linear mixed effects models* (Version 3.0–1) [R]. https://CRAN.R-project.org/package=lmerTest

Landau, M. J., Solomon, S., Greenberg, J., Cohen, F., Pyszczynski, T., Arndt, J., Miller, C. H., Ogilvie, D. M., & Cook, A. (2004). Deliver us from evil: The effects of mortality salience and reminders of 9/11 on support for President George W. Bush. *Personality and Social Psychology Bulletin*, *30*(9), 1136–1150. https://doi.org/10.1177/0146167204267988

Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. *Journal of Neuroendocrinology*, *28*(4). https://doi.org/10.1111/jne.12384

Lenth, R. V. (2016). Least-squares means: The R package lsmeans. *Journal of Statistical Software*, *69*(1), 1–33. https://doi.org/10.18637/jss.v069.i01

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thosand Oaks, CA: Sage.

Ma, D. S., & Devos, T. (2013). Every heart beats true, for the red, white, and blue: National identity predicts voter support. *Analyses of Social Issues and Public Policy*, *14*(1). https://doi.org/10.1111/asap.12025

McGuire, W. J. (2013). An additional future for psychological science. *Perspectives on Psychological Science*, *8*(4), 414–423. https://doi.org/10.1177/1745691613491270

Nosek, B. A. (2005). Moderators of the relationship between implicit and explicit evaluation. *Journal of Experimental Psychology: General*, *134*(4), 565–584. https://doi.org /10.1037/0096-3445.134.4.565

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of*

Experimental Social Psychology, *45*(4), 867–872. https://doi.org/10.1016/j.jesp.2009.03.009

Pereira, T. V., Patsopoulos, N. A., Salanti, G., & Ioannidis, J. P. A. (2010). Critical interpretation of Cochran's Q test depends on power and prior assumptions about heterogeneity. *Research Synthesis Methods*, *1*(2), 149–161. https://doi.org/10.1002/jrsm.13

Rydell, R. J., Hamilton, D. L., & Devos, T. (2010). Now they are American, now they are not: Valence as a determinant of the inclusion of African Americans in the American identity. *Social Cognition*, *28*(2), 161–179. https://doi.org/10.1521/soco.2010.28.2.161

Schmidt, K., & Nosek, B. A. (2010). Implicit (and explicit) racial attitudes barely changed during Barack Obama's presidential campaign and early presidency. *Journal of Experimental Social Psychology*, *46*(2), 308–314. https://doi.org/10.1016/j.jesp.2009.12.003

Schröder, T., & Thagard, P. (2013). The affective meanings of automatic social behaviors: Three mechanisms that explain priming. *Psychological Review*, *120*(1), 255–280. https://doi.org/10.1037/a0030972

Sibley, C. G., Hoverd, W. J., & Duckitt, J. (2011). What's in a flag? Subliminal exposure to New Zealand national symbols and the automatic activation of egalitarian versus dominance values. *Journal of Social Psychology*, *151*(4), 494–516. https://doi.org/10.1080/00224545.2010.503717

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Skitka, L. J. (2005). Patriotism or nationalism? Understanding post-September 11, 2001, flag-display behavior. *Journal of Applied Social Psychology*, *35*(10), 1995–2011. https://doi.org/10.1111/j.1559-1816.2005.tb02206.x

Steele, S. (2008, November 5). Obama's post-racial promise. *Los Angeles Times*. http://www.latimes.com/opinion/opinion-la/la-oe-steele5-2008nov05-story.html

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine*, *22*(13), 2113–2126. https://doi.org/10.1002/sim.1461

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viechtbauer, W., & Cheung, M. W. L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, *1*(2), 112–125. https://doi.org/10.1002/jrsm.11

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods*, *20*(3), 360–374. https://doi.org/10.1037/met0000023

Wicherts, J., Veldkamp, C., Augusteijn, H., Bakker, M., van Aert, R., & van Assen, M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. https://doi.org/10.1037/0033-2909.116.1.117

Wilson, T. D., & Capitman, J. A. (1982). Effects of script availability on social behavior. *Personality and Social Psychology Bulletin*, *8*(1), 11–19. https://doi.org/10.1177/014616728281002

Wyer, R. S., & Srull, T. K. (1986). Human cognition in its social context. *Psychological Review*, *93*(3), 322–359. https://doi.org/10.1037/0033-295X.93.3.322