Achieving Statistical Significance with Control Variables and without Transparency

Gabriel S. Lenz[*]
Alexander Sahn[†]

May 27, 2020

Abstract

How often do articles depend on suppression effects for their findings? How often do they disclose this fact? By suppression effects, we mean control-variable-induced increases in estimated effect sizes. Researchers generally scrutinize suppression effects as they want reassurance that researchers have a strong explanation for them, especially when the statistical significance of the key finding depends on them. In a re-analysis of observational studies from a leading journal, we find that over 30% of articles depend on suppression effects for statistical significance. Although increases in key effect estimates from including control variables are of course potentially justifiable, none of the articles justify or disclose them. These findings may point to a hole in the review process: journals are accepting articles that depend on suppression effects without readers, reviewers, or editors being made aware.

[*]glenz@berkeley.edu. Travers Department of Political Science, University of California, Berkeley. 210 Barrows Hall, Berkeley, CA 94720-1950

[†]asahn@berkeley.edu. Travers Department of Political Science, University of California, Berkeley. 210 Barrows Hall, Berkeley, CA 94720-1950

Imagine you are reading a study that reports a statistically significant finding and includes several control variables in the regression model. You learn that this finding only emerges with the addition of these control variables and that adding them increases the effect size estimate, shifting it from nonsignificance to significance. After learning this fact, you will want reassurance that researchers have a strong explanation for the control-variable-induced increase in estimated effect size, sometimes called a suppression effect. You will especially want reassurance when the statistical significance of the key finding depends on this suppression effect. To quote an introductory statistics textbook: "In fact, it is pretty much standard operating procedure that when suppressors arise most researchers dismiss the finding as a statistical artifact unless there is a very strong theoretical explanation for the result" (Bobko, 2001, 254). Likewise, Andrew Gelman and his co-authors write: "Suffice it to say that, generally, suppression effects are considered statistical artifacts unless there is a strong theoretical explanation for their occurrence" (Crede, Gelman and Nickerson, 2016).

In this paper, we investigate how often studies depend on suppression effects. While many definitions exist, we define suppression effects as increases in the key coefficient estimate that result from adding control variables. This definition corresponds with Conger's (1974, 36-37) expansive and generally accepted definition of a suppressor variable: "a variable which increases the predictive validity of another variable (or set of variables) by its inclusion in a regression equation," where predictive validity is assessed by the magnitude of the regression coefficient (Tzelgov and Henik, 1991; MacKinnon, Krull and Lockwood, 2000). This definition includes reciprocal suppression or cooperative suppression, which is a term some researchers use for the most familiar form of suppression, where control variables increase the magnitude of the key effect but do not change the sign. It also includes the somewhat less common case where the key coefficient switches signs, sometimes called negative or net suppression.[1] Finally, it includes the rare classical suppression, where the suppressor variable is nearly uncorrelated with the dependent variable (Horst, 1941). Several sources formally define and discuss types of suppression (Nickerson and Brown, 2019; Conger, 1974; Lewis and Escobar, 1986; Tzelgov and Henik, 1991).

Although suppression effects should be treated with scrutiny, they can be justified when researchers have a strong explanation for their existence, that is, a strong explanation for why their key effect estimate is suppressed. Tesler and Sears (2010, chapter 6) provide an example in the 2008 US presidential election. To their surprise, they found that white Democrats with "gender traditional" attitudes tended to choose Hillary Clinton over Barack Obama in the Democratic primary, even though Hillary Clinton was a feminist icon. When they controlled for racial resentment, however, the relationship between gender traditionalism and vote choice switched: now the gender traditionalists were less likely to support Clinton. Why? They show that gender traditionalism and racial resentment correlated and that racial resentment much more strongly predicted vote choice. Racial resentment therefore suppressed the positive association between gender traditionalism and opposing Clinton. When controlled for, the expected relationship emerged. In this case, the authors provided a plausible explanation for suppression.

Even though suppression effects can be justified, they deserve greater scrutiny for several reasons. The first is robustness. When researchers have a strong bivariate relationship, one that holds up

[1] We acknowledge that negative suppression may not always fall under Conger's definition, such as when a coefficient estimate changes signs but is smaller in absolute magnitude. We continue to call such cases suppression.

with every control they can throw at it, they feel confident about it—it seems robust. When a finding depends on suppression effects, however, it is not robust to alternative model specifications—it is by definition acutely sensitive to the specification, as in the Tesler and Sears example. When readers are evaluating statistical findings, they generally want to know which of these two worlds they're in: generally robust to model specification or generally not. If the latter, they require much greater confidence in the model specification that generates the findings.

A second reason for skepticism of suppression effects is that they can introduce bias that favors the authors in a nontransparent way. Consider an example of a data generating process with one key variable and 10 control variables, the first five of which increase the key effect estimate, and the remaining five decreases the key effect estimate. As a researcher adds these controls sequentially to a model, the first five will increasingly bias their key effect estimate upwards, as these are omitted suppressors. They induce bias because they unmask bias from the still omitted five control variables. Only when the researcher adds all 10 will she estimate an unbiased effect. In this example, a researcher could include the first five control variables, but leave out some of the second five, biasing the effect estimate upwards. Since readers never know the "true" data generating process, they may not realize that control variables have been excluded. Researchers have strong incentives to publish statistically significant findings and, given the numerous forking paths they face when making research decisions about control variables, they may have many opportunities to intentionally or unintentionally take advantage of suppression effects. While all control variables deserve scrutiny, suppression effects can induce bias that favors the authors, and do so in a way that may be hidden from readers, and so arguably deserve greater scrutiny. In contrast, when control variables reduce estimated effect sizes, they work against authors' publishing incentives and so may require less scrutiny.[2]

A third reason for skepticism is that control variables can introduce bias through opaque mechanisms. The hypothetical example above illustrates the potential for bias that favors authors from confounding. But control variables can also introduce bias that favors authors through amplification bias or through mediation, to name just two (Middleton et al., 2016; Pearl, 2010; MacKinnon, Krull and Lockwood, 2000). Given the complexity of some of these effects, readers need to know if findings could potentially depend on them.

As a reader, you therefore want to be alerted to suppression effects, especially when the statistical significance of the key finding depends on them. As a field, we want to know how often leading journals are publishing such articles and especially how often they are doing so without alerting readers, reviewers, and editors.

So, how often do articles depend on undisclosed suppression effects for statistical significance? In this paper, we investigate this question by analyzing replication data from a leading journal. We examine whether findings depend on suppression effects by asking whether the main estimate presented in the articles, which includes control variables, is larger an absolute value than a bivariate estimate, which excludes those controls. Specifically, we look for how often statistical

[2]We quote additional authors on why researchers should scrutinize suppression effects in the supporting information. The idea that controls can introduce biases is unintuitive to some readers. This hypothetical example illustrates that they can when some but not all controls are included. Numerous articles have discussed bias introduced by controls (Bhattacharya and Vogt, 2012; Clarke, 2005; Cole et al., 2010; Middleton et al., 2016; Pearl, 2010, 2011, 2009; Wooldridge, 2016; Wyss et al., 2014).

significance of the articles key result depends on the increase in estimated effect size induced by control variables, that is, suppression effects. We find a startling result: over 30% of observational studies depend on suppression effects for the statistical significance of their findings. Moreover, we find that none of these studies disclose this fact.

These findings may point to a gap in the review process: journals publish articles that depend on suppression effects for statistical significance without the awareness of readers, reviewers, or editors. They also reinforce concerns about the potential for researcher discretion with control variables, concerns that have existed for decades (Leamer, 1983). Replication efforts and meta-analyses suggest that a reasonably large fraction of studies are false positives or report much larger effects than actually exist (Ioannidis, 2005; Ioannidis, Stanley and Doucouliagos, 2017; Klein et al., 2014, 2018). They have also pointed to suspicious patterns in test statistics (Gerber et al., 2010; Gerber and Malhotra, 2008; Brodeur et al., 2016). Undisclosed suppression effects may be one source of these patterns. Many solutions exist for this problem, some of which we discuss in our conclusion. The simplest solution, however, is for reviewers and editors to ask for a tad more transparency.

1 Data

To conduct this analysis, we replicate and reanalyze studies published in the *American Journal of Political Science* (AJPS). AJPS was one of the earliest social science journals to adopt a firm data transparency policy, enforcing the posting of replication data and code for articles published beginning in 2013. Following our pre-analysis plan,[3] we analyze articles from *AJPS* 2013-2015 that focus mainly on establishing a single causal claim and have a standard statistical model with at least one control variable (see Table S1 in the supporting information for exclusion reasons). 64 of 163 articles in these years met these criteria. 49 of these are observational and 15 are experimental. The topics of these studies range widely, from the effects of judges having daughters on their rulings to whether United Nations peacekeepers succeed in protecting civilians. On average, these studies include 9 control variables in the fully specified model.

We successfully reproduce the main findings in each of these articles, including the coefficient estimates and the standard errors, as Figure 1 shows. Given the difficulties researchers have faced with reproducibility (King, 1995), this result is reassuring. In a handful of cases, the figure reveals that we estimate different standard errors.

2 Frequency and Disclosure of Suppression Effects in Observational Studies

We first examine how often researchers use control variables to achieve statistical significance through suppression effects in observational studies (we examine experimental studies later).

[3]See this link: https://bit.ly/2luoV3E. The plan specifies how we collected the data. We did not pre-specify the key analyses in this paper (Figure 3). We present the analyses we did specify in the supporting information and the results are consistent with our main findings.

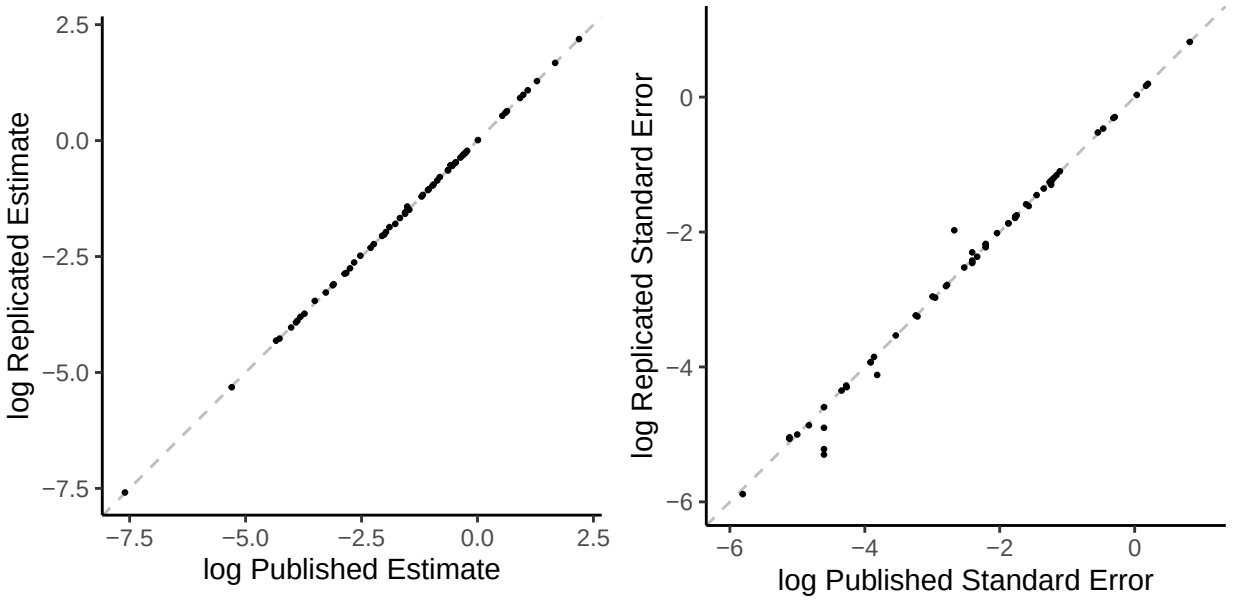

Figure 1: Replication of Key Estimates and Standard Errors for Full Specification Models

For each article, we examine the effects of control variables on the estimated effect of the key explanatory variable, the variable about which the article attempts to establish a causal claim. To do so, we compare the p-value of the key explanatory variable estimate in the full model (all control variables) to a bivariate model. In cases where the research design or identification strategy requires covariates (such as the main effects of an interaction term, lagged dependent variables, and fixed effects), we include these and count this as a bivariate specification. About half of studies include some covariates in the "bivariate specification," but our main finding holds up when we look only at studies that did not require them (see Figure 6 below and Figures S1, S2, and S4 in the supporting information).

Given the nature of our analysis—examining the effects of whatever control variables authors used—our definition of suppression effects based on Cogner is expansive. Our definition includes all types of suppression effects generated through any mechanism. Conger's definition refers to a single suppressor but we are examining the effects of one or more control variables—whatever the original authors used—and therefore apply the concept to the general linear model originally suggested by Holling (1983). Although our analysis does not examine the types or mechanisms giving rise to suppression, researchers' analysis should depend on the nature of the suppression effect, a topic we return to later.

Since we are especially interested in how often findings depend on undisclosed suppression effects, we examine the effect of controls separately among articles that disclose or do not disclose a bivariate estimate for their key effect. When researchers reveal a bivariate estimate, readers can determine whether the key result depends on suppression effects, justified or otherwise. To classify whether researchers disclosed, we count studies as revealing the potential presence of suppression effects if they report a bivariate estimate for their key explanatory variable. We count this if they report it in a statistical model or in some other form, such as a scatterplot, cross-tab, or difference in means. We count these regardless of whether they occur in figures, tables, the text, or in notes.

Figures 2 and 3 present the main finding of this paper.[4] Figure 2 begins by showing the effect of control variables on the p-values of the key effect estimates and how often they disclose this fact. The figure presents arrows that start at the p-value of the bivariate specification and end at the p-value of the full specification. When control variables lower p-values of the key estimate, the arrows end pointing downwards. When control variables increase p-values, the arrows end pointing upwards. The figure shows a large number of downward arrows. When researchers include control variables, their p-values drop a lot, many from well above 0.05 to below 0.05. In four cases, the introduction of controls switches the estimated effect direction. To convey this switch, the figure shows the arrows increasing from the bivariate specification p-value, to p=1, and then back down to the p-value for the full specification.

These drops in p-values only represent suppression effects if they result from increases in the absolute value of coefficients (as the decreases could also result from smaller standard errors). Figure 3 adds the additional information of how much of these p-value drops result from coefficient changes. It presents the key result of this paper. To show how much of these p-value drops results from coefficient changes, Figure 3 alters the arrows so that the solid portion now shows the changes in the p-value attributable only to coefficient changes, that is, the change attributable to suppression effects. The dotted component of the arrows shows the p-value changes attributable to only standard error changes. As the solid portion of the arrows makes clear, suppression effects drive much of the p-value decreases, though both components contribute. Based on a careful examination of Figure 3, suppression effects helped the authors lower their p-values towards 0.05 in about 20 of the 49 observational articles. Not all of these studies achieved statistical significance at p<0.05 in the full specification, but the decreases in p-values may have been necessary for publication.

Figures 2 and 3 split the results into the left panel showing studies that did not present a bivariate specification and the right panel showing studies that did. Only a single one of the 20 studies revealed the presence of suppression effects.

In sum, undisclosed suppression effects are common. We cannot know how many of the 19 studies that did not disclose suppression would have been published without the contribution from suppression effects, but it seems possible that many or even all would not have. So, undisclosed suppression effects may have contributed to the publication of as much as ($19/49 \approx$) 40% of observational studies. A conservative estimate would be ($15/49 \approx$) 30%. Another way of describing this finding is that, when reading an observational article that does not show a bivariate estimate in some form, readers should assume as high as a ($19/34 \approx$) 55% chance that it depends on suppression effects for significance, a high probability.

The effect of suppression on p-values is considerable, especially in studies that do not reveal the presence of suppression effects. In studies that are transparent by showing a bivariate, p-values actually slightly increase (by 0.001). In studies that are not transparent, p-values decrease by 0.33 on average. The difference in p-value changes between those that do and do not is highly statistically significant (p=0.0009, Wilcoxon rank-sum test). If we limit the p-value changes to

[4] Given the complexity of the findings we present, we show the same results with several alternative figure designs in the supporting information. See Figures S6 and S7a-S7d (we thank an anonymous reviewer for the suggestions).

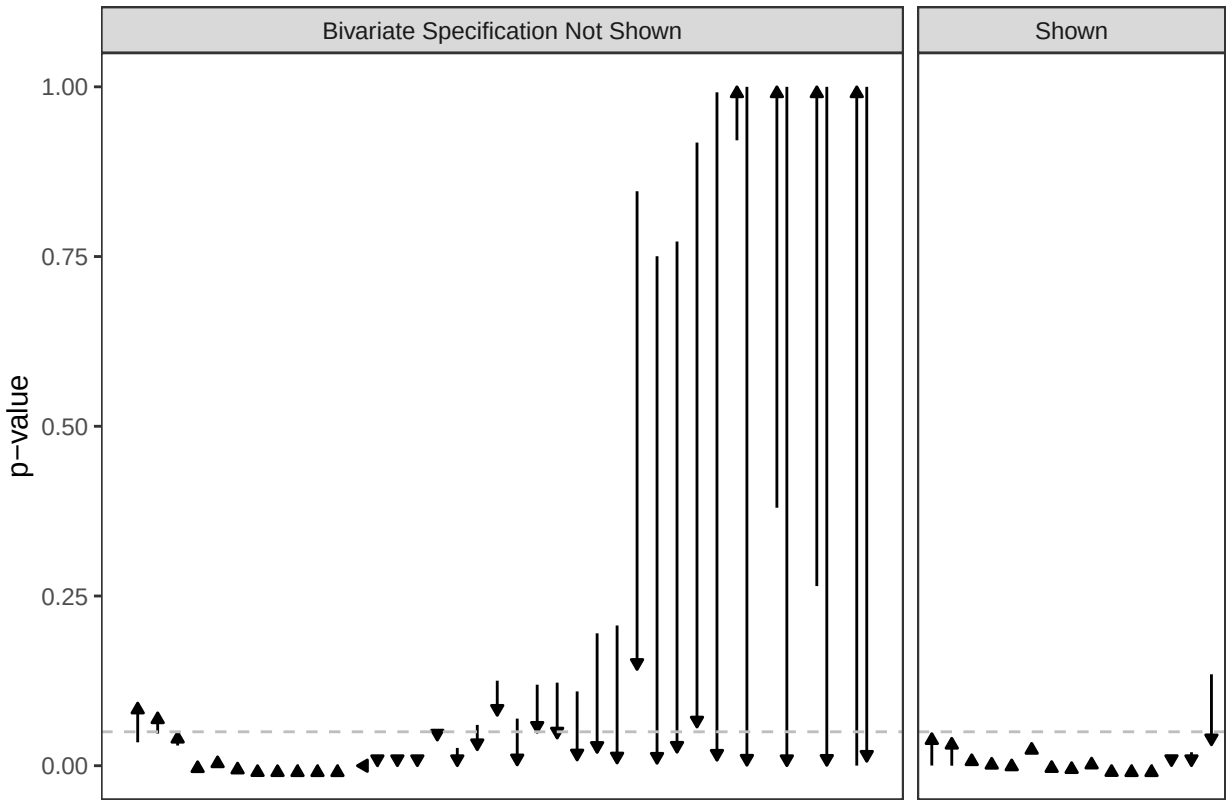

Figure 2: P-Value Changes in Observational Studies. Arrows for each study show the total p-value changes from the bivariate to the full specification (all controls). The figure shows that, when articles did not present a bivariate specification, control variables lowered p-values for the articles' key effect estimate. The four up and down arrows reflect cases where the estimate changed sign from the bivariate to the full specification. The next figure shows how much of the p-value drops result from changes in coefficients from suppression effects versus changes in standard errors.

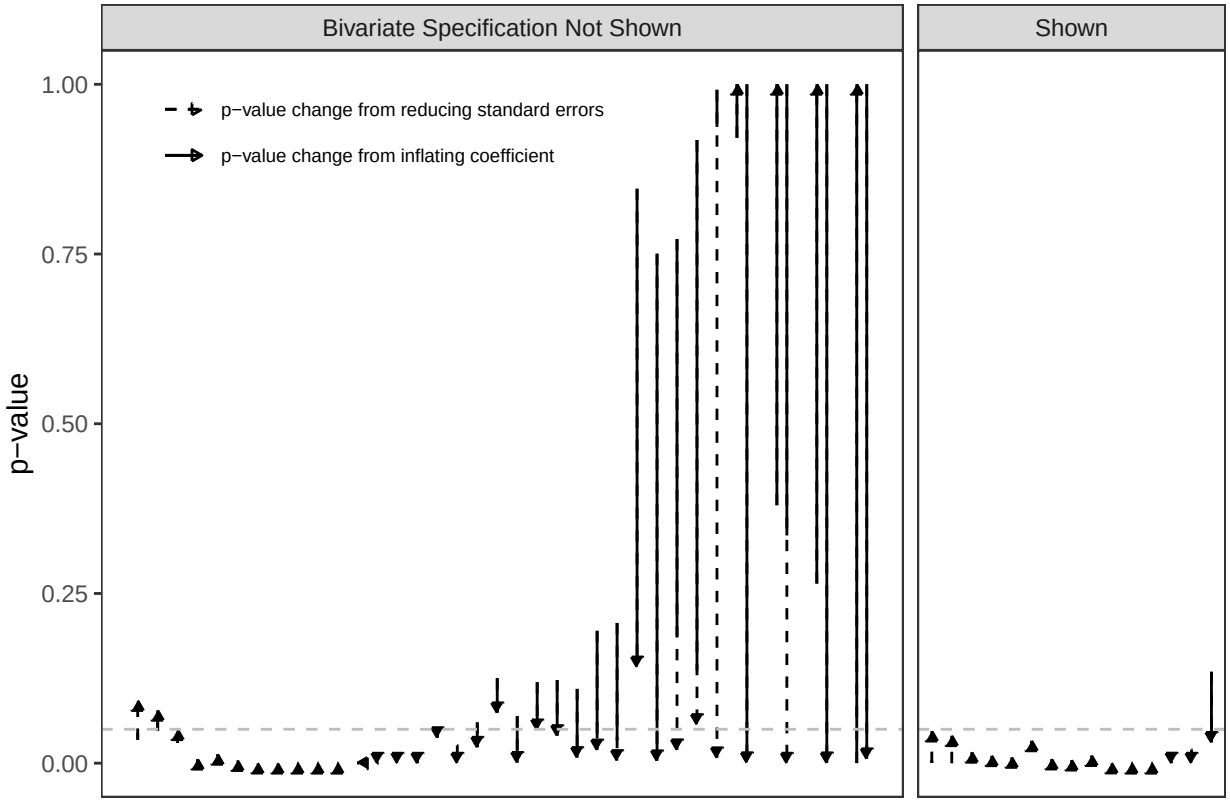

Figure 3: P-Value Changes in Observational Studies from Coefficient and Standard Error Change. Arrows for each study show the total p-value changes from the bivariate to the full specification (all controls). The solid part of the arrows shows the p-value changes from only coefficient estimate changes, while the dotted part shows the remaining p-value changes from standard error changes. The figure shows that, when articles failed to present a bivariate specification, they often depend on suppression effects to achieve statistical significance. The four up and down arrows reflect cases where the estimate changed sign from the bivariate to the full specification. Figure 6 and Figures S1 and S2 in the supporting information present robustness checks.

only that due to coefficient changes, non-transparent studies decrease them by 0.29 on average and the difference in p-value changes remains highly statistically significant (p=0.0027, Wilcoxon rank-sum test).

To show how control variables give rise to the p-value changes in Figures 2 and 3, Figure 4 presents the log change in coefficients and standard errors of key variables from the bivariate to the multivariate. It uses log changes because they approximate percent changes (while avoiding the symmetry and additivity problems of percent changes).[5] Each dot shows either the log change in coefficients or in standard errors, and we sort dots by the p-value changes, as in Figures 2 and 3. Figure 4 makes clear that suppression effects occur in several cases where they have no impact on the statistical significance of studies. In such cases, suppression effects, even if not justified, are less problematic since statistical significance does not depend on them. The figure also makes clear that even small changes in effect sizes can have large effects on statistical significance. Finally, this figure illustrates the degree to which increased precision, a frequent justification of the inclusion of control variables, isn't terribly common.

How many of the undisclosed cases of suppression represent false positives where? That is, how many of these are cases where researchers are relying on suppression effects, intentionally or not, to publish nonexistent effects? This is a hard question. We spent considerable time reading and reanalyzing replication data for these articles, trying to determine whether suppression effects could be justified, that is, whether researchers could have provided a good explanation for their key effect estimates being suppressed.

In some cases, we think they do. One example is Davenport (2015), who examines the effect of casualties and low draft numbers on parents' turnout during the Vietnam War. She finds that parents whose children are at high risk of being drafted (low lottery numbers) and who live in towns that have casualties, are more likely to turn out. One can tell a simple story about suppressor variables: poor regions of the country have lower turnout and disproportionately contribute soldiers to combat roles in Vietnam, so are hit disproportionately by casualties. Socio-economic status may therefore suppress Davenport's key interaction estimate. Indeed, in reanalyzing Davenport's data, we find that prior turnout (which likely captures individual socio-economic status) and town measures of socio-economic status increase the size of her key interaction estimate, lowering its p-value from 0.11 in the bivariate specification to 0.01 in the full specification. So, Davenport can provide a good explanation for why her finding depends on suppression effects.

For many articles, however, we cannot find an obvious justification for suppression effects, though it is possible that authors could provide one. In several of these articles, unusual control variables were the key suppressors. These include variables that arguably should not be in the models, such as post-treatment variables, and variables that change the interpretation of the finding in a way that seems unintended by the authors, such as including a lagged dependent variable when the article is not about explaining change. These potentially questionable control variables would have been harder to miss in the review process if authors had alerted readers to the presence of

[5] To calculate the log change, we code the multivariate and bivariate estimates to positive. In the four cases where the signs flip between the bivariate and the multivariate, we rescale so that the bivariate estimate is zero and we add the bivariate to the multivariate. We then add one to all the estimates before taking the log. We also add one before calculating the log change in standard errors to keep the scale for coefficients and standard errors similar. Adding constants other than one doesn't substantively change the findings.

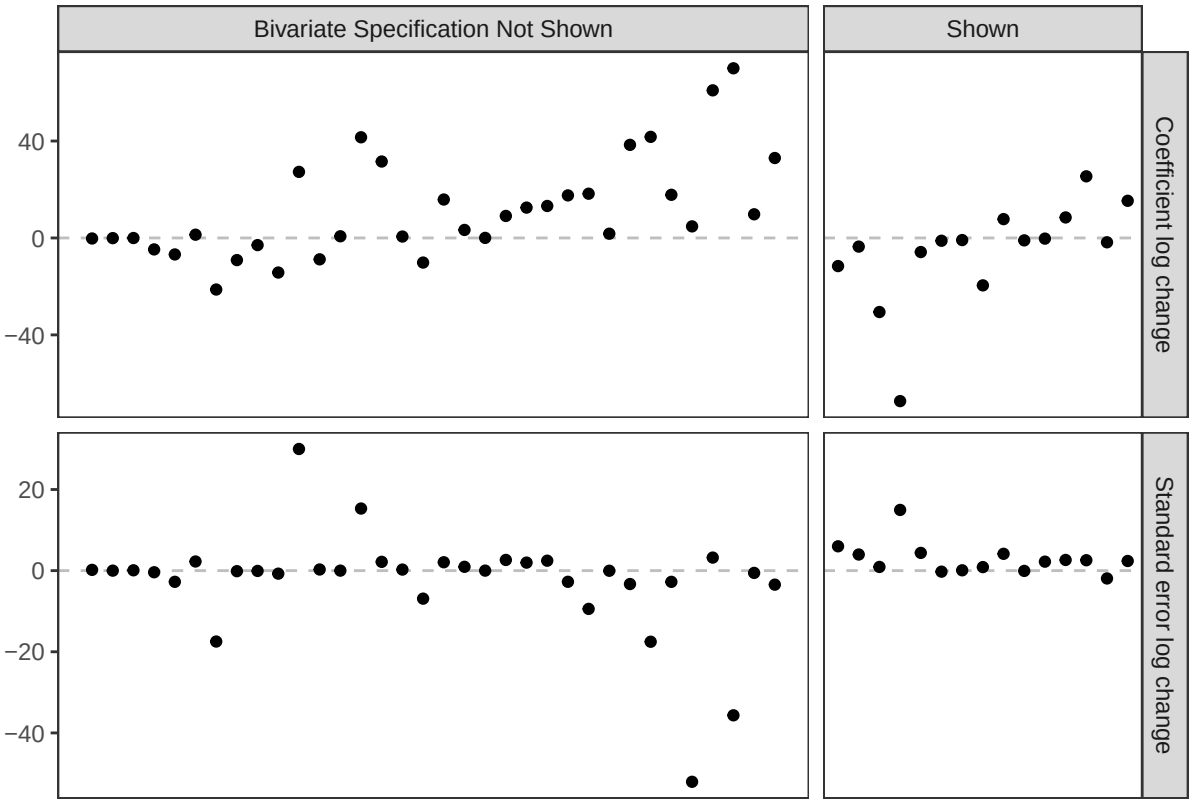

Figure 4: Key Coefficient and Standard Error Change in Observational Studies. Each dot shows either the log change in coefficients (top) and standard errors (bottom). The dots are sorted by the p-value changes, the exact same order as in Figures 2 and 3.

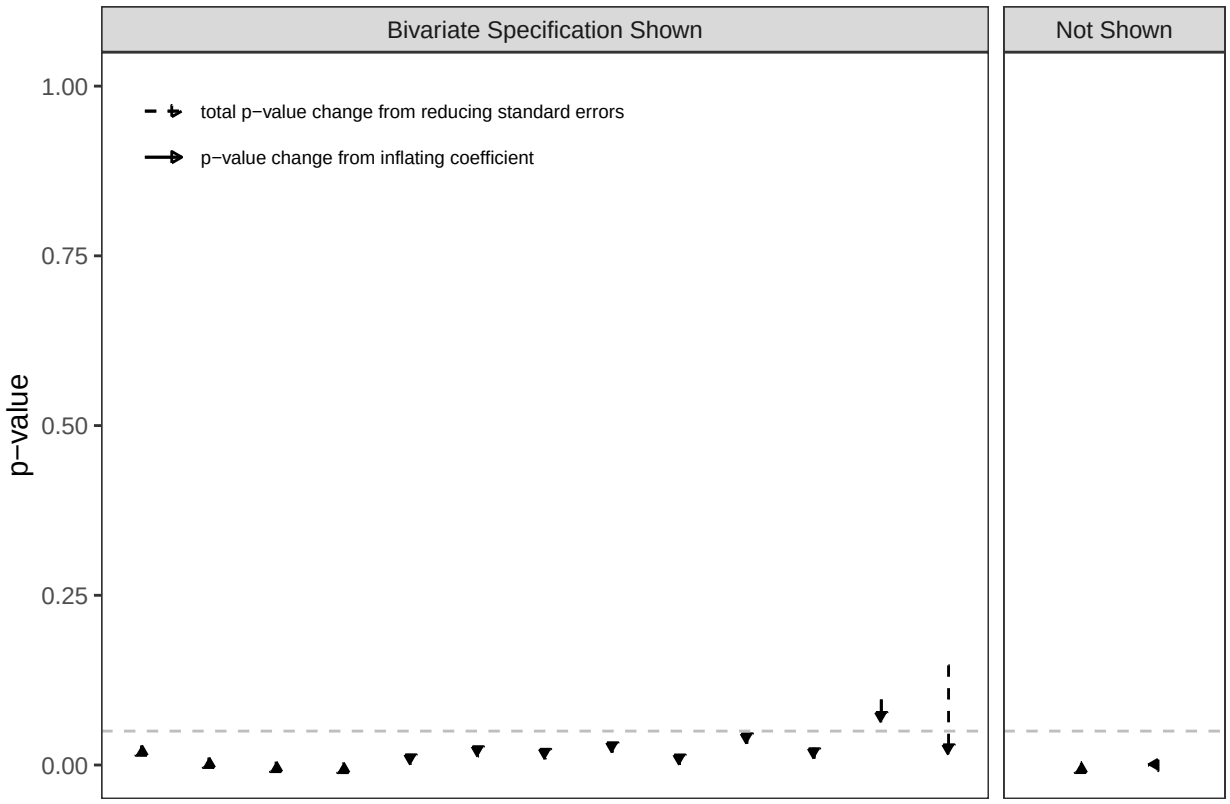

Figure 5: P-Value Changes in Experimental Studies. For each article, the arrows show the total p-value changes from the bivariate to the full specification. The solid part of the arrows shows the p-value changes from only coefficient estimate changes, while the dotted part shows p-value changes from standard error changes.

suppression effects.

3 Frequency and Disclosure of Suppression Effects in Experimental Studies

Suppression effects should be less common in experimental studies that randomly assign treatment in large samples. In observational studies, controls can change the key variable's estimate because they can correlate with the dependent variable and key independent variable. When researchers randomly assign the treatment in large samples, by contrast, the correlation with controls is minimal, especially as the sample size increases. Therefore, we would expect minimal sensitivity to control variable choice. All the experiments we examine here used random assignment and have relatively large Ns; the median N is 868 and smallest is 156. Consistent with this expectation, Figure 5 reveals little sign of p-value decreases in experimental studies with controls. Randomized experiments have many advantages—one is less vulnerability to discretion in control variable choice.

4 Robustness and Potential Objections

The main finding of this paper, shown in Figures 2 and 3, is robust: when researchers don't disclose a bivariate, adding control variables drops p-values consistently across various subsets of the data, as shown in Figure 6. P-values drop when researchers use fixed effects or don't use fixed effects and when we exclude the cases where the sign of the effects flip. They drop in each of the three volumes of the *AJPS*, in each of the three empirical subfields (American, comparative, and international relations), and when the bivariate/minimal specification has no covariates versus when we must include covariates for the minimal specification to make sense. Finally, they also drop in studies with less than the median number of controls, which is eight, and in studies with more than the median number of controls (see also Figures S1 and S2 in the supporting information).

One potential alternative explanation for our key finding is that researchers may not show the bivariate specification when controls favor them because they are working on topics where suppressor variables are well known. To address this possibility and to examine whether researchers justified suppression effects, we carefully read the sections on controls in these articles but found no acknowledgment or justification. We present each articles' text relating to controls in the supporting information (for articles that did not present a bivariate specification).

Another account for our key finding comes from the incentives of the journal review process. Researchers may believe that, if they disclose suppression effects, reviewers will get hung up on them. The pattern we observe—only one published study relies on suppression effects for significance and discloses this fact—is consistent with this concern. If this is authors' motivation, however, it points to a hole in the review process, as reviewers should be assessing the plausibility of control-induced increases in effect sizes.

Some scholars have argued that showing multiple specifications in regression tables, as researchers often do, is enough, especially when estimates generally seem stable across the multiple specifications. We believe, however, that researchers may underappreciate the degree to which estimates vary across combinations of controls not shown. Researchers often present a table of several regression specifications, sequentially adding controls, but these represent a handful of thousands of possible specifications. We show the range of effects researchers could have produced in Figure 7, which presents the distribution of t-statistics across all possible control combinations (randomly sampled in cases with many control) for the observational studies. We show t-statistics to scale the coefficients across these studies. As before, the figure breaks the studies into those that show a bivariate specification and those that do not. It also shows the t-statistic for the bivariate specification (b) and the full multivariate specification (m). The figure shows a large range of estimates in many of the studies across control combinations, with estimates often crossing the threshold for statistical significance and sometimes producing effects of opposite signs. This figure understates the potential for discretion since it only reflects the controls the authors chose to condition on. And, of course, control variable choice is just one of several sources of modeling discretion.

Another objection we have encountered is that readers do not need bivariate specifications to assess the effects of control variable discretion because they can merely look at the control variables

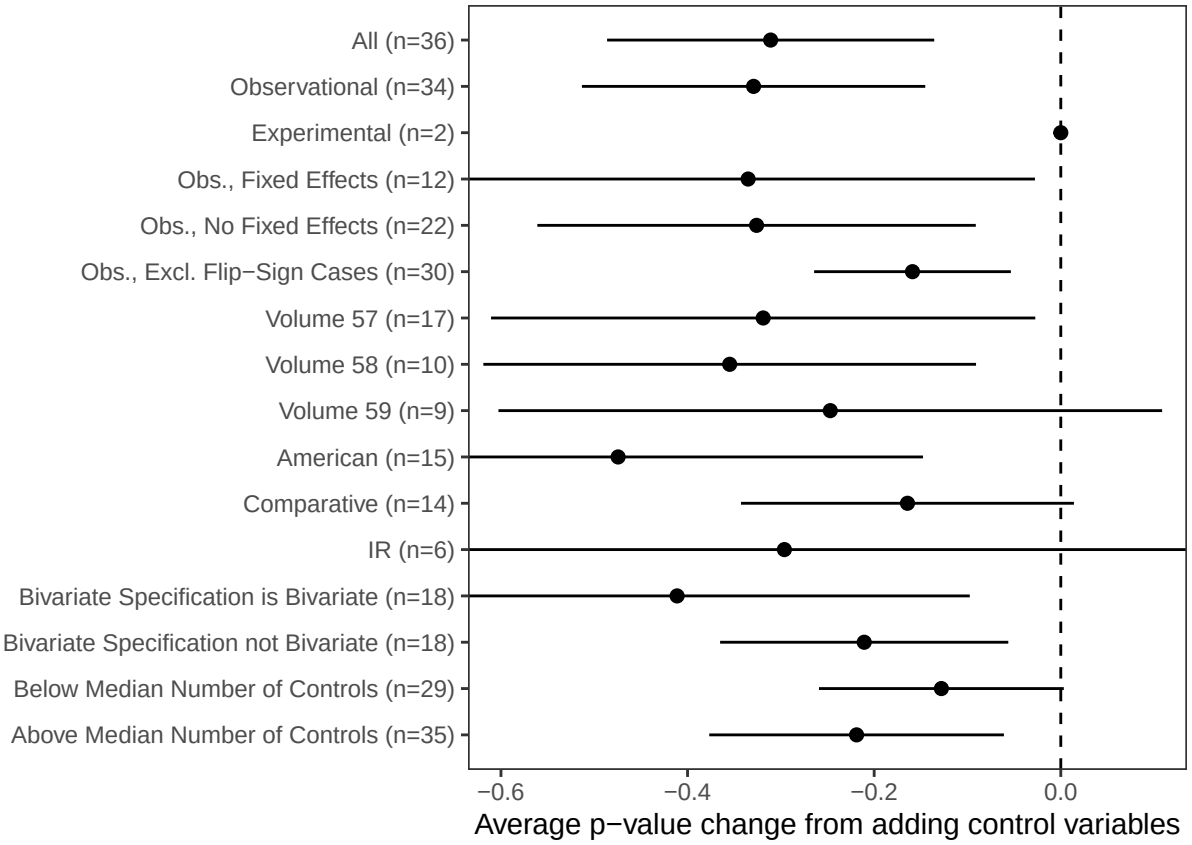

Figure 6: P-value Changes in Studies that Do Not Show Bivariate. This plot shows the average change in p-values from the bivariate to the full specification (with 95% confidence intervals) for observational and experimental studies that did not show a bivariate specification. The mean number of controls in these studies is nine and the median is eight. Figures S1 and S2 in the supporting information present similar plots for all studies and for p-value changes only from coefficient changes.

themselves. They can assess, for instance, whether researchers have left out controls that would work against their key finding, suppressing its effect. Undoubtedly, readers can do this in some cases. In our reanalysis of these articles, however, some of the controls responsible for increasing estimated effect sizes are unexpected, and we doubt experts would have anticipated them.

Another objection is that researchers may be unable to justify the effects of controls on their estimates because they are too complicated to explain in a multidimensional context. Although this point has merit, we think that researchers will often have priors about the reasonableness of the net effect of controls. We also find that control variables largely have similar effects on their own when conditioning on other control variables (see Figure S3 in the supporting information). More importantly, if researchers cannot be required to provide justifications, then discretion will be left partly unchecked.

Finally, it's worth emphasizing that the p-value drops we observe from adding controls do not generally result from large changes in the key coefficient estimates. Although scaling changes across studies is difficult, the key coefficients increase on average by only about 20%, calculated with log $+1$ changes, but those increases translate into large p-value decreases (see Figures 4 and S4-S5 in the supporting information). Researchers, therefore, are benefiting from suppression, but relatively weak suppression, enough to shift their p-values below 0.05. Consistent with this pattern, multicollinearity does not seem especially high in these studies, nor does it vary with p-value changes. In the supplemental information, we present plots of the average correlation between the key variable and control variables for all studies (see supporting information Figure S8).

5 Discussion and Conclusion

Based on a reanalysis of political science studies, we find that statistical significance depends on suppression effects in a large fraction of observational studies. Almost none of these studies are transparent about this fact. By increasing the estimates of key effect sizes, undisclosed suppression effects may have contributed to the publication of over 30% of observational studies. When reading an observational article that does not show a bivariate estimate in some form, readers should assume nearly a 60% chance that the article's key effect depends on suppression for statistical significance. Suppression effects in these articles may be well justified, but authors do not acknowledge, let alone justify them. Without disclosure, readers, reviewers, and editors cannot subject them to scrutiny, asking whether the authors have a good justification for suppression.

It is important to be clear about the argument of this paper. We are in no way claiming that suppression effects necessarily introduce bias. They can be an important part of observational research, capturing data generating processes whose outcomes would otherwise go unnoticed. Instead, we are arguing that, because suppression effects can introduce bias that favors authors, and do so in a nontransparent way, readers typically want to be aware of them. We are also in no way claiming that these 30-40% of articles that rely on suppression effects to achieve statistical significance are false positives. Instead, we are simply pointing out that editors and reviewers would likely have scrutinized the control variables more closely if they knew these articles

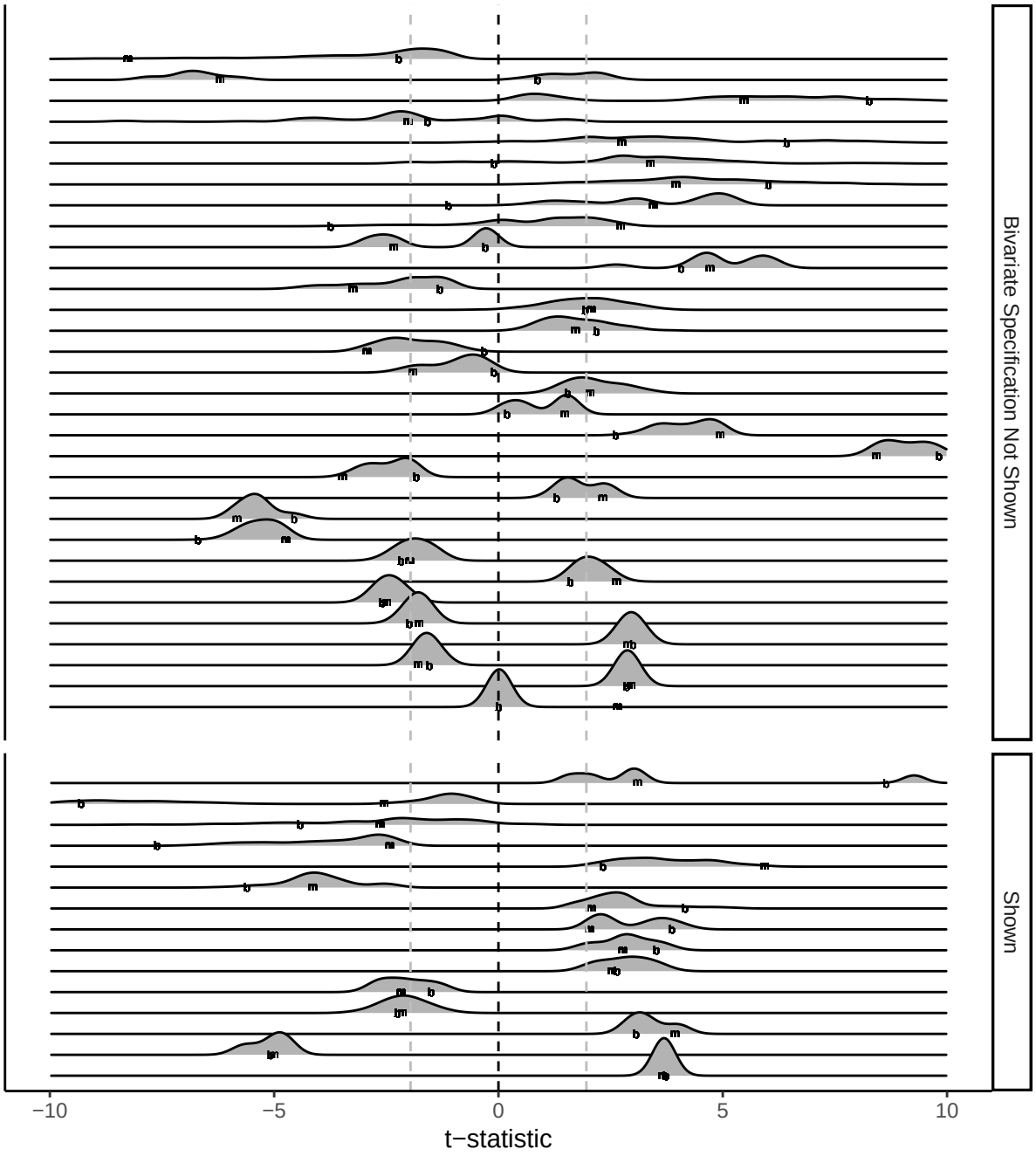

Figure 7: Distribution of t-Statistics across All Possible Control Variable Combinations for Observational Studies (randomly sampled in cases with more than 10,000 combinations, sorted by variance). 'm' shows the full multivariate t-statistic and 'b' the bivariate t-statistic. The grey dotted lines mark t-statistics at -1.96 and 1.96, the $p = 0.05$ thresholds for conventional statistical significance. We exclude two studies here because of the inordinate time required to estimate each specification.

depended on suppression effects for their statistical significance. That additional scrutiny may or may not change the publication outcomes for these articles. We are also in no way arguing that bivariate estimates are less biased or preferable. Instead, we are arguing that requiring disclosure of bivariates alerts readers to suppression effects and the extent of those effects. Readers want to know about their presence, we think, because they mean that estimates are, by definition, not robust to model specification, more vulnerable to hidden discretion by authors in their choice of controls, and potentially produced by opaque mechanisms such as amplification bias. In contrast, control variables that reduce the magnitude of key estimates can also introduce bias, but this bias works against authors.

The degree to which we find studies depending on suppression effects highlights the problems with research oriented around p-value thresholds (McShane et al., 2019). When researchers must report findings with p-values below a certain threshold to publish, incentives for them to "find a way" become strong. This leads to numerous perverse incentives, including model selection based on p-values.

As a field, we can limit discretion. Authors can assess robustness to a wider array of control variable choices with Bayesian Model Averaging (Bartels, 1997; Montgomery and Nyhan, 2010; Leamer, 2016) and related techniques. They can also use specification-curve analysis in which they report their key effect estimate across all theoretically justifiable model specifications (Simonsohn, Simmons and Nelson, 2015). Researchers can also limit the effects of model discretion by matching before they analyze their data (Hainmueller, 2012; Sekhon, 2011; Imai and Ratkovic, 2014), reducing model extrapolation (Ho et al., 2007). They could use hold out samples and apply newly developed estimators for cases where the number of control variables is large (Athey, Imbens and Wager, 2018; Ning, Peng and Imai, 2017). They can also preregister control variables before data collection as part of a pre-analysis plan (Casey, Glennerster and Miguel, 2012; Humphreys, Sanchez de la Sierra and van der Windt, 2013).

Most simply, researchers could disclose the bivariate specification to allow reviewers and readers to assess the effects of adding control variables.[6] If the bivariate specification departs noticeably from other specifications, authors need to explain why. This is difficult in cases with more than a handful of controls (Achen, 2002), though see Gelbach (2016) for a formalization. By disclosing the bivariate, readers can assess whether control variables could be introducing bias that favors the authors. Given the high rate of false positives in published research, readers should know this basic fact.

Besides limiting the potential for discretion, researchers should also be attuned to the source of the suppression effects. In some cases, researchers have strong justifications for them, such as in the Tesler and Sears example (2010) and the Davenport (2015) examples discussed earlier. In other cases, however, researchers should consider excluding the suppressor variables, such as when the source is classical suppression, mediation, or amplification bias.

[6]Presenting bivariate relationships is important for yet another reason: they may help researchers infer the level of confounding (Peters, Bühlmann and Meinshausen, 2016; Oster, 2016; Altonji, Elder and Taber, 2005).

6 Authors' Note

Funding for this research was provided by the Berkeley Initiative for Transparency in the Social Sciences (BITSS), a program of the Center for Effective Global Action (CEGA), with support from the Laura and John Arnold Foundation. For helpful comments, we thank participants at the Midwest Political Science Association 2017 Meeting, the Berkeley Initiative for Transparency in the Social Sciences (BITSS) 2016 Annual Meeting, the Berkeley Research Workshop in American Politics, and Jasjeet Sekhon's methods seminar. We also thank Larry Bartels, Martin Bisgaard, Allan Dafoe, Vicky Fouka, Donald Green, Andy Hall, Jens Hainmueller, Daniel Hopkins, Macartan Humphreys, Greg Huber, Kosuke Imai, Josh Kalla, Martin Vinaes Larsen, Yotam Margalit, John Bullock, Seth Hill, John Marshall, Andrew McCall, David Nield, Mark Pickup, Jasjeet Sekhon, Laura Stoker, and the anonymous reviewers. Replication materials are available from Lenz and Sahn (2020).

References

Achen, C. H. 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5:423–450.

Altonji, Joseph G., Todd E. Elder and Christopher R. Taber. 2005. "An Evaluation of Instrumental Variable Strategies for Estimating the Effects of Catholic Schooling." *The Journal of Human Resources* 40(4):791–821.

Athey, Susan, Guido W. Imbens and Stefan Wager. 2018. "Efficient Inference of Average Treatment Effects in High Dimensions via Approximate Residual Balancing." *Journal of the Royal Statistical Society-Series B* 80(4):597–623.

Bartels, Larry M. 1997. "Specification Uncertainty and Model Averaging." *American Journal of Political Science* 41:641–674.

Bhattacharya, Jay and William B Vogt. 2012. "Do Instrumental Variables Belong in Propensity Scores?" *International Journal of Statistics & Economics* 9(A12):107–127.

Bobko, Philip. 2001. *Correlation and Regression: Applications for Industrial Organizational Psychology and Management*. SAGE.

Brodeur, Abel, Mathias Lé, Marc Sangnier and Yanos Zylberberg. 2016. "Star Wars: The Empirics Strike Back." *American Economic Journal: Applied Economics* 8(1):1–32.

Casey, Katherine, Rachel Glennerster and Edward Miguel. 2012. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan." *The Quarterly Journal of Economics* 127(4):1755–1812.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science* 22(4):341–352.

Cole, Stephen R, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson and Charles Poole. 2010. "Illustrating Bias Due to Conditioning on a Collider." *International journal of epidemiology* 39(2):417–420.

Conger, Anthony J. 1974. "A Revised Definition for Suppressor Variables: A Guide to Their Identification and Interpretation." *Educational and psychological measurement* 34(1):35–46.

Crede, Marcus, Andrew Gelman and Carol Nickerson. 2016. "Questionable Association between Front Boarding and Air Rage." *Proceedings of the National Academy of Sciences* 113(47):E7348.

Davenport, Tiffany C. 2015. "Policy-Induced Risk and Responsive Participation: The Effect of a Son's Conscription Risk on the Voting Behavior of His Parents." *American Journal of Political Science* 59(1):225–241.

Gelbach, Jonah B. 2016. "When Do Covariates Matter? And Which Ones, and How Much?" *Journal of Labor Economics* 34(2):509–543.

Gerber, Alan and Neil Malhotra. 2008. "Do Statistical Reporting Standards Affect What Is Published? Publication Bias in Two Leading Political Science Journals." *Quarterly Journal of Political Science* 3(3):313–326.

Gerber, Alan S., Neil Malhotra, Conor M. Dowling and David Doherty. 2010. "Publication Bias in Two Political Behavior Literatures." *American Politics Research* 38(4):591–613.

Hainmueller, Jens. 2012. "Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies." *Political Analysis* 20(1):25–46.

Ho, Daniel E., Kosuke Imai, Gary King and Elizabeth A. Stuart. 2007. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15(3):199–236.

Holling, Heinz. 1983. "Suppressor Structures in the General Linear Model." *Educational and Psychological Measurement* 43(1):1–9.

Horst, Paul. 1941. "The Role of Predictor Variables Which Are Independent of the Criterion." *Social Science Research Council* 48(4):431–436.

Humphreys, Macartan, Raul Sanchez de la Sierra and Peter van der Windt. 2013. "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." *Political Analysis* 21(1):1–20.

Imai, Kosuke and Marc Ratkovic. 2014. "Covariate Balancing Propensity Score." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76(1):243–263.

Ioannidis, John PA. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8):e124.

Ioannidis, John PA, Tom D Stanley and Hristos Doucouliagos. 2017. "The Power of Bias in Economics Research." *The Economic Journal* 127:F236–F265.

King, Gary. 1995. "Replication, Replication." *PS: Political Science and Politics* 28(3):444–452.

Klein, Richard A, Kate A Ratliff, Michelangelo Vianello, Reginald B Adams Jr, Štěpán Bahník, Michael J Bernstein, Konrad Bocian, Mark J Brandt, Beach Brooks and Claudia Chloe Brumbaugh. 2014. "Investigating Variation in Replicability." *Social psychology* 14(3):142–52.

Klein, Richard A, Michelangelo Vianello, Fred Hasselman, Byron G Adams, Reginald B Adams Jr, Sinan Alper, Mark Aveyard, Jordan R Axt, Mayowa T Babalola and Štěpán Bahník. 2018. "Many Labs 2: Investigating Variation in Replicability across Samples and Settings." *Advances in Methods and Practices in Psychological Science* 1(4):443–490.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *The American Economic Review* 73:31–43.

Leamer, Edward E. 2016. "S-Values: Conventional Context-Minimal Measures of the Sturdiness of Regression Coefficients." *Journal of Econometrics* 193(1):147–161.

Lenz, Gabriel S. and Alexander Sahn. 2020. "Replication Data for: Achieving Statistical Significance with Control Variables and without Transparency." *Harvard Dataverse* V1.
URL: *https://doi.org/10.7910/DVN/XIEJCR*

Lewis, Jerry W and Luis A Escobar. 1986. "Suppression and Enhancement in Bivariate Regression." *Journal of the Royal Statistical Society: Series D (The Statistician)* 35(1):17–26.

MacKinnon, David P, Jennifer L Krull and Chondra M Lockwood. 2000. "Equivalence of the Mediation, Confounding and Suppression Effect." *Prevention Science* 1(4):173–181.

McShane, Blakeley B, David Gal, Andrew Gelman, Christian Robert and Jennifer L Tackett. 2019. "Abandon Statistical Significance." *The American Statistician* 73(sup1):235–245.

Middleton, Joel A, Marc A Scott, Ronli Diakow and Jennifer L Hill. 2016. "Bias Amplification and Bias Unmasking." *Political Analysis* 24(3):307–323.

Montgomery, Jacob M. and Brendan Nyhan. 2010. "Bayesian Model Averaging: Theoretical Developments and Practical Applications." *Political Analysis* 18(2):245–270.

Nickerson, Carol A and Nicholas JL Brown. 2019. "Simpson's Paradox Is Suppression, but Lord's Paradox Is Neither: Clarification of and Correction to Tu, Gunnell, and Gilthorpe (2008)." *Emerging Themes in Epidemiology* 16(1):5–16.

Ning, Yang, Sida Peng and Kosuke Imai. 2017. "High Dimensional Propensity Score Estimation via Covariate Balancing." https://imai.fas.harvard.edu/research/files/hdCBPS.pdf.

Oster, Emily. 2016. "Unobservable Selection and Coefficient Stability: Theory and Evidence." *Journal of Business & Economic Statistics* pp. 1–18.

Pearl, Judea. 2009. *Causality*. Cambridge, NY: Cambridge University Press.

Pearl, Judea. 2010. On a Class of Bias-Amplifying Variables That Endanger Effect Estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in ARtificial Intelligence*. Corvallis, OR: AUAI pp. 417–424.

Pearl, Judea. 2011. "Invited Commentary: Understanding Bias Amplification." *American Journal of Epidemiology* 174(11):1223–1227.

Peters, Jonas, Peter Bühlmann and Nicolai Meinshausen. 2016. "Causal Inference by Using Invariant Prediction: Identification and Confidence Intervals." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78(5):947–1012.

Sekhon, Jasjeet S. 2011. "Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching Package for r." *Journal of Statistical Software* 42.

Simonsohn, Uri, Joseph P. Simmons and Leif D. Nelson. 2015. Specification Curve: Descriptive and Inferential Statistics on All Reasonable Specifications. SSRN Scholarly Paper 2694998 Social Science Research Network Rochester, NY: .

Tesler, Michael and David O Sears. 2010. *Obama's Race: The 2008 Election and the Dream of a Post-Racial America*. Chicago: University of Chicago Press.

Tzelgov, Joseph and Avishai Henik. 1991. "Suppression Situations in Psychological Research: Definitions, Implications, and Applications." *Psychological Bulletin* 109(3):524.

Wooldridge, Jeffrey M. 2016. "Should Instrumental Variables Be Used as Matching Variables?" *Research in Economics* 70(2):232–237.

Wyss, Richard, Mark Lunt, M Alan Brookhart, Robert J Glynn and Til Stürmer. 2014. "Reducing Bias Amplification in the Presence of Unmeasured Confounding through Out-of-Sample Estimation Strategies for the Disease Risk Score." *Journal of Causal Inference* 2(2):131–146.