

*The Significance of Differences Interval: Assessing the Statistical or Substantive Difference between Two Quantities of Interest** †

MARIUS RADEAN‡
University of Essex

ABSTRACT

The standard practice in discussing results has shifted from coefficients to substantive quantities of interest. Hypothesis testing nowadays entails computing substantive effects with the 95% CI for alternative scenarios and comparing them. Absent an evaluation of the difference in estimates, current practice often cannot provide a definitive answer. When CIs overlap, the estimates *may or may not* be statistically different. This ambiguity invites mistakes, as analysts turn to ill-advised conjectures to infer whether estimates are distinct. My literature survey indicates this is a widespread problem, with more than half of the articles not providing the evidence required to assess significance of differences. One practical solution is to report instead significance of differences intervals (SDIs), which can be used for direct comparisons. I expand the SDI method to accommodate unpaired sample data, asymmetric distributions, and, for substantive significance, differences larger than zero. I also provide an easy-to-use software to compute SDIs.

*I thank Andreas Beger, Daina Chiba, Kristian Skrede Gleditsch, Daniel W. Hill, Jr., Lucas Leemann, Sara M. T. Polo, Jonathan Slapin, and participants at the workshop on “Best Practices in Academic Research” at University of Essex for their helpful comments and suggestions.

†This work was supported by the British Academy [SRG1819\191241].

‡Lecturer, Department of Government, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom, mradean@essex.ac.uk

1 Introduction

The standard practice in discussing research results has gradually shifted from coefficients to substantive quantities of interest (King, Tomz and Wittenberg, 2000). The norm today is to compute substantive effects for competing scenarios and compare them. To convey the uncertainty around specific estimates, researchers typically report the standard (usually 95%) confidence interval (CI). When comparing two quantities of interest, though, the use of the standard CI can be misleading and may lead to incorrect inferences. The problem is that when two CIs overlap, the associated estimates *may or may not* be statistically different. Simply put, on its own, the CI does not provide the information necessary to ascertain whether the compared estimates are, in fact, distinct. This is a most consequential limitation, as researchers often examine a given quantity of interest in relation to other estimates, not in isolation. Indeed, hypothesis testing typically entails assessing whether two quantities of interest are statistically different. Whether we compare two sample means (e.g., the treated and untreated experiment groups), or results from either simple (e.g., the probability of y in the presence and absence of a given factor) or more complex models (e.g., the conditional effect of x at different values of the interacting variable z), judging significance of differences lies at the heart of empirical analysis. Upon showing that the current practice is not fit for purpose and should be revised, I introduce a comprehensive solution to the overlap problem.

One common mistake associated with the CI overlap is to conclude that the compared estimates are statistically indistinguishable. This approach is more conservative than standard tests, and, as a result, the estimates may still be statistically different. When this is the case, researchers fail to reject the null hypothesis (type II error). One may argue that this is mainly a theoretical rather than a practical problem, since analysts are familiar with and know how to interpret the standard CI. This assumption is overly optimistic. In the health sciences, a nonexhaustive search identified more than 60 articles, in 22 different peer-reviewed journals, that either use or recommend the overlap method to examine significance of differences (Schenker and Gentleman, 2001,

182). Similarly, in a survey of researchers who had published in journals in psychology, behavioral neuroscience, and medicine, Belia et al. (2005) find that over 30% misinterpret overlapping confidence intervals. Even if researchers were aware of this problem, it is unlikely that the larger audience (e.g., policy makers, nontechnical readers) is equally well informed. Practically, the “use of a single visualization with overlapping and nonoverlapping confidence intervals leads many to draw such [wrong] conclusions, despite the best efforts of statisticians toward preventing users from reaching such conclusions” (Wright, Klein and Wieczorek, 2019, 165).

More sophisticated researchers, who know better than to equate overlapping CIs with statistical insignificance, use at times conjectures or rules of thumb to infer whether the compared estimates are statistically distinct. This can lead to wrongly accepting the research hypothesis (type I error). The two most oft-used heuristics to judge significance of differences are (i) whether the coefficient on the factor whose effect we evaluate is statistically significant, and (ii) whether only one (not both) of the estimates is significant. Both reasonings are problematic. In non linear models, whether a substantive effect is significant cannot be inferred reliably from the significance status of the associated coefficient (Greene, 2009).¹ Similarly, just because one of the compared estimates is indistinguishable from zero and the other one is not, does not mean they are necessarily different *from each other* (Gelman and Stern, 2006; Gill, 1999).

Given the recent shift to substantive effects in the discussion of empirical results, the problem of the CI overlap, or a lack thereof, will likely amplify. As this study makes clear, when assessing substantive differences (i.e., differences larger than zero), the standard CI is never informative – not even on the off chance that there is no overlap. Practically, two estimates may be substantively indistinguishable even if the associated 95% CIs do not overlap. This is a counterintuitive finding that may take by surprise even seasoned researchers, who are used to associate a lack of CI overlap

¹ Specifically, for any variable k there is no “guarantee that both the estimated coefficient, θ_k , and the associated partial effect, δ_k [e.g., $\frac{\partial \Pr(y)}{\partial k}$], will both be statistically significant, or statistically insignificant” (Greene, 2009, 487).

with statistical significance. I am not aware of any work that alerts analysts not to infer substantive differences from *nonoverlapping* CIs.

There are two general solutions to circumvent the CI overlap problem.² One solution is to assess the difference in estimates (DE). If the first or second difference is statistically significant (insignificant), the original estimates are distinct (indistinguishable).³ The other solution is to report significance of differences intervals (SDIs)—a different type of uncertainty interval that, unlike the standard CI, can be used for direct comparisons (Afshartous and Preston, 2010; Goldstein and Healy, 1995; Schenker and Gentleman, 2001; Tryon, 2001). While the CI captures the uncertainty around a single estimate, the SDI reflects the uncertainty around both compared estimates as well as the dependence between them (i.e., whether they are independent, or either positively or negatively correlated). In effect, SDIs are *relational* intervals in the sense that they convey *between* estimates information. SDIs are designed such that when they do not overlap the compared estimates are distinct, even if the standard 95% CIs overlap.

The SDI approach is increasingly popular among political scientists, with many recent studies adopting this technique (e.g., Adams, Ezrow and Wlezien, 2016; Arceneaux et al., 2016; Chiba, Johnson and Leeds, 2015; Fulton and Dhima, 2021; Johns and Davies, 2019; Karpowitz, Monson and Preece, 2017; Komisarchik, Sen and Velez, Forthcoming; Radean, 2019). However, they all employ a generic solution that makes several simplifying assumptions. Specifically, the SDI level employed in the respective analyses presupposes the compared estimates are normally distributed, independent, and have identical standard errors. As these are very restrictive and unrealistic assumptions, these studies may have drawn incorrect inferences from the data. I provide a compre-

² This is the case within frequentist framework, as Bayesian methods provide other solutions.

³ I use the term first difference to indicate the difference between two expected or predicted values, e.g., $\Pr(y|(x + 1)) - \Pr(y|x)$. The second difference captures the difference between two first differences, such as the conditional effect of x in the presence and absence of the moderating variable z , e.g., $[\Pr(y|(x + 1), z = 1) - \Pr(y|x, z = 1)] - [\Pr(y|(x + 1), z = 0) - \Pr(y|x, z = 0)]$.

hensive solution that relaxes these assumptions, allowing the SDI to be used in all settings.

For a broader assessment of the practice in the discipline, I examine all 2016 articles published in *American Journal of Political Science* (AJPS) and *The Journal of Politics* (JOP). The survey reveals that the overlap problem is widespread. In 55% of the studies, the estimates' CIs overlap and no additional clarifying information is provided (e.g., tests examining significance of differences). As a result, in a majority of cases we cannot tell whether the empirics support the research hypothesis or supplementary analysis. Taken together with the similar results from other fields (e.g., Belia et al., 2005; Schenker and Gentleman, 2001), this finding suggests that the overlap problem is not a discipline specific issue, but one that affects social sciences generally.

The starkest finding of the survey is that political scientists do not make good use of existing techniques to examine the difference in estimates. Specifically, most studies only report the compared estimates with the 95% CI, without conducting significance of differences tests. This may create more confusion than clarity, as oft-used heuristics (e.g., overlapping CIs, whether the associated coefficient is significant, or only one of the estimates is different from zero) cannot and should not replace standard tests. In sum, the current practice is not fit for purpose and should be revised. While I compare the DE and SDI methods and show the latter's advantage, ultimately either approach can be used to judge if two point estimates are distinct. Choosing one over the other may be a matter of taste, but employing one of the two ought not to be optional.

This study makes several contributions. First, I derive a universal formula that can assess significance of differences from any given value, not just zero. Focusing on differences from zero retains the flaws associated with the null hypothesis significance testing, in the sense that it devotes excessive attention to statistical significance at the expense of substantive or practical significance. Second, I expand the SDI method to accommodate comparisons between samples with unequal number of observations (e.g., treated and untreated experimental groups, different survey waves), and between skewed distributions (e.g., the household income in the U.S., probability of civil war onset). More specifically, I introduce an original technique to compute *empirical* SDIs, which are

derived numerically based on the estimates' actual values or percentiles. To date, technical studies have considered only formula-based solutions to the overlap problem (i.e., standard error-based SDIs). Analytical solutions, however, are not applicable when researchers (i) employ percentile intervals, or (ii) compare unpaired and unequal samples. Third, I provide an easy-to-use Stata software that automatically computes SDIs.⁴

2 Assessing significance of differences

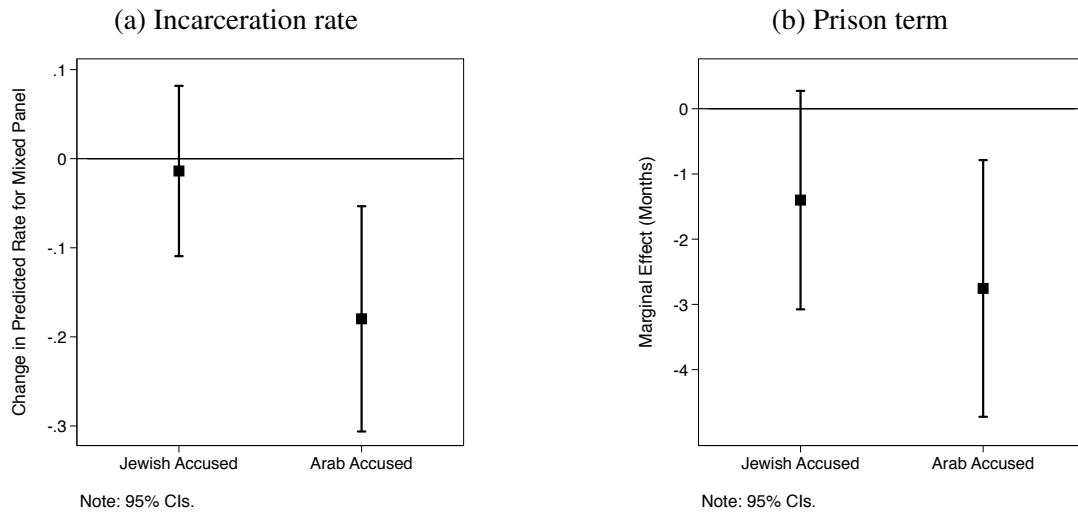
Theoretically, the limitation of overlapping CIs is straightforward: absent additional significance of differences tests, we cannot reach any conclusion. Put differently, we can neither accept nor reject the research hypothesis. Practically, however, this ambiguity leads many practitioners to incorrectly conclude that the estimates are either statistically similar or distinct. By directly comparing two estimates with overlapping CIs, some researchers deduce that the two are statistically indistinguishable. Other analysts indirectly compare such estimates by assessing whether they are both distinct from a third value, typically zero. This usually happens when the researcher compares two marginal effects and concludes they are distinct if only one (not both) is different from zero (i.e., statistically significant). Neither of these conclusions is warranted.

To illustrate the problem, let us consider two analyses from Grossman et al. (2016) about the effect of ethnicity-based court composition on judicial outcomes in Israel. Figure 1a and 1b reproduce their incarceration rate and prison term results, conditional on the ethnicity of the accused.⁵ Summarizing the findings (p. 44), the authors conclude that

⁴ Afshartous and Preston (2010) also provide a computer code, but with limited capabilities. Specifically, their program cannot accommodate samples with unequal number of observations, or asymmetric distributions. Moreover, it can assess only differences from zero and no other value, rendering it impractical for substantive significance. My software has none of these limitations.

⁵ These are the results from Figure 2 Model 4 (p. 59), which is the model specification with the full set of covariates and the best AIC model fit (Grossman et al., 2016, 55).

Figure 1: The effect of courts' ethnic composition on judicial outcomes



Note: Figure 1a and 1b reproduce the results from Figure 2 Model 4, in Grossman et al. (2016, 59). They illustrate the impact of mixed court panels on incarceration rate and prison term, respectively, conditional on the accused's ethnicity.

[...] appeal outcomes for Jewish defendants are independent of panels' ethnic composition. By contrast, panel composition is highly consequential for Arab defendants, who receive more lenient punishments when their case is heard by a panel that includes at least one Arab judge, compared to all-Jewish panels. The magnitude of these effects is sizable: a 14–20% reduction in incarceration and a 15–26% reduction in prison sentencing.

Per the authors' assessment, both incarceration rate and prison sentencing analyses support the theoretical expectation that the effect of mixed court panels is conditional on the defendant's ethnicity. The overlap between the 95% CIs around the estimated effects, however, should give us pause in agreeing with the study's conclusion. As I show below, in one of the analyses the effect of judicial diversity is not statistically different between the two ethnic groups. Despite the fact that only the estimate for Arab defendant is significant (but not for Jewish defendant), the substantive effects are, in fact, *not* distinct. This practically means that the effect of mixed panels on judicial outcomes may actually be homogenous across ethnic cleavages, not heterogeneous as the theory implies. Since the two analyses are seemingly equivalent, which one supports the theory, and which one does not? From the information provided one cannot tell.

This example concerns first differences (i.e., differences between predicted values), but if we

were to consider studies with overlapping CIs comparing different quantities of interest (such as estimated coefficients (e.g., Lowande and Augustine Potter, 2021), expected values (e.g., Kitchens, 2021), observed means or percent rates of different groups (e.g., Jackman, 2014), etc., the problem would be the same. Regardless of how the estimates are computed or measured, overlapping CIs render informed readers unable to determine whether the empirics supports the hypothesis, and may mislead nontechnical readers (Wright, Klein and Wieczorek, 2019). My literature survey reveals this is a widespread problem, with most practitioners reporting the compared estimates with the 95% CI instead of the difference (for more details see *A survey of the literature* section).

2.1 The significance of differences interval

Do we face an impossible situation? On the one hand, we are encouraged to discuss substantive effects rather than coefficients (King, Tomz and Wittenberg, 2000). Since “the point estimate is simply a best *guess*,” we also need to acknowledge the uncertainty around estimated effects via CIs (McCaskey and Rainey, 2015, 89, emphasis added). On the other hand, the use of the standard CI around the estimates to be compared can be misleading and may lead to incorrect inferences.

To circumvent the CI overlap problem, one can employ either the DE or the SDI method. In this section I first review the lesser-known SDI approach, and then compare the two alternative solutions. Next I introduce an original technique to compute empirical SDIs, thereby extending the SDI method to asymmetric distributions and unpaired sample data. Lastly, I present a new procedure to compute SDIs that can convey substantive significance.

Standard error-based SDIs

Standard error-based intervals are typically used when the (asymptotic) distribution of a given quantity of interest is normal. Alternative solutions for normal distributions have already been derived (see Afshartous and Preston, 2010; Schenker and Gentleman, 2001; Tryon and Lewis, 2009), but I revisit this case for several reasons. The first reason is to provide the intuition behind the overlap problem. Under the normality assumption, the mean and standard error of the difference

distribution can be expressed in terms of the corresponding elements of the two original distributions (see Eq. (1)). As a result, it is easier to trace the root of the problem. Second, the current understanding of the issue offers a baseline against which to evaluate my new technique to assess nonzero differences, and the technique to compute empirical SDIs. Third, I formally derive the possible range of the standard error-based SDIs, which is a novel piece of information.

Let us say we have two quantities of interest, Q_1 and Q_2 , and want to know whether they are statistically distinct. One solution is to check whether the standard CI of the difference, $(Q_1 - Q_2)$, includes zero. To simplify the notation by not having to consider both CI bounds, let us assume that $Q_1 > Q_2$. In this case Q_1 and Q_2 are statistically different if

$$(Q_1 - Q_2) - z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} > 0$$

$$(Q_1 - Q_2) > z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2}, \quad (1)$$

where SE_1 and SE_2 are the standard errors, z is the standard score, and ρ is the correlation level. This is the formula used by standard tests to check for significance of differences. By contrast, with the overlap method, the two quantities of interest are deemed statistically different if the lower CI bound of the higher statistic does not overlap with the upper CI bound of the lower statistic

$$Q_1 - zSE_1 > Q_2 + zSE_2$$

$$(Q_1 - Q_2) > z(SE_1 + SE_2). \quad (2)$$

The difference between standard tests and the overlap method is reflected in the unequal width of the respective $(Q_1 - Q_2)$ intervals. Since the left-hand sides in Eq. (1) and (2) are equal, but the right-hand sides are not, the two methods are bound to produce different results. To have the intervals around the compared estimates correctly indicate significance of differences, the difference between Q_1 's lower interval bound and the upper Q_2 bound must match the value of the lower CI bound of $(Q_1 - Q_2)$. To ensure results equivalence, for Q_1 and Q_2 we need to use a significance of differences z_d -score that satisfies the following equality

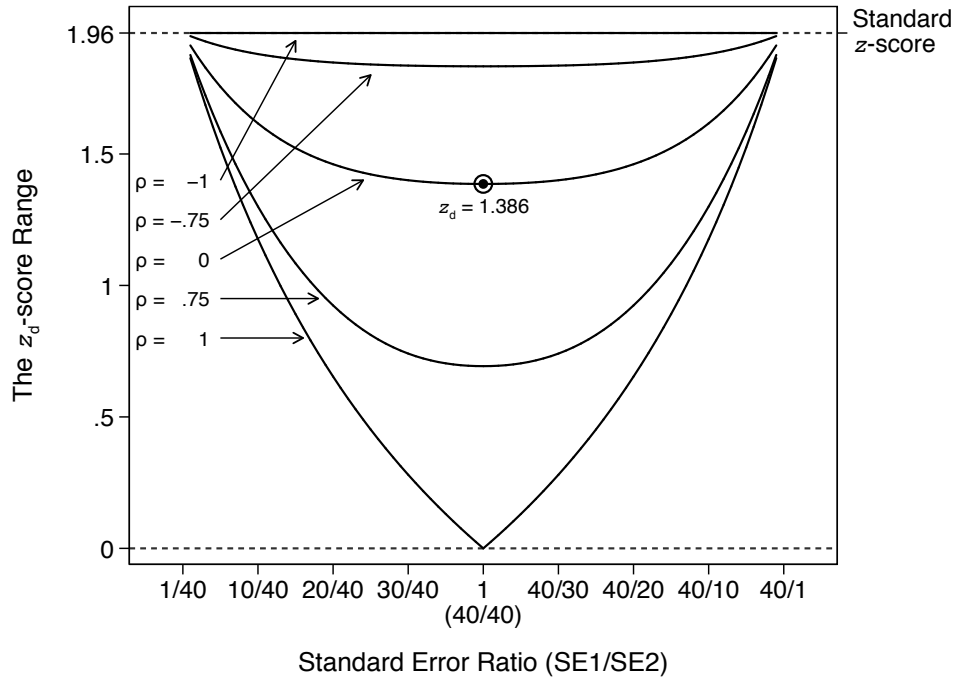
$$\begin{aligned}
(Q_1 - z_d SE_1) - (Q_2 + z_d SE_2) &= (Q_1 - Q_2) - z \sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} \\
(Q_1 - Q_2) - z_d(SE_1 + SE_2) &= (Q_1 - Q_2) - z \sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} \\
-z_d(SE_1 + SE_2) &= -z \sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} \\
z_d &= \frac{z \sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2}}{SE_1 + SE_2}. \tag{3}
\end{aligned}$$

Thus, the z_d -score is a function of four distinct parameters, and its value cannot be determined from the standard z -score alone. Figure 2 illustrates the non-monotonic relation between the z_d - and z -score. Specifically, it shows the z_d -score required to indicate significance of differences at the 0.05 level, for different combinations between five correlation levels ($\rho(Q_1, Q_2) = \{-1, -0.75, 0, 0.75, 1\}$), and standard error ratios, $\frac{SE_1}{SE_2}$, ranging from 1/40 to 40/1.⁶ As a reference point, the dashed line at the very top of the plot indicates the value of the standard z -score associated with the 0.05 significance level: 1.96.

The flat solid line at the top of the graph indicates that, irrespective of the standard error ratio, when $\rho = -1$, z_d equals the standard z -score. This is z_d 's highest possible value. Conversely, z_d reaches its minimum value, 0, when $\rho = 1$ and the two quantities of interest have identical standard errors, $\frac{SE_1}{SE_2} = 1$. As the value of the standard error ratio moves away from 1, the z_d level increases and approaches the standard z -score value. The dot in the middle of the graph identifies the z_d -score associated with the 83.5% interval, the generic level employed in empirical research (e.g., Adams, Ezrow and Wlezien, 2016; Arceneaux et al., 2016; Chiba, Johnson and Leeds, 2015; Fulton and Dhima, 2021; Johns and Davies, 2019; Karpowitz, Monson and Preece, 2017; Komisarchik, Sen and Velez, Forthcoming; Radean, 2019). This interval level corresponds to a scenario where Q_1 and Q_2 are independent, $\rho = 0$, and have equal standard errors, $\frac{SE_1}{SE_2} = 1$. The isolated value clearly illustrates the limited applicability of the solution adopted by many practitioners.

⁶ To graph this in a two dimensional space we need to reduce the number of parameters. In Online Appendix A.1 I show that the z_d -score can also be written as a function of the $\frac{SE_1}{SE_2}$ ratio.

Figure 2: The z_d -score required to indicate significance of differences at the 0.05 level



Note: The z_d -score required to indicate significance of differences at the 0.05 level, for different combinations between five correlation levels, $\rho(Q_1, Q_2) = \{-1, -0.75, 0, 0.75, 1\}$, and standard error ratios, $\frac{SE_1}{SE_2}$, ranging from 1/40 to 40/1.

More generally, the possible range for the z_d -score is $[0, z]$.⁷ When $z_d = 0$ no SDIs are required to indicate significance of differences; Q_1 and Q_2 are statistically distinct provided that the point estimates are different, $Q_1 \neq Q_2$. When $z_d = z$, the SDI has the same width as the standard CI. This is the only case when the overlap method and standard significance of differences tests provide the same answer.

All values within the z_d range are possible in practice since quantities of interest can be either positively or negatively correlated. Positive correlations are more likely when there is an overlap between the observations used to compute the compared estimates, but alternative scenarios can lead to “either positive or negative correlations, even where there is no overlap in the sets of elements used for the two estimates” (Schenker and Gentleman, 2001, 185).

⁷ In Online Appendix A I present the formal proof by optimizing the z_d function, Eq. (3).

The formal definition of the SDI

The above discussion focused primarily on the SDI's practical ability to indicate significance of differences. Next I formally define this type of uncertainty interval, along the lines of the respective concept for an equivalent CI. The SDIs are relational intervals in the sense that they convey *between* estimates information, namely, whether the point estimates are statistically different. In this sense, the SDI conveys information about the difference in estimates parameter, $Q_{diff} = (Q_1 - Q_2)$. Since the SDI level is computed such that the CI bounds of the difference can be expressed in terms of the SDI bounds of the compared estimates, we have

$$P(L_{Q_{diff}[1, \dots, \infty]} < Q_{diff} < U_{Q_{diff}[1, \dots, \infty]}) = 1 - \alpha$$

$$P((L_{Q_1[1, \dots, \infty]}^{SDI} - U_{Q_2[1, \dots, \infty]}^{SDI}) < Q_{diff} < (U_{Q_1[1, \dots, \infty]}^{SDI} - L_{Q_2[1, \dots, \infty]}^{SDI})) = 1 - \alpha, \quad (4)$$

where $L_{Q_{diff}}$ and $U_{Q_{diff}}$ indicate the lower and upper CI limits of Q_{diff} . $[1, \dots, \infty]$ indicates that the $(1 - \alpha)\%$ CI is calculated repeatedly in an (infinitely) long sequence of valid applications. $L_{Q_*}^{SDI}$ and $U_{Q_*}^{SDI}$ represent the lower and upper SDI limits of the respective quantities of interest. Lastly, P is the probability of Q_{diff} falling within the respective bounds, and α is the significance level.

Given the probability in Eq. (4), we can interpret the SDI as follows. Under (infinitely) many repeated applications, $(1 - \alpha)\%$ of the $[(L_{Q_1}^{SDI} - U_{Q_2}^{SDI}), (U_{Q_1}^{SDI} - L_{Q_2}^{SDI})]$ intervals will contain, on average, the true value of the difference in estimates Q_{diff} .

2.2 Conveying statistical significance via SDIs

Before looking at practical applications, it is important to consider the performance of the SDI method. To check whether SDIs perform in practice as the theory implies, I conduct a series of simulations detailed in Online Appendix B. The Monte Carlo simulations vary the sample size and the difference in means between the compared estimates. As expected, precision increases as either the number of observations or the difference in estimates grows larger. For normal samples with 1,000 cases and over, the accuracy is extremely high (100%) as the SDI method is able to pick up even on small differences. When comparing distributions (e.g., the distribution of postestimation

quantities of interest such as predicted values or marginal effects), accuracy is all but guaranteed since sample size is not an issue and results are directly derived from Eq. (3).

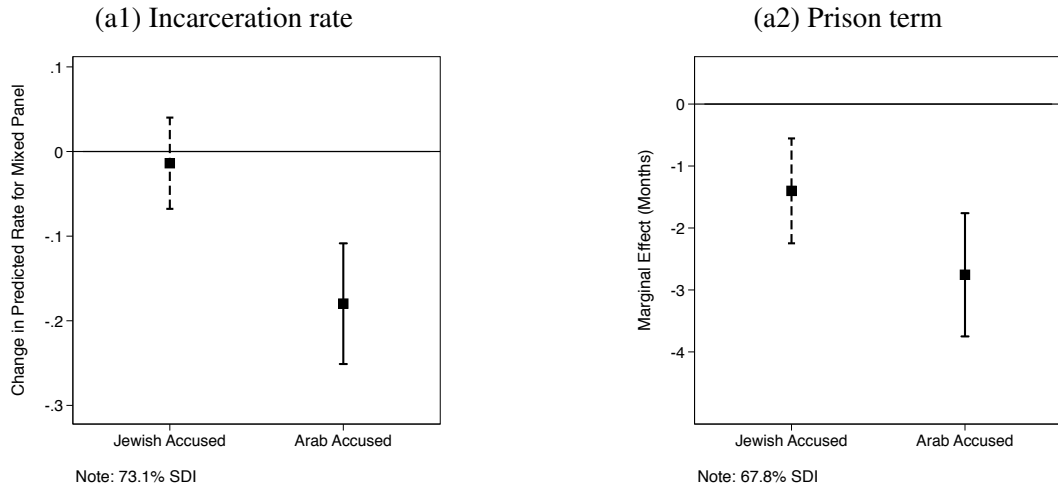
To illustrate how SDIs can be used to judge significance of differences, I revisit the example from Grossman et al. (2016) about the effect of judicial diversity in multiethnic societies. Figure 3a1 and 3a2 show the effect of mixed panels with SDIs. Since the required SDI level is case specific, the two analyses have different interval levels, that is, 73.1% and 67.8%. Given the SDIs' properties, significance of differences can be inferred directly from their overlap, or the lack thereof. In Figure 3a1 the associated SDIs do not overlap, so the estimates are statistically different at the 0.05 level. By contrast, the SDIs overlap in Figure 3b2. This means that mixed court panels are likely to impose similar prison sentences regardless of the ethnicity of the accused. This finding does not square with the claim that judicial diversity has a positive effect in multiethnic societies.

By using different line patterns, SDI graphs can also outline whether individual estimates are statistically significant. For example, in Figure 3a I use solid lines for the SDI if a particular effect is significant at the 0.05 level (e.g., the estimate when the defendant is Arab), and dashed lines otherwise (e.g., the estimate for Jewish defendant). Thus, the SDI *level* indicates whether the estimates are statistically different from each other, whereas the SDI *pattern* indicates whether they are different from zero. However, one ought not to report this additional piece of information by default. To the contrary, in order to be effective graphs should display the minimum amount of information required to get the point across. Knowing whether the individual estimates are statistically significant is neither necessary nor sufficient to assess the hypothesis.

Thus, unlike the compared estimates with the 95% CI results (Figure 1), the SDI results suggest the two analyses are not equivalent. To illustrate how one would calculate “by hand” the SDI that led to the different conclusion, I use the actual results from the prison term analysis to demonstrate the process. The first step is to calculate the point estimate and standard error of the effect of mixed court panels, for both Arab and Jewish defendants. Following Grossman et al., I use the `margins` command for this task. The respective quantities of interest are $Q_J = -1.40107$

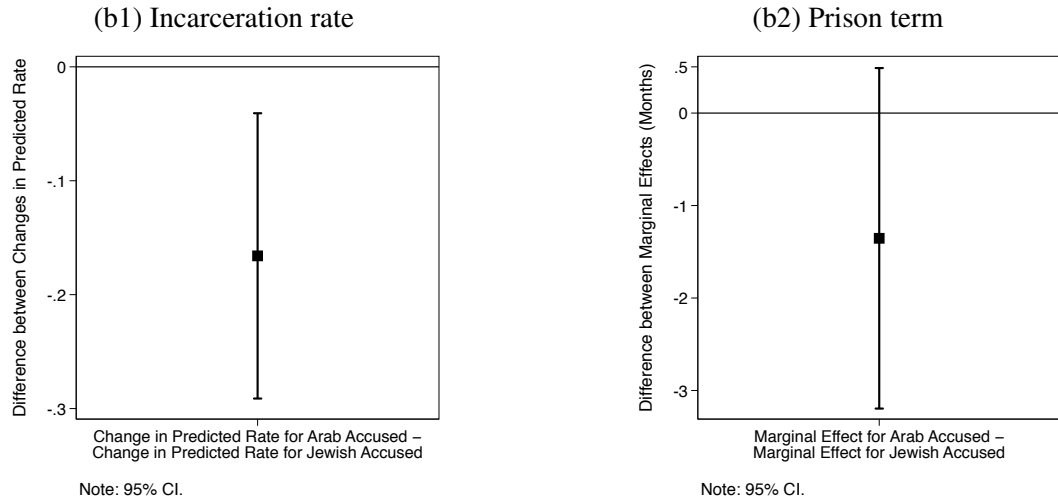
Figure 3: The effect of courts' ethnic composition on judicial outcomes

(a) The SDI method

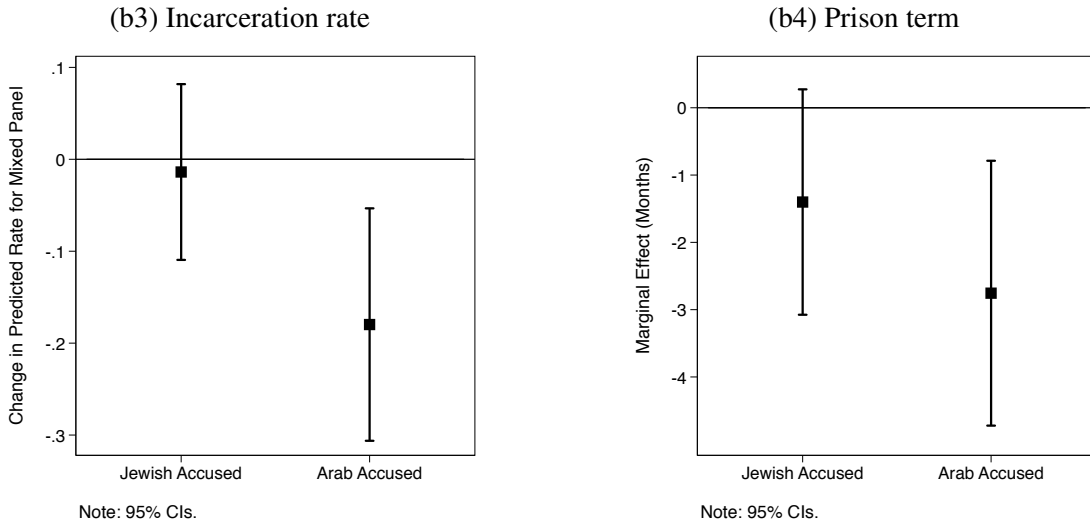


(b) The DE approach

(b1–b2) The difference in estimates with the standard CI



(b3–b4) The compared estimates with the standard CI



and $SE_J = 0.854486$, and $Q_A = -2.755442$ and $SE_A = 1.004028$, for a Jewish and Arab accused. Using the formula $Q_* \pm zSE_*$, I next compute the standard CI. For a 95% interval z equals 1.96, and the respective CIs are $[-3.08, 0.27]$ for Q_J , and $[-4.72, -0.79]$ for Q_A (see Figure 1b). Since Q_J 's CI contains 0 this estimate is not statistically significant, whereas Q_A is. But this does not mean that Q_J and Q_A are statistically different.

To directly compare Q_J and Q_A we require SDIs, for which we need to calculate the adjusted z_d -score. Using the formula from Eq. (3) we have

$$\begin{aligned} z_d &= \frac{z\sqrt{SE_J^2 + SE_A^2 - 2\rho SE_J SE_A}}{SE_J + SE_A} \\ &= \frac{1.96\sqrt{0.854486^2 + 1.004028^2 - 2 \times 0.498764 \times 0.854486 \times 1.004028}}{0.854486 + 1.004028} \\ &= 0.990663, \end{aligned} \tag{5}$$

where $\rho(Q_J, Q_A) = 0.498764$, a statistic most softwares would calculate. The respective z_d -score corresponds to a 67.8% SDI level (i.e., $(1 - 2(1 - \Phi(z_d))) \times 100 = 67.8$, where Φ is the cumulative standard normal distribution). Using the familiar $Q_* \pm z_d SE_*$ formula, we can now compute the significance of differences intervals. The respective SDIs are $[-2.27, -0.55]$ for Q_J , and $[-3.75, -1.76]$ for Q_A (see Figure 3a2). Since the associated SDIs overlap, we conclude that mixed court panels are likely to impose similar prison sentences regardless of the ethnicity of the accused.

2.3 The SDI vs. the DE approach

Technically, the SDI and DE approach provide the same answer on significance of differences. Parsimony and susceptibility to misinterpretation are two common criteria to discriminate between techniques that yield the same result, with the SDI method having a comparative advantage.

Specifically, the SDI method is more informative than the bare-bone DE variant (where just the difference is reported), and more parsimonious than the comprehensive DE approach (where both the estimates and their difference are reported). This is because the SDI method encodes three pieces of information into a single set of results: (1) the sign and (2) size of the compared estimates

(expressed on the scale of the variable of interest), as well as (3) whether the estimates are statistically distinct. The SDI results also invite fewer mistakes since the interpretation of this uncertainty interval is consistent and not context-specific (i.e., when SDIs do not overlap the compared estimates are distinct, and when they do overlap the estimates are statistically indistinguishable).

Figure 3b1 and 3b2 illustrate the bare-bones DE alternative, and outline the (second) difference with the standard CI for the incarceration rate and prison term analyses. As expected, the DE results convey the same information as their SDI counterparts. In Figure 3b1, the 95% CI of the difference does not cross the zero line, and therefore the associated predicted probabilities are statistically different. In contrast, the CI of the difference contains zero in Figure 3b2. Substantively, this means that Jewish and Arab defendants receive statistically similar prison sentences.

The bare-bones DE variant can be underwhelming. By itself, the difference does not reveal the values used to create it, which is necessary to assess the rate of change. A difference of 1 can represent either a 100% or a 1% increase in y depending on whether the start value is 1 or 100. Also, placing the difference in estimates at the forefront of the empirical analysis weakens the link between theory and empirics. Our theories and hypotheses are generally framed around the estimates to be compared (e.g., the probability of y is higher (lower) in the presence (absence) of x), and typically do not engage with the difference (e.g., whether it has a substantive meaning). The difference is often just an empirical tool for comparison, one measured on a different metric “with a different mean, standard deviation, and standard error” (Tryon and Lewis, 2009, 182). This is why, for many researchers, the theoretical quantity of interest are the estimates not the difference.⁸

⁸ Conditional hypotheses, for instance, typically take the following form: “The marginal effect of X on Y is positive at all values of Z; this effect is strongest when Z is at its lowest and declines in magnitude as Z increases” (Berry, Golder and Milton, 2012, 659). Notably, the typical formulation does not explicitly acknowledge that, to be conditional on Z, the effect of X at low and high Z values must be statistically different. This explains in part why virtually all authors report interaction effect graphs that outline the effect of X with the 95% CI at various levels of Z. But only a minority

To convey all this information, the researcher needs to additionally report the compared estimates. For the prison term analysis, instead of the reporting solely the SDI results in Figure 3a2, one would need to graph Figure 3b2 *and* 3b4. Since it entails two sets of results, the comprehensive DE variant is less parsimonious than the SDI method. Having two sets of empirical evidence creates redundancy (with the analyst needing to clarify what each set of empirical evidence contributes to answering the research question), and may lead to confusion. This is particularly the case when overlapping CIs send mixed signals to less technical readers. For example, the CIs of the compared marginal effects overlap in Figure 3b4. Ideally, readers should know that this piece of information does not elucidate whether the two effects are distinct, and would sift through the other results for a definitive answer. In practice, however, the CI overlap “leads many to draw such [wrong] conclusions, despite the best efforts of statisticians” (Wright, Klein and Wieczorek, 2019, 165). The SDI has the comparative advantage of reporting a single set of results, where even overlapping intervals have an unambiguous interpretation (i.e., the estimates are not statistically distinct). As a result, it is less likely to lead readers astray.

One may argue that the DE approach has the advantage of employing the standard CI, an interval type analysts are already familiar with. Yet surveys from multiple disciplines reveal that the CI’s widespread use is not necessarily correlated with accurate understanding or correct usage. For instance, Greenland et al. (2016, 337) note that “[m]isinterpretation and abuse of statistical tests, confidence intervals [...] remain rampant” even among professionals. The problem is certainly more acute among researchers with a nontechnical background or practitioners. Another issue is that the interpretation of the CI for the difference and compared estimates may vary. While the CI of the first difference can be used to assess whether the effect is statistically significant, the CI around the original estimates (e.g., means of different groups, expected or predicted values), cannot serve this function. Case in point, the CI of predicted probabilities cannot contain negative

directly assess whether the effect of X varies with Z, by examining the difference in the effect of X *between* specific values of Z.

values. This adds to the challenge of a consistent and correct interpretation of the standard CI.

Having a specialized interval designed to judge significance of differences ought to signal that employing other types of intervals (e.g., the CI) for this task is inappropriate. This could also encourage researchers to engage more meaningfully with the standard CI. Practitioners rarely discuss or interpret the CI⁹ – except to assess statistical significance. But this is not the main feature of the CI, and is not even generally applicable. Put differently, acknowledging the information CIs cannot convey, may spur practitioners to discuss what the reader should infer from the reported CI.

Theoretical considerations aside, a practical issue with the DE method is that researchers have to calculate the second difference by hand, which is taxing and error-prone. The second difference is necessary, for instance, to compare conditional effects (e.g., the effect of x in the presence and absence of z). Yet many easy-to-use softwares do not calculate this quantity of interest (e.g., Clarify (King, Tomz and Wittenberg, 2000), SPost (Long and Freese, 2014)).

3 Empirical SDIs

To my knowledge, existing technical studies have considered only analytical, formula-based solutions to the overlap problem (Afshartous and Preston, 2010; Goldstein and Healy, 1995; Schenker and Gentleman, 2001; Tryon, 2001). However, numerical solutions are required for specific types of analyses. This is the case, for instance, when researchers (i) employ percentile- rather than standard error-based intervals, or (ii) compare unpaired and unequal samples. In what follows I introduce an original technique to calculate empirical SDIs for scenarios where formula solutions

⁹ The technical definition of the CI is fairly abstract and rarely invoked. Strictly speaking, the CI contains information that can be used to infer the frequency with which a very large number of CIs include the unknown population parameter. Specifically, if one calculates the 95% CI repeatedly in a long sequence of valid applications, 95% of them will contain, on average, the true effect size. Any single CI, however, either contains the true effect or not, and the probability of that being the case is either one or zero (Greenland et al., 2016, 344).

are not appropriate. These numerical procedures are canned in my easy-to-use software.

3.1 Percentile-based SDIs

For various reasons, researchers sometimes report percentile- rather than standard error-based CIs. This is particularly the case when dealing with skewed distributions where, instead of being a center value, the mean can be (very) close to one of the end points of the distribution's range (see Online Appendix C for a practical example). Because standard error intervals are symmetric around the point estimate, the CI may contain unrealistic extrapolations on one side of the mean. Specifically, the bounds can “include values that exceed the range of the statistic being estimated (e.g., a bound for a predicted probability could be negative or greater than one)” (Xu and Long, 2005, 539). Percentile intervals are more flexible than their symmetric standard error counterparts, and cannot exceed the range of the statistic. That said, it is beyond the scope of this study to compare the merits of percentile- and standard error-based intervals. I simply note that the former are fairly common in practice, and are the default choice of several softwares. As examples, *Clarify* reports only percentile CIs and *SPost* automatically reports bootstrapped percentile CIs.

When employing percentile CIs, one cannot use the z_d -score formula to calculate the required SDI level. Importantly, this is the case regardless of whether the compared distributions are normal or skewed. There are two main reasons for this. First, if there are theoretical reasons to calculate percentile CIs for the compared distributions, these reasons likely apply to the difference distribution as well. Yet, when using the z_d formula, we assume a standard error-based CI for the difference. Second and more importantly, there is no guarantee that the difference between the lower and upper percentile bounds of Q_1 and Q_2 , respectively, is not greater than the lower standard error CI bound of Q_{diff} . When the percentile difference is larger, one may wrongly conclude that the compared estimates are distinct (type I error).¹⁰

¹⁰ Take the simple case of two independent normal distributions with equal standard errors, for which the required SDI level is 83.5%. For this interval level, the associated z_d -score is 1.386,

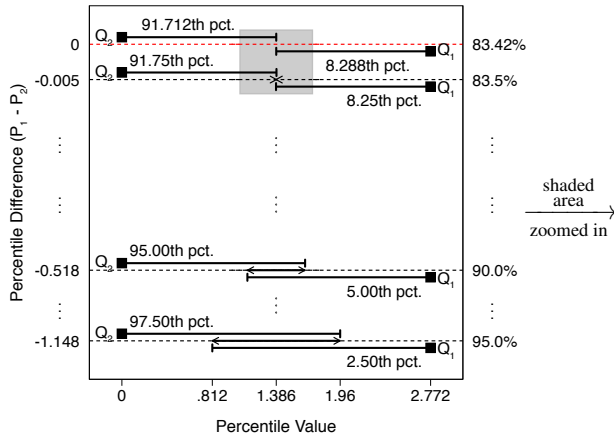
While using a standard error-based technique to determine the interval level for percentile bounds is far from ideal, currently there is no alternative solution. Indeed, all three applications that employ narrower *percentile* intervals to indicate significance of differences use a formula-derived interval level (see Adams, Ezrow and Wlezien, 2016; Arceneaux et al., 2016; Chiba, Johnson and Leeds, 2015). Next I present a new technique to compute SDIs based solely on the percentiles of the estimates' sampling distributions.

The obstacle in deriving a purely percentile-based solution is that there is no one-to-one correspondence between the percentiles of two random distributions, Q_1 and Q_2 , and the percentiles of the difference distribution, Q_{diff} . Consequently, when using the distribution percentiles to calculate SDIs we cannot employ a predetermined formula. Alternatively, we can reverse-engineer the required percentile level for Q_1 and Q_2 by obtaining the Q_{diff} distribution first. We then calculate its 2.5th percentile, which is Q_{diff} 's lower 95% CI bound. Next, we need to empirically identify the pair of Q_1 and Q_2 percentiles whose difference matches that value. We achieve this by calculating the lower interval bound for Q_1 (i.e., the k^{th} percentile value of Q_1) and the upper bound for Q_2 (i.e., the $(100-k)^{\text{th}}$ percentile), for candidate values of k . As we increase k from the theoretical minimum value of 2.5, we will eventually find a value k such that the difference between the k^{th} percentile value of Q_1 and the $(100-k)^{\text{th}}$ percentile value of Q_2 matches the value of Q_{diff} 's 2.5th percentile value.

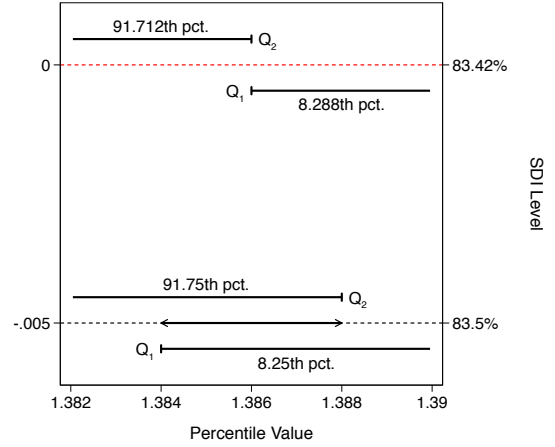
 and the upper Q_1 and lower Q_2 interval bounds are, as a percentage, 8.25 and 91.75 (i.e., $(100 - 83.5)/2$). While $(Q_1 - 1.386 \times SE_1) - (Q_2 + 1.386 \times SE_2) = Q_{diff} - 1.96 \times SE_{Q_{diff}}$ by definition, the difference between the respective percentiles does not necessarily satisfy even the weaker condition $(8.25^{\text{th}} \text{ pct.}_{Q_1} - 91.75^{\text{th}} \text{ pct.}_{Q_2}) \leq Q_{diff} - 1.96 \times SE_{Q_{diff}}$. In the simulations I ran for this relatively liberal test with normal distributions, the false positive rate was 15% (i.e., the percent of cases where the difference in percentiles was larger). The issue is that we try to match observed values (i.e., Q_1 's and Q_2 's percentiles) with a computed value (i.e., the analytically derived CI bound of Q_{diff}), without there being a direct link. A foolproof percentile-based solution requires using observed values for the difference distribution as well.

Figure 4: An illustration of the percentile-based method

(a) Evaluating candidate (P_1, P_2) percentile pairs



(b) Shaded area from panel a) zoomed in



Note: Finding the z_d -score for Q_1 and Q_2 by evaluating candidate (P_1, P_2) percentile pairs. The goal is to find the pair whose difference matches the value of the 2.5th percentile of the difference distribution ($Q_1 - Q_2$).

percentile. This k gives us the required SDI level as $(100 - 2k)\%$.

Let us consider a practical example. For ease of comparison to the standard error-based results, let Q_1 and Q_2 follow a particular bivariate normal distribution: $(Q_1, Q_2) \sim \mathcal{N}((1.96 \times \sqrt{2}), 0, 1, 1; 0)$. Given these specific values, Q_{diff} 's 2.5th percentile is exactly zero,¹¹ and the required SDI level is approximately 83.5%. Figure 4a plots the mean of Q_1 and Q_2 , which are indicated by solid square marks, as well as the lower interval bound of the higher statistic, Q_1 , and the upper bound of Q_2 (the solid horizontal lines). The percentile difference, $P_1 - P_2$, is indicated on the left-hand side of the y -axis. The SDI level associated with a given (P_1, P_2) pair is displayed on the right-hand side.

For example, at the bottom of the plot are graphed Q_1 's 2.5th and Q_2 's 97.5th percentiles ($k = 2.5$), which corresponds to a 95% interval. The percentile difference associated with the 95%

¹¹ $2.5^{\text{th}} \text{pct.}_{Q_{diff}} = Q_{diff} - zSE(Q_{diff}) = [(1.96 \times \sqrt{2}) - 0] - 1.96\sqrt{1^2 + 1^2 - 2 \times 0 \times 1 \times 1} = 0$. $(Q_1, Q_2) \sim \mathcal{N}(\mu_{Q_1}, \mu_{Q_2}, \sigma_{Q_1}, \sigma_{Q_2}; \rho(Q_1, Q_2))$ indicates that Q_1 and Q_2 follow a bivariate normal distribution with the respective means, μ , standard errors, σ , and correlation level, ρ .

interval, -1.148 , is really off from the 0 target. Since the 95% interval is too wide, we need to calculate the $(P_1 - P_2)$ difference for smaller and smaller intervals until we obtain a percentile difference very close to 0. As the interval level decreases, so does the difference between the corresponding percentiles. Around the 83% level, the associated P_1 and P_2 values are so close that we need to zoom in to see whether the percentiles overlap (Figure 4b). The percentile difference is 0 at the 83.42% SDI level (i.e., $k = 8.288$). However, multiple digit precision intervals are seldom necessary in practice. The one digit precision SDI level associated with a percentile difference no larger than 0, is the 83.5% interval.

As this is only a schematic illustration, in Online Appendix C I present the procedure more systematically using an example with non normal distributions. I also discuss how to increase the level of precision when multiple digit SDIs are required. The procedure to compute percentile SDIs is analogous to the one concerning t distributions, which I introduce in the next section.

3.2 The case of t distributions for unpaired and unequal samples

In many political science applications scholars make inferences from (sub)samples of the population. When the sample size is relatively small or the standard deviation of the population is unknown, the t distribution is advised rather than the normal distribution. With appropriate substitutions, the z_d formula in Eq. 3 can be used to calculate the analogous significance of differences t_d -score as long as we have matched or paired observations (e.g., patient data over time where the researcher has multiple readings on the same subjects). However, we need a different procedure to compare unpaired samples with unequal number of observations (e.g., different survey waves, treated and untreated experimental groups). The problem is that in the aforementioned formula there is a single unknown, that is, the t_d -score. When the compared samples have different degrees of freedom, the t_{d1} - and t_{d2} -score are distinct unknowns for the same SDI level. As a result, there is no analytical solution to the equation

$$(Q_1 - t_{d1}SE_1) - (Q_2 + t_{d2}SE_2) = Q_{diff} - t_{diff}SE_{diff}$$

$$\begin{aligned}
(Q_1 - Q_2) - (t_{d1}SE_1 + t_{d2}SE_2) &= Q_{diff} - t_{diff}SE_{diff} \\
t_{d1}SE_1 + t_{d2}SE_2 &= t_{diff}SE_{diff} \\
T\left(df_1, \frac{1}{2}\alpha_d\right)SE_1 + T\left(df_2, \frac{1}{2}\alpha_d\right)SE_2 &= T\left(df_{diff}, \frac{1}{2}\alpha\right)SE_{diff}, \tag{6}
\end{aligned}$$

where $T(df, \frac{1}{2}\alpha)$ is the function for the survivor Student's t distribution, df is the degrees of freedom, and α is the significance level.

Analogous to the procedure to compute percentile SDIs, we can solve numerically for the required SDI level for Q_1 and Q_2 , by obtaining first the lower 95% CI bound of Q_{diff} , $Q_{diff} - T\left(df_{diff}, \frac{1}{2} \times 0.05\right)SE_{diff}$. Next, we need to empirically identify the pair of Q_1 and Q_2 interval bounds whose difference matches that value. We achieve this by calculating the lower bound for Q_1 (i.e., the k^{th} interval) and the upper bound for Q_2 (i.e., the $(100-k)^{\text{th}}$ interval), for candidate values of k . k is the significance level in percentage points, $\frac{1}{2}\alpha_d \times 100$, and α_d is the significance level for the SDI. By iteratively increasing k , we will eventually find the required k value.

For a concrete example, let us say we have two samples, $Q_1 \sim (60, 10, 2)$ and $Q_2 \sim (40, 5, 4)$. The numbers in parentheses are the sample's number of observations, mean, and standard deviation, respectively. Q_{diff} is distributed as Student's t with $n_1 + n_2 - 2$ degrees of freedom, mean $\bar{Q}_{diff} = \bar{Q}_1 - \bar{Q}_2$, and standard error $SE_{diff} = \left\{ \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} \right\}^{1/2} \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}$ (Gosset [Student, pseud.], 1908). n_* and s_* represent the original samples' number of observations and standard deviations, respectively. The formulas for the standard error of the difference and its degrees of freedom presuppose that the underlying populations represented by the two samples have equal variances. This modeling decision is meant to keep the example relatively simple. However, my easy-to-use software can compute SDIs for applications with unequal standard deviations, in which case the researcher has the option to calculate the approximate degrees of freedom using either the Satterthwaite's (1946) or Welch's (1947) formula.

Table 1 outlines, step-by-step, the procedure to compute standard error-based SDIs for samples with different degrees of freedom. Recall that the procedure requires to identify the pair of

Table 1: Computing standard error-based SDIs for unpaired and unequal samples

| Q_{diff} | | Q_1 | | Q_2 | | Bound Difference | SDI Level |
|--------------|---|-------------|-----------------|-------------|---------------------|------------------|---------------|
| Bound Value | Percent ($\frac{\alpha}{2} \times 100$) | Bound Value | Percent (k) | Bound Value | Percent ($100-k$) | $B_1 - B_2$ | $(100-2k)$ |
| 2.949 | 0.05% | 9.106 | 0.05% | 7.250 | 99.95% | 1.856 | 99.9 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 3.800 | 2.50% | 9.483 | 2.50% | 6.279 | 97.50% | 3.204 | 95.0 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 4.191 | 9.20% | 9.653 | 9.20% | 5.855 | 90.80% | 3.798 | 81.6 % |
| 4.193 | 9.25% | 9.654 | 9.25% | 5.853 | 90.75% | 3.801 | 81.5 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 4.999 | 49.95% | 10.000 | 49.95% | 5.001 | 50.05% | 4.999 | 0.1% |

Note: $Q_1 \sim (60, 10, 2)$ and $Q_2 \sim (40, 5, 4)$ are two samples with the respective number of observations, mean, and standard deviation. The first six columns show, in increments of 0.05 percentage points, the interval bounds and corresponding percentages (i.e., $k = \frac{1}{2}\alpha_d \times 100$) for Q_{diff} , Q_1 , and Q_2 . The last two columns indicate the bound difference, $B_1 - B_2$, and the SDI level associated with a given (B_1, B_2) pair, respectively.

$k_{Q_1}^{th}$ and $(100-k)_{Q_2}^{th}$ interval bounds whose difference matches the value of the difference distribution's lower CI bound. The first two columns show Q_{diff} 's lower bound values and corresponding percentages, ranging from 0.05% to 49.95%, in increments of 0.05 percentage points.¹² The increment step used here is suited to compute one digit precision SDIs. We can increase the level of precision by employing smaller increments. Since we are interested in significance of differences at the 0.05 level, the Q_{diff} 's value of interest is the one associated with the 2.5 percent, 3.800. This value is highlighted in the table. The following four columns present the same information for Q_1 and Q_2 . The only difference is that Q_2 's interval bounds are sorted in descending order.

The last step is to calculate the difference for all (B_1, B_2) interval bound pairs, and then

¹² To keep things concise, only a limited number of values are shown in the table. The upper interval value, $k = 49.95\%$, is a theoretical limit. Since the SDI level equals $(100-2k)\%$, a higher number would lead to either zero or negative values. In practice this means that for any $k \geq 50$, Q_1 and Q_2 are statistically different as long as the point estimates are distinct, $Q_1 \neq Q_2$.

search for the one with a value matching Q_{diff} 's lower 95% CI bound. The last two columns capture the bound difference and the associated SDI level. For example, on the second row the 3.204 difference corresponds to the $(2.5^{th}_{Q_1}, 97.5^{th}_{Q_2})$ bound pair, as indicated by the percentages in the fourth and sixth columns. Given that in this case $k = 2.5$, the associated SDI level is $(100 - 2k) = 95\%$. The difference that is the closest to but not larger than 3.800, is the one associated with the 81.6% SDI. The respective value, 3.798, and SDI level are also highlighted in the table.

4 The SDI for substantive significance

Not to make big claims about small effects that are inconsequential in practice, researchers are encouraged to check whether the reported estimates are also substantively meaningful, not just statistically significant. To test for real-world significance one should first determine a range of substantively insignificant effects, $[-m, m]$, where $|m|$ is a theoretically derived value for the smallest meaningful effect. An estimate is substantively meaningful if its entire confidence interval lies outside of the insignificant effects region (Gross, 2015; Rainey, 2014). This estimation-based CI test is related to equivalence tests or tests of design, which are used to check whether the data are consistent with the identification assumptions or theory (Hartman and Hidalgo, 2018; Lakens, 2017). While there are some differences when it comes to statistical power in small sample, these tests yield similar results in most cases and “are effectively indistinguishable with reasonable sample sizes” (Hartman and Hidalgo, 2018, 1004). This means there are no differences when comparing postestimation quantities of interest (e.g., predicted values, marginal effects), as sample size is not an issue when dealing with distributions.

The SDI procedure for meaningful differences draws on the estimation-based CI approach. When judging substantive significance, the two quantities of interest are meaningfully different if the Q_{diff} 's lower CI bound is greater than m , or $(Q_{diff} - zSE_{Q_{diff}}) - m > 0$. To have the intervals around the compared estimates correctly indicate *substantive* differences, we need to use a significance of differences score for Q_1 and Q_2 that accounts for the meaningful value m .

Specifically, $z_{d|m}$ must satisfy the following equality

$$\begin{aligned} (Q_1 - z_{d|m}SE_1) - (Q_2 + z_{d|m}SE_2) &= \left[(Q_1 - Q_2) - z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} \right] - m \\ z_{d|m} &= \frac{z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} + m}{SE_1 + SE_2}. \end{aligned} \quad (7)$$

Practically, the z_d -score discussed in the literature is a special case of the $z_{d|m}$ -score, where $m = 0$. The range of the $z_{d|m}$ -score is $\left[\frac{m}{SE_1 + SE_2}, \left(z + \frac{m}{SE_1 + SE_2} \right) \right]$.¹³ Notably, compared to the statistical significance case where $z_d \leq z$, $z_{d|m}$ can be larger than z . As a result, the SDI may be *wider* than the standard CI. This is consequential because it means that when outlining substantive significance, confidence intervals are never informative, not even on the off chance that they do not overlap. Specifically, two estimates may be substantively insignificant even if their CIs do not overlap, as wider SDIs might. I am not aware of any work that alerts analysts not to infer substantive differences from *nonoverlapping* CIs.¹⁴

5 A survey of the literature

Some may argue that the overlap problem is a well known issue, of which people are both aware of and know how to address. While the limitations of standard CIs are well understood in the methodology literature, the practice at large does not reflect this knowledge. For a more systematic assessment of the current practice in the discipline, I examine all 2016 AJPS and JOP articles.¹⁵

¹³ In Online Appendix D I present the formal proof for this.

¹⁴ In Online Appendix D.1 I present a practical example on employing substantive SDIs.

¹⁵ An article was included in the survey if either the hypothesis entails a comparison (e.g., interaction effects), or it engages in estimates comparison in the empirical section. The latter captures studies that directly compare expected or predicted values, as well as articles that report the first or second difference. Articles that focus solely on coefficient interpretation and do not compute post-estimation quantities of interest are not included. This is because in this scenario the CI overlap problem cannot arise. Political theory and game theoretical studies without an empirical analysis

Table 2: A survey of the 2016 AJPS & JOP Articles

| | <i>Standard CI</i> | | | <i>Narrower Intervals</i> | <i>No Intervals</i> | Total |
|--|--------------------|---------------------|------------------|---------------------------|---------------------|-----------|
| | Overlap Method | No Additional Tests | Additional Tests | | | |
| All Articles | 8 (8%) | 49 (51%) | 27 (28%) | 2 (2%) | 11 (11%) | 97 (100%) |
| Cannot Ascertain Significance of Differences | 6 (6%) | 37 (38%) | 0 (0%) | 2 (2%) | 8 (8%) | 53 (55%) |

Note: The numbers indicate the frequency for each category, and, in parentheses, the overall percentage.

Table 2 provides a summary of the literature survey. In the first column, the top value indicates that out of a total of 97 articles, eight (8%) employ the overlap method to judge significance of differences. In two of these analyses the 95% CIs do not overlap, and, therefore, the conclusion that the estimates are statistically distinct is technically correct. The remaining six articles, though, reference the CI overlap as evidence that the corresponding estimates are not statistically different. This conclusion is not warranted.

The second category represents the bulk of the articles, roughly 51%, which do not invoke the CI overlap—or a lack therefore—as evidence. The problem is that they do not reference any other test results either to back the conclusion. The aforementioned Grossman et al. (2016) study falls into this category. Practically, these studies simply point out that one estimate is nominally higher than another estimate, or that only one of the two is statistically significant. Crucially, they do not present evidence to clarify whether or not the compared estimates are distinct. If the analyst did conduct significance of differences tests but did not mention them, the problem is that the reader cannot independently assess the strength of undisclosed evidence. If no such tests were conducted, the conclusion rests on ill-advised conjectures and the inferences may be inaccurate.

In the third column are the studies that compute and report the difference in estimates, about _____ are also excluded. Lastly, the AJPS Workshop articles are not included because they focus on methodological issues rather than theory building and hypothesis testing.

28%. It turns out that a popular approach is to present two sets of findings: the compared estimates with the 95% CI, as well as the associated first or second difference. Among these articles, Rueda and Stegmueller's (2016, 483) study stands out as being the only one that explicitly advises readers that it is not "correct to infer the significance of the difference from our (non-)overlapping confidence intervals." For this information, they refer readers to a separate in-text table. By reporting the difference in estimates, many more articles implicitly acknowledge that one cannot evaluate the hypothesis by examining only the compared estimates and their CIs. Having two sets of empirical evidence, however, creates redundancy and may lead to confusion.

Two articles, acknowledged in the fourth column, employ narrower intervals to convey significance of differences: Adams, Ezrow and Wlezien (2016), and Arceneaux et al. (2016). Specifically, they both report 83.5% SDIs to indicate significance of difference at the 0.05 level. This interval level, however, requires that the compared estimates are normally distributed, independent, and have identical standard errors. Yet neither article mentions whether all three conditions are simultaneously met to warrant the use of this particular level.

The fifth column captures the articles that report only the point estimates. The decision not to report the CI at all, may be indicative of a growing realization that this interval type conveys tangential information that often cannot be used to answer the research question. While a few of these studies use the text to clarify whether the compared estimates are statistically different, most of them simply call attention to the fact that one estimate is nominally higher than another estimate.

The starkest finding of the survey is that political scientists do not make good use of existing techniques to convey significance of differences. The numbers in the second row of Table 2 indicate the frequency and the overall percentage of articles for which one cannot determine whether the compared quantities of interest are statistically distinct. An article was included in this category if, in at least one of the analyses, the CIs overlap and there is no additional clarifying information.¹⁶

¹⁶ I employ a liberal approach to coding these articles. Besides the studies that expressly mention having had conducted formal significance of differences tests, the articles that describe estimates

Among the five categories, only the articles that examine the difference in estimates are foolproof. Overall, informed readers cannot ascertain whether there is empirical support for the research hypothesis or supplementary analysis in 55% of the cases. More problematically, these studies can mislead nontechnical readers (Wright, Klein and Wieczorek, 2019). Thus, while the CI overlap may seem like an unlikely problem for methodologists, it is a widespread problem in practice.

6 Conclusion

Researchers automatically report the standard CI around point estimates because we are taught to always acknowledge the uncertainty around estimated effects. Yet, different types of intervals (e.g., confidence intervals, prediction intervals, tolerance intervals) convey different information and they are *not* interchangeable. Consequently, researchers must ensure they report the appropriate interval type for their analysis. Linking the reported uncertainty back to the research question is crucial because hypotheses are rarely expressed in terms of the level of uncertainty (e.g., the uncertainty around \hat{y} is larger at lower values of x , but it becomes smaller as x increases). When comparing estimates, the use of the standard CI can be misleading and may lead to incorrect inferences. For direct comparisons, significance of differences intervals are advised. SDIs are designed such that when they do not overlap the estimates are distinct, even if the 95% CIs overlap.

Confidence intervals are so ubiquitous in empirical research that we tacitly assume everyone knows when and how to employ them. This assumption in turn makes it easy to dismiss the CI overlap problem as a fringe issue, which it is not. A survey of the health sciences literature identifies scores of articles that misuse the CI (Schenker and Gentleman, 2001, 182), and my survey reveals that it is a widespread problem in political science as well. Specifically, *more than half* of the articles do not provide the evidence required to assess whether the compared estimates are statistically different. The large share of problematic studies aside, it is worth noting that

as either “statistically” or “significantly” different are also excluded.

they are published in a wide range of outlets. Schenker and Gentleman count 22 different peer-reviewed journals, and both journals I surveyed published such articles. Thus, absent a conscious reevaluation, the review process alone is unlikely to properly address this problem.

To conclude, the practical problem with reporting CIs around compared estimates is that when there is overlap, the estimates may or may not be statistically different. As a result, we can neither accept nor reject the research hypothesis. While nontechnical readers are particularly likely to draw wrong conclusions from overlapping CIs (Wright, Klein and Wieczorek, 2019), the lack of a definitive answer invites mistakes even among professionals. The latter often turn to ill-advised conjectures or heuristics to judge significance of differences. What is more, given the recent shift in results discussion to meaningful effects, the problem will likely amplify. Since the SDI for substantive significance can be wider than the equivalent CI, the latter is never informative when assessing meaningful differences, not even on the off chance that the associated CIs do *not* overlap. One practical solution to circumvent these problems is to employ the SDI method. I expand the SDI method to accommodate unpaired samples, asymmetric distributions, and, for substantive significance, differences larger than zero. I also provide an easy-to-use Stata software that automatically computes SDIs.

References

- Adams, James, Lawrence Ezrow and Christopher Wlezien. 2016. "The Company You Keep: How Voters Infer Party Positions on European Integration from Governing Coalition Arrangements." *American Journal of Political Science* 60(4):811–823.
- Afshartous, David and Richard A. Preston. 2010. "Confidence Intervals for Dependent Data: Equating Non-overlap with Statistical Significance." *Computational Statistics & Data Analysis* 54:2296–2305.
- Arceneaux, Kevin, Martin Johnson, René Lindstädt and Ryan J. Vander Wielen. 2016. "The Influence of News Media on Political Elites: Investigating Strategic Responsiveness in Congress." *American Journal of Political Science* 60(1):5–29.
- Belia, Sarah, Fiona Fidler, Jennifer Williams and Geoff Cumming. 2005. "Researchers Misunderstand Confidence Intervals and Standard Error Bars." *Psychological Methods* 10(4):389–396. <https://doi.org/10.1177/0022343313476529>.
- Berry, William D., Matt Golder and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *Journal of Politics* 74(3):653–671.
- Chiba, Daina, Jesse C. Johnson and Brett Ashley Leeds. 2015. "Careful Commitments: Democratic States and Alliance Design." *The Journal of Politics* 77(4):968–982.
- Fulton, Sarah A. and Kostanca Dhima. 2021. "The Gendered Politics of Congressional Elections." *Political Behavior* 43:1611–1637. <https://doi.org/10.1007/s11109-020-09604-7>.
- Gelman, Andrew and Hal Stern. 2006. "The Difference Between "Significant" and "Not Significant" is not Itself Statistically Significant." *The American Statistician* 60(4):328–331. <https://doi.org/10.1198/000313006X152649>.
- Gill, Jeff. 1999. "The Insignificance of Null Hypothesis Significance Testing." *Political Research Quarterly* 52(3):647–74.
- Goldstein, Harvey and Michael J. R. Healy. 1995. "The Graphical Presentation of a Collection of Means." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158(1):175–177.
- Gosset [Student, pseud.], W. S. 1908. "The Probable Error of a Mean." *Biometrika* 6:1–25. <https://doi.org/10.2307/2331554>.
- Greene, William. 2009. Discrete Choice Modelling. In *Palgrave Handbook of Econometrics: Applied Econometrics, Volume 2*, ed. Terence C. Mills and Kerry Patterson. Hampshire, U.K.: Palgrave Macmillan pp. 473–556.

- Greenland, Sander, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman and Douglas G. Altman. 2016. “Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations.” *The European Journal of Epidemiology* 31:337–350. <https://doi.org/10.1007/s10654-016-0149-3>.
- Gross, Justin H. 2015. “Testing What Matters (If You Must Test At All): A Context-Driven Approach to Substantive and Statistical Significance.” *American Journal of Political Science* 59(3):775–788. <https://doi.org/10.1111/ajps.12149>.
- Grossman, Guy, Oren Gazal-Ayal, Samuel D. Pimentel and Jeremy M. Weinstein. 2016. “Descriptive Representation and Judicial Outcomes in Multiethnic Societies.” *American Journal of Political Science* 60(1):44–69. <https://doi.org/10.1111/ajps.12187>.
- Hartman, Erin and F. Daniel Hidalgo. 2018. “An Equivalence Approach to Balance and Placebo Tests.” *American Journal of Political Science* 62(4):1000–1013. <https://doi.org/10.1111/ajps.12387>.
- Jackman, Molly C. 2014. “Parties, Median Legislators, and Agenda Setting: How Legislative Institutions Matter.” *The Journal of Politics* 76(1):259–272. <https://doi.org/10.1017/S0022381613001291>.
- Johns, Robert and Graeme A. M. Davies. 2019. “Civilian Casualties and Public Support for Military Action: Experimental Evidence.” *Journal of Conflict Resolution* 63(1):251–281. <https://doi.org/10.1177/0022002717729733>.
- Karpowitz, Christopher F., J. Quin Monson and Jessica Robinson Preece. 2017. “How to Elect More Women: Gender and Candidate Success in a Field Experiment.” *American Journal of Political Science* 61(4):927–943. <https://doi.org/10.1111/ajps.12300>.
- King, Gary, Michael Tomz and Jason Wittenberg. 2000. “Making the Most of Statistical Analyses: Improving Interpretation and Presentation.” *American Journal of Political Science* 44(2):347–361.
- Kitchens, Karin E. 2021. “Exit or Invest: Segregation Increases Investment in Public Schools.” *The Journal of Politics* 83(1):71–86. <https://doi.org/10.1086/708916>.
- Komisarchik, Mayya, Maya Sen and Yamil Velez. Forthcoming. “The Political Consequences of Ethnically Targeted Incarceration: Evidence from Japanese-American Internment During WWII.” *The Journal of Politics*. <https://doi.org/10.1086/717262>.
- Lakens, Daniël. 2017. “Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses.” *Social Psychological and Personality Science* 8(4):355–362. <https://doi.org/10.1177/1948550617697177>.
- Long, J. Scott and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata*. 3rd ed. College Station, TX: Stata Press.

- Lowande, Kenneth and Rachel Augustine Potter. 2021. "Congressional Oversight Revisited: Politics and Procedure in Agency Rulemaking." *The Journal of Politics* 83(1):401–408. <https://doi.org/10.1086/709436>.
- McCaskey, Kelly and Carlisle Rainey. 2015. "Substantive Importance and the Veil of Statistical Significance." *Statistics, Politics, and Policy* 6(1-2):77–96.
- Radean, Marius. 2019. "Sometimes You Cannot Have It All: Party Switching and Affiliation Motivations as Substitutes." *Party Politics* 25(2):140–152. <https://doi.org/10.1177/1354068816688363>.
- Rainey, Carlisle. 2014. "Arguing for a Negligible Effect." *American Journal of Political Science* 58(4):1083–1091. <https://doi.org/10.1111/ajps.12102>.
- Rueda, David and Daniel Stegmueller. 2016. "The Externalities of Inequality: Fear of Crime and Preferences for Redistribution in Western Europe." *American Journal of Political Science* 60(2):472–489.
- Satterthwaite, F. E. 1946. "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2:110–114. <https://doi.org/10.2307/3002019>.
- Schenker, Nathaniel and Jane F. Gentleman. 2001. "On Judging the Significance of Differences by Examining the Overlap Between Confidence Intervals." *The American Statistician* 55(3):182–186.
- Tryon, Warren W. 2001. "Evaluating Statistical Difference, Equivalence, and Indeterminacy Using Inferential Confidence Intervals: An Integrated Alternative Method of Conducting Null Hypothesis Statistical Tests." *Psychological Methods* 6(4):371–386.
- Tryon, Warren W. and Charles Lewis. 2009. "Evaluating Independent Proportions for Statistical Difference, Equivalence, Indeterminacy, and Trivial Difference Using Inferential Confidence Intervals." *Journal of Educational and Behavioral Statistics* 34(2):171–189.
- Welch, B. L. 1947. "The Generalization of 'Student's' Problem when Several Different Population Variances Are Involved." *Biometrika* 34:28–35. <https://doi.org/10.2307/2332510>.
- Wright, Tommy, Martin Klein and Jerzy Wieczorek. 2019. "A Primer on Visualizations for Comparing Populations, Including the Issue of Overlapping Confidence Intervals." *The American Statistician* 73(2):165–178. <https://doi.org/10.1080/00031305.2017.1392359>.
- Xu, Jun and J. Scott Long. 2005. "Confidence Intervals for Predicted Outcomes in Regression Models for Categorical Outcomes." *The Stata Journal* 5(4):537–559. <https://doi.org/10.1177/1536867X0500500405>.

Online Appendix A: Standard error-based SDIs

In the main text, when discussing the range of the significance of differences z_d -score, I only note the minimum and maximum possible values. Here I present step-by-step the z_d function optimization procedure, which allows one to formally derive the z_d 's extreme values. Recall that the z_d -score formula is

$$z_d = \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}}{\sigma_1 + \sigma_2}, \quad (\text{A1})$$

where σ_1 and σ_2 are the standard errors of the two quantities on interest, Q_1 and Q_2 , ρ is the correlation parameter, and z is the standard score. Moreover, $\sigma_1 > 0$, $\sigma_2 > 0$, and $-1 \leq \rho \leq 1$.

To find the minimum and maximum of z_d , we set the first partial derivatives with respect to ρ , σ_1 , and σ_2 to 0, treating the standard z -score as a constant.

$$\begin{aligned} \frac{\partial}{\partial \rho} z_d &= 0 \\ \frac{\partial}{\partial \rho} \left[\frac{z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] &= 0 \\ \frac{z}{\sigma_1 + \sigma_2} \times \frac{\partial}{\partial \rho} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} &= 0 \\ \frac{\partial}{\partial \rho} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} &= 0 \\ \left[\frac{\partial}{\partial(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \right] \left[\frac{\partial}{\partial \rho} (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) \right] &= 0 \\ \left[\frac{1}{2\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}} \right] \left[\frac{\partial \sigma_1^2}{\partial \rho} + \frac{\partial \sigma_2^2}{\partial \rho} - \frac{\partial 2\rho\sigma_1\sigma_2}{\partial \rho} \right] &= 0 \\ [0 + 0 - 2\sigma_1\sigma_2] &= 0 \\ -2\sigma_1\sigma_2 &= 0 \\ \sigma_1\sigma_2 &= 0 \quad (\text{A2}) \end{aligned}$$

$$\begin{aligned}
& \frac{\partial}{\partial \sigma_1} z_d = 0 \\
& \frac{\partial}{\partial \sigma_1} \left[\frac{z \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] = 0 \\
& z \times \frac{\partial}{\partial \sigma_1} \left[\frac{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] = 0 \\
& \frac{\partial}{\partial \sigma_1} \left[\frac{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] = 0 \\
& \frac{\frac{\partial}{\partial \sigma_1} \left[\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \right] (\sigma_1 + \sigma_2) - \frac{\partial}{\partial \sigma_1} \left[(\sigma_1 + \sigma_2) \right] \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{(\sigma_1 + \sigma_2)^2} = 0 \\
& \frac{\partial}{\partial \sigma_1} \left[\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \right] (\sigma_1 + \sigma_2) - \frac{\partial}{\partial \sigma_1} \left[(\sigma_1 + \sigma_2) \right] \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} = 0 \\
& \left[\frac{\partial}{\partial (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} \right] \left[\frac{\partial}{\partial \sigma_1} (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) \right] (\sigma_1 + \sigma_2) \\
& \quad - \left[\frac{\partial \sigma_1}{\partial \sigma_1} + \frac{\partial \sigma_2}{\partial \sigma_1} \right] \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} = 0 \\
& \left[\frac{1}{2\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}} \right] \left[\frac{\partial \sigma_1^2}{\partial \sigma_1} + \frac{\partial \sigma_2^2}{\partial \sigma_1} - \frac{\partial 2\rho\sigma_1\sigma_2}{\partial \sigma_1} \right] (\sigma_1 + \sigma_2) - [1 + 0] \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} = 0 \\
& \left[\frac{1}{2\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}} (2\sigma_1 + 0 - 2\rho\sigma_2) \right] (\sigma_1 + \sigma_2) - \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} = 0 \\
& \frac{\sigma_1 - \rho\sigma_2}{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}} (\sigma_1 + \sigma_2) - \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} = 0 \\
& \frac{(\sigma_1 - \rho\sigma_2)(\sigma_1 + \sigma_2) - (\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}} = 0 \\
& \sigma_1^2 + \sigma_1\sigma_2 - \rho\sigma_1\sigma_2 - \rho\sigma_2^2 - \sigma_1^2 - \sigma_2^2 + 2\rho\sigma_1\sigma_2 = 0 \\
& \sigma_1\sigma_2 - \sigma_2^2 + \rho\sigma_1\sigma_2 - \rho\sigma_2^2 = 0 \\
& \sigma_2[\sigma_1 - \sigma_2 + \rho(\sigma_1 - \sigma_2)] = 0 \\
& (\sigma_1 - \sigma_2)(1 + \rho) = 0
\end{aligned} \tag{A3}$$

Given the symmetry of the expression in Eq. (A1), $\frac{\partial}{\partial \sigma_2} z_d = 0 \implies (\sigma_2 - \sigma_1)(1 + \rho) = 0$.

In summary,

$$\begin{array}{lll}
\frac{\partial}{\partial \sigma_1} z_d = 0 & \frac{\partial}{\partial \sigma_2} z_d = 0 & \frac{\partial}{\partial \rho} z_d = 0 \\
(\sigma_1 - \sigma_2)(1 + \rho) = 0 \quad (\text{A4}) & (\sigma_2 - \sigma_1)(1 + \rho) = 0 \quad (\text{A5}) & \sigma_1 \sigma_2 = 0 \quad (\text{A6}) \\
\implies \frac{\partial}{\partial \sigma_1} z_d = 0 \text{ when} & \implies \frac{\partial}{\partial \sigma_2} z_d = 0 \text{ when} & \implies \frac{\partial}{\partial \rho} z_d = 0 \text{ when} \\
\text{either } \sigma_1 = \sigma_2 \text{ or } \rho = -1. & \text{either } \sigma_1 = \sigma_2 \text{ or } \rho = -1. & \text{either } \sigma_1 = 0 \text{ or } \sigma_2 = 0.
\end{array}$$

The first partial derivative with respect to ρ , Eq. (A6), leads to unfeasible values, as neither σ_1 nor σ_2 can be 0. This means that z_d 's extrema lie on the boundary. Since both the numerator and denominator in Eq. (A1) are functions of the same variables, σ_1 and σ_2 , whether z_d reaches its minimum or maximum depends on the values of the numerator's extra parameter, ρ . In particular, z_d is maximized when the numerator reaches its maximum. This requires $-2\rho\sigma_1\sigma_2$ to be at its largest. Given that both σ_1 and σ_2 are positive, this is the case when ρ takes its lowest negative value, -1 .

$$\begin{aligned}
z_d &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2 \times (-1) \times \sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{(\sigma_1 + \sigma_2)^2}}{\sigma_1 + \sigma_2} \\
&= z
\end{aligned} \tag{A7}$$

Thus, $z_d = z$ whenever $\rho = -1$.

Conversely, the numerator reaches its minimum, and, consequently, z_d is minimized, when

$-2\rho\sigma_1\sigma_2$ is at its lowest. This is the case when ρ takes its highest positive value, 1.

$$\begin{aligned}
z_d &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2 \times 1 \times \sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2}}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{(\sigma_1 - \sigma_2)^2}}{\sigma_1 + \sigma_2}
\end{aligned} \tag{A8}$$

Since $\sqrt{(\sigma_1 - \sigma_2)^2}$ cannot be negative, its minimum value is 0. This is the case when $\sigma_1 = \sigma_2$.

Therefore, $z_d = 0$ when $\rho = 1$ and $\sigma_1 = \sigma_2$.

A.1 The z_d -score as a function of the standard error ratio

In the main text I note that to graph the correspondence between the z - and z_d -score in a two dimensional space, we need to consolidate the number of parameters. Let r be the ratio between the estimates' standard errors such that $r = \frac{SE_1}{SE_2}$. Base on this ratio, $SE_1 = rSE_2$ and the z_d -score formula becomes

$$\begin{aligned}
z_d &= \frac{z\sqrt{(rSE_2)^2 + SE_2^2 - 2\rho \times rSE_2 \times SE_2}}{rSE_2 + SE_2} \\
&= \frac{z\sqrt{r^2SE_2^2 + SE_2^2 - 2\rho rSE_2^2}}{SE_2(1 + r)} \\
&= \frac{z\sqrt{SE_2^2(1 - 2\rho r + r^2)}}{SE_2(1 + r)} \\
&= \frac{zSE_2\sqrt{1 - 2\rho r + r^2}}{SE_2(1 + r)} \\
&= \frac{z\sqrt{1 - 2\rho r + r^2}}{1 + r}.
\end{aligned} \tag{A9}$$

Online Appendix B: Simulations on the performance of the SDI method

To check whether the SDI method performs well in practice, I conduct a series of simulations to assess the performance of standard error- and percentile-based SDIs (for normal and skewed distributions, respectively). As a reference point I also report the corresponding DE results. All in all there are 64 scenarios, where I vary the sample size ($N \in \{10000, 1000, 100, 10\}$) and the difference in means between the compared estimates. Q_1 and Q_2 are randomly drawn from a bivariate normal distribution, $(Q_1, Q_2) \sim \mathcal{N}((1.96 \times \sqrt{2} + d), 0, 1, 1; 0)$ ¹, where $d \in \{0, 0.3, 0.5, 1\}$. The last three values of d represent thirty percent, a half, and one full standard error ($\sigma = 1$). When $d = 0$, the lower 95% CI bound of the difference distribution is exactly 0, $Q_{diff} - zSE(Q_{diff}) = [(1.96 \times \sqrt{2} + 0) - 0] - 1.96\sqrt{1^2 + 1^2 - 2 \times 0 \times 1 \times 1} = 0$. When $d = 1$, the value of the CI bound is 1. In effect, d represents the distance between the lower CI bound and 0, with positive values indicating that Q_1 and Q_2 are statistically different. To obtain skewed distributions, I exponentiate Q_1 and Q_2 .

Table B1 shows the results of 10,000 simulations for each combination of N and d values. The test results represent the percent of cases in which Q_1 and Q_2 are found to be statistically distinct. The first thing to note is that, as expected, the DE and SDI results are the same. Thus, there are no accuracy costs associated with adopting the SDI method. Generally, precision increases as either the number of observations or the difference between estimates grows larger.

With sample sizes of 1,000 cases and over, the tests' accuracy is extremely high (100%) as they are able to pick up even on small differences ($d = 0.3$). Importantly, when the lower CI bound of the difference distribution is 0 ($d = 0$), the test results are slightly above 0 half of the times (e.g., 50.27% when $N = 10,000$), and just below 0 the other half. The 50-50 ratio indicates that

¹ $(Q_1, Q_2) \sim \mathcal{N}(\mu_{Q_1}, \mu_{Q_2}, \sigma_{Q_1}, \sigma_{Q_2}; \rho(Q_1, Q_2))$ indicates that Q_1 and Q_2 follow a bivariate normal distribution with the respective means, μ , standard errors, σ , and correlation level, ρ .

Table B1: Simulations on the performance of the SDI method

| | | (a) Normal distributions | | | | (b) Skewed distributions | | | |
|------------|--------|----------------------------------|-------|-------|-------|--------------------------|-------|-------|-------|
| | | <i>The relative difference d</i> | | | | | | | |
| | N | 0 | 0.3 | 0.5 | 1 | 0 | 0.3 | 0.5 | 1 |
| DE method | 10,000 | 50.27 | 100 | 100 | 100 | 50.32 | 100 | 100 | 100 |
| SDI method | | 50.27 | 100 | 100 | 100 | 50.32 | 100 | 100 | 100 |
| DE method | 1,000 | 51.15 | 99.98 | 100 | 100 | 51.01 | 99.18 | 100 | 100 |
| SDI method | | 51.15 | 99.98 | 100 | 100 | 51.01 | 99.18 | 100 | 100 |
| DE method | 100 | 51.58 | 89.61 | 98.01 | 100 | 55.37 | 81.88 | 91.26 | 99.22 |
| SDI method | | 51.58 | 89.61 | 98.01 | 100 | 55.37 | 81.88 | 91.26 | 99.22 |
| DE method | 10 | 54.47 | 69.21 | 76.61 | 91.11 | 77.28 | 86.64 | 90.03 | 96.14 |
| SDI method | | 54.47 | 69.21 | 76.61 | 91.11 | 77.28 | 86.64 | 90.03 | 96.14 |

Note: The values represent the percent of cases in which the tests report that Q_1 and Q_2 are statistically distinct. Standard error-based intervals are employed for normal distributions, and percentile intervals for skewed distributions.

the tests are not biased either for or against finding statistical significance when there is none. This is the hardest test among the four d -value scenarios. For small samples, the tests' performance substantially improves as the difference grows larger. When $N = 10$, the tests picks up on the distinction between Q_1 and Q_2 in more than 90% of the cases once d equals 1.

With a reasonable sample size, the results are substantively similar for skewed distributions. When both the sample size and difference between estimates are small, though, the results may be off. For example, when $N = 10$ and $d = 0$, the tests report statistical significance in 77.28% of the cases (this value is 50.32% for $N = 10,000$). This is to be expected because, when using observed values as interval bounds, we need many observations to capture subtle differences. With a sample of 10, for instance, all uncertainty intervals higher than 90% are practically indistinguishable as 90.1th = 90.2th, . . . , = 99.9th percentile. Only with samples over 1,000 observations we can have distinguishable, 1-digit precision percentiles.

Online Appendix C: Percentile-based SDIs

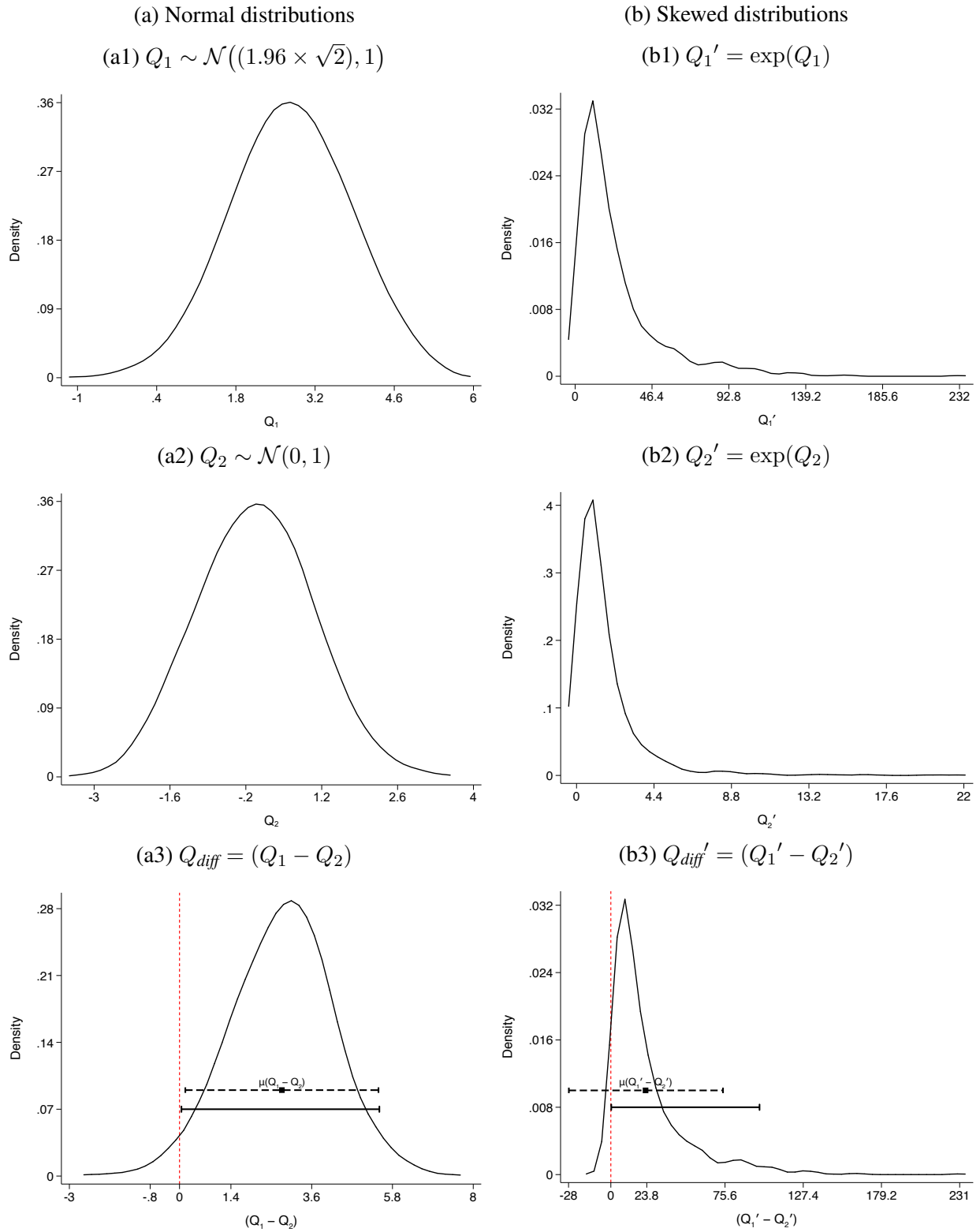
In the main text I formally outline the steps required to compute SDIs based on the quantities of interest's percentiles. Moving from theory to practice, I also present a schematic illustration of the process employing two normally distributed quantities of interest. The choice of the specific distributional form was for practical reasons to facilitate the comparison to the standard error-based results. Here I discuss the procedure more systematically using an example that relaxes the normality assumption. Specifically, first I randomly draw Q_1 and Q_2 from the following bivariate normal distribution, $(Q_1, Q_2) \sim \mathcal{N}((1.96 \times \sqrt{2}), 0, 1, 1; 0)$,² with 1,000 observations each. I then exponentiate Q_1 and Q_2 to obtain right-skewed distributions, $Q_1' = \exp(Q_1)$ and $Q_2' = \exp(Q_2)$. The kernel density plots in Figure C1 outline the resultant distributions along with the respective difference distributions.

On the left hand side, Figure C1a, the normal distributions have the familiar curved bell shape. Moreover, the percentile- and standard error-based CIs of the normal difference distribution are very similar, Figure C1a3. As neither crosses zero (i.e., the red vertical line), they both indicate that the difference is statistically significant. The results, however, are quite different for the skewed difference distribution, Figure C1b3. Specifically, the lower bound of the standard error CI, which is symmetric around the mean μ , crosses the zero line. More problematically, it includes values that Q_{diff}' does not actually contain. In contrast, the percentile CI includes only observed values and it does not cross zero. In effect, it suggests that the difference is statistically significant, i.e., Q_1' and Q_2' are distinct. For skewed distributions, percentile intervals are generally advised.

Table C1 outlines, step-by-step, the procedure to compute percentile-based SDIs using the actual values of Q_1' , Q_2' and Q_{diff}' . Recall that the procedure requires to identify the pair of $k_{Q_1}'^{\text{th}}$ and $(100-k)_{Q_2}'^{\text{th}}$ percentiles whose difference matches the value of the difference distribution's

² $(Q_1, Q_2) \sim \mathcal{N}(\mu_{Q_1}, \mu_{Q_2}, \sigma_{Q_1}, \sigma_{Q_2}; \rho(Q_1, Q_2))$ indicates that Q_1 and Q_2 follow a bivariate normal distribution with the respective means, μ , standard errors, σ , and correlation level, ρ .

Figure C1: Percentile- and standard error-based CIs for normal and skewed distributions



Note: In Figure C1a3 and C1b3, μ indicates the mean of the difference distributions. The dashed horizontal line represents the standard error-based 95% CI around the mean, while the solid horizontal line the percentile-based 95% CI. The vertical line indicates the zero threshold. For statistical significance, the 95% CI should not cross the zero line.

Table C1: Computing percentile-based SDIs

| Q_{diff}' | | Q_1' | | Q_2' | | Percentile Difference | SDI Level |
|------------------|---------|------------------|-----------------|------------------|---------------------|-----------------------|--------------|
| Percentile Value | Percent | Percentile Value | Percent (k) | Percentile Value | Percent ($100-k$) | $P_1' - P_2'$ | ($100-2k$) |
| -11.770 | 0.05% | 0.525 | 0.05% | 21.578 | 99.95% | -21.053 | 99.9 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 0.283 | 2.50% | 2.204 | 2.50% | 7.023 | 97.50% | -4.819 | 95.0 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 2.512 | 8.05% | 4.139 | 8.05% | 3.862 | 91.95% | 0.277 | 83.9 % |
| 2.513 | 8.10% | 4.162 | 8.10% | 3.856 | 91.90% | 0.306 | 83.8 % |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 14.334 | 49.95% | 15.875 | 49.95% | 1.001 | 50.05% | 14.874 | 0.1% |

Note: The first six columns show, in increments of 0.05 percentage points, the percentiles and corresponding percentages for Q_{diff}' , Q_1' , and Q_2' . The last two columns indicate the percentile differences, $P_1' - P_2'$, and the SDI level associated with a given (P_1', P_2') percentile combination, respectively.

2.5th percentile. The first two columns show the Q_{diff}' percentile values and corresponding percentages, ranging from 0.05% to 49.95%, in increments of 0.05 percentage points.³ The increment step used here is suited to compute one digit precision SDIs. We can increase the level of precision by employing smaller increments.⁴ Since we are interested in significance of differences at the 0.05

³ To keep things concise, only a limited number of values are shown in the table. The upper percentile value, $k = 49.95\%$, is a theoretical limit. Since the SDI level equals $(100-2k)\%$, a higher number would lead to either zero or negative values. In practice this means that for any $k \geq 50$, Q_1' and Q_2' are statistically different as long as the point estimates are distinct, $Q_1' \neq Q_2'$. Compared to the standard error-based approach, this corresponds to a scenario where $z_d = 0$.

⁴ Realistically, the level of precision is conditional on the sampling distribution's number of observations, N . One simple formula to compute the p^{th} percentile, denoted as x_i , is $i = p/100 \times N$. x_i refers to the x in ascending order for $i = 1, 2, \dots, N$. When $N = 100$, for instance, only the integer percentiles are differentiated, e.g., 2nd \neq 3rd percentile, but 2.1th = 2.2th, \dots , = 2.9th percentile. Practically, with Clarify's default output (i.e., 1,000 realizations of a given quantity of

level, the value of interest is Q_{diff}' 's 2.5th percentile, 0.283, which is highlighted in the table. The following four columns present the same information for the two original distributions, Q_1' and Q_2' . The only difference is that Q_2' 's percentiles are sorted in descending order.

The last step is to calculate the percentile difference for all (P_1', P_2') pairs, and then search for the one with a value matching the 2.5th $_{Q_{diff}'}$ percentile. The last two columns capture the percentile difference and the associated SDI level. For example, on the second row the -4.819 difference corresponds to the $(2.5^{\text{th}}_{Q_1'}, 97.5^{\text{th}}_{Q_2'})$ percentile pair, as indicated by the percentages in the fourth and sixth columns. Given that in this case $k = 2.5$, the associated SDI level is $(100 - 2k) = 95\%$. The percentile difference that is the closest to but not larger than 0.283, is the one associated with the 83.9% SDI. The respective value, 0.277, and SDI level are also highlighted in the table.

interest), we can meaningfully compute only one digit precision percentiles.

Online Appendix D: SDIs for substantive significance

In the main text, when discussing the range of the $z_{d|m}$ -score for meaningful differences, I only note the minimum and maximum possible values. Here I present step-by-step the $z_{d|m}$ function optimization procedure, which allows one to formally derive the z_d 's extreme values. Recall that the $z_{d|m}$ -score formula is

$$z_{d|m} = \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2}, \quad (\text{D1})$$

where σ_1 and σ_2 are the standard errors of the two quantities on interest, Q_1 and Q_2 , ρ is the correlation parameter, z is the standard score, and m is the minimum value for a meaningful effect. Moreover, $\sigma_1 > 0$, $\sigma_2 > 0$, and $-1 \leq \rho \leq 1$.

To find the minimum and maximum of $z_{d|m}$, we set the first partial derivatives with respect to ρ , σ_1 , σ_2 , and m to 0, treating the standard z -score as a constant.

$$\begin{aligned} \frac{\partial}{\partial \rho} z_{d|m} &= 0 \\ \frac{\partial}{\partial \rho} \left[\frac{z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} + m}{\sigma_1 + \sigma_2} \right] &= 0 \\ \frac{1}{\sigma_1 + \sigma_2} \times \frac{\partial}{\partial \rho} [z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} + m] &= 0 \\ \frac{\partial}{\partial \rho} [z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}] + \frac{\partial}{\partial \rho} m &= 0 \\ z \times \frac{\partial}{\partial \rho} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} + 0 &= 0 \\ \frac{\partial}{\partial \rho} \sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)} &= 0 \quad (\text{the same as line four in Eq. (A2)}) \\ &\vdots \\ \sigma_1\sigma_2 &= 0 \end{aligned} \quad (\text{D2})$$

$$\begin{aligned}
& \frac{\partial}{\partial \sigma_1} z_{d|m} = 0 \\
& \frac{\partial}{\partial \sigma_1} \left[\frac{z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) + m}}{\sigma_1 + \sigma_2} \right] = 0 \\
& \frac{\partial}{\partial \sigma_1} \left[\frac{z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] + \frac{\partial}{\partial \sigma_1} m = 0 \\
& z \times \frac{\partial}{\partial \sigma_1} \left[\frac{\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}}{\sigma_1 + \sigma_2} \right] + 0 = 0 \quad (\text{the same as line three in Eq. (A3)}) \\
& \vdots \\
& (\sigma_1 - \sigma_2)(1 + \rho) = 0 \tag{D3}
\end{aligned}$$

Given the symmetry of the expression in Eq. (D1), $\frac{\partial}{\partial \sigma_2} z_{d|m} = 0 \implies (\sigma_2 - \sigma_1)(1 + \rho) = 0$.

$$\begin{aligned}
& \frac{\partial}{\partial m} z_{d|m} = 0 \\
& \frac{\partial}{\partial m} \left[\frac{z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) + m}}{\sigma_1 + \sigma_2} \right] = 0 \\
& \frac{1}{\sigma_1 + \sigma_2} \times \frac{\partial}{\partial m} [z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2) + m}] = 0 \\
& \frac{\partial}{\partial m} [z\sqrt{(\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2)}] + \frac{\partial}{\partial m} m = 0 \\
& 1 = 0 \tag{D4}
\end{aligned}$$

In summary,

$$\frac{\partial}{\partial \sigma_1} z_{d|m} = 0$$

$$(\sigma_1 - \sigma_2)(1 + \rho) = 0 \tag{D5}$$

$\implies \frac{\partial}{\partial \sigma_1} z_{d|m} = 0$ when either

$$\sigma_1 = \sigma_2 \text{ or } \rho = -1.$$

$$\frac{\partial}{\partial \sigma_2} z_{d|m} = 0$$

$$(\sigma_2 - \sigma_1)(1 + \rho) = 0 \tag{D6}$$

$\implies \frac{\partial}{\partial \sigma_2} z_{d|m} = 0$ when either

$$\sigma_1 = \sigma_2 \text{ or } \rho = -1.$$

$$\begin{aligned} \frac{\partial}{\partial \rho} z_{d|m} = 0 & & \frac{\partial}{\partial m} z_{d|m} = 0 \\ \sigma_1 \sigma_2 = 0 & \quad (D7) & 1 = 0 & \quad (D8) \end{aligned}$$

$$\begin{aligned} \implies \frac{\partial}{\partial \rho} z_{d|m} = 0 \text{ when either} & & \implies \text{unfeasible values.} \\ \sigma_1 = 0 \text{ or } \sigma_2 = 0. & & \end{aligned}$$

The first partial derivatives with respect to ρ , Eq. (D7), and m , Eq. (D8), lead to unfeasible values. This means that $z_{d|m}$'s extrema lie on the boundary. Since both the numerator and denominator in Eq. (D1) are functions of the same variables, σ_1 and σ_2 , whether $z_{d|m}$ reaches its minimum or maximum depends on the values of the numerator's extra parameter, ρ . Both z and m are exogenously determined. In particular, $z_{d|m}$ is maximized when the numerator reaches its maximum. This requires $-2\rho\sigma_1\sigma_2$ to be at its largest. Given that both σ_1 and σ_2 are positive, this is the case when ρ takes its lowest negative value, -1 .

$$\begin{aligned} z_{d|m} &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\ &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2 \times (-1) \times \sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\ &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\ &= \frac{z\sqrt{(\sigma_1 + \sigma_2)^2} + m}{\sigma_1 + \sigma_2} \\ &= \frac{z\sqrt{(\sigma_1 + \sigma_2)^2}}{\sigma_1 + \sigma_2} + \frac{m}{\sigma_1 + \sigma_2} \\ &= z + \frac{m}{\sigma_1 + \sigma_2} \end{aligned} \tag{D9}$$

Thus, $z_{d|m} = z + \frac{m}{\sigma_1 + \sigma_2}$ whenever $\rho = -1$.

Conversely, the numerator reaches its minimum, and, consequently, $z_{d|m}$ is minimized, when

$-2\rho\sigma_1\sigma_2$ is at its lowest. This is the case when ρ takes its highest positive value, 1.

$$\begin{aligned}
z_{d|m} &= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2 \times 1 \times \sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{\sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2} + m}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{(\sigma_1 - \sigma_2)^2} + m}{\sigma_1 + \sigma_2} \\
&= \frac{z\sqrt{(\sigma_1 - \sigma_2)^2}}{\sigma_1 + \sigma_2} + \frac{m}{\sigma_1 + \sigma_2}
\end{aligned} \tag{D10}$$

Since $\sqrt{(\sigma_1 - \sigma_2)^2}$ cannot be negative, its minimum value is 0. This is the case when $\sigma_1 = \sigma_2$. Therefore, $z_{d|m} = \frac{m}{\sigma_1 + \sigma_2}$ when $\rho = 1$ and $\sigma_1 = \sigma_2$.

D.1 An application to interaction effects between continuous variables

Building on an analysis from Rainey (2014), in what follows I illustrate how SDIs can be employed to indicate substantive significance.⁵ The original analysis concerns the effect of electoral permis-

⁵ The SDI for substantive significance is similar to the interval for statistical equivalence (ISE) that has been developed to examine whether, at the beginning of an experiment, the experimental and control groups are matched (Tryon and Lewis, 2009, 178-179). There are at least three notable differences between the ISE and the SDI. First, the two types of intervals serve different purposes. The ISE is designed to gauge whether the *upper* CI bound of the difference distribution is *lower* than a given value, i.e., Δ in their notation (see the first full paragraph on page 179, and Figure 3 on page 177). The SDIs, however, are meant to capture whether the *lower* CI bound of the difference is *higher* than the meaningful value m . Second, the ISE does not incorporate the external value Δ (see Eq. (15), p. 179). In contrast, the $z_{d|m}$ -score is a function of the estimates' standard errors and correlation level *plus* the m value. Third, the ISE cannot be larger than the standard CI. The SDI for substantive significance, however, can be either smaller or larger than the CI.

siveness (measured by the size of district magnitude) on the effective number of political parties. Based on a U.S.–U.K. comparison, Rainey first determines the range of substantively insignificant effects to be $[-0.62, 0.62]$. He next examines the effect of increasing district magnitude from one to seven in countries with a single ethnic group (a proxy for social homogeneity). Since the estimate's CI crosses into the insignificant effects region (the right panel in Figure 1, p. 1089), the effect of electoral permissiveness may be not substantively meaningful.

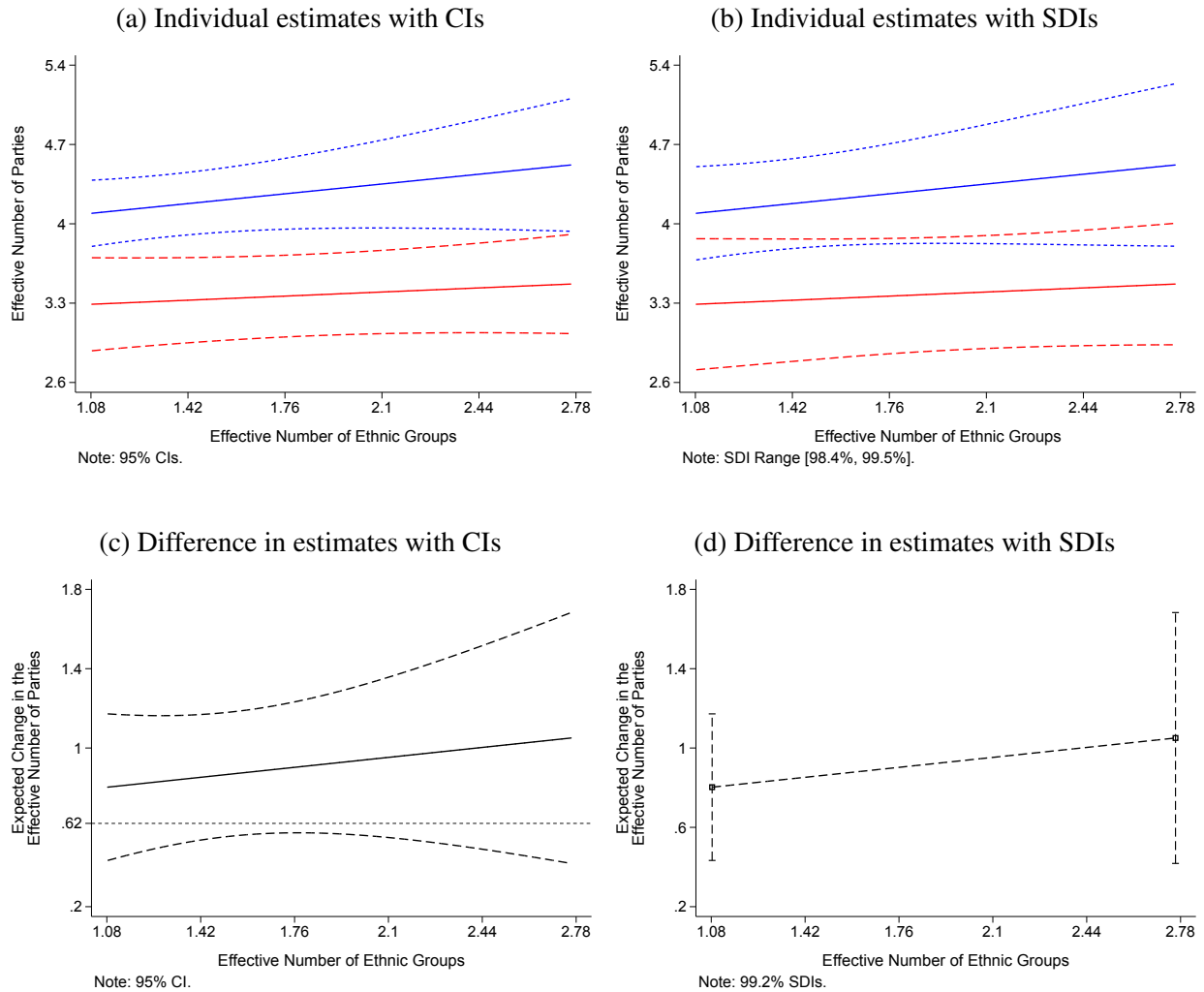
As the number of ethnic groups is set to one, this effect represents only a snapshot of the impact of *District Magnitude* on the effective number of parties. Given that the *Number of Ethnic Groups* and *District Magnitude* variables are interacted in the model, there are theoretical reasons to believe that the effect of *District Magnitude* changes with the values of the modifying variable. If there were empirical support for this expectation, the effect of increasing *District Magnitude* should be substantively higher in heterogenous societies with many ethnic groups.

Figure D1a graphs the expected number of parties across the range of *Number of Ethnic Groups* (specifically between its 10th and 90th percentiles), when *District Magnitude* is at one (in red with dashed CIs) and at seven (in blue with dotted CIs). While the 95% CIs do not overlap at all, this conveys no relevant information when assessing substantive significance. To examine significance of meaningful differences, we need to graph the expected values with SDIs, Figure D1b. As the SDIs overlap throughout it means that, in spite of the 95% CIs not overlapping, the quantities of interest may be not meaningfully different. This conclusion is confirmed by the difference in estimates plot, Figure D1c. Here the lower bound of the 95% CI of the first difference is always below the 0.62 line, which delineates the region of insignificant effects.

Lastly, it is worth noting that, in spite of its widespread use, the standard interaction plot is not well equipped to examine interaction effects (i.e., whether or not the effect of x varies with the values of the interacting variable). Take, for example, Figure D1c. The information it conveys simply reflects that increasing district magnitude from one to seven may be associated with substantively insignificant effects at all values of ethnic homogeneity. More formally, it says that

Figure D1: Indicating substantive significance with SDIs

The Effect of increasing *District magnitude* from 1 to 7 across *Number of ethnic groups*



Note: For Figure D1a and D1b, in red with dashed CI or SDI, respectively, is the expected number of parties when *District Magnitude* is at one across the range of *Number of Ethnic Groups*. In blue with dotted CI or SDI is the expected number of parties when *District Magnitude* is at seven. The SDIs are customized for each $E(y|x = 7; v = [\min, \dots, \max]) - E(y|x = 1; v = [\min, \dots, \max])$ pairwise comparison. Here y is the expected number of effective parties, x is *District Magnitude*, and v is *Number of Ethnic Groups*. Figure D1c graphs the difference between the expected values with the 95% CI. Figure D1d graphs the same difference in estimates with SDIs to visually compare $[E(y|x = 7; v = \min) - E(y|x = 1; v = \min)]$ vs. $[E(y|x = 7; v = \max) - E(y|x = 1; v = \max)]$. For the difference in estimates with SDIs plot, Figure D1d, solid lines indicate that the effect is statistically significant, whereas dashed lines indicate that the effect is not statistically different from 0 at the 0.05 level.

$[E(y|x = 7; v) - E(y|x = 1; v)] < 0.62$ for each $v \in [\min_v, \max_v]$.⁶ None of these comparisons,

⁶ y is the expected number of effective parties, x is the variable of interest (*District Magnitude*),

however, addresses the question whether electoral permissiveness has a substantively different effect in homogenous versus heterogenous societies. To answer this question, we need to evaluate the following second difference: $\{ [E(y|x = 7; v = \max) - E(y|x = 1; v = \max)] - [E(y|x = 7; v = \min) - E(y|x = 1; v = \min)] \} \stackrel{?}{>} 0.62$. To indicate a 0.62 meaningful difference between the effect of *District Magnitude* