Political Science Research and Methods Vol 6, No. 1, 33–56 January 2018

doi:10.1017/psrm.2015.74

Assessing Threats to Inference with Simultaneous Sensitivity Analysis: The Case of US Supreme Court Oral Arguments*

JEFFREY BUDZIAK AND DANIEL LEMPERT

Political scientists relying on observational data face substantial challenges in drawing causal inferences. A particularly problematic threat to inference is the unobserved confounder. As a means to assess this threat, we introduce simultaneous sensitivity analysis to the political science literature. As an application, we consider the potentially confounded relationship between Supreme Court justice voting and oral argument quality. We demonstrate that this relationship is sensitive to the presence of a confounder, to a degree that threatens inference, and explore the confounder both theoretically and empirically. More generally, we show how sensitivity analysis can guide inquiry related to a covariate that cannot be directly measured.

Social science researchers are fundamentally interested in inferring causal relationships. In some disciplines, this is primarily achieved by experimental manipulation. By randomly assigning units to treatment conditions, experimentation allows researchers to isolate the effect of the treatment on the observed outcome. This allows researchers to have confidence that any differences in the outcome demonstrated by observations in the treated and control conditions were in fact caused by the treatment. While the value of randomized experimentation for causal inference is beyond debate, not all social science questions permit experimental manipulation. This problem is particularly acute for political science, where the nature of the subject matter frequently forecloses traditional randomized experimentation. Granted, political scientists can address certain questions with "natural" or "quasi-experiments" that replicate some (but not all) of the benefits of randomized experimentation (Sehkon 2009; Gerber and Green 2012; Sekhon and Titiunik 2012). However, for many research questions, political scientists are forced to rely on exclusively observational data.

Political scientists are nevertheless just as interested in demonstrating causal relationships as are social scientists whose subject matter lends itself to greater control over experimental conditions. The demands of drawing appropriate causal inferences has spurred recent work in political science that has pointed to difficulties associated with inferring causality in observational studies where experimental manipulation is untenable (see Green and Gerber 2002; Cook, Shadish and Wong 2008; Imai et al. 2011; Keele and Minozzi 2013). Though some of these difficulties have been addressed by methodological advances reviewed in Morgan and Winship (2007) and Rosenbaum (2010), confounding by an unobserved variable remains a particularly vexing threat to inference. To interpret a statistical association between X and Y as causal,

* Jeffrey Budziak is an Assistant Professor in the Department of Political Science, Western Kentucky University, 1906 College Heights Blvd., 300 Grise Hall, Bowling Green, KY 42101 (jeffrey.budziak@wku.edu). Daniel Lempert is an Assistant Professor in the Department of Politics, SUNY Potsdam, 44 Pierrepont Avenue, Potsdam, NY 13676 (lemperds@potsdam.edu). A version of this paper was presented at the 2010 Conference on Empirical Legal Studies at Yale University. For helpful comments, discussion, and suggestions, the authors thank Larry Baum, Greg Caldeira, William Minozzi, and Dylan Small. The authors thank Timothy Johnson, James Spriggs, and Paul Wahlbeck for making their data available. To view supplementary material for this article, please visit http://dx.doi.org/10.1017/psrm.2015.74

a researcher must be confident that there is no confounder that affects both X and Y. In observational studies, such confidence is often unwarranted. After all, the characteristics of confounders are, by nature, difficult to assess: scholars typically lack either the theoretical foresight necessary to anticipate a confounder's presence, the tools necessary to measure its impact, or both. This limitation poses a fundamental problem for any causal claim using observational data.

Here, we introduce to the political science literature *simultaneous sensitivity analysis*—a tool specifically designed to help overcome this limitation. Simultaneous sensitivity analysis allows scholars to assess an inference's sensitivity to unobserved confounders (Small et al. 2009). This is true, even absent any *a priori* expectations regarding the nature of the relationship between a potential confounder and the variables of theoretical interest. This tool is thus among recent methodological advances (see, e.g., Rosenbaum 2010) that have the potential to drastically improve political scientists' confidence in drawing causal inferences using observational data.

The article proceeds as follows. First, we situate simultaneous sensitivity analysis within a family of methodological tools designed to explore the robustness of causal inferences to confounding by unobserved covariates. Second, we sketch some basic theory of simultaneous sensitivity analysis and of the methodological context within which it is applied. Next, we apply the method to assess the sensitivity of an observed relationship between the quality of oral argument presentation and Supreme Court justice voting. We then present two empirical tests that indicate the presence of a confounder that, the sensitivity analysis suggests, threatens inference; in addition, we make a theoretical case for *legal quality* as a candidate confounder. We conclude with methodological and substantive implications.

SENSITIVITY ANALYSIS: OVERVIEW AND INTUITION

With greater emphasis on drawing causal inferences has come more attention to the impact of unobserved covariates on such inferences. Various tools often subsumed under the label "sensitivity analysis" serve to quantify the impact of a hypothesized confounder—that is, an unobserved covariate that is theorized to be associated with a treatment and the response—on a putative causal effect's statistical significance or size. The approach we use below is one of the so-called "Rosenbaum-style" sensitivity analyses.[1] Sensitivity analyses in this class are typically based on randomization inference[2] conducted after stratifying observations into sets within which observations are matched on observed covariates, and allow the analyst to specify at least one of two parameters, indicating a hypothesized confounder's relationship to treatment (the cause of interest) and/or response (the outcome of interest).[3] Given this hypothesized confounder, the output from the sensitivity analysis is usually the maximum probability that the null holds, a point estimate for the minimum effect size, or an associated confidence interval. Figure 1 depicts a generic case in which the sensitivity analysis is useful.

[1] Liu, Kuramoto and Stuart (2013) is a recent overview that describes and discusses implementation of some other, less closely related approaches to sensitivity analysis; see also citations in Small et al. (2013, 1462), in particular, Imbens (2003).

[2] For an introduction to randomization inference, see, for example, Rosenbaum (2002, 19–70) or Rosenbaum (2010, 21–63); we describe some of the intuition below.

[3] Though the approach we describe is applied after stratification into matched sets, we note that the essential assumptions for causal inference are identical, or at least closely analogous, across matching and regression-based methods (Morgan and Winship 2007, Chapter 5.3; Hosman, Hansen and Holland 2010, 849–50; Keele and Minozzi 2013, 193). For a recent approach to sensitivity analysis after ordinary least squares regression, see Hosman, Hansen and Holland (2010).

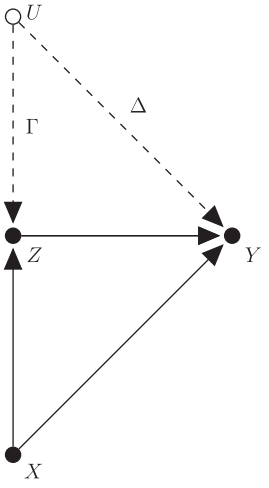

Fig. 1. An unobserved covariate U confounds the effect of treatment Z on outcome Y, despite adjustment for measured covariates X
Note: In a simultaneous sensitivity analysis, analyst-specified parameters Γ and Δ quantify the strength of the hypothesized confounding relationships, thereby allowing inference about the true effect of *Z* on *Y*.

We now discuss some specifics in an informal manner; technical details are available in the cited sources and reviewed as needed for our application in third section. The Supporting Information describes relevant software.

A simple (though somewhat unsubtle) way to classify Rosenbaum-style sensitivity analyses is by the sensitivity parameters that are specified by the analyst, by the type of treatments and responses (i.e., ordinal/continuous or binary) that are admissible, and by the type of matching (e.g., pair, 1:*k*, or full) after which it is employed. We sketch basic developments, focusing on these distinctions. Foundational work in sensitivity analysis is reviewed in Rosenbaum (2002, 105–70) and focuses on *primal* sensitivity analysis for matched pairs. In a primal sensitivity analysis, the hypothesized confounder is assumed to have a near-perfect association with response, and the analyst specifies a single sensitivity parameter that indicates how much more likely an observation is to be assigned to treatment (and not control) due to the unobserved confounder.[4] Thus, the analyst specifies only a single parameter. Primal sensitivity analysis for matched pairs is probably the best known of the Rosenbaum-style sensitivity analyses; applications in political science include Hainmueller and Hangartner (2013) and Sen (2014). Though these studies, like most applications, consider a binary treatment, ordinal or continuous treatments can also be analyzed: Rosenbaum (1989) gives clear and detailed discussion. Gastwirth, Krieger and Rosenbaum (2000) extends primal sensitivity analysis beyond pair matching. The case of matching (each treated subject) with multiple controls is detailed, though analysis after full matching is closely analogous. The authors consider a binary treatment with a continuous response.

Closely parallel to primal sensitivity analysis is *dual* sensitivity analysis (see Gastwirth, Krieger and Rosenbaum 1998). As in a primal sensitivity analysis, a single parameter is specified, but in a dual sensitivity analysis the analyst-specified parameter indicates the degree of association between the confounder and *response*, and it is the relationship between the confounder and treatment that is assumed to be nearly perfect. Though dual sensitivity analysis

[4] Intuitively, for a continuous or ordinal treatment dose, this parameter indicates how likely, for a typical pair of observations, it is that the observation with the higher value of the confounder receives the higher dose.

is relatively rarely used, Gastwirth, Krieger and Rosenbaum (1998, 918) discusses designs for which it is appropriate, and cites an application.

Simultaneous sensitivity analysis for matched pairs is introduced by Gastwirth, Krieger and Rosenbaum (1998). In a simultaneous sensitivity analysis, the analyst specifies two parameters: one (Γ) describes the association between a hypothesized confounder and treatment, and the other (Δ) the association between confounder and response. Thus, primal and dual sensitivity analysis can be thought of as special cases of simultaneous sensitivity analysis. Gastwirth, Krieger and Rosenbaum (1998, sec. 3) presents applications with binary treatments, but the method can be used with any combination of binary or ordinal/continuous treatments and responses (Gastwirth, Krieger and Rosenbaum 1998, 908). Small et al. (2009) extends simultaneous sensitivity analysis to 1:k matching, matching with a variable number of controls, and full matching. The case of binary treatment and continuous/ordinal response is developed.[5]

SIMULTANEOUS SENSITIVITY ANALYSIS AFTER FULL MATCHING: NOTATION, ASSUMPTIONS, BASIC THEORY

In this section, we specify some of the basic theory of simultaneous sensitivity analysis. We draw very heavily on Gastwirth, Krieger and Rosenbaum (1998) and Small et al. (2009), to which we refer the reader for additional results and important details. Because, in our application, we use full matching, we present only the result from Small et al. (2009). However, the discussion is largely applicable to other Rosenbaum-style sensitivity analyses as well. Specifically, the "Randomization Inference and Hypothesis Testing" and "A Model for Simultaneous Sensitivity Analysis" sections are wholly consistent with simultaneous sensitivity analysis after pair matching (Gastwirth, Krieger and Rosenbaum 1998).[6]

Full Matching

Full matching was introduced by Rosenbaum (1991), and is implemented in software written by Hansen (2007); Hansen (2004) gives an application of full matching, and describes some of its key properties for applied research. Full matching constructs a collection of matched sets containing either of one treated unit and any positive number of controls, or one control and any positive number of treated units (Hansen 2004, 613). Unlike 1:1 matching, full matching in general does not require that the analyst discard observations. Unlike fixed ratio matching (or any other matching), full matching guarantees that the weighted average distance between treated and control units is minimized, for any reasonable weight, for any match of the same size (Hansen 2004, 613).[7] Simply put, "a stratification that makes treated subjects and controls as similar as possible is always a full matching (Rosenbaum 2010, 183)."

Randomization Inference and Hypothesis Testing

A full matching, then, constructs I matched sets, $i = 1, \ldots, I$ each containing $n_i \geq 2$ units, of which m_i are treated and $n_i - m_i$ are controls. In each set, $n_i = 1$ or $m_i - n_i = 1$, or both. Let

[5] With careful interpretation of the parameters, the Small et al. (2009) result can be applied as well to the case of continuous/ordinal treatment and binary response; that is, a case-referent (Rosenbaum 2002, 7–8) study where strata are constructed after full matching (such as Guan et al. 2009).

[6] Recall also that primal and dual sensitivity analyses can be considered special cases of the simultaneous sensitivity analysis, with Δ and Γ, respectively, approaching infinity.

[7] The *size* of a match is an ordered pair indicating the number of treated units and the number of control units that are matched (Hansen 2004, 613).

$Z_{ij} = 1$ if unit j in set i is treated and $Z_{ij} = 0$ if it is control. r_{Cij} is the response that unit j in set i would exhibit if placed in the control group—we will call r_{Cij} "response under control." r_{Tij} the response that would be exhibited if unit j in set i were exposed to treatment. Only one of r_{Cij} or r_{Tij} is observed for a given unit: call this R_{ij}—response under treatment assignment actually received (e.g., Rosenbaum 2010, 21–63). Let $\mathbf{r}_{Ci}$ be the n_i-dimensional vector of responses under control in group i, and $\mathbf{r}_C = [\mathbf{r}_{C1}^T, \ \ldots, \ \mathbf{r}_{CI}^T]^T$. Define $\mathbf{Z}_i$ and $\mathbf{Z}$ analogously.

Fisher's (1935) null hypothesis of no treatment effect states that $r_{Tij} = r_{Cij}$, $\forall\, i, j$ (implying $r_{Cij} = R_{ij}$). If treatment assignment within sets is random, this hypothesis can be assessed via randomization inference. In an observational study, there is no random assignment; however, by matching on all relevant pre-treatment covariates $\mathbf{x}$, we assume that the probability of assignment to treatment within matched groups is equal (Rosenbaum 2010, 65–90). Randomization inference can then proceed by "shuffling" treatment assignment, taking every permutation of each $\mathbf{Z}_i$ while holding r_{Cij} fixed, and calculating a test statistic $t(\mathbf{Z}, \mathbf{r}_C)$, for each of these equiprobable shuffles, thereby determining a null distribution of the test statistic. The null hypothesis is tested by calculating a significance level: the proportion of the time $t(\mathbf{Z}, \mathbf{r}_C)$ is greater than or equal to the observed test statistic. For larger data sets, calculating $t(\mathbf{Z}, \mathbf{r}_C)$ for every possible treatment assignment is impractical, so computational shortcuts to calculate significance levels have been developed (e.g., Rosenbaum 2010, 21–63).

A Model for Simultaneous Sensitivity Analysis

Randomization inference requires that all relevant pre-treatment covariates have been taken into account when constructing the matched sets. How can inference proceed when this requirement is not met? We are concerned here with a confounder that was not considered in the matching, and thus may be associated with both treatment and response, thereby violating the requirement that treatment assignment be random within strata.

More specifically, consider a binary confounder u_{ij}. (This assumption is a conservative one, and not as limiting as it might seem at first glance.[8]) Assume that u_{ij} is the relevant confounder, in that, for every subject, treatment assignment is conditionally independent of response under control, given the observed covariates and u_{ij}. Let $\mathbf{u}_i$ be the n_i-dimensional vector of u_{ij} in group i, and $\mathbf{u} = [\mathbf{u}_1^T, \ldots, \mathbf{u}_I^T]^T$.

In a simultaneous sensitivity analysis, two sensitivity parameters are specified. The sensitivity parameter γ describes the strength of the relationship between u_{ij} and assignment to treatment, and the sensitivity parameter δ describes the relationship between u_{ij} and response under control, as follows. Assume the following holds in the population before matching:

$$\Pr(Z_{ij} = 1 \,|\, \mathbf{x}_{ij}, u_{ij}) = \frac{\exp(\beta_i + \gamma u_{ij})}{1 + \exp(\beta_i + \gamma u_{ij})}, \tag{1}$$

and $\Pr(Z_{ij} = 0 \,|\, \mathbf{x}_{ij}, u_{ij}) = 1 - \Pr(Z_{ij} = 1 \,|\, \mathbf{x}_{ij}, u_{ij})$.

[8] For $0 \le u_{ij} \le 1$, Wang and Krieger (2006) show that for matched pairs, $n_i = 2$, the values of u_{ij} that maximize the null distribution of the test statistic are $u_{ij} = 0$ or 1. Thus, for matched pairs, assuming that u_{ij} is binary is just as conservative as assuming $0 \le u_{ij} \le 1$. The restriction that $0 \le u_{ij} \le 1$ is just a restriction on the scale of the unobserved covariate, a restriction needed if the numerical scale on the sensitivity parameters γ and δ is to have any meaning (Rosenbaum 2002, 109). Small et al. (2009, 208–9) show a result related to that of Wang and Krieger (2006): that for matched sets with three subjects, $n_i = 3$, the values of u_{ij} that maximize the null distribution of the test statistics are $u_{ij} = 0$ or 1. Further, Small et al. (2009) conjecture that this result holds for sets of any size.

Assume further that in the population, letting $\zeta_i(u_{ij})$ be a normalizing constant and $\kappa_i(r)$ an unknown function,

$$\Pr(r_{Cij} = r \,|\, \mathbf{x}_{ij}, u_{ij}) = \exp\{\zeta_i(u_{ij}) + \kappa_i(r) + \delta r u_{ij}\}. \quad (2)$$

Any number of outcome models, including the logit, multinomial logit, Poisson, normal, and gamma can be written in the form of (2) (see Gastwirth, Krieger and Rosenbaum 1998, 909). Note that β_i and κ_i vary with i because they are functions of $\mathbf{x}_i$.

The most directly relevant implications of this model are as follows. First, $\Gamma \equiv e^{\gamma}$ is, for any pair of observations with the same values of the observed covariates, but with different values of the confounder, the maximum ratio of the odds that one is treated, to the odds that the other is treated (e.g., Rosenbaum 2002, 106–9). Also, if responses are binary, $\Delta \equiv e^{\delta}$ is, for any pair of observations with the same values of the observed covariates, but with different values of the confounder, the maximum ratio of that odds that one has higher response, to the odds that the other has the higher response. If responses are not binary, and one unit has response r and the other $r^* > r$, the maximum ratio is $e^{(r^* - r)\delta}$ (see the discussion in Gastwirth, Krieger and Rosenbaum 1998, 909–10, 916).

Simultaneous Sensitivity Analysis After Full Matching

Our ultimate quantity of interest is the maximum probability of obtaining the observed test statistic $T = t(\mathbf{Z}, \mathbf{r}_C) = \mathbf{Z}^T\mathbf{q}$ under the null, for specified γ and δ. In other words—given a hypothesized confounder—we seek the maximum probability that we would obtain a test statistic associated with the treatment effect that is at least as large as the one present in the observational data, *even if in fact the treatment has zero effect on the outcome*. The key step in approximating this probability is finding, for each stratum i, the values of the confounder u_{ij} that maximize μ_i, the expectation of the test statistic's null distribution (Gastwirth, Krieger and Rosenbaum 2000; Small et al. 2009). (When more than one possible $\mathbf{u}_i$ maximizes μ_i, then—of such $\mathbf{u}_i$—the one that maximizes σ_i^2, the variance of the null distribution, is used in approximating the maximum probability (for theoretical justification, see Gastwirth, Krieger and Rosenbaum 2000, sec. 4).) A formal presentation of the approximation, from Small et al. (2009, 205–6), is given in the Appendix.

AN APPLICATION: ORAL ARGUMENT QUALITY AND SUPREME COURT JUSTICE VOTING

As an application, we use simultaneous sensitivity analysis to reexamine the relationship between quality of oral argument and Supreme Court justice voting. Johnson, Wahlbeck and Spriggs (2006) recently demonstrated a statistical relationship between oral argument quality and justice vote choice. This finding was contrary to the (then-)conventional wisdom that largely dismissed the influence of oral arguments on judicial behavior. Given the relatively limited role of legal argumentation in theories of Supreme Court decision making, this result has led many scholars of judicial politics to substantially reevaluate how they approach the study of judicial behavior and has spurred an active line of research, including examinations of how oral arguments reflect justices' emotions (Black et al. 2011), indicate issue salience (Black, Sorenson and Johnson 2013), and influence coalition formation (Black, Johnson and Wedeking 2012).

