The Psychophysiology of Political Ideology: Replications, Reanalyses, and Recommendations

Mathias Osmundsen, Aarhus University
David J. Hendry, London School of Economics and Political Science
Lasse Laustsen, Aarhus University
Kevin B. Smith, University of Nebraska
Michael Bang Petersen, Aarhus University

This article presents a large-scale, empirical evaluation of the psychophysiological correlates of political ideology and, in particular, the claim that conservatives react with higher levels of electrodermal activity to threatening stimuli than liberals. We (1) conduct two large replications of this claim, using locally representative samples of Danes and Americans; (2) reanalyze all published studies and evaluate their reliability and validity; and (3) test several features to enhance the validity of psychophysiological measures and offer a number of recommendations. Overall, we find little empirical support for the claim. This is caused by significant reliability and validity problems related to measuring threat sensitivity using electrodermal activity. When assessed reliably, electrodermal activity in the replications and published studies captures individual differences in the physiological changes associated with attention shifts, which are unrelated to ideology. In contrast to psychophysiological reactions, self-reported emotional reactions to threatening stimuli are reliably associated with ideology.

Going back to the 1950s, studies in both political science and political psychology have proposed that individual differences in political ideology do not just reflect differences in narrow political considerations but, rather, express broader sets of individual differences pertaining to personality, basic values, or broader social outlooks (Hibbing, Smith, and Alford 2014; Jost 2006). In particular, a common argument has been that a conservative political ideology is likely to be endorsed by individuals motivated to reduce threats in their daily lives. In this view, so-called threat-sensitive individuals find the order inherent in a conservative ideology attractive. A large range of studies support this basic assertion using diverse methods, including assessing differences between liberals and conservatives in self-reported need to reduce insecurity (e.g., Jost, Federico, and Napier 2009), observing the living spaces of liberals and conservatives (e.g., Carney et al. 2008), assessing personality differences between liberals and conservatives (e.g., Gerber et al. 2010), and investigating the impact of threatening events (e.g., terrorist attacks) on public endorsement of conservative policies (e.g., Merolla and Zechmeister 2009).

Recently, the literature on the broader underpinnings of ideological differences has turned toward their potential biological roots. A consistent finding across studies using methods from behavioral genetics (primarily, but not exclusively, twin studies) is that individual differences in political ideology are

Mathias Osmundsen (m.osmundsen@ps.au.dk) is an assistant professor at Aarhus University, Aarhus, DK-8000 Aarhus C., Denmark. David J. Hendry (d.hendry@lse.ac.uk) is an assistant professor at the London School of Economics, London, WC2A 2AE, United Kingdom. Lasse Laustsen (ll@ps.au.dk) is an associate professor at Aarhus University, Aarhus, DK-8000 Aarhus C, Denmark. Kevin B. Smith (ksmith1@unl.edu) is a professor at the University of Nebraska, Lincoln, NE 68588-0328, United States. Michael Bang Petersen (michael@ps.au.dk) is a professor at Aarhus University, Aarhus, DK-8000 Aarhus C, Denmark.

This research involves human participants and was conducted in compliance with relevant laws. It was approved by an appropriate institutional ethics committee and was conducted in line with the ethical standards contained in the 1964 Declaration of Helsinki and its amendments. Support for this research was provided by the Velux Foundation grant VELUX33267 to Michael Bang Petersen. An appendix with supplementary material is available at https://doi.org/10.1086/714780. Data and supporting materials necessary to reproduce the numerical results in the article are available in the *JOP* Dataverse (https://dataverse.harvard.edu/dataverse/jop).

Published online November 24, 2021.

The Journal of Politics, volume 84, number 1, January 2022. © 2021 Southern Political Science Association. All rights reserved. Published by The University of Chicago Press for the Southern Political Science Association. https://doi.org/10.1086/714780

genetically heritable (Hatemi and McDermott 2012). Furthermore, psychologists have used techniques from neuroscience (such as fMRI) to identify neural differences between liberals and conservatives that correspond to differences in threat sensitivity, especially relating to the structure and function of the amygdala, a brain region involved in the processing of fearful, threatening, and otherwise emotionally vivid stimuli (for a review, see Jost and Amodio 2012).

While psychologists have turned toward neuroscience in understanding the biological underpinnings of political ideology, political scientists have turned toward techniques from psychophysiology. Whereas measures obtained via neuroscience methods are expensive and require extensive training, psychophysiological studies are far less costly to conduct (Soroka 2019). In particular, this work has focused on the measure of skin conductance or electrodermal activity (EDA), an index of sympathetic nervous system arousal obtained by measuring microscopic changes in sweat production via electrodes on the fingertips (Figner and Murphy 2011). The seminal finding was established by Oxley and colleagues (2008), who found that conservatives responded with higher EDA than liberals when viewing images of diverse threats such as spiders, maggots, and guns. Since then, a number of follow-up articles have been published, all using psychophysiology to shed light on the psychological underpinnings of political ideology (Dodd et al. 2012; Knoll, O'Daniel, and Cusato 2015; Smith et al. 2011).

The aim of this article is to establish the first large-scale, empirical evaluation of the literature on the psychophysiological correlates of political ideology and, in particular, of the claim that relative to liberals, conservatives react with higher levels of EDA responses to threatening visual stimuli. We evaluate the evidence in favor or against this claim by, first, conducting a large replication effort, fielding laboratory experiments based on locally representative samples of Danes and Americans with a combined sample size of 348 (over seven times the number of participants in the original Oxley and colleagues 2008 study. Second, we reanalyze all published studies with the specific aim of establishing their reliability and validity. Third, we examine several coding features to enhance the validity of the utilized measures and, on this basis, offer a number of recommendations.

Now, a little more than a decade after the psychophysiological study of political ideology was initiated, we believe it is of critical importance to evaluate the introduction of psychophysiological methods into political science. First of all, claims based on these methods have been viewed as controversial within recent political science research, sparking debates on, for example, the changeability of political views (for an overview, see Hibbing 2013). Second, the evidence from the existing studies is mixed, with several studies observing supportive evidence and one study failing to replicate the finding (Knoll et al. 2015).[1] Relatedly, psychophysiological measures are shaped by "a large number of nuisance variables" (Tomarken 1995, 390), but little has been done to evaluate the consequences of the properties of psychophysiological measures in political science. This is particularly noteworthy because a lack of attention to measurement properties is also present in the psychophysiological field itself (Ogorevc et al. 2013; Tomarken 1995) and because there are many ways to collect and code psychophysiological measures (Figner and Murphy 2011). Third, because psychophysiological methods are cheap and appear easy to administer, there has recently been a proliferation of interest in using these methods outside the study of political ideology. Studies on issue attitudes (Aarøe, Petersen, and Arceneaux 2017), political communication effects (Coe et al. 2017), and party cue effects (Petersen, Giessing, and Nielsen 2015) all rely on measures of electrodermal activity. From the larger field of political behavior research, it is of key importance to pause and ask: How well do these measures capture the constructs we are interested in?

Our main contribution is to raise significant methodological concerns about the use of psychophysiological methods in political science. We are only able to replicate the original Oxley and colleagues (2008) finding in the United States, not in Denmark. Furthermore, our reanalyses show that our replications and many past studies used measures that by conventional standards are unreliable (at least as measures of the target constructs) and that the available data, to a larger extent than previously recognized, do not support the existence of an association between physiological threat sensitivity and political ideology. Still, our recommendations point a way forward for the use of psychophysiological measures in political science. We explore several possible protocols for addressing issues related to reliability and measurement validity and we identify the coding decisions that will most likely yield reliable measures. However, as we discuss the psychophysiological literature and additional analyses of the present data, these reliable measures do not capture individual differences in threat sensitivity. Instead, they are better seen as capturing individual differences in the physiological activity associated with attention shifts. This suggests that measures of electrodermal activity could be better incorporated into political science research when they are firmly anchored in theoretical frameworks

[1] In the process of revising this article, a preprint of another large-scale replication effort became available. Bakker et al. (2019) field two conceptual replications, as well a preregistered direct replication of Oxley et al. (2008). All of these efforts fail to replicate the results. We encourage readers to consult Bakker et al. (2019), which is aligned with and reinforces the conclusions of the present article.

that explicitly address how such individual differences are relevant to political attitudes and behavior.

METHODOLOGICAL CONCERNS: REPLICABILITY, RELIABILITY, AND VALIDITY

The aim of this article is to provide a methodologically thorough assessment of the evidence for an association between political conservatism and threat sensitivity measured as individual variation in EDA when processing images with negative (e.g., threatening) content. We examine this association using four central criteria for scientific contributions: (1) the replicability of the association, (2) the reliability of EDA as a psychophysiological measure, (3) the measurement validity of both EDA and measures of political conservatism, and (4) the external validity of the association.

Replicability is a hallmark of science, and the social sciences increasingly recognize the value of replication studies (Open Science Collaboration 2015). Still, few have attempted to replicate the original Oxley and colleagues (2008) finding. Teams associated with the original author set have reported two successful follow-up studies on the association (Dodd et al. 2012; Smith. et al. 2011). However, while not disclosed in Dodd and colleagues (2012), the EDA analyses in that article are based on the same data set as Oxley and colleagues (2008), with slightly different operationalizations.[2] Furthermore, an independent replication attempt failed to identify the association in a sample of American undergraduates (Knoll et al. 2015). While subsequent work has identified a number of key differences in the procedures between the original study and the independent replication (Peterson, Smith, and Hibbing 2016), the lack of replicability from independent labs raises concerns. In this regard, it is important to note that a number of studies have recently been published with relevant psychophysiological data (Aarøe et al. 2017; Coe et al. 2017; Petersen et al. 2015). While these studies focus on different research questions, they all include measures of both political ideology and EDA measures of threat sensitivity, although they do not report the associations between the two variables. Consequently, there exists a pool of additional data that can be used to examine the replicability of the association between physiological responses to threat and political ideology.

Another key criterion of research concerns the reliability of the measures used. As argued by Tomarken (1995, 389), "Psychophysiological measures are only useful to the degree that they meet the same psychometric criteria that are commonly invoked for self-report and behavioral measures." In this regard, it is relevant that EDA is influenced by a range of factors that are likely to vary randomly and arbitrarily across individuals (Figner and Murphy 2011; Tomarken 1995). These factors include outside noises, deep breaths, coughs, room temperature, bodily movements, thickness of the skin of the fingertips, preexperiment arousal (e.g., from having biked to the lab), and so forth. Yet, there has been a lack of attention to the measurement properties of EDA measures. In an assessment of the broader psychophysiological literature, Tomarken (1995, 389) concludes that "despite its evident importance, the reliability of psychophysiological measures recorded on a single occasion is rarely assessed or reported." In a similar assessment, Manuck and colleagues (1989, 368) note that "few investigators have examined the reproducibility of psychophysiological responses over multiple experimental sessions." These assessments were echoed as recently as in 2013: "Almost all papers discussing skin conductance measurements describe the measurement results in absolute terms using an appropriate measurement unit . . . , but accuracy and consequently reliability of reported measurement results is seldom questioned and investigated" (Ogorevc et al. 2013, 2994). These remarks certainly fit the studies using psychophysiology within political science. For example, published studies rarely report standard tests of reliability, such as Cronbach's alpha. Thus, we simply do not know if measures of EDA in political science research are empirically reliable.

In addition to concerns about the reliability of the measures used in past studies, it is also relevant to note concerns about their measurement validity. First, as noted above, EDA was originally validated as a measure of physiological arousal. By specifically examining EDA responses to negative visual stimuli, studies in political science have sought to obtain discrete measures of threat sensitivity. At the same time, these studies lack regular validity tests such as tests of convergent validity (i.e., do EDA responses to different threatening images converge?) and discriminant validity (i.e., do EDA responses to threatening images differ from EDA responses to, say, positive images?). These are particularly relevant questions, as some researchers have recently argued that physiological differences between liberals and conservatives relate less to threat and more to individual differences in general arousal, with conservatives being more easily aroused than liberals (Tritt, Inzlicht, and Peterson 2014). Second, there is ambiguity about the

[2] For example, Dodd et al. (2012) focus on EDA reactions to three "aversive" images ("a spider on a man's face," "an open wound with maggots in it," and "a crowd fighting with a man") and Oxley et al. (2008) focus on EDA reactions to three "threatening" images ("a very large spider on the face of a frightened person, a dazed individual with a bloody face, and an open wound with maggots in it"). The image of "a crowd fighting a man" is not mentioned in Oxley et al. (2008) and, hence, the reason it is not considered a "threatening" image is not discussed. It should also be noted that Dodd et al. (2012) include analyses of data beyond Oxley et al. (2008), including analyses of physiological reactions to political images and a separate study using eye tracking.

nature of the stimuli used to measure physiological threat sensitivity. In the original study by Oxley and colleagues (2008), the negative stimuli included both images of threats to physical safety and images that provoked disgust. In a subsequent article by Smith and colleagues (2011), the negative stimuli exclusively focused on disgust. Other physiological studies in political science have also used disgust- and threat-related stimuli to varying extents. There are strong theoretical reasons and ample empirical support for an association between self-reported measures of threat sensitivity and disgust sensitivity, on the one hand, and political ideology on the other (Terrizzi, Shook, and McDaniel 2013). But at present, no studies have directly compared and discriminated between these two forms of stimuli in the context of psychophysiology.

A question about measurement validity can also be raised regarding measures of ideology and which ideological dimensions are associated with threat sensitivity. Oxley and colleagues (2008) used items from a Wilson-Patterson political attitude scale as their ideological measure and found that higher threat sensitivity was associated specifically with preferences for "socially protective policies." Yet, ideological differences extend beyond the domain of social attitudes: attitudes in the economic domain (e.g., relating to government redistribution) are also important. At present, however, we do not know whether physiological measures of threat sensitivity are only associated with the social components of ideology, as prior ideological measures have not included economic components. This question is made salient by the fact that, in the broader literature on the psychology of ideology, there is a debate about whether the social and economic components of ideology constitute a single liberal-conservative dimension or two separate dimensions (e.g., Jost et al. 2009; Malka, Lelkes, and Soto 2019).

Finally, existing studies raise concerns about the external validity of the association between threat sensitivity as measured by EDA and ideology. Existing studies have all been conducted in the United States with nonrepresentative samples. As emphasized by Hibbing and colleagues (2014, 303), "Additional studies are needed . . . because much of the extant physiological work is based on small, geographically constrained samples and much of the psychological work relies on college undergraduates who may have yet to form stable political attitudes." Even if the original findings hold, we remain ignorant of whether the association between physiological markers of threat sensitivity and ideological orientations generalizes to populations outside the United States. Looking to the broader literature on the psychology of ideology, two contrasting expectations emerge. On the one hand, it is possible that support for socially protective policies is the universal output of psychological mechanisms for threat management across countries and cultures (see, e.g., Aarøe et al. 2017). On the other hand, some studies suggest that there is contextual variation in the link between ideology and psychological measures related to threat sensitivity such as feelings of uncertainty (Malka et al. 2019). One explanation for this is that the output of threat management mechanisms is not support for particular kinds of policies but rather support for the status quo (often referred to as "system justification"; Jost, Banaji, and Nosek 2004). In this case, physiological measures of threat sensitivity would be associated with conservatism in conservative countries and contexts, but with liberalism in liberal countries and contexts. In order to examine this, we need cross-national replications in countries that differ in the ideological profile of their political systems.

REPLICATIONS: CROSS-NATIONAL LABORATORY STUDIES

To provide an initial examination of the issues of replicability, reliability, measurement validity, and external validity, we conducted two cross-national and well-powered conceptual replication studies of Oxley and colleagues (2008).

Sampling

We executed parallel laboratory studies in Aarhus, a midsized university town in Denmark, and Lincoln, Nebraska, a midsized university town in the Midwest of the United States. The site of the US study is the same as in Oxley and colleagues (2008), and the Danish sample provides leverage in establishing external validity. Thus, Denmark provides a "liberal" political context, both in terms of economic and social issues, with a large, universalistic welfare state and liberal policies and public opinion regarding, for example, abortion and the rights of homosexuals.

In Denmark, the construction of the samples and the invitations to participate were carried out by the YouGov survey agency from September 2015 to November 2015; in the United States, recruitment was carried out by an agency at the University of Nebraska from June 2016 to October 2016. The Danish sample consisted of 172 participants, while 154 participants took part in the US study. The number of participants in each sample is three to four times higher than in the original Oxley and colleagues (2008) study.[3] Thus, our much larger sample sizes allow us to zoom in on the ideological extremes

[3] One difference between the sampling strategy of our replication studies and the original study, however, was that Oxley et al. (2008) specifically sampled individuals with "strong political convictions." To assess whether this difference in sampling strategy affects the findings from the replications, app. sec. 5C reproduces the present analysis while removing participants with weak political convictions. This does not change the conclusions reported here.

after, rather than before, collecting data. The composition of the samples was chosen to be representative of the broader populations of the two cities with respect to gender, age, and education. In the Danish sample, 52% of participants were female, the average age was 43 years (SD = 15, Min = 18, Max = 70), 10% of the participants had no high school diploma, 42% were high school graduates or similar, 6% had less than two years of college, 28% had three to four years of college, and 15% had more than four years of college. The median household income was $45,000–$54,999. In the US sample, 58% were female, the average age was 50 years (SD = 15, Min = 20, Max = 85), 9% had no high school diploma, 25% were high school graduates or similar, 14% had less than two years of college, 31% had three to four years of college, 21% had more than four years of college, and the median household income was $55,000–$64,999.

Measures

In both the Danish and American laboratory studies, we recorded participants' EDA while they viewed a series of images on a computer screen. While our images were not identical to the images in Oxley and colleagues (2008), we chose them in close consultation with authors of that study. The participants viewed 24 images in random order, where each stimulus image was shown once for eight seconds, and was preceded by an interstimulus interval (ISI), a blank screen lasting six seconds. To allow us to examine issues of measurement validity, we chose images that tapped into four distinct emotions. Six images elicited feelings of threat (e.g., a man with a knife, a man pointing a gun toward the screen), six elicited disgust (e.g., a man eating maggots, a baby with an open wound), six had positive emotional content (e.g., a waterfall, a couple kissing), and six were neutral (e.g., an umbrella, a dustpan).

To obtain our measure of EDA, we followed the "log-and-subtract" procedure from Oxley and colleagues (2008). Specifically, we first took the average logged EDA response during exposure to the stimulus image and subtracted from that the average logged EDA response drawn from the preceding ISI. This procedure allows us to isolate the EDA response to a specific image corrected for between-subject baseline variations in EDA.[4] We then combined the changes in EDA for the six images within each of the four emotion categories to produce an overall mean EDA response within that image category (e.g., we created an overall measure of EDA response to the six threatening images). Below we assess the reliability and measurement validity of these measures. We removed one outlier in the American data with EDA responses 15 standard deviations above the mean EDA response.[5]

We used four measures of political ideology to address concerns about measurement validity. Following Oxley and colleagues (2008), the first measure was a Wilson-Patterson 20-item policy issue battery ($\alpha_{DK} = .82$; $\alpha_{US} = .92$). To assess the potential distinction between social and economic components of ideology, our second and third measures were a five-item social conservatism scale ($\alpha_{DK} = .82$; $\alpha_{US} = .73$) and a six-item economic conservatism scale ($\alpha_{DK} = .82$; $\alpha_{US} = .60$) taken from Slothuus and colleagues (2010).[6] Our final measure was a single-item Ideological Self-Placement, where participants placed themselves from "Most liberal" ("Extremely left-wing" in the Danish sample) to "Most conservative" ("Extremely right-wing" in Denmark; see app. sec. 2B for detailed descriptions; appendix is available online). Higher values on all scales indicate greater conservatism.

In all models, we control for gender, age, educational level, and income (cf. Oxley et al. 2008). Gender is a binary indicator for female (female = 1, else = 0), while age (measured in years) and education level and income (measured on ordinal scales) are standardized to have a mean of 0 and standard deviation of 1.

Assessing reliability and validity of the EDA measures

We first address concerns about the reliability and validity of physiological measures of threat sensitivity, focusing on participants' EDA responses to the threatening and disgusting images. We examine reliability in two ways. First, we calculate Cronbach's α for EDA responses to the negative images (i.e., threatening and disgusting images). If our physiological measures are reliable, Cronbach's α should be high. Second, we investigate whether the study design provides a strong measurement signal. As argued above, prior studies isolate the EDA signal by taking the difference between EDA responses to threatening images ($\text{EDA}_{\text{Stimulus}}$) and EDA responses when viewing the preceding black screen ($\text{EDA}_{\text{Interstimulus}}$). As a measure of reliability, we thus

[4] We also explore a novel alternative correction strategy: using reactions to the neutral images as baseline. We report these analyses in app. sec. 5E. The findings using this alternative method are in line with the results presented in the main text.

[5] We also obtained measures of electromyography over the corrugator supercilii muscle. This measure was not utilized in Oxley et al. (2008), and we report the findings for this measure in app. sec. 5B.

[6] As these items originate in the Danish National Election survey, app. sec. 5F provides a replication using another measure of social conservatism, Right-Wing Authoritarianism, and a different economic conservatism measure, Social Dominance Orientation, both of which may fit better in a US context. Appendix sec. 5F also includes the Society Works Best scale, which constitutes another measure of political ideology.

examine the correlations between $\text{EDA}_{\text{Stimulus}}$ and $\text{EDA}_{\text{Interstimulus}}$ responses to the threatening and disgusting images, respectively. If exposure to these images produce a strong signal, $\text{EDA}_{\text{Stimulus}}$ should be much larger than $\text{EDA}_{\text{Interstimulus}}$, and thus they should not correlate highly. In contrast, the signal is weak if $\text{EDA}_{\text{Stimulus}}$ and $\text{EDA}_{\text{Interstimulus}}$ are almost identical. If a measurement tool is precise and measured without noise, a weak signal would not necessarily constitute a problem. Yet, because EDA is influenced by multiple confounding factors, a weak signal would be a cause of concern here.[7]

To identify the measurement validity of our EDA measures, we first examine convergent validity by testing whether participants have similar EDA responses to different threatening images. If EDA responses to diverse sets of negative images reflect the same latent trait, they should correlate positively. To next examine the measures' discriminant validity, we test whether EDA responses to threatening images correlate with EDA responses to other image types (e.g., positive and neutral images). If EDA responses to different image types reflect distinct latent traits, the correlations across image categories should be low.

Turning first to our reliability tests, we find that our measures are very unreliable. Cronbach's α for EDA responses to threatening images are low in both our samples (Denmark: $\alpha_{\text{Threat Images}} = .11$; United States: $\alpha_{\text{Threat Images}} = .14$); indeed, EDA responses to disgusting images were so unreliable that we could not calculate scale reliability coefficients. Further, participants' $\text{EDA}_{\text{Stimulus}}$ responses to both threatening and disgusting images correlated extremely highly with their $\text{EDA}_{\text{Interstimulus}}$ responses (r values $> .99$ in both Denmark and the United States). The high correlation between $\text{EDA}_{\text{Stimulus}}$ and $\text{EDA}_{\text{Interstimulus}}$ implies that both measures reliably capture the same common quantity. As we return to in the conclusion, this common quantity most likely reflects individual differences in baseline physiological reactivity. But the high correlation also implies that the real quantity of interest—EDA responses to threatening images—is at best only very weakly captured. The combination of a very high correlation and numerous potential confounds for EDA responses suggests that the tiny difference between $\text{EDA}_{\text{Stimulus}}$ and $\text{EDA}_{\text{Interstimulus}}$ could reflect noise rather than a signal of EDA reactivity to threat. This interpretation is bolstered by another result: difference-of-means tests reveal that participants do not have stronger EDA responses for negative images than during the interstimulus intervals.[8] If each of $\text{EDA}_{\text{Stimulus}}$ and $\text{EDA}_{\text{Interstimulus}}$ largely reflects some common quantity plus an error term, the computed difference score is essentially just random error.

Our measure of convergent validity also performs poorly. The average interitem correlations between the six threatening images and among the six disgusting images are very low in both countries, suggesting that they do not reflect the same underlying traits (Denmark: $r_{\text{Threat Images}} = .02$; $r_{\text{Disgust Images}} = .10$; United States: $r_{\text{Threat Images}} = .03$; $r_{\text{Disgust Images}} = .06$). Given the unreliable measures and low convergent validity, it is unsurprising that our final divergent validity test shows that EDA responses to threatening and disgusting images do not correlate with EDA responses to positive and neutral images. In the United States, the correlations between image categories varied between $r = -.16$ and $r = .16$ and, in Denmark, they varied between $r = -.15$ and $r = .08$. In most circumstances, this would suggest that divergent validity is high but, given the other measurement issues, it is difficult to interpret these correlations.

We return to a thorough discussion of these signs of extreme unreliability in subsequent sections.[9] For now, we set aside the measurement problems and instead test whether we are able to replicate the main finding in Oxley and colleagues (2008) that EDA responses to threatening images correlate positively with political conservatism.

Results

We present the results in figure 1, which displays estimated regression coefficients from models where we regressed our

[7] In their discussion of the use of psychophysiology as a measure of individual differences, Manuck et al. (1989, 367) write: "Because correlations of baseline measurements with both task values and arithmetic change scores are rarely, if ever, perfect, individuals' physiologic states during periods of stimulus presentation exhibit residual variability that cannot be accounted for by a knowledge of baseline values alone. It is this residual variability that might best be considered as capturing variability as indicative of the psychophysiologic 'reactivity' of individuals." However, in contrast to the present data, the example referred to in this discussion exhibited a correlation of only $r = .55$ between physiological reactions during the interstimulus interval and the stimulus interval.

[8] In both Denmark and the United States, we find no statistically significant difference between EDA responses for threatening images versus EDA responses for the preceding interstimulus interval (Denmark: $t = -1.11$, $p = .27$; United States: $t = -1.44$, $p = .15$). In Denmark, participants had stronger EDA responses during the interstimulus interval than during exposure to disgusting images ($t = -3.53$, $p < .001$); the difference was insignificant in the United States ($t = -.94$, $p = .35$).

[9] It is relevant to note that all processing of the raw data was initially conducted in one laboratory. To ensure that the reliability and measurement validity problems did not reflect miscodings, the other laboratory visually inspected and subsequently reprocessed and recoded all raw data. We did note some differences, and we report analyses of the reprocessed data in app. sec. 5D. Taken together, however, the reprocessing of the data did not substantially change the results.

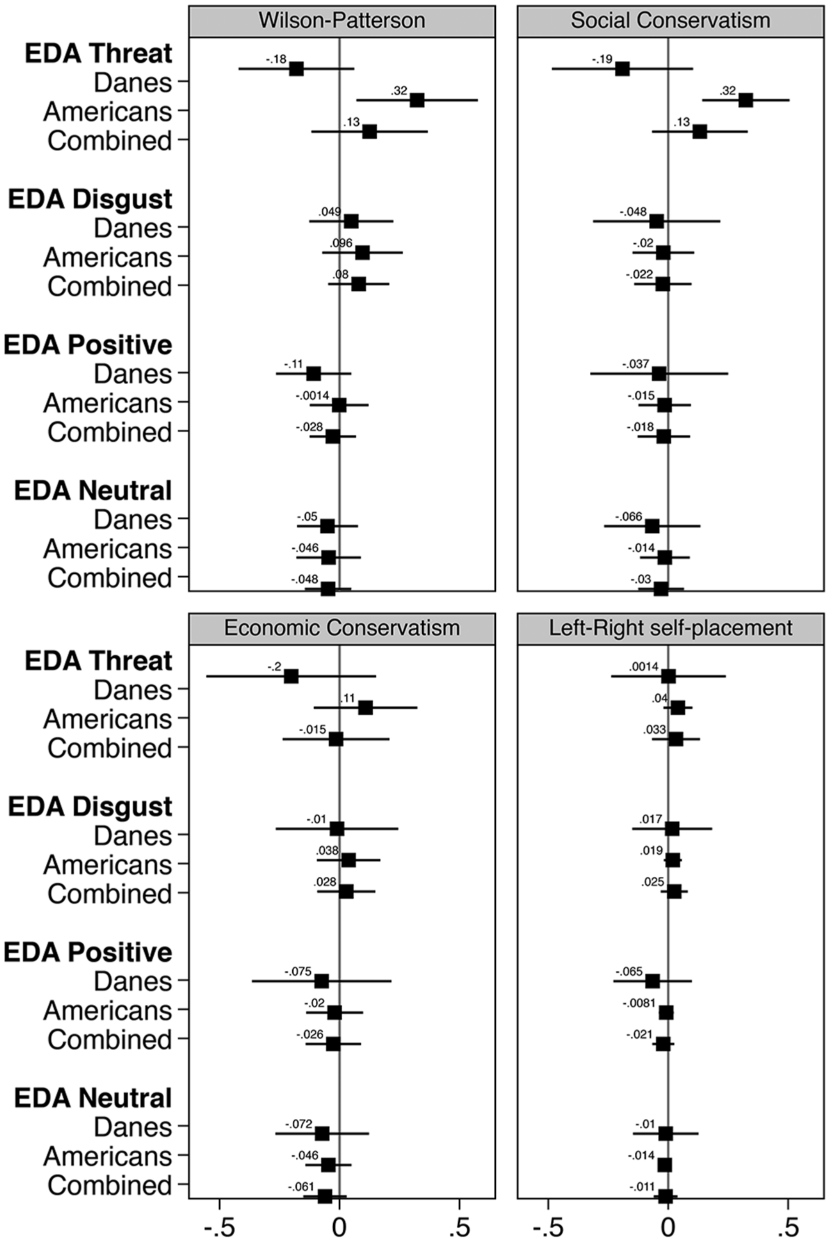

Figure 1. Coefficient estimates from ordinary least squares regression analyses of the associations among ideology measures and electrodermal activity when viewing threatening, disgusting, positive, and neutral images among Danes, Americans, and the combined samples (N = 155, 152, and 307, respectively). Models for EDA responses to each of the four image categories estimated separately; models for Denmark and the United States estimated separately, as well as for the combined sample. Binary indicator for gender. Age, education level, and household income standardized to have a mean of 0 and a standard deviation of 1. See appendix 2 for alternative specifications using interactions between country and treatment.

four measures of political ideology on EDA responses to threatening, disgusting, neutral, and positive images.[10] We estimated models for EDA responses to each of the four image categories separately, and we estimated the models for Denmark and the United States separately, as well as for the combined sample. Each model included the control variables discussed above, and in the combined sample, we additionally controlled for country. To compare the sizes of the estimated coefficients, we standardized all the measures. Horizontal bands display 95% confidence intervals.

[10] In app. sec. 5G, we estimate the relationship between EDA responses and political ideology with random effects models in which we treat participants' reactions to each image as the unit of analysis. The results do not differ appreciably from those presented here. Future studies might increase the reliability of estimates by exposing participants to many more images and then use this multilevel approach.

For the United States, the results largely support the original Oxley and colleagues (2008) study: individuals displaying higher EDA responses to threatening images were more likely to support conservative policies on the Wilson-Patterson battery ($b = .32$, $p = .013$) and the Social Conservatism scale ($b = .32$, $p = .001$). On the other hand, EDA responses to threat did not correlate significantly with the two other ideological measures in the US sample (Economic Conservatism: $b = .11$, $p = .280$; Left-Right Self-Placement: $b = .04$, $p = .181$). While this might indicate a difference in associations between threat sensitivity, on the one hand, and social and economic conservativism, on the other, further analyses show that this difference is not in itself statistically significant ($p = .118$). In the United States, reactions to disgusting images were also generally associated with conservatism, although the associations are appreciably weaker and not significant at conventional levels. We did not detect any systematic relationship between EDA responses to neutral and positive images on the one hand and political ideology on the other.

Results differed markedly among the Danish participants. In general, we did not detect a systematic relationship between EDA responses to any of the four image types and political ideology. If anything, stronger EDA responses to threatening images were associated with more *liberal* preferences, although not significantly for any of our measures. Further, examining the interaction between EDA responses and country, we found that the relationships between EDA responses to threat and the Wilson-Patterson battery and the Social Conservatism scale were statistically different from one another in the two countries ($p_{\text{Wilson-Patterson}} = .007$; $p_{\text{Social Conservatism}} = .003$). Because of these country differences in the direction of the relationship between EDA responses and political ideology, none of the combined results for Denmark and the United States were statistically significant at conventional levels.

ASSESSING THE ASSOCIATION WITH MORE RELIABLE MEASURES

Past studies have supported the existence of an association between threat sensitivity and political conservatism using other, often self-reported, measures. At the same time, there is a debate about whether this association depends on the ideological context. To assess whether the lack of an association in the Danish replication study reflects a true effect or a false negative resulting from measurement error, we examine an additional measure that we collected in the replication studies: self-reported emotional reactions to the images.

We note that it may seem curious to some readers that we validate psychophysiological results using self-reported measures, given that psychophysiological measures are explicitly employed to move beyond self-reports. We are not the first to do so: key validations of psychophysiology as a measure of arousal are themselves based on self-reports (see Lang et al. 1993).

Materials and methods

In the Danish and American laboratory samples, we asked participants to rate their self-reported emotional reactions to a subset of the 24 images previously shown on two dimensions—*valence* and *arousal*. Specifically, we presented participants with two images from each of the four categories: two threatening images (a man with a knife, a snake), two positive images (a skydiver, a romantic couple kissing), two disgusting images (a baby with a tumor, worms) and two neutral images (a plate, a mug). For each image, we measured valence by asking on nine-point scales whether participants responded with "Happy, positive feelings" or "Unhappy, negative feelings" when viewing that image. We measured arousal by asking, on nine-point scales, whether participants had "No reaction" or a "Strong reaction" when viewing the image. We then constructed indexes for each of the four image categories by combining responses to the two images from that image category; scaled so higher values indicated more negative reactions and higher arousal, respectively. As outlined in appendix section 2C, all indexes show satisfactory levels of reliability. Section 2C also shows that the self-reported ratings do not generally correlate with physiological reactions to those same images. The only exceptions are the correlations between self-reported valence ratings and physiological reactions to threatening images in both the United States and Denmark ($r_{\text{United States}} = .16$; $r_{\text{Denmark}} = .31$), and between self-reported valence rating and physiological reactions to disgusting images in the United States ($r = .20$, $p = .02$).

In the analyses that follow, we relied on the same four measures of political ideology as our dependent variables: Wilson-Patterson, Social Conservatism, Economic Conservatism, and Left-Right Self-Placement. Again, we use measures scored with means of 0 and standard deviations of 1.

Results

Here we focus on the valence ratings, which directly measure sensitivity to the threatening nature of the stimuli (see app. sec. 2E for similar analyses and results using the arousal ratings). We present the findings in figure 2. As before, the figure displays estimated regression coefficients, but this time from models where we regress our four measures of political ideology on self-reported valence reactions to the four image categories. In the figure, positive coefficients indicate that negative evaluations of the images are associated with conservatism. We estimated the models separately for Denmark and the United States, as well as for the combined sample. In all

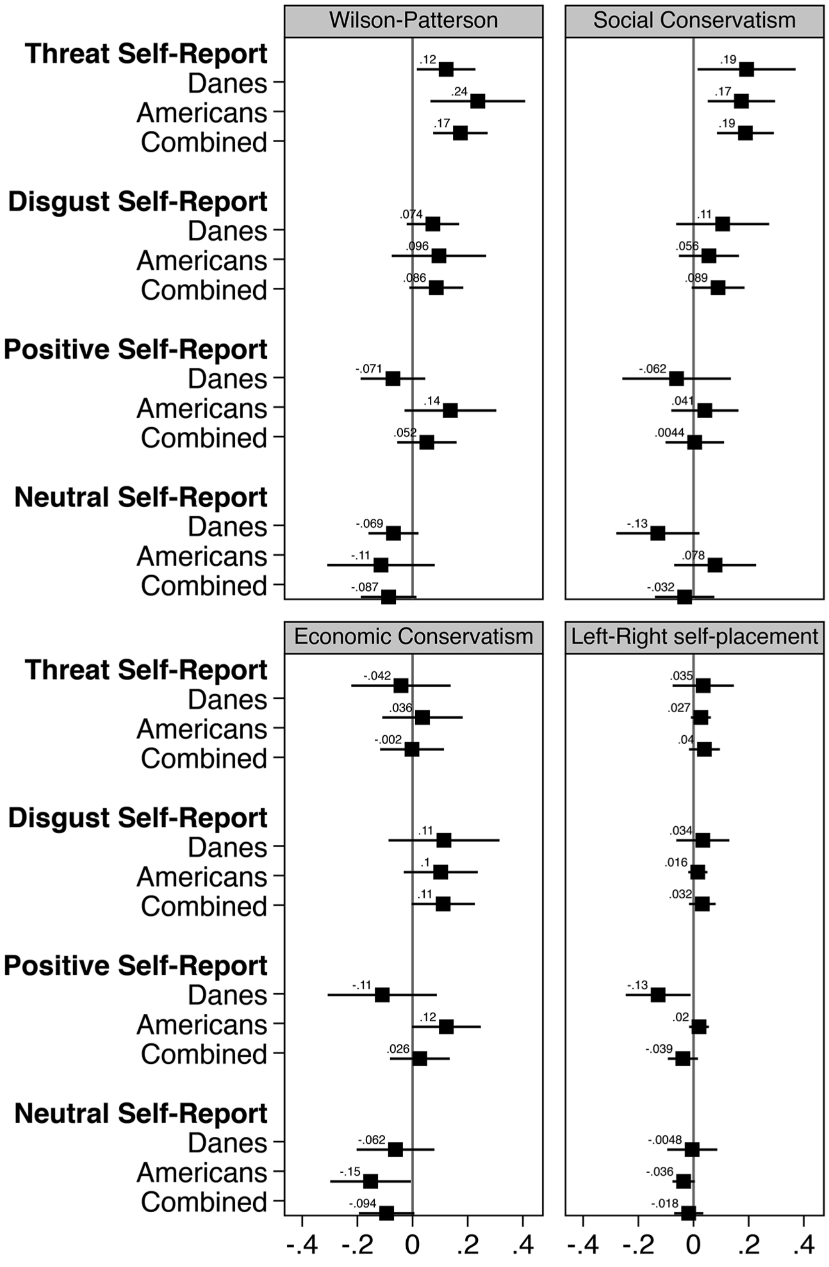

Figure 2. Coefficient estimates from ordinary least squares regression analyses of the associations among ideology measures and self-reported valence reactions when viewing threatening, disgusting, positive, and neutral images among Danes, Americans, and the combined samples (N = 155, 152, and 307, respectively). Models for self-reported responses to each of the four image categories modeled separately; models for Denmark and the United States estimated separately, as well as for the combined sample. Binary indicator for gender. Age, education level, and household income standardized to have a mean of 0 and a standard deviation of 1. See appendix 2 for alternative specifications using interactions between country and treatment.

models, we included the same set of covariates described above in the context of figure 1.

In contrast to physiological responses to threat, self-reported reactions to the threatening images were associated in both Denmark and the United States with more conservative beliefs, but only significantly so for the two measures that arguably reflect social conservatism: the Wilson-Patterson scale ($b_{\mathrm{DK}} = .12, p_{\mathrm{DK}} = .025; b_{\mathrm{US}} = .24, p_{\mathrm{US}} = .007$) and the Social Conservatism scale ($b_{\mathrm{DK}} = .19, p_{\mathrm{DK}} = .034; b_{\mathrm{US}} = .17, p_{\mathrm{US}} = .006$). In other words, participants who rated the threatening images as more negative were more socially conservative. In Denmark, the difference in strength of the association for the economic and socially conservative measures was statistically significant ($p = .006$) and, in the United States, it was marginally significant ($p = .10$). Participants' reactions to the disgusting images were also associated with more conservative preferences, but not significantly so for any of the ideology measures. We did not detect a systematic relationship between participants' evaluations of the positive and neutral images, on the one hand, and their ideological orientations, on the other hand.

In sum, these analyses support the argument that the failure to identify an association in the Danish data could result from the measurement properties of the psychophysiological measures rather than a contextual difference between Denmark and United States. This underscores the need to validate findings that employ psychophysiology using measures with more desirable measurement properties.

EXPLORING POSSIBLE METHODS FOR INCREASING THE MEASUREMENT PROPERTIES OF PHYSIOLOGICAL REACTIONS

The findings from the self-reported test suggest that physiological reactions to threatening pictures ought to be associated with political conservatism in both Denmark and United States. Yet, to identify such an association, we need to counterbalance the documented issues in the measurement of threat sensitivity. Above, we noted the general lack of associations between self-reported ratings for images and physiological reactions to the images and, although associations were significant for threatening pictures, they were at the same time relatively low. This suggests that we cannot assume that all threatening images are equally threatening in general or for each individual respondent. In this section, we explore whether it is possible to take these issues into account by combining self-reported and physiological measures.

Specifically, the method we explore relies on ratings of each picture used in our replication studies on those specific dimensions that are key for differentiating from a measurement perspective (e.g., threat and arousal). We then analyze the data at the level of reactions to a specific picture as the unit of analysis. This allows us to model interactive effects between the properties of the image (as defined by the ratings) and physiological reactions to the image on measures of political ideology. This could increase, first, measurement validity, as we no longer assume that all threatening pictures are equally threatening or that all nonthreatening pictures are equally nonthreatening. Instead, we directly obtain verifications of how threatening each picture is and therefore can test whether physiological reactions become more strongly associated with ideology with continuous increases in the degree of threat in the stimuli. Because the effect of the reactions is modeled separately for each image, this approach also does not rely on the assumption that the reactions emerge from a common latent trait. Second, this method could increase reliability as it expands the number of data points substantially. Thus, it is now possible to model the associations on the basis of reactions to many images rather than just a small handful of pictures (as is the case in both our replications and in existing studies). At the same time, it is worth mentioning that this modelling approach certainly does not solve all issues identified above. In particular, it cannot circumvent the fact that the measures of each reaction are themselves extremely noisy given the sizeable correlations between reactions to the image and the preceding ISI.

Materials and methods

We examine in two ways whether self-reported ratings can enrich psychophysiological data: first, we rely on the participants' own self-reported ratings of the eight images, which were also utilized in the previous test. In this analysis, we thus interact physiological reactivity to a picture with the participants' own ratings of the very same picture. This analysis allows us to test whether associations depend on whether participants' self-reported and physiological reactions to a specific image align. Second, we collected a dedicated rating survey where we had all images rated on relevant dimensions. In this analysis, we thus interact physiological reactivity to a picture with the average rating of the picture by external raters.

The rating survey was fielded as an approximately representative online survey with 450 participants, collected in Denmark in June 2018. In the survey, we asked all participants to evaluate all the 24 images from our laboratory replication studies. We sought to obtain a fine-grained measure of image evaluations asking the participants to evaluate the images on five dimensions. Thus, for each image, we asked participants to state whether they disagreed or agreed with five statements about their emotional reactions to the images: "I have a strong emotional reaction" (Emotion Strength), "I feel uncomfortable" (Uncomfortable), "I feel happy" (Happy), "I feel threatened" (Threatened), and "I feel disgusted" (Disgusted). We standardized all variables to have mean 0 and standard deviation 1.[11]

To test whether the link between physiological reactions and political ideology depends on the underlying characteristics of the images, we combined our laboratory data on physiological readings and political ideology with the two types of ratings of the different images from our survey sample. To this end, we estimated a series of models where we regressed our measures of political ideology on physiological reactions to the images, the ratings of the images on the separate dimensions,

[11] The survey also included two measures of political ideology: a five-item Social Conservatism scale ($\alpha = .83$) and a five-item Economic Conservatism scale ($\alpha = .80$). In app. sec. 3D, we replicate the analyses from the previous test and find that in Denmark, self-reported perceptions of threat in images are a significant positive predictor of social, but not economic, conservatism, and the difference in these associations is itself significant.

and the interactions between the two variables. To make maximal use of the available data, our key units of analysis are responses to a specific image. Because each of our participants had multiple responses, we cluster the standard errors by subject to correct for within-participant autocorrelation. In addition, we control for the same set of covariates as in the previous tests.

Results

We present the key statistical analyses in appendix section 4 and summarize the findings here. Turning first to the analyses where we focus on participants' own self-reported image ratings, we find no evidence of an interaction effect. In both Denmark and the United States, the association between EDA responses to the images and political ideology does not depend on self-reported valence or arousal ratings of the images (see tables 4A.a through 4A.d and figs. 4A.a and 4A.b in app. sec. 4). Thus, even when we take into account the fine-grained properties of the images (i.e., self-reported valence and arousal ratings), we fail to obtain a relationship between physiological reactions and ideology. When we turn to the analyses with the even more detailed survey ratings of the images—that is, where survey participants rated the images on five dimensions—we obtain essentially similar results. We find occasional hints that the association between EDA responses and political ideology is stronger among more negatively rated images in the United States (see app. sec. 4). But in Denmark, and when we examine the combined Danish and US samples, the interaction effects between physiological responses and self-reported ratings are insignificant. And when we examine the three-way EDA response × Self-reported rating × Country interactions, we find that the differences between the results from Denmark and the United States are not statistically different (all p's $> .11$).

The hope was that an integration of, on the one hand, data about (1) self-reported reactions to pictures and (2) the properties of individual pictures with, on the other hand, data about the physiological reactions to these pictures, would increase the ability to detect associations between political ideology and physiological reactions. Overall, the inconclusive nature of the findings suggest that the reliability issues identified in the replication studies are severe and cannot in any simple way be counterbalanced through an increase in the measurement validity of the utilized measures.

A META-ANALYSIS OF ALL PUBLISHED STUDIES

Based on the above conclusions, an important question is the extent to which the identified issues of reliability and validity are study specific or method specific. In other words, are these issues specific to the present study or do they also plague previous studies? To asses this, our final test is a meta-analysis of all published studies that allow us to assess an association between physiological measures of threat sensitivity and political ideology, examining not only these associations but also the properties of the utilized physiological measures.

Sampling

As discussed in appendix section 1, we identified seven existing studies of laypeople that included measures of EDA responses to negative images and political ideology: Aarøe and colleagues (2017), Coe and colleagues (2017), Dodd and colleagues (2012), Knoll and colleagues (2015), Oxley and colleagues (2008), Petersen and colleagues (2015), and Smith and colleagues (2011). Table 1 provides an overview of the studies, including their country location, sample size, types of images included (and those used to generate their measure of threat sensitivity), the specific method for analyzing the physiological data, the included ideological measures, and our tests of the reliability and validity of the measures. In assessing the totality of the evidence, it is crucial to note that Oxley and colleagues (2008) and Dodd and colleagues (2012) are based on the same underlying data but differ in terms of analytical choices (i.e., the images and political ideological measures they chose to include in the analyses, the way they calculated physiological reactions to the images). Also, as noted in Smith and colleagues (2011), nine individuals from the Oxley and colleagues study (2008) were also invited to participate in this project (equaling 18% of the total sample).

To obtain physiological measures, all studies follow a template similar to Oxley and colleagues (2008), but they differ in their specific methods for estimating changes in EDA in response to images. This will turn out to be important. Knoll and colleagues (2015), Oxley and colleagues (2008), Smith and colleagues (2011), and Coe and colleagues (2017) use the log-and-subtract method described above. Dodd and colleagues (2012) use a similar approach but index the proportion rather than the difference. To facilitate comparison, we have here recoded their data to follow the original setup of Oxley and colleagues (2008). Finally, Aarøe and colleagues (2017) and Petersen and colleagues (2015) followed a recent recommendation (Figner and Murphy 2011, 167) and calculate the area bounded by the phasic curve, measured between one second after stimulus onset to stimuli offset. Because this approach does not rely on ISIs to correct for baseline variations in EDA, the data sets from these studies did not include ISI measures.

Assessment of reliability and measurement validity

To examine the reliability of the physiological measures used in existing studies, we first examine whether participants' EDA responses to threatening and disgusting images correlated with

Table 1. Overview and Measurement Properties of Published Studies with Ideological and Skin Conductance Measures

						Correlation		
Study	Country	*N*	Image Type	Method	Ideology Measure	Image and ISI	Negative Images and Other Image Types	α and Average Interitem Correlation
Aarøe et al. (2017)	Denmark	42	6 highly disgusting*, 8 mildly disgusting, 10 positive, 4 neutral	Area under the curve	Left-Right Self-Placement, Social Conservatism	NA	$\text{Corr}_{\text{Negative} - \text{Positive}} = .86$; $\text{Corr}_{\text{Negative} - \text{Neutral}} = .75$	$\alpha_{\text{High Disgust}^{\dagger}} = .71$; $\text{Corr}_{\text{High Disgust}} = .29$
Oxley et al. (2008)	United States	46	3 threatening*, 3 positive	Log and subtract	Social Conservatism	$\text{Corr}_{\text{Threat}} = .998$	$\text{Corr}_{\text{Negative} - \text{Positive}} = -.10$	$\alpha_{\text{Threat}} = .29$; $\text{Corr}_{\text{Threat}} = .12$
Dodd et al. (2012)	United States	46	3 threatening*, 3 positive	Log and subtract[‡]	Social Conservatism, Left-Right Self-Placement	$\text{Corr}_{\text{Threat}} = .998$	$\text{Corr}_{\text{Negative} - \text{Positive}} = -.04$	$\alpha_{\text{Threat}} = .04$; $\text{Corr}_{\text{Threat}} = .01$
Smith et al. (2011)	United States	51	3 disgusting*	Log and subtract	Social Conservatism, Left-Right Self-Placement	$\text{Corr}_{\text{Threat}} = .999$	NA	$\alpha_{\text{Threat}} = .40$; $\text{Corr}_{\text{Threat}} = .18$
Petersen et al. (2015)	Denmark	58	2 negative*, 6 positive/ neutral	Area under the curve	Left-Right Self-Placement, Social Conservatism, Economic Conservatism	NA	$\text{Corr}_{\text{Negative} - \text{Nonnegative}} = .71$	$\alpha_{\text{Threat}} = .74$; $\text{Corr}_{\text{Threat}} = .59$
Coe et al. (2017)	United States	182	6 threatening*	Log and subtract	Left-Right Self-Placement	$\text{Corr}_{\text{Threat}} = .995$	NA	$\alpha_{\text{Threat}} = .57$; $\text{Corr}_{\text{Threat}} = .18$
Knoll et al. (2015)	United States	63	6 threatening*, 3 nonthreatening	Log and subtract	Social Conservatism, Economic Conservatism	$\text{Corr}_{\text{Threat}} = .995$	$\text{Corr}_{\text{Negative} - \text{Nonnegative}} = -.01$	$\alpha_{\text{Threat}} = .19$; $\text{Corr}_{\text{Threat}} = .04$

Note. ISI = interstimulus interval; NA = not applicable. See app. sec. A for details on sampling.

* Images used for threat sensitivity.

[†] Before calculating α and average interitem correlations, we standardized the items to account for differences in standard deviations.

[‡] Original study uses ratio (stimulus/interstimulus) instead of log and subtract.

their EDA responses from the preceding interstimulus intervals. Because Aarøe and colleagues (2017) and Petersen and colleagues (2015) used the area bounded by a curve and hence did not include data on ISIs, it is impossible to carry out this specific test for these two studies. The tests reported in table 1 demonstrate that participants' reactions to negative images largely resemble their baseline responsiveness. In all five studies, participants' reactions to negative images closely mirrored their reactions during the previous interstimulus interval, all r's $> .99$. Thus, the signal captured by the difference score is at best very weak. As noted previously, this is especially problematic for EDA measures for which the risk of a low signal-to-noise ratio is already high, given the many potential confounding factors. We also calculate Cronbach's α for scales consisting of changes in EDA during exposure to the various threatening images for each of the seven studies. As displayed in table 1, the α coefficients range from .04 in Dodd and colleagues (2012) to .74 in Petersen and colleagues (2015). While Dodd and colleagues (2012) and Oxley and colleagues (2008) rely on the same participants and have in common two of the three images, the α coefficient in Dodd and colleagues (2012) is much lower than in Oxley and colleagues (2008) because their last image (i.e., a man fighting a crowd) correlates negatively with the two others. Finally, an interesting observation is that the two studies using the area-bounded-by-the-curve method have markedly higher reliability (Aarøe et al. 2017; Petersen et al. 2015).

To speak to convergent validity, we examined the average interitem correlations for reactions to individual threatening images. Again, these vary considerably from one study to another, ranging from .04 in Knoll and colleagues (2015) to .59 in Petersen and colleagues (2015). Again, we observe that the two studies utilizing the area-bounded-by-the-curve method have markedly higher correlations.

To assess discriminant validity, we examine the average interitem correlations between reactions to images of different types. In Oxley and colleagues (2008) and Dodd and colleagues (2012), we compare negative images to the same three positive images. In Aarøe and colleagues (2017), we compare reactions to negative images to reactions to both positive and neutral images. Finally, in the Petersen and colleagues (2015) and Knoll and colleagues (2015) studies, we compare reactions to negative images to reactions to nonnegative images (i.e., a mix of both positive and neutral images). Coe and colleagues (2017) and Smith and colleagues (2011) only provided information on negative images, and we therefore cannot carry out this validity test for their studies. The first main finding to emerge is that in the studies that rely on the log-and-subtract method, the correlations between negative images and other image types are very low: ranging from $-.10$ in Oxley and colleagues (2015) to $-.01$ in Knoll and colleagues (2015). This suggest that people react differently to negative images than they do to other types of images. The second main finding is that EDA responses to negative images correlate highly with reactions to other image types in Aarøe and colleagues (2017) and Petersen and colleagues (2015), the two studies that rely on the area-bounded-by-the-curve approach. Thus, while these studies have satisfactory convergent validity, they do not appear to have high degrees of discriminant validity.

Overall, this assessment of the published literature suggests the reliability and validity issues identified in the replications are method specific rather than something produced in the present replications. In essence, studies using versions of the log-and-subtract method have not relied on measures with satisfactory degrees of reliability or measurement validity. This naturally limits the weight of the empirical evidence provided by these studies on the existence of an association between physiological measures of threat sensitivity and political ideology. Studies using the area-bounded-by-the-curve method fare better, but their failure to discriminate between physiological reactions to threatening and positive images suggests that these measures are essentially indicators of individual differences in arousal and physiological reactivity rather than individual differences in sensitivity to particular valences, like threat. This is exactly what EDA has been validated in the psychophysiological literature to measure reliably (Frith and Allen 1983; Lang et al. 1993). At the same time, it suggests that efforts to measure threat sensitivity specifically using EDA will be fraught with difficulty. This is a key issue and we return to it in the discussion.

Results

Setting aside these nontrivial measurement concerns, we use ordinary least squares regression to estimate the relationship between the measure of threat sensitivity and political ideology for each of the seven studies separately and for each available ideology measure. To compare the strength of association across studies, we standardized our physiological and ideological measures to have a mean of 0 and a standard deviation of 1. Finally, we note that Oxley and colleagues' (2008) preferred model includes controls for income, education, age, and gender. While not all of the data we obtained from the studies in the meta-analysis include those same variables, for each study, we include as many from the list as possible. Results are presented in figure 3.

First, we find a positive association between EDA responses to negative images and right-wing ideology in seven out of 14 tests, but only five of these are statistically significant at the .05 level. (On social conservatism, see Dodd et al. 2012 and Oxley et al. 2008; on left-right self-placement, see Dodd et al. 2012 and Smith et al. 2011 [in fig. 3].). In contrast, we find a

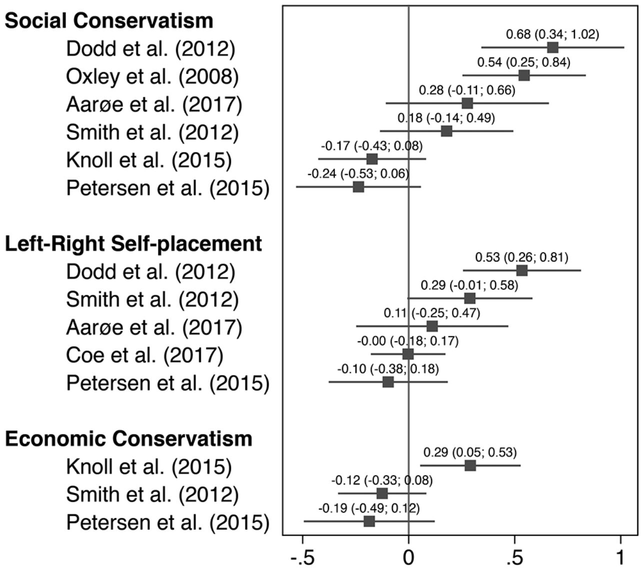

Figure 3. Coefficient estimates from ordinary least squares regression analyses of the associations among measures of political ideology and electrodermal activity responses to negative images in seven published studies. All variables scored with a mean of 0 and a standard deviation of 1. For study details, see table 1. Note that participants in Dodd et al. (2012) and Oxley et al. (2008) are identical, just as the studies rely on partially the same images and ideological measures.

negative relationship between EDA responses and conservatism in six tests, but none of these are statistically distinguishable from zero. Second, the findings are similar across our three measures of political ideology in that the different studies contain both negative and positive associations, though the relationship between EDA and social conservatism shows the most consistent pattern: Here four out of six coefficient estimates are positive, although it is important to emphasize again that the two significant estimates (Dodd et al. 2012; Oxley et al. 2008) come from studies that rely on the same samples and partially the same images. Third, the estimated coefficients vary considerably, from −.24 in Petersen and colleagues (2015) to .68 in Dodd and colleagues (2012).

It is relevant to observe that neither the unreliable measures of individual differences in threat sensitivity nor the more reliable measures of individual differences in arousal (i.e., from Petersen et al. 2015 and Aarøe et al. 2017 using the area-bounded-by-the-curve approach) show consistent associations with ideology. Hence, even without considering the reliability and validity issues, the available data do not produce consistent evidence for conservatives being either higher in threat sensitivity or higher in arousal, as measured by EDA. (Because many of the associations in fig. 1 come from the same studies and are correlated with one another, we opted to not combine the results to produce an "overall meta-analytical effect" (e.g., Harbord and Higgins 2008) across the different tests.)

To examine these claims further, we recoded the EDA measures from our own replication studies following the area-bounded-by-the-curve approach. Consistent with the findings from the meta-analysis, we also find that this method yields reliable measures ($a_{\text{Threat Images, United States}} = .90$; $a_{\text{Threat Images, Denmark}} = .84$).[12] We then reexamine the main effects of these measures on political ideology, as well as the interaction effects with the ratings of the images. We present the analyses in appendix section 5A. The analyses provide little evidence that these reliable measures are consistently related to political ideology, whether economic or social in nature, or in the United States or Denmark. Furthermore, we conducted additional analyses to assess which particular feature of the area-bounded-by-the-curve approach yielded the increase in reliability. We show the analyses in appendix section 5A.1. Briefly, the results show that it is specifically the transformation of the EDA signal from a tonic signal to a phasic signal, which involves the removal of between-participant differences in baseline EDA levels and within-participant drift in the EDA signal by imposing a so-called high pass filter (Figner and Murphy 2011).

CONCLUSION

In this article, we have undertaken the most thorough assessment to date of the association between individual differences

[12] In app. tables 5A.c through 5A.d, we show that our area-bounded-by-the-curve measures do not correlate strongly with the log-and-subtract measures presented earlier in the manuscript, nor do they correlate with self-reported reactions to the images.

in skin conductance reactions to negative images and individual differences in political ideology. The existence of such an association has emerged as a key finding in recent research on political behavior and has paved the way for the increased use of physiological measures outside the study of political ideology. Consequently, it is important to pause and take stock of current methods and findings.

Our focus has been methodological rather than theoretical. The theory that self-reported experiences of threat, uncertainty, and negativity are associated with political conservatism has been subject to repeated tests over several decades using a large number of different methods and is, in our view, well supported. Hence, our research question is whether studies using psychophysiological measures designed to capture EDA when processing threatening pictures are reliably able to reproduce this association. Specifically, we asked about the replicability, reliability, measurement validity, and external validity of this association when assessed using EDA.

By combining all available published data on psychophysiological reactions and political ideology among lay individuals with two novel, large-scale replication studies, we found limited replicability of the ideology-physiology link. The association between heightened physiological reactions to negative or threatening images and conservatism has been identified in three existing analyses of samples from Nebraska (if we include the Dodd et al. 2012 article that relies on the same participants as the original study in Oxley et al. 2008). We were able to replicate this association in our own sample collected from the same population. However, other existing studies from other US locations and Denmark could not identify the association, nor were we able to replicate it with our new Danish replication study. This suggests that the association has limited external validity. Importantly, tests using self-reported emotional reactions to images suggested that this does not reflect a lack of association between the strength of reactions to threatening pictures and political ideology outside the context of the original study by Oxley and colleagues (2008). In both the Danish and US replication samples, we found consistent and significant associations in the expected direction using these self-reported measures.

In our analyses, we have consistently compared potentially distinct reactions in the form of feelings of disgust and feelings of threat. The measurement properties of the physiological measures make it difficult to draw any strong inferences, but if we include the findings using self-reported measures, the totality of the evidence suggests that if there is an association between political ideology and reactions to images, then this association is more reliably related to feelings of threat than feelings of disgust. We have also consistently compared associations for economic and social conservatism. With the same caveats in mind, there is some evidence from the self-reported reactions for stronger associations between threat sensitivity and social conservatism compared to economic conservatism.

Overall, our analyses suggest that the fickle nature of the association between psychophysiological measures of threat sensitivity and political ideology reflects poor reliability and measurement validity. Proponents of the most widespread extraction method, log and subtract, argue it allows researchers to examine EDA responses to specific stimuli (e.g., threatening images). But data from our reanalyses of existing studies and two replication studies suggest that estimates from the log-and-subtract method are extremely noisy and fall far below conventional standards of reliability and convergent validity. An alternative method, the area-bounded-by-the-curve method based on the phasic EDA signal, yielded reliable measures. Yet, these measures do not seem to track individual differences in threat sensitivity. Instead, they seem to track individual differences in general psychophysiological reactivity. This is consistent with recent studies in psychophysiology. For example, Bulteel and colleagues (2014, 39) conclude, "There is large idiographic variation in individuals' physiological responses to emotional events, and hence, in the nature of response patterning that will be observed in different individuals." In other words, the individual differences in general physiological responsiveness are so large that it becomes extremely difficult to capture individual-level variation in responsiveness to particular types of stimuli, like threatening images. In this way, the identified issues have less to do with psychophysiological measures themselves and more to do with the use of these measures within political science. In essence, electrodermal activity has been utilized to measure something that it cannot, in fact, measure.

There is general agreement that stronger skin conductance responses to images track how arousing these images are (Lang et al. 1993). It therefore seems natural that the phasic (and reliable) individual difference measure of general physiological responsiveness would track individual differences in arousal. If valid, the lack of correlation between the phasic measure and measures of political ideology in both United States and Denmark becomes relevant for current debates in political psychology. Thus, Tritt and colleagues (2014) suggest that individual differences in arousal are related to political ideology, such that conservatives are more easily aroused. The present analyses, however, provide little supporting evidence for this claim.

At the same time, the relationship between EDA and the "somewhat vague entity arousal" is not straightforward (Frith and Allen 1983, 35). Thus, Frith and Allen (1983) conclude that EDA does not simply track emotional arousal but rather "short term changes" (38) in "the general engagement of attention during performance of any task" (35), which could occur as a result of changes in arousal. Consistent with this, the phasic measures of skin conductance reactivity are highly associated across image types including neutral (and, presumably, nonarousing) images.[13] Thus, one possible interpretation is that the phasic, reliable individual difference measure extracted in the present analyses is a measure of individual differences in psychophysiological reactivity associated with attentional shifts. In the context of the present studies, these shifts occurred when attention was turned toward an image following a blank screen. In support of this, difference-of-means t-tests reveal that phasic physiological reactions are stronger during exposure to the images than during the interstimulus intervals in both our Danish ($t = 9,35, p < .001$) and US ($t = 7,75, p < .001$) replications.

All in all, the present findings demonstrate beyond any reasonable doubt that political scientists should show extraordinary care when collecting, analyzing, and interpreting psychophysiological measures. Furthermore, the present analyses provide a clear set of recommendations for researchers committed to such an endeavor. On the basis of the present findings, we first recommend that political scientists always subject psychophysiological measures to standard tests of reliability and measurement validity and that any findings be interpreted in light of the properties of the measurement. Second, we recommend that political scientists always collect alternative measures with higher levels of measurement reliability and validity and validate the conclusions from psychophysiological measures with these more robust measures, including self-reports. Third, we recommend that political science researchers take significant steps to improve the measurement properties of physiological measures. One way forward is to mirror more closely the protocols used when electrodermal activity is deployed as a diagnosis tool in clinical studies. This includes, for example, periods of relaxation before data acquisition and a minimization of external influences such as noises (see also Figner and Murphy 2011; Tomarken 1995). But given that widespread approaches, such as the log-and-subtract method, yield extremely noisy measures, it also means that political scientists should consider alternatives. On the basis of the present findings, we echo Figner and Murphy (2011) and recommend that researchers instead use the area-under-the-curve approach, or other approaches relying on the phasic signal, for extracting the measures.

As we have suggested, the phasic measures do not seem to capture individual differences in threat sensitivity, but more general individual differences in the psychophysiological correlates of attentional shifts. Our most basic recommendation is to integrate this directly into the studies of the political correlates of psychophysiological individual differences. Rather than try to measure something that cannot be reliably measured with psychophysiology equipment, political scientists interested in psychophysiology should build hypotheses directly from theories of the type of psychophysiological activity that can, in fact, be reliably captured. Gruszczynski and colleagues (2013), for example, show that individuals with higher electrodermal reactivity tend to participate more in politics. It is indeed plausible that people who react more at the physiological level to contextual shifts are more drawn to the hustle and bustle of politics. Further studies along such lines could hold more promise for the use of psychophysiology in political science research.

ACKNOWLEDGMENTS

We thank Lene Aarøe and John Hibbing for valuable help with the design and collection of the replication studies. For valuable advice and comments in the preparation of the manuscript, we thank Kevin Arceneaux, Bert Bakker, Andrew Gelman, Frank Gonzales, Dan Mønster, John Peterson, and Gijs Schumacher. For insights on the difference between the different extraction methods, we are, in particular, indebted to Stuart Soroka. Finally, we thank all authors of the analyzed studies for generously sharing their data with us. Previous versions of this manuscript were presented at the 2016 annual meetings of the American Political Science Association and the Danish Political Science Association and at the 2017 annual meeting of the Midwest Political Science Association.

REFERENCES

Aarøe, Lene, Michael B. Petersen, and Kevin Arceneaux. 2017. "The Behavioral Immune System Shapes Political Intuitions: Why and How Individual Differences in Disgust Sensitivity Underlie Opposition to Immigration." *American Political Science Review* 111 (2): 277–94.

Bakker, Bert N., Gijs Schumacher, Claire Gothreau, and Kevin Arceneaux. 2019. "Conservatives and Liberals Have Similar Physiological Responses to Threats." *Nature Human Behavior* 4:613–21.

Bulteel, Kirsten, Eva Ceulemans, Renee J. Thompson, Christian E. Waugh, Ian H. Gotlib, Francis Tuerlinckx, and Peter Kuppens. 2014. "DeCon: A Tool to Detect Emotional Concordance in Multivariate Time Series Data of Emotional Responding." *Biological Psychology* 98:29–42.

Carney, Dana R., John T. Jost, Samuel D. Gosling, and Jeff Potter. 2008. "The Secret Lives of Liberals and Conservatives: Personality Profiles, Interaction Styles, and the Things They Leave Behind." *Political Psychology* 29 (6): 807–840.

Coe, Chelsea M., Kayla S. Canelo, Kau Vue, Matthew V. Hibbing, and Stephen P. Nicholson. 2017. "The Physiology of Framing Effects: Threat Sensitivity and the Persuasiveness of Political Arguments." *Journal of Politics* 79 (4): 1465–68.

Dodd, Michael D., Amanda Balzer, Carly M. Jacobs, Michael W. Gruszczynski, Kevin B. Smith, and John R. Hibbing. 2012. "The Political Left Rolls with the Good and the Political Right Confronts the Bad: Connecting Physiology and

[13] See also app. sec. 5A, which shows that EDA responses are not associated with particular types of ratings of pictures.

Cognition to Preferences." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1589): 640–49.

Figner, Bernd, and Ryan O. Murphy. 2011. "Using Skin Conductance in Judgment and Decision Making Research." In Michael Schulte-Mecklenbeck, Anton Kuehberger, and Joseph G. Johnson, eds., *A Handbook of Process Tracing Methods for Decision Research*. New York: Routledge, 163–84.

Frith, Christopher D., and Heidelinde A. Allen. 1983. "The Skin Conductance Orienting Response as an Index of Attention." *Biological Psychology* 17 (1): 27–39.

Gerber, Alan S., Gregory A. Huber, David Doherty, Conor M. Dowling, and Shang E. Ha. 2010. "Personality and Political Attitudes: Relationships across Issue Domains and Political Contexts." *American Political Science Review* 104 (1): 111–33.

Gruszczynski, M. W., A. Balzer, C. M. Jacobs, K. B. Smith, and J. R. Hibbing. 2013. "The Physiology of Political Participation." *Political Behavior* 35 (1): 135–52.

Harbord, R. M., and J. P. Higgins. 2008. "Meta-regression in Stata." *Stata Journal* 8 (4): 493–519.

Hatemi, Peter K., and Rose McDermott. 2012. "The Genetics of Politics: Discovery, Challenges, and Progress." *Trends in Genetics* 28 (10): 525–33.

Hibbing, John R. 2013. "Ten Misconceptions Concerning Neurobiology and Politics." *Perspectives on Politics* 11 (2): 475–89.

Hibbing, John R., Kevin. B. Smith, and John. R. Alford. 2014. "Differences in Negativity Bias Underlie Variations in Political Ideology." *Behavioral and Brain Sciences* 37 (3): 297–307.

Jost, John T. 2006. "The End of the End of Ideology." *American Psychologist* 61 (7): 651.

Jost, John T., and David M. Amodio. 2012. "Political Ideology as Motivated Social Cognition: Behavioral and Neuroscientific Evidence." *Motivation and Emotion* 36 (1): 55–64.

Jost, John T., Mahzarin R. Banaji, and Brian A. Nosek. 2004. "A Decade of System Justification Theory: Accumulated Evidence of Conscious and Unconscious Bolstering of the Status Quo." *Political Psychology* 25 (6): 881–919.

Jost, John T., Christopher M. Federico, and Jaime L. Napier. 2009. "Political Ideology: Its Structure, Functions, and Elective Affinities." *Annual Review of Psychology* 60:307–37.

Knoll, Benjamin R., Tyler J. O'Daniel, and Brian Cusato. 2015. "Physiological Responses and Political Behavior: Three Reproductions Using a Novel Dataset." *Research and Politics* 2 (4): 2053168015621328.

Lang, Peter J., Mark K. Greenwald, Margaret M. Bradley, and Alfons O. Hamm. 1993. "Looking at Pictures: Affective, Facial, Visceral, and Behavioral Reactions." *Psychophysiology* 30 (3): 261–73.

Malka, Ariel, Yphtach Lelkes, and Christopher J. Soto. 2019. "Are Cultural and Economic Conservatism Positively Correlated? A Large-Scale Cross-National Test." *British Journal of Political Science* 49 (3): 1045–69.

Manuck, Stephen, Alfred L. Kasprowicz, Scott M. Monroe, Kevin T. Larkin, and Jay R. Kaplan. 1989. "Psychophysiologic Reactivity as a Dimension of Individual Differences." In Neil Schneiderman, Stephen M. Weiss, and Peter G. Kaufmann, eds., *Handbook of Research Methods in Cardiovascular Behavioral Medicine*. Boston: Springer, 365–82.

Merolla, Jennifer L., and Elizabeth J. Zechmeister. 2009. *Democracy at Risk: How Terrorist Threats Affect the Public*. Chicago: University of Chicago Press.

Ogorevc, Jaka, Gregor Geršak, Domen Novak, and Janko Drnovšek. 2013. "Metrological Evaluation of Skin Conductance Measurements." *Measurement* 46 (9): 2993–3001.

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349 (6251): aac4716.

Oxley, Douglas R., Kevin B. Smith, John R. Alford, Matthew V. Hibbing, Jennifer L. Miller, Mario Scalora, Peter K. Hatemi, and J. R. Hibbing. 2008. "Political Attitudes Vary with Physiological Traits." *Science* 321 (5896): 1667–70.

Petersen, Michael B., Ann Giessing, and Jesper Nielsen. 2015. "Physiological Responses and Partisan Bias: Beyond Self-Reported Measures of Party Identification." *PloS One* 10 (5): e0126922.

Peterson, Jonathan C., Kevin B. Smith, and John R. Hibbing. 2016. "Physiology and Political Beliefs: A Response to Knoll, O'Daniel, and Cusato." *Research and Politics* 3 (3): 2053168016662892.

Slothuus, Rune, Rune Stubager, Kasper M. Hansen, Michael B. Petersen, and Morten Pettersson. 2010. *Måling af politiske værdier og informationsbearbejdning: Nye indeks for fordelingspolitik, værdipolitik og "Need to Evaluate" blandt danske vælgere*. Research note, Department of Political Science and Government, University of Aarhus, Denmark.

Smith, Kevin B., Douglas Oxley, Matthew V. Hibbing, John R. Alford, and John R. Hibbing. 2011. "Disgust Sensitivity and the Neurophysiology of Left-Right Political Orientations." *PloS One* 6 (10): e25552.

Soroka, Stuart. 2019. "Skin Conductance in the Study of Politics and Communication." In Gigi Foster, ed., *Biophysical Measurement in Experimental Social Science Research*. Cambridge, MA: Academic Press, 85–104.

Terrizzi, John A., Jr., Natalie J. Shook, and Michael A. McDaniel. 2013. "The Behavioral Immune System and Social Conservatism: A Meta-analysis." *Evolution and Human Behavior* 34 (2): 99–108.

Tomarken, Andrew J. 1995. "A Psychometric Perspective on Psychophysiological Measures." *Psychological Assessment* 7 (3): 387.

Tritt, Shona M., Michael Inzlicht, and Jordan B. Peterson. 2014. "Confounding Valence and Arousal: What Really Underlies Political Orientation?" *Behavioral and Brain Sciences* 37 (3): 330–31.