Johnson, Wahlbeck and Spriggs (2006) demonstrates this relationship by making use of novel data: oral argument "grades" assigned by Justice Harry Blackmun. Throughout his tenure on the Supreme Court, Justice Blackmun assigned each attorney presenting oral arguments either a numeric or letter grade. These grades, Johnson, Wahlbeck and Spriggs (2006, 100, 111)

explains, offer unique insight into Justice Blackmun's substantive evaluation of the quality of the oral argument presented by each attorney. After standardizing these grades onto a common metric, the authors create an "oral argument grade" measure by subtracting the appellees's standardized grade from the appellant's (for additional details concerning the Blackmun grades, see Johnson, Wahlbeck and Spriggs 2006, 104). The authors then include this measure, along with a host of control variables, in a regression model predicting the likelihood of a party's victory before the Court. The results indicate that parties represented by attorneys receiving relatively "better" oral argument grades were significantly more likely to win in the final vote.

Johnson, Wahlbeck and Spriggs (2006, 100, 111) invites the reader to infer a causal effect from the statistical relationships presented. (In particular: "this analysis is the first study to demonstrate a causal relationship between oral arguments and justices' votes [… and offers] systematic evidence of exactly how justices evaluate these arguments and whether they *directly* influence decisions" (Johnson, Wahlbeck and Spriggs 2006, 100).) However, as we have suggested above, statistical relationships consistent with causal effects do not imply them. To assert that the quality of an attorney's relative oral argument presentation affects justice votes, we, at a minimum, must address the threat posed by an unobserved covariate that affects both relative oral argument quality and justice voting.[9] We now assess the extent of this threat with simultaneous sensitivity analysis.

The specific steps are as follows. First, we discuss the basic design of our observational study, and then explain and justify our approach to matching. After evaluating balance (and finding it satisfactory), we present a series of sensitivity analyses based on the matched data.

Data

Treatment. We are interested in comparing cases similar in terms of observed characteristics, but differing in terms of parties' relative oral argument quality. We therefore define the following treatment and control groups. In every analysis we present, the control cases are those in which the plaintiff and the respondent attorneys had the same oral argument grade. In one set of analyses, we consider the treated group to be the cases in which the petitioner's attorney had the better oral argument. In the other, we consider the cases in which the respondent's attorney had the better oral argument to be the treated group. To compactly refer to these analyses, we write "petitioner-better" and "respondent-better," respectively.

We further consider three thresholds for categorization as a treated case. According to the first, any case in which the petitioner (respondent) had a better argument grades is considered treated. Per the second, only cases in which grades differed by at least a half of a standard deviation are considered treated (and cases with smaller, but non-zero, differences are dropped). According to the third, only cases in which the difference in grades was at least 1 SD are considered treated. In short hand, we will call these analyses, respectively, "positive-difference," "medium-difference," and "large-difference."

Response. The response (dependent variable) is the proportion of votes in favor of the party with the better oral argument. That is, in the petitioner-better analyses, the response is the proportion of votes in favor of the petitioner. In the respondent-better analyses, the response is

[9] Johnson, Wahlbeck and Spriggs (2006, 108) do recognize, but ultimately dismiss, one potentially confounding variable—better "case facts"—as such a threat. Yet, the danger of confounding remains, as subsequent commentary has recognized (see Sides and Lax 2012).

the proportion of votes in favor of respondent. Thus, treated cases are expected to have greater responses than controls.

Observed covariates. We consider the set of covariates in the second column of table 3 in Johnson, Wahlbeck and Spriggs (2006, 109), with one exception. Because we conduct a court-level—not a justice-level—analysis, instead of the justice-level *Ideological Compatibility with Appellant*, we balance on *Court Median Ideology*—as measured by the median justice's Martin and Quinn (2002) score—and whether there was a *Liberal Decision Below* (Spaeth 2007).

Matching

We use full matching, matching cases—within propensity score calipers—on the Mahalanobis distance, with the optmatch package for R (Hansen and Klopfer 2006).[10] In constructing the matched sets, we are mindful of two, potentially competing, considerations: not discarding data, and achieving good balance on observed covariates between the control and treated groups. To these ends, we specify two ways of constructing matched sets.

Matching Specification 1. The first specification emphasizes matching all observations. The caliper is one-fifth of the standard deviation of the propensity scores. If two observations' propensity scores differ by more than this amount, a large penalty (1000) is added to the Mahalanobis distance between the observations (e.g., Small et al. 2009, 207). In practice, this means that any treated observation can be matched to any control observation, but observations whose propensity scores differ by more than the caliper will be matched only if no match within the caliper is feasible, given the constraints of the problem. Following the recommendation in Hansen (2004, 614), we use a *thinning cap* of 1/2 and a *thickening cap* of 2: informally, we allow "the number of treated subjects divided by the number of controls [to range in a set] from about half up to about twice what that ratio is in the sample as a whole."

Matching Specification 2. The second specification favors better balance, at the cost of discarding observations. Again, a penalty is added to the Mahalanobis distance between two observations that differ in propensity scores by more than one-fifth of the standard deviation. However, matches between two observations that differ in propensity score by more than three-tenths of the standard deviation are absolutely barred (for a similar approach, see Haviland et al. 2008).[11] As well, we impose a weak common support requirement, such that any treated observation with a greater propensity score than the control with the greatest propensity score, and any control with a smaller propensity score than the treated observation with the smallest propensity score, is dropped from the analysis. Lastly, we relax the restriction on the sizes of matched sets, allowing up to eight observations in a matched set.

Balance. Balance on observed covariates is evaluated in the supplementary material; we use the R package RITools (Hansen and Bowers 2008; Bowers, Fredrickson and Hansen 2010; Koenker and Ng 2012). We report two statistics for each covariate, pre- and post-matching: the

[10] This is a standard approach in the literature (see e.g., Rosenbaum 2010, 163–86). If $\mathbf{x}$ is the matrix of covariate values, and $\hat{\Sigma}$ the sample covariance matrix of $\mathbf{x}$, the Mahalanobis distance between observations j and k is $(\mathbf{x}_j - \mathbf{x}_k)^T \hat{\Sigma}^{-1} (\mathbf{x}_j - \mathbf{x}_k)$ (e.g., Rosenbaum 2010, 170). To calculate the Mahalanobis distance between cases, we use the Mahascores package for Stata (Kantor n.d.). The propensity score is defined as the predicted probability of being in the treated group, based on a logit regression of treatment status on observed covariates, including linear terms for all covariates and quadratic terms for *Complexity* and *Experience*.

[11] Formally, this is done by setting the Mahalanobis distance between such observations to ∞.

standardized difference of means and a p-value associated with a randomization inference-based test for conditional independence of treatment status and the covariate within a matched set.[12] Referring the reader to the supplementary material for details, we make a few observations. Every full matching does a credible job of improving balance. However (as expected), balance is better for the set of matchings in Specification 2 (where we are not constrained to keep all observations). Under Specification 2, every covariate in every analysis (i.e., petitioner-better, respondent-better, positive-, medium-, and large-difference) is balanced, according to conventional interpretations of our balance statistics. Under Specification 1, post-matching balance tends to be marginal for at least one covariate; in particular, *US Appellant*, *US Appellee*, *SG Appellant*, *SG Appellee*, and *Relative Experience* are occasionally problematic.

Sensitivity Analysis

Having constructed matched sets, we conduct a series of simultaneous sensitivity analyses (Small et al. 2009). In particular, we are interested in the maximum probability of obtaining the observed test statistic $\mathbf{Z}^T\mathbf{q}$, given that the null hypothesis of no treatment effect holds, for specified levels of unobserved confounding. Here, $\mathbf{Z}^T\mathbf{q}$ is a version of the Hodges–Lehmann test statistic that is constructed by taking the sum of the aligned ranks for each treated observation's response, and dividing this sum by half the maximum of the aligned ranks (Small et al. 2009, 208).[13] Defining $\mathbf{Z}^T\mathbf{q}$ in this way, and replacing each response with its aligned rank for the purposes of the sensitivity analysis, allows for a convenient interpretation of the sensitivity parameter $\Delta = \exp(\delta)$ (Small et al. 2009, 208).[14] In particular, Δ can be interpreted as the maximum factor increase in the odds that, for a typical treatment-control pair of observations in the same matched set, the observation with the greater value of the unobserved covariate u_{ij} also has the higher response r_{Cij}. For example, a hypothesized confounder that, in a typical pair of observations, doubled the odds of having a greater response would correspond to $\Delta = 2$.[15] We interpret $\Gamma = \exp(\gamma)$ as the maximum factor by which odds of assignment to treatment can vary for any two observations in a matched set. For example, a confounder that tripled the odds of being treated would correspond to $\Gamma = 3$.

[12] Pre-matching, "the standardized difference of covariate means is the mean in the treatment group minus the mean in the control group, divided by the sd (standard deviation) in the same variable estimated by pooling treatment and control group sds on the same variable (Bowers, Fredrickson and Hansen 2010, 7)." After matching, "the denominator of the standardized difference remains the same but the numerator is a weighted average of within-stratum differences in means on the covariate (Bowers, Fredrickson and Hansen 2010, 7)." Following Hansen and Bowers (2008, 12), we weight each set in proportion to the harmonic mean of the number of treated and control cases in the set. Standardized differences smaller than 0.2 (e.g., Haviland et al. 2008) or 0.25 (e.g., Stuart and Green 2008) are routinely taken to indicate acceptable balance. Calculation of the randomization inference-based p-value involves the construction of a null distribution by randomly permuting treatment assignment for all observations (pre-matching) or within matched sets (post-matching); then, the observed unstandardized difference of covariate means is compared with the differences of means under this null distribution (Bowers, Fredrickson and Hansen 2010, 8). In our application, the two statistics usually, but not always, lead to the same inference about covariate balance.

[13] An aligned rank is calculated by subtracting from each observation's response the mean response in the matched set containing the observation, and ranking the residuals (Lehmann 1975).

[14] Formally, using the notation from fourth section, we replace each $\boldsymbol{r}_{ij}$ with its aligned rank, divide each rank by half the maximum of the ranks, and let $\mathbf{q}$ be the identity function, so that $\mathbf{Z}^T\mathbf{q}(\mathbf{r}, \mathbf{m})$ is the sum, over all treated observations, of this modified aligned rank (Small et al. 2009, 208). On the merits of this approach, see also Gastwirth, Krieger and Rosenbaum (1998, 915–6).

[15] A "typical" pair of observations entails one observation with response at the 25th percentile of responses, and the other with the response at the 75th percentile of responses; thus, their difference in responses is "typical" (Small et al. 2009, 208).

TABLE 1 *Simultaneous Sensitivity Analysis for Selected Values of Δ and Γ, Matching Specification 2*

Γ	Δ = 1	Δ = 1.1	Δ = 1.2	Δ = 1.5	Δ = 2	Δ = 2.5	Δ = 3	Δ = ∞
1	0.0966	0.0966	0.0966	0.0966	0.0966	0.0966	0.0966	0.0966
1.1	0.0966	0.1002	0.1037	0.1127	0.1241	0.1324	0.1386	0.1823
1.2	0.0966	0.1037	0.1105	0.1289	0.1536	0.1722	0.1864	0.2917
1.5	0.0966	0.1128	0.1292	0.1770	0.2470	0.3025	0.3455	0.6435
2	0.0966	0.1250	0.1555	0.2507	0.3954	0.5064	0.5872	0.9400
2.5	0.0966	0.1345	0.1768	0.3139	0.5170	0.6594	0.7518	0.9937
3	0.0966	0.1421	0.1944	0.3668	0.6117	0.7642	0.8515	0.9995
∞	0.0966	0.2175	0.3790	0.8167	0.9937	0.9999	1	1

Note: petitioner-better, positive-difference cases are considered treated; balance is evaluated in Supplemental Table 1. See text for details.

TABLE 2 *Simultaneous Sensitivity Analysis for Selected Values of Δ and Γ, Matching Specification 2*

Γ	Δ = 1	Δ = 1.1	Δ = 1.2	Δ = 1.5	Δ = 2	Δ = 2.5	Δ = 3	Δ = ∞
1	0.0914	0.0914	0.0914	0.0914	0.0914	0.0914	0.0914	0.0914
1.1	0.0914	0.0950	0.0984	0.1071	0.1179	0.1256	0.1313	0.1706
1.2	0.0914	0.0984	0.1051	0.1229	0.1462	0.1635	0.1765	0.2719
1.5	0.0914	0.1073	0.1233	0.1697	0.2364	0.2886	0.3287	0.6079
2	0.0914	0.1191	0.1487	0.2415	0.3811	0.4873	0.5646	0.9214
2.5	0.0914	0.1282	0.1691	0.3026	0.5010	0.6394	0.7302	0.9895
3	0.0914	0.1353	0.1858	0.3534	0.5937	0.7456	0.8338	0.9989
∞	0.0914	0.2035	0.3544	0.7860	0.9900	0.9998	1	1

Note: petitioner-better, medium-difference cases are considered treated; balance is evaluated in Supplemental Table 2. See text for details.

We implement Small et al.'s (2009) sensitivity analysis in arsimsens, our user-written program for Stata (included with the replication materials).[16] Tables 1–3 present the results of the petitioner-better analyses for Matching Specification 2, for a range of Γ and Δ; the other analyses are relegated to Supplemental Tables 7–15 in the supplementary material. For each combination of Γ and Δ, the tables give a p-value: the maximum probability that the null holds, given the specified Γ and Δ. We highlight some key trends. For most specifications, the inference of a non-zero effect is sensitive to an unobserved confounder, even one that has only a weak-to-moderate relationship to both oral argument quality and Court voting. Generally, we can be more confident in the inference that oral argument has an effect when it is the petitioner who has the stronger oral argument. In particular, as Table 3 shows, if we compare *only* cases in which the petitioner had a *much* better argument than the respondent to cases where the parties' oral arguments were of equal quality, the inference that oral arguments—in this subset of cases—affected justice voting is fairly robust. Nonetheless, even this inference would be sensitive to an unobserved confounder that has either (1) a moderately strong relationship to oral arguments and a very strong relationship to justice voting or (2) a very strong relationship to oral arguments and a moderate relationship to justice voting. (For example, $p > 0.05$ for $\Gamma = \infty$, $\Delta = 1.5$, and for $\Gamma = 1.5$, $\Delta = \infty$.)

[16] For detailed description, see Lempert (2015).

TABLE 3 *Simultaneous Sensitivity Analysis for Selected Values of Δ and Γ, Matching Specification 2*

Γ	Δ = 1	Δ = 1.1	Δ = 1.2	Δ = 1.5	Δ = 2	Δ = 2.5	Δ = 3	Δ = ∞
1	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027	0.0027
1.1	0.0027	0.0029	0.0030	0.0033	0.0038	0.0042	0.0044	0.0067
1.2	0.0027	0.0030	0.0033	0.0040	0.0051	0.0060	0.0068	0.0139
1.5	0.0027	0.0034	0.0040	0.0063	0.0103	0.0143	0.0179	0.0663
2	0.0027	0.0039	0.0052	0.0106	0.0226	0.0362	0.0497	0.2663
2.5	0.0027	0.0043	0.0063	0.0153	0.0380	0.0655	0.0938	0.5198
3	0.0027	0.0046	0.0072	0.0199	0.0549	0.0990	0.1448	0.7261
∞	0.0027	0.0087	0.0217	0.1296	0.4967	0.8046	0.9425	1

Note: petitioner-better, large-difference cases are considered treated; balance is evaluated in Supplemental Table 3. See text for details.

How might one assess whether a particular combination of Γ and Δ is plausible? Given the nature of confounding, direct empirical assessment will not be available. Rather, theory- and subject-specific knowledge (Does the literature suggest the presence of a relevant unmeasured covariate?) is required; ideally, this can be supplemented with indirect empirical tests used to give some sense of a confounder's characteristics. Continuing with the case of oral arguments, we next suggest how such an inquiry could proceed, examining a possible confounder theoretically and empirically.

EVALUATING THE CONFOUNDER: THEORY AND EVIDENCE

The results of our sensitivity analysis prompt two broad questions. First: What could be a confounder? Below, we offer an some theoretically informed speculation, based on our reading of existing scholarship on judicial decision making, that *legal quality* is a candidate confounder. Second: Can we detect the presence of a confounder that threatens inference about the effect of oral arguments? We do so by presenting two empirical tests suggesting an affirmative answer.

It is crucial to distinguish between the means by which we address these questions. Our empirical tests will show that a confounder exists, but can say nothing about what the confounder is. We rely on a review of relevant literature to make a theoretical case for legal quality as an intuitively plausible confounder, but our argument is only speculative. We must be clear that we are in no position to make causal claims about legal quality.[17]

Theoretical Expectations

Dominant theories of judicial behavior agree that Supreme Court justices are motivated by their policy preferences when casting votes (Rohde and Spaeth 1976; Segal and Spaeth 1993; Epstein and Knight 1998; Maltzman, Spriggs and Wahlbeck 2000; Segal and Spaeth 2002). According to these perspectives, legal factors—which represent the unique considerations associated with the jurisprudential nature of judicial decision making—play at most a minor role in constraining the choices of justices (for discussion, see Segal and Spaeth 2002, 92–6). But as noted above, a recently resurgent literature in the field of judicial politics examines the possible influence these

[17] The nebulous status of the confounder will be typical in studies in which sensitivity analysis is useful. After all, the only convincing way to show the presence of a specific confounder is to validly measure it and adjust for it—at which point it of course ceases to be a confounder, thereby vitiating the need for a sensitivity analysis.

legal factors (oral arguments among them) on Supreme Court decision making (Gillman 2001; Richards and Kritzer 2002; Friedman 2005; Bailey and Maltzman 2008; Hansford and Spriggs 2008; Bartels 2009; Lax and Rader 2010; Bailey and Maltzman 2011).

As difficult as it is to evaluate *whether* legal factors matter for Supreme Court decision making (Segal and Spaeth 2002, 48–53), assessing *how* legal factors matter poses a distinct set of complications. To effectively address either of these questions, it is important to distinguish conceptually between the different types of legal factors. Our theoretical framework is inspired by discussion in Baum (1997, 72–6), which considers legal influence in general, and offers some guidance in distinguishing between legal factors. Our framework examines, in turn, three types of legal factors: legal doctrine, case facts, and legal argumentation.[18]

The possible influence of legal doctrine is an outgrowth of the common law nature of the American legal system. In such a system, policy develops not only through legislation but through the aggregation of judicial decisions. This process creates a *corpus juris* that evolves slowly and is fundamentally resistant to radical change (Stone 1936). Judges operating within this type of legal system are therefore expected to demonstrate strict fidelity to legal precedents. Scholars attempting to verify the influence of legal doctrine have examined a number of potential mechanisms, including jurisprudential regimes (Richards and Kritzer 2002, but see Lax and Rader 2010), precedent (Spaeth and Segal 1999; Bailey and Maltzman 2008; Hansford and Spriggs 2008; Bartels 2009; Bailey and Maltzman 2011) and theories of constitutional interpretation (Howard and Segal 2002).

A related, but distinct, set of legal factors stem from the nature of the case itself. Case facts are not the result of choices made by judges or the parties to the case; rather, they are the independent set of facts that existed before the dispute rose to the status of a legal challenge. Scholars have examined the possible influence of case facts in a variety of circumstances, such search and seizure (Segal 1984),[19] the death penalty (George and Epstein 1992), establishment of religion (Ignagni 1994), and obscenity (McGuire 1990). Legal doctrine is frequently dependent on case facts. Most legal doctrines compel future judges to approach different fact situations through the lens of a test or rule laid down in a previous decision(s). Therefore, while conceptually distinct, legal doctrine and case facts are typically dependent on one another.

The final set of legal factors, legal argumentation, describes the possible influence of external players on the choices of judges. Legal argumentation is distinct from both case facts and legal doctrine. While the latter two are a function—at least in part—of the case itself, the former is entirely a function of the advocates before the court. The effect of legal argumentation can be observed when advocates, presenting case facts and legal doctrine in a manner favorable to their preferred outcome, persuade a court previously disinclined to adopt that preferred outcome to do so. There are a variety of opportunities for this type of persuasion, and scholars have examined how considerations like lawyer experience (Galanter 1974; McGuire 1995;

[18] Baum (1997), in his discussion of legal factors, examines "precedent" (72), "legal arguments" (74), and "case facts" (75) separately. His conceptualization of case facts is identical to what we present here. While he focuses specifically on precedent, precedent represents one component of the broader concept of legal doctrine. And while his discussion of "legal arguments" is specific to the types of arguments made by justices, we believe the logic generalizes to the types of arguments made by parties (both direct and indirect) to the case. Therefore, while Baum's discussion does not mirror ours exactly, we believe it provides a useful framework upon which to build.

[19] Segal (see Segal and Spaeth 2002, 314–20) argues that a case fact-based analysis is entirely consistent with attitudinal voting and therefore should not be characterized as definitive evidence of a "legal" influence. However, scholars who suggest that the law operates differently than a mere external constraint argue that results of case fact analyses are more consistent with a "legal" influence on judicial behavior (e.g., Gillman 2001, 473).

McAtee and McGuire 2007), briefs on the merits (Corley 2008; Wedeking 2010), on petition for certiorari, and as amici (Spriggs and Wahlbeck 1997; Collins 2004; Corley 2008), influence the decision-making process. Clearly, oral arguments (Johnson 2001; Johnson 2004; Johnson, Wahlbeck and Spriggs 2006) fall in the category of legal argumentation as well—thus, we can think of oral arguments as one element of one set (legal argumentation) of legal factors.

We urge that scholars examining the influence of legal factors (either generally or specifically) consider their potentially interdependent nature. If certain (or even most or all) legal factors are inherently inseparable, this presents a (potentially unsolvable) challenge for drawing casual inferences concerning the unique influence of a single factor. (This threat looms especially large, of course, when not all relevant factors can be measured.) We suggest that the extent to which one party has a more viable legal position is unlikely to be driven by any one legal factor alone. One might reasonably expect that the individual elements of legal argumentation (e.g., oral arguments, merits briefs, cert briefs) are at least moderately associated with one another. And because legal argumentation is dependent on case facts and governing legal doctrine, it may well be that evaluations of oral argument quality will be related to an evaluations of the overall legal quality of the party's position.[20] Assuming we are right about this, it is useful to conceptualize *legal quality* as lying in the intersection of a set of closely related legal factors.

The relationship between legal quality and justice voting should be weaker than the relationship between oral arguments and legal quality. Many theories of Supreme Court decision making allow for the possibility that legal factors could influence judicial choices. In perhaps the most famous formulation, Gibson notes that "judges' decisions are a function of what they prefer to do, tempered by what they think they out to do, but constrained by what they perceive is feasible to do" (1983, 9). Legal factors are theorized to potentially limit the choices justices make, either because they believe that legal factors compel them to behave in a certain manner or constrain them in pursuing their policy goals.[21]

The above discussion is inherently speculative in nature. We must emphasize again that theory, by itself, can never definitively answer whether a confounder is present, and if so, what that confounder is. But theory is nonetheless useful (perhaps even necessary) for interpreting the results of a sensitivity analysis—that is, assessing whether a confounder with the characteristics that, according to the sensitivity analysis, threaten inference is likely to exist. In our application, we will be able to present ancillary empirical tests that indicate the presence of a confounder and give some sense of its characteristics. But such an exercise will not always be possible, and it is then, especially, that subject-specific theory becomes indispensable for evaluating the confounding threat.

A Test at the Certiorari Stage

Our first test designed to probe for a confounder begins in what is, at first glance, an unlikely place: the agenda-setting stage of the judicial decision-making process. The certiorari (cert) vote affords a critical test for the independence of oral arguments from legal quality (or some other

[20] The example of search and seizure is instructive. The well-demonstrated importance of case facts (Segal 1984) for search and seizure cases exists because the Court has established a set of legal doctrines that are heavily fact-dependent. Participants (as parties or as amici) attempt to persuade the Court to apply a legal doctrine favorable to their preferred outcome based on the facts of the case.

[21] It should be noted that many theorists of judicial behavior question just how constraining legal factors are in practice, particularly in light of the unique institutional position of the Court, which seems to allow for justices to pursue policy preferences nearly unconstrained (Segal and Spaeth 2002, 92–6).

factor observable to the justices at the agenda-setting stage). This is because of the temporal ordering of the events—cert voting occurs before oral arguments. Therefore, if oral argument grades are truly independent of all potential confounders—including other unmeasured legal factors—there should be no relationship between the grades and the likelihood of a justice voting to grant cert. However, if oral argument grades are confounded, we may observe such a relationship at the agenda-setting stage.

Data and measurement. To test our hypotheses, we analyze individual justice cert and jurisdictional votes in cases heard by the Burger Court (1970–1985), for which oral argument grades are publicly available.[22] Our dependent variable is whether a justice votes to hear a case (=1) or not (=0). The key independent variable is *oral argument grade*. When the petitioner receives a higher score than the respondent, we expect the justices to be more likely to vote for a grant (i.e., in favor of the petitioner). If our intuition is correct, oral argument grade should be positively signed.

We include a number of other variables that have been demonstrated in the agenda-setting literature to influence certiorari voting (Caldeira and Wright 1988; Caldeira, Wright and Zorn 1999). All of the variables, unless otherwise noted, are drawn from the Expanded Burger Court Judicial Database (Spaeth 2007). Several variables stem from the actions of the lower courts. These include *Conflict*, which takes the value of one if the Court reported to hear a case to resolve a conflict between or among circuits, *Unconstitutional*, which takes the value of one if a city, state, or federal law was declared unconstitutional by any court below, *Intermediate Reversal*, which takes the value of one if the court being reviewed by the Supreme Court reversed a lower court decision and *Dissent*, which takes the value of one when the Supreme Court noted a dissent in the court whose decision it reviewed. Variables capturing characteristics of legal parties have also been demonstrated to affect cert voting. These include *US Petitioner*, which takes the value of one if the appellant was the US government and *Amicus Brief*, which is coded one if at least a single amicus brief was filed in a court below. Further, we control for case characteristics. Specifically, as there have been suggestions that civil liberties cases may be more likely to be heard (Tanenhaus et al. 1963; Caldeira, Wright and Zorn 1999), we include a variable that indicates whether a case involves *Civil Liberties*. Finally, we include a control variable that accounts for a justice's ideological relationship to the parties. We follow the modification of Martin and Quinn (2002) (MQ) scores made by Johnson, Wahlbeck and Spriggs (2006). MQ scores are a vote-based ideal point estimate scaled so that the mean score is near 0, and the more conservative a justice, the higher his score. We rescale these scores to indicate *Ideological Affinity* with the petitioner: for conservative lower court decisions (where the petitioner presumably seeks a liberal outcome at the Supreme Court), we take the negative of each justice's MQ score, and for conservative lower decisions, we utilize the unmodified MQ score.[23] Because the dependent

[22] We are limited to this period because data on certiorari votes are not available after the end of the Burger Court. A further limitation, of course, is that oral argument grades exist only for cases receiving the necessary number of votes to be granted. This restricts the size of our sample, and limits our ability to generalize about the Court's agenda-setting, but this limitation is unavoidable, and does not bear on the results that are relevant for the present purposes.

[23] Note that this operationalization assumes that the more conservative a justice is, the more likely he is to vote to hear a liberal lower court decision, and the less likely he is to vote to review all conservative lower court decisions. This probably oversimplifies the ideological considerations at play, for at least two reasons. First, justices may behave strategically at the cert stage, violating the sincerity assumption implicit in the variable (Caldeira, Wright and Zorn 1999). Second, the unidimensional Martin and Quinn scores may not capture the relevant ideological dimension for all cases to come before the Court.

TABLE 4 *Relationship Between Oral Argument Grade and Justice Cert Voting*

Covariates	Coefficient
Oral argument grade	0.125***
	(0.038)
Conflict	0.102
	(0.126)
Unconstitutional	0.713*
	(0.286)
Ideological affinity	0.293***
	(0.038)
Intermediate reversal	−0.074
	(0.130)
Civil liberties	−0.240
	(0.070)
US petitioner	0.315*
	(0.146)
Amicus brief	0.233
	(0.125)
Dissent	−0.233
	(0.091)
Constant	0.892
	(0.152)

Note: dependent variable: Did justice vote to hear case? (1 = yes). Logit coefficients; standard errors in parentheses, clustered on justice. $N = 2843$.
*p < 0.05, **p < 0.01, ***p < 0.001.

variable is dichotomous, logistic regression (logit) techniques are employed. To account for possible non-independence of errors, standard errors are clustered by justice.

Results. The maximum likelihood estimates from a logit model predicting individual justices' votes to hear a case are presented in Table 4.[24] Most of the control variables behave as expected. Most importantly, oral argument grade is positively signed and statistically significant. All things being equal, justices are more likely to vote to grant a petition for cert when the attorney for the petitioner ultimately earns a higher grade in oral argument than the attorney for the respondent. Although the association is relatively modest when a justice and the petitioner are ideologically aligned, when a justice is only "somewhat close" or "very far" from the petitioner, changing the relative oral argument grade from its minimum to its maximum is associated with a 20 percentage points increase in the likelihood of a vote to grant. The result is robust to a number of model specifications. Even if we add to our baseline model the full suite of control variables from Johnson, Wahlbeck and Spriggs (2006, 109) designed to control for lawyer quality, the relationship remains.

How should we interpret the demonstration of a significant relationship between oral argument grades and the decision to grant cert? Plainly, it cannot be oral argument performance that affects the cert vote. Rather, the results can only be explained by the existence of a confounder that is significantly associated with oral arguments and observable to the justices before the cert vote. Thus, we think that legal quality remains a plausible candidate

[24] We utilize a more conventional, and simpler, regression technique in this analysis because here, we are not interested in the sensitivity of a casual inference. To the contrary, because oral arguments take place after the vote on cert, we are certain the two are not causally related.

confounder: cases in which the petitioner occupies the preferable legal position are more likely to be granted cert, and the strength of the parties' respective legal positions is known to the justices before cert voting. We hasten to add, though, that all the empirics show here is that a confounder observable to justices at the cert stage exists.

As relevant to our sensitivity analyses, we thus draw the following conclusions. There is an unobserved confounder that is—at least—substantially related to oral argument quality; moreover, this confounder has an appreciable impact on voting. And though this test shows the relationship between the confounder and cert voting, not merits voting, the association between these two sets of votes has been well demonstrated (see e.g., Caldeira, Wright and Zorn 1999). In the next section, we explore further the relationship between the confounder and oral arguments, presenting a test that suggests that this relationship is very strong indeed.

A Test Comparing Two Votes at the Merits Stage

What can we say about the structure of the relationship between oral argument grades and our confounder? We proceed by comparing two stages of merits voting: the justices' preliminary, conference vote on the merits and justices' final, report vote on the merits. Here, we hypothesize that if oral arguments are appreciably distinct from our confounder (say, legal quality) oral argument grades should have a greater impact on the conference vote than on the final report vote. The logic is straightforward. Psychologists have long demonstrated (Ebbinghaus 1913) a strong correlation between memory and recency of exposure to information. Because oral argument occurs at a fixed point in time, its impact should be greatest at temporally proximate stages of the decision-making process (the conference vote) and fade overtime (the report vote). On the other hand, if the effect of oral argument is strongly confounded with that of legal quality or some other factor that consistently influences justices throughout the period between the two votes, we would expect the strength of the correlation between oral argument grade and voting to remain similar in the conference and final votes.

Data and measurement. To test this hypothesis, we estimate separate logit models predicting conference votes and final report votes. In addition to oral argument grade, we include the control variables specified by Johnson, Wahlbeck and Spriggs (2006, table 1, column 2) in their original model of oral argument influence on merits votes.

Results. Column 1 of Supplemental Table 16, in the supplementary material, presents the results for votes cast at conference, and Column 2, the results for the final, report vote. The results are quite similar across both models. Oral argument grades correlate with merits votes both at the conference and report vote stages. However, we are chiefly interested in the substantive magnitude of the (putative) effect of oral argument grade. If the effect of oral argument grade on the probability of voting for the petitioner is appreciably stronger in the conference vote model, this would offer indirect evidence of the unique contribution of oral arguments to the decision-making process. Thus, in Figure 2 we plot the relationship between oral argument grades and the conference and report votes.[25] Figure 2 suggests there is very little appreciable difference in the impact across votes. The slopes of the predicted probabilities as

[25] The graphic is generated with spost13 (Long and Freese 2014). We set Complexity and Ideological Affinity at their mean, set Washington Elite Appellant and Washington Elite Appellee at 1, and the remaining variables at 0.

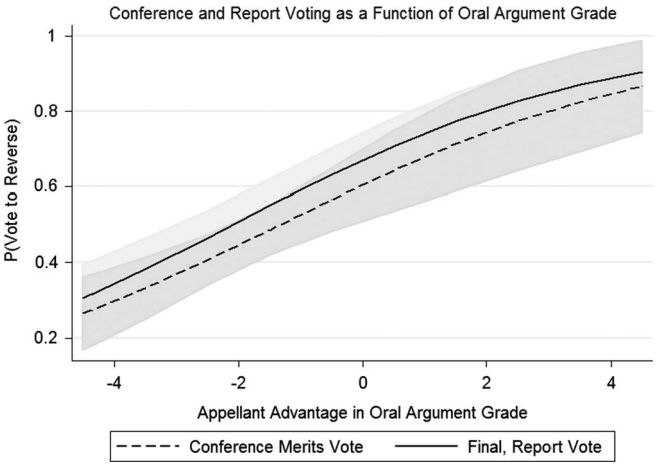

Fig. 2. Probability of voting to reverse, at conference, and at the final, report vote, as a function of oral argument grade

they range from the minimum to the maximum values of oral argument grade are almost perfectly parallel for the two votes.[26]

This result lends itself to two potential interpretations. First, it is theoretically possible that the estimated impact is consistent across decision-making stages because the effect of oral argument does not fade: it is not attenuated by the introduction of other influences on the process that are likely to be consulted between the conference vote and the report vote. In practice, this seems unlikely. At the latter stage, the effect of oral arguments should decrease as these arguments fade from memory; and it is exactly between the two stages that the justices begin to bargain over, persuade, and influence the votes of other justices (and the content of the majority opinion).[27] Scholars have consistently demonstrated that this collection of behaviors, frequently described as "the collegial game," exerts a substantial influence on the choices justices make (Maltzman, Spriggs and Wahlbeck 2000). Throughout the game, justices can readily refer to other potential influences, including the lower court opinions, merits and amicus briefs, and legal argumentation made by other justices through inter-chamber memoranda. This should attenuate the effect of oral argument.

A second, more likely, interpretation is that the putative effect of oral arguments is confounded by some other factor prominently before the Court between votes. That is, the observed consistency of the effect is driven not by the unique influence of oral arguments, but the effect of a confounder strongly associated with oral argument grades. We again speculate (but cannot prove) that legal quality is a plausible candidate confounder: a party who has the better oral argument is also more likely to be defending a position of higher legal quality as

[26] For the report vote, the change in predicted probabilities over the range plotted is 0.597; for the conference vote, 0.601. A more formal test for significance of the second difference, $\Delta\Delta[Pr(Y)] = [Pr(Y \mid \text{Grade} = 4.5, \text{Vote} = \text{Report}) - Pr(Y \mid \text{Grade} = -4.5, \text{Vote} = \text{Report})] - [Pr(Y \mid \text{Grade} = 4.5, \text{Vote} = \text{Conference}) - Pr(Y \mid \text{Grade} = -4.5, \text{Vote} = \text{Conference})]$ (see Berry, DeMeritt and Esarey 2010), as implemented in CLARIFY (Tomz, Wittenberg and King 2003), gives a 95 percent confidence interval of [–0.142, 0.163].

[27] It can be argued that this process actually occurs during oral argument (see Black, Johnson and Wedeking 2012). Even if this is correct, conventional understanding of the Court's decision-making process suggest that the greatest degree of interaction among the justices regarding the disposition occurs after oral argument rather than beforehand.

reflected in records, briefs, and memoranda. As justices refer to these documents, the inclination to favor the party with stronger legal support is reinforced.

But whatever the confounder, we believe that our result here (that the impact of oral argument grade essentially does not diminish between the two sets of merits votes) has clear implications for the relationship between the confounder and oral argument grades: that the relationship between two is very, very strong. In the language of our sensitivity analysis, it suggests a very high value of Γ. And our sensitivity analysis has shown that, given such a relationship between treatment and confounder, even a moderate relationship between the confounder and outcome threatens inference.

Can Some Confounders be Ruled Out?

We have argued that the preceding analyses are at least not inconsistent with the interpretation that legal quality is confounding the relationship between oral arguments and justice voting. But, given the unobserved nature of confounders, we cannot make an empirical case for any *specific* confounder. Acknowledging this reality, can we nonetheless narrow future inquiry by ruling out some potential confounders?

The relationship between cert voting and oral argument grades implies that some of the confounding is due to a variable, or variables, observable to justices *before* the cert vote. In principle, the confounder could be any single legal factor (or other covariate) then observable to the justices. But empirical assessment of the possibilities presents serious challenges. First, consider legal doctrine: it exists before the cert stage, but developing a test for its unique influence has proven difficult. Scholars have disagreed on the proper way to assess the constraining influence of precedent (see the debate between Spaeth and Segal (1999) and Gillman (2001)) or doctrine (see the debate among Bartels (2009), Richards and Kritzer (2002), and Lax and Rader (2010)).[28] A second potentially confounding legal factor that exists before cert are the facts of the case itself. We attempted to specifically test the role of case facts by examining whether oral argument grades would continue to predict final votes on the merits once we controlled for case facts. To do so, we focused on our analysis exclusively on search and seizure cases. We did so because of the well-developed case facts model Segal (1984) created to predict votes on search and seizure cases. Unfortunately, oral argument grade fails to predict final votes in this subset of cases *before* the inclusion of any case fact variables. As a result, no significance can be imputed to a null finding for oral argument's effects after controlling for case facts, and so we are left unable to foreclose the possibility that case facts could account for the confounding we observe. A final potential factor is legal argumentation, which includes argumentation contained directly in the cert and reply briefs (and indirectly reflected in lower court opinions) and lawyer quality (as reflected in the preparation of briefs and the identity of the lawyers filing briefs). Because the relationship between oral argument grades and cert voting is robust to inclusion of controls for lawyer quality, we believe that the characteristics of the lawyers cannot be the confounder.

Thus, we are limited in what we can say about the factors that account for the confounding. We cannot offer a conclusive series of tests eliminating legal doctrine, case facts, or alternative forms of legal argumentation as factors individually responsible for the results. Indeed, as we have noted, it is possible that a covariate heretofore uncontemplated by scholars is the confounder.

[28] One potential alternative is to assess the relevance of precedent in the manner developed in Hansford and Spriggs (2008). This approach is beyond the scope of this analysis, but appears to be an avenue for future research.

Our case for thinking about the confounder as legal quality is essentially theoretical. We have argued that the components of legal quality may well be closely inter-related. After all, it is difficult to understand the relevance of case facts independent of the legal doctrine that structures how they should be considered, just as it is difficult to consider the proper application of legal doctrine in the absence of relevant case facts. And all of this has the potential to be shaped by the manner through which it is presented to a court. If we are right, it is appropriate to give consideration to legal quality as the confounder we have demonstrated. But no stronger claim can be made until a measure of legal quality is developed, its effect is isolated by adjusting for the other potentially relevant covariates, and that influence is found reasonably robust to confounding.

CONCLUSION

To recap, we have introduced and then applied a method of sensitivity analysis to assess the characteristics that a confounder would have to have, for it to threaten the inference that oral arguments affect justice voting. Next, we made a theoretical case for legal quality as a candidate confounder. We then presented two tests that indicated the presence of a confounder (plausibly, our candidate confounder)—one that appears to influence the judicial decision-making process both before "A Test at the Certiorari Stage section" and after "A Test Comparing Two Votes at the Merits Stage section" oral arguments; more to the point, the tests suggested that the relationship of this confounder to treatment and outcome are of a character that the sensitivity analysis had indicated threaten inference.

Substantively, the nature of the confounding relationship we uncover makes drawing inferences about the effect of oral arguments on justice voting problematic. Suppose momentarily, for the sake of concreteness, that we are correct to speculate that legal quality is the confounder. To validly conclude that oral arguments *affect* changes in justice votes, one must be confident that justices are influenced significantly through oral arguments, instead of through the variety of other mechanisms by which they may come to understand the parties' relative legal quality. Moreover, even if such a conclusion could be validly drawn, it leaves unaddressed what is, arguably, the truly interesting counterfactual: would justice votes would be different if there were no oral arguments? To show this *counterfactual dependence*, in addition to causality, (e.g., Hitchcock 2007) one must show that a justice *would not* be influenced by the legal quality of a party's position in the absence of oral arguments. Demonstrating this lack of influence would be difficult, both theoretically and practically.[29]

In sum, given our empirical tests, we believe there is yet no warrant for the claim that oral arguments affect justice votes, let alone the claim that the votes are counterfactually dependent on oral arguments. But a door to further inquiry is opened. Identifying the confounder, perhaps by operationalizing and disentangling heretofore unmeasured legal factors, will no doubt require researchers to confront serious measurement and statistical challenges. But we anticipate that the importance of this line of research will be commensurate with its difficulty.

Methodologically, we have introduced to the political science literature a method of quantifying a putative causal effect's sensitivity to unobserved confounding that is particularly useful when an analyst has a reasoned basis for approximating the relationship between a

[29] Theoretically, doing so requires analysis of an unobservable counterfactual. Practically, evidence, including the results presented here, suggests that justices *are* influenced by components of legal quality (Collins 2004; Corley 2008; Wedeking 2010). And even if we disregard this evidence, it is clear that the *means* for justices to learn about legal quality are readily available.

proposed confounder and the outcome of interest, and the confounder and the independent variable of interest. We then proposed two indirect tests for assessing the strength of these relationships; although our exact approach will not generalize to all similar problems, these tests may serve to motivate appropriate exercises in other applications. It is worth emphasizing that we went beyond the typical approach to sensitivity analysis, in that we actually attempted to empirically assess characteristics of the confounder. Additionally, note that our application is one that is particularly well suited for a sensitivity analysis: we have an unobserved confounder we cannot measure—and thus cannot directly adjust for—but we can still say something about the confounder's strength of association with outcomes and (putative) causes of interest. In such situations, a simultaneous sensitivity analysis has much to recommend it. But we encourage that it be used to ensure that results are not fragile, even in cases where an unobserved confounder is merely a theoretical or hypothetical concern.

REFERENCES

Bailey, Michael A., and Forrest Maltzman. 2008. 'Does Legal Doctrine Matter? Unpacking Law and Policy Preferences on the U.S. Supreme Court'. *American Political Science Review* 102(3):369–84.

Bailey, Michael A., and Forrest Maltzman. 2011. *The Constrained Court: Law, Politics and the Decisions Justices Make*. Princeton, NJ: Princeton University Press.

Bartels, Brandon L. 2009. 'The Constraining Capacity of Legal Doctrine on the U.S. Supreme Court'. *American Political Science Review* 103(3):474–95.

Baum, Lawrence. 1997. *The Puzzle of Judicial Behavior*. Ann Arbor, MI: University of Michigan Press.

Berry, William D., Jacquline H. R. DeMeritt, and Justin Esarey. 2010. 'Testing for Interactions in Binary Logit and Probit Models: Is a Product Term Essential?'. *American Journal of Political Science* 54 (1):248–66.

Black, Ryan C., Marion W. Sorenson, and Timothy R. Johnson. 2013. 'Toward an Actor-Based Measure of Supreme Court Case Salience: Information-Seeking and Engagement During Oral Argument'. *Political Research Quarterly* 66(4):804–18.

Black, Ryan C., Sarah A. Truel, Timothy R. Johnson, and Jerry Goldman. 2011. 'Emotions, Oral Arguments, and Supreme Court Decision Making'. *Journal of Politics* 73(2):572–81.

Black, Ryan C., Timothy R. Johnson, and Justin Wedeking. 2012. *Oral Arguments and Coalition Formation on the U.S. Supreme Court: A Deliberate Dialogue*. Ann Arbor, MI: University of Michigan Press.

Bowers, Jake, Mark Fredrickson, and Ben B. Hansen. 2010. 'RItools: Randomization Inference Tools'. R Package Version 0.1-11.

Caldeira, Gregory, and John Wright. 1988. 'Organized Interests and Agenda Setting at the U.S. Supreme Court'. *American Political Science Review* 82(4):1109–127.

Caldeira, Gregory A., John R. Wright, and Christopher J.W. Zorn. 1999. 'Sophisticated Voting and Gate-Keeping in the Supreme Court'. *Journal of Law, Economics, and Organization* 15(3):549–72.

Collins, Paul M. Jr. 2004. 'Friends of the Court: Examining the Influence of Amicus Curiae Participation in U.S. Supreme Court Litigation'. *Law and Society Review* 38(4):807–32.

Cook, Thomas D., William R. Shadish, and Vivian C. Wong. 2008. 'Three Conditions Under Which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-Study Comparisons'. *Journal of Policy Analysis and Management* 27(4):724–50.

Corley, Pamela. 2008. 'The Supreme Court and Opinion Content: The Influence of Parties' Briefs'. *Political Research Quarterly* 61(3):468–78.

Ebbinghaus, Hermann. 1913. *Memory. A Contribution to Experimental Psychology*. New York, NY: Teachers College Columbia University.

Epstein, Lee, and Jack Knight. 1998. *The Choices Justices Make*. Oxford: Oxford University Press.

Fisher, Ronald A. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Friedman, Barry. 2005. 'The Politics of Judicial Review'. *Texas Law Review* 84(2):257–337.
Galanter, Marc. 1974. 'Why the "Haves" Come Out Ahead: Speculations on the Limits of Legal Change'. *Law and Society Review* 9(1):95–160.
Gastwirth, Joseph L., Abba M. Krieger, and Paul R. Rosenbaum. 1998. 'Dual and Simultaneous Sensitivity Analysis for Matched Pairs'. *Biometrika* 85(4):907–20.
Gastwirth, Joseph L., Abba M. Krieger, and Paul R. Rosenbaum. 2000. 'Asymptotic Separability in Sensitivity Analysis'. *Journal of the Royal Statistical Society, Series B* 62(3):545–55.
George, Tracey E., and Lee Epstein. 1992. 'On the Nature of Supreme Court Decision Making'. *American Political Science Review* 86(2):323–37.
Gerber, Alan S., and Donald P. Green. 2012. *Field Experiments: Design, Analysis, and Interpretation*. New York, NY: W.W. Norton.
Gibson, James L. 1983. 'From Simplicity to Complexity: The Development of Theory in the Study of Judicial Behavior'. *Political Behavior* 5(1):7–49.
Gillman, Howard. 2001. 'What's Law Got to Do With It? Judicial Behavorialists Test the "Legal Model" of Judicial Decision Making'. *Law and Social Inquiry* 26(2):465–504.
Green, Donald P., and Alan S. Gerber. 2002. 'Reclaiming the Experimental Tradition in Political Science'. In Ira Katznelson and Helen V. Milner (eds), *Political Science: The State of the Discipline*. 805–32. New York, NY: W.W. Norton.
Guan, Weihua, Liming Liang, Michael Boehnke, and Gonalo R. Abecasis. 2009. 'Genotype-Based Matching to Correct for Population Stratification in Large-Scale Case-Control Genetic Association Studies'. *Genetic Epidemiology* 33(6):508–17.
Hainmueller, Jens, and Dominik Hangartner. 2013. 'Who Gets a Swiss Passport? A Natural Experiment in Immigrant Discrimination'. *American Political Science Review* 107(1):159–87.
Hansen, Ben B. 2004. 'Full Matching in an Observational Study of Coaching for the SAT'. *Journal of the American Statistical Association* 99(467):609–18.
Hansen, Ben B. 2007. 'Optmatch: Flexible, Optimal Matching for Observational Studies'. *R News* 7: 18–24.
Hansen, Ben B., and Jake Bowers. 2008. 'Covariate Balance in Simple, Stratified and Clustered Comparative Studies'. *Statistical Sciences* 23(2):219–36.
Hansen, Ben B., and Stephanie Olsen Klopfer. 2006. 'Optimal Full Matching and Related Designs Via Network Flows'. *Journal of Computational and Graphical Statistics* 15(3):609–27.
Hansford, Thomas G., and James F. Spriggs II. 2008. *The Politics of Precedent on the U.S. Supreme Court*. Princeton, NJ: Princeton University Press.
Haviland, Amelia, Daniel S. Nagin, Paul R. Rosenbaum, and Richard E. Tremblay. 2008. 'Combining Group-Based Trajectory Modeling and Propensity Score Matching for Causal Inferences in Nonexperimental Longitudinal Data'. *Developmental Psychology* 44(2):422–35.
Hitchcock, Christopher. 2007. 'Prevention, Preemption, and the Principle of Sufficient Reason'. *Philosophical Review* 116(4):494–532.
Hosman, Carrie A., Ben B. Hansen, and Paul W. Holland. 2010. 'The Sensitivity of Linear Regression Coefficients' Confidence Limits to the Omission of a Confounder'. *The Annals of Applied Statistics* 4(2):849–70.
Howard, Robert M., and Jeffrey A. Segal. 2002. 'An Original Look at Originalism'. *Law and Society Review* 36(1):113–38.
Ignagni, Joseph A. 1994. 'Explaining and Predict Supreme Court Decision Making: The Burger Court's Establishment Clause Decisions'. *Journal of Church and State* 36(2):301–27.
Imai, Kouske, Luke Keele, Dustin Tingley, and Teppei Yamamoto. 2011. 'Unpacking the Black Box of Causality: Learning About Causal Mechanisms from Experimental and Observational Studies'. *American Political Science Review* 105(4):766–88.
Imbens, Guido W. 2003. 'Sensitivity to Exogeneity Assumptions in Program Evaluation'. *American Economic Review* 2(93):126–32.
Johnson, Timothy R. 2001. 'Information, Oral Arguments and Supreme Court Decision Making'. *American Politics Research* 29(4):331–51.

Johnson, Timothy R. 2004. *Oral Arguments and Decision Making on the U.S. Supreme Court*. Albany, NY: State University of New York Press.

Johnson, Timothy R., Paul J. Wahlbeck, and James F. Spriggs. 2006. 'The Influence of Oral Arguments on the U.S. Supreme Court'. *American Political Science Review* 100(1):99–113.

Kantor, David. n.d. 'Mahascores'. Accessed 23 January 2013.

Keele, Luke, and William Minozzi. 2013. 'How Much is Minnesota Like Wisconsin? Assumptions and Counterfactuals in Causal Inference with Observational Data'. *Political Analysis* 21(2):193–216.

Koenker, Roger, and Pin Ng. 2012. 'SparseM: Sparse Linear Algebra'. R Package Version 0.96.

Lax, Jeffrey R., and Kelly T. Rader. 2010. 'Legal Constraints on Supreme Court Decision Making: Do Jurisprudential Regimes Exist?'. *Journal of Politics* 72(1):273–84.

Lehmann, Erich Leo. 1975. *Nonparametrics: Statistical Methods Based on Ranks*. Oakland, CA: Holden Day.

Lempert, Daniel. 2015. 'Simultaneous Sensitivity Analysis in Stata: Arsimsens and Pairsimsens'. *Observational Studies* 1(1):74–90.

Liu, Weiwei, Janet Kuramoto, and Elizabeth A. Stuart. 2013. 'An Introduction to Sensitivity Analysis for Unobserved Confounding in Nonexperimental Prevention Research'. *Prevention Science* 14(6): 570–580.

Long, J. Scott, and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*, 3rd ed., College Station, TX: Stata Press.

Maltzman, Forrest, James F. Spriggs, and Paul J. Wahlbeck. 2000. *Crafting Law on the Supreme Court: The Collegial Game*. Cambridge: Cambridge University Press.

Martin, Andrew D., and Kevin M. Quinn. 2002. 'Dynamic Ideal Point Estimation Via Markov Chain Monte Carlo for the US Supreme Court, 1953–1999'. *Political Analysis* 10(2):134–53.

McAtee, Andrea, and Kevin T. McGuire. 2007. 'Lawyers, Justices, and Issue Salience: When and How do Legal Arguments Affect the U.S. Supreme Court'. *Law and Society Review* 41(2): 259–78.

McGuire, Kevin T. 1990. 'Obscenity, Libertarian Values, and Decision Making in the Supreme Court'. *American Politics Research* 18(1):47–67.

McGuire, Kevin T. 1995. 'Repeat Players in the Supreme Court: The Role of Experienced Lawyers in Litigation Success'. *The Journal of Politics* 57(1):187–96.

Morgan, Stephen L., and Christopher Winship. 2007. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Richards, Mark J., and Herbert M. Kritzer. 2002. 'Jurisprudential Regimes in Supreme Court Decision Making'. *American Political Science Review* 96(2):305–20.

Rohde, David, and Harold J. Spaeth. 1976. *Supreme Court Decision Making*. New York, NY: W.H. Freeman.

Rosenbaum, Paul R. 1989. 'Sensitivity Analysis for Matched Observational Studies with Many Ordered Treatments'. *Scandinavian Journal of Statistics* 16(3):227–36.

Rosenbaum, Paul R. 1991. 'A Characterization of Optimal Designs for Observational Studies'. *Journal of the Royal Statistical Society* 53(3):597–610.

Rosenbaum, Paul R. 2002. *Observational Studies*, 2nd ed., New York, NY: Springer.

Rosenbaum, Paul R. 2010. *Design of Observational Studies*. New York, NY: Springer.

Segal, Jeffrey A. 1984. 'Predicting Supreme Court Cases Probablistically: The Search and Seizure Cases, 1962-1981'. *American Political Science Review* 78(4):891–900.

Segal, Jeffrey A., and Harold J. Spaeth. 1993. *The Supreme Court and the Attitudinal Model*. New York, NY: Cambridge University Press.

Segal, Jeffrey A., and Harold J. Spaeth. 2002. *The Supreme Court and the Attitudinal Model Revisited*. New York, NY: Cambridge University Press.

Sehkon, Jasjeet S. 2009. 'Opiates for the matches: Matching Methods for Causal Inference'. *Annual Review of Political Science* 12:487–508.

Sekhon, Jasjeet, and Rocio Titiunik. 2012. 'When Natural Experiments are Neither Natural Nor Experiments'. *American Political Science Review* 106(1):35–57.

Sen, Maya. 2014. 'How Judicial Qualification Ratings May Disadvantage Minority and Female Candidates'. *Journal of Law and Courts* 2(1):33–65.

Sides, John, and Jeffrey R. Lax. 2012. 'More on the Persuasiveness of Oral Arguments'. The Monkey Cage, March 27. Available at http://themonkeycage.org/2012/03/27/more-on-the-persuasiveness-of-oral-arguments/, accessed 1 December 2015.

Small, Dylan, Jing Cheng, M. Elizabeth Halloran, and Paul R. Rosenbaum. 2013. 'Case Definition and Design Sensitivity'. *Journal of the American Statistical Association* 108(504):1457–468.

Small, Dylan, Joseph L. Gastwirth, Abba M. Krieger, and Paul R. Rosenbaum. 2009. 'Simultaneous Sensitivity Analysis for Observational Studies Using Full Matching or Matching with Multiple Controls'. *Statistics and its Interface* 2(2):203–11.

Spaeth, Harold J. 2007. 'The Judicial Research Initiative—United States Supreme Court Judicial Databases'. Available at http://artsandsciences.sc.edu/poli/juri/sct.htm, accessed 1 December 2015.

Spaeth, Harold J., and Jeffrey A. Segal. 1999. *Majority Rule or Minority Will: Adherence to Precedent on the U.S. Supreme Court*. Cambridge: Cambridge University Press.

Spriggs, James F. II, and Paul J. Wahlbeck. 1997. 'Amicus Curiae and the Role of Information at the Supreme Court'. *Political Research Quarterly* 50(2):365–86.

Stone, Harlan F. 1936. 'The Common Law in the United States'. *Harvard Law Review* 50(1):4–26.

Stuart, Elizabeth A., and Kerry M. Green. 2008. 'Using Full Matching to Estimate Causal Effects in Nonexperimental Studies: Examining the Relationship Between Adolescent Marijuana Use and Adult Outcomes'. *Developmental Psychology* 44(2):395–406.

Tanenhaus, Joseph, Marvin Schick, Matthew Muraskin, and Daniel Rosen. 1963. 'The Supreme Court's Certiorari Jurisdiction: Cue Theory'. In Glendon Schubert (ed.), *Judicial Decision Making*. 111–32. New York, NY: Free Press.

Tomz, Michael, Jason Wittenberg, and Gary King. 2003. 'CLARIFY: Software for Interpreting and Presenting Statistical Results. Version 2.1'. Available at http://gking.harvard.edu/clarify, accessed 1 December 2015.

Wang, Liangsheng, and Abba M. Krieger. 2006. 'Causal Conclusions are Most Sensitive to Unobserved Binary Covariates'. *Statistics in Medicine* 25(13):2257–271.

Wedeking, Justin. 2010. 'Supreme Court Litigants and Strategic Framing'. *American Journal of Political Science* 54(3):617–31.

APPENDIX

The maximum probability of obtaining the observed test statistic, under the null, given confounding as specified by the model for simultaneous sensitivity analysis (A Model for Simultaneous Sensitivity Analysis section), is formally given as follows. Let m_i be the number of treated units in set i, and $\mathbf{m} = [m_1, \ldots, m_I]^T$. For any vector $\mathbf{w}$, define Orb($\mathbf{w}$) as the set containing every vector that can be obtained by permuting the coordinates of $\mathbf{w}$. Let $\vec{\mathbf{r}}_{Ci}$ be the vector of $\mathbf{r}_{Ci}$'s coordinates arranged in increasing order. In set i, for $k = 0, \ldots, n_i$, define $\tilde{\mathbf{u}}_k$ as the vector with k zeros followed by $n_i - k$ ones. Small et al. (2009, 205–6), relying in part on Gastwirth, Krieger and Rosenbaum (2000), shows

$$\mu_{ik} = \sum_{\mathbf{z}_i \in \text{Orb}(\mathbf{Z}_i)} \sum_{\mathbf{r}_i \in \text{Orb}(\vec{\mathbf{r}}_{Ci})} \mathbf{z}_i^T \mathbf{q}_i(\mathbf{r}, \mathbf{m}) \frac{\exp(\gamma \tilde{\mathbf{u}}_k^T \mathbf{z}_i)}{\sum_{\mathbf{b}_i \in \text{Orb}(\mathbf{Z}_i)} \exp(\gamma \tilde{\mathbf{u}}_k^T \mathbf{b}_i)} \frac{\exp(\delta \tilde{\mathbf{u}}_k^T \mathbf{r}_i)}{\sum_{\mathbf{w}_i \in \text{Orb}(\vec{\mathbf{r}}_{Ci})} \exp(\delta \tilde{\mathbf{u}}_k^T \mathbf{w}_i)}$$

and

$$\sigma_{ik}^2 = \sum_{\mathbf{z}_i \in \text{Orb}(\mathbf{Z}_i)} \sum_{\mathbf{r}_i \in \text{Orb}(\vec{\mathbf{r}}_{Ci})} \left\{ \mathbf{z}_i^T \mathbf{q}_i(\mathbf{r}, \mathbf{m}) - \mu_{ik} \right\}^2 \frac{\exp(\gamma \tilde{\mathbf{u}}_k^T \mathbf{z}_i)}{\sum_{\mathbf{b}_i \in \text{Orb}(\mathbf{Z}_i)} \exp(\gamma \tilde{\mathbf{u}}_k^T \mathbf{b}_i)} \frac{\exp(\delta \tilde{\mathbf{u}}_k^T \mathbf{r}_i)}{\sum_{\mathbf{w}_i \in \text{Orb}(\vec{\mathbf{r}}_{Ci})} \exp(\delta \tilde{\mathbf{u}}_k^T \mathbf{w}_i)},$$

with $r = [\vec{\mathbf{r}}_{C1}, \ldots, \vec{\mathbf{r}}_{C,i-1}, \mathbf{r}_i, \vec{\mathbf{r}}_{C,i+1}, \ldots, \vec{\mathbf{r}}_{CI}]^T$. Call

$$\mu_{imax} = \max_{k \in \{0,1,\ldots,n_i\}} \mu_{ik}$$

$$A_i = \{k : \mu_{ik} = \mu_{imax}\}$$

$$\sigma^2_{imax} = \max_{k \in A_i} \sigma^2_{ik}.$$

Then, the maximum probability of obtaining test statistic $T \geq s$, under the null, given confounding, is approximated by:

$$\Pr(T \geq s \,|\, \vec{\mathbf{r}}_C, \mathbf{m}, \mathbf{X}, \mathbf{u}) = 1 - \Phi\left(\frac{s - \sum_{i=1}^{I} \mu_{imax}}{\sqrt{\sum_{i=1}^{I} \sigma^2_{imax}}}\right), \tag{3}$$

for any number s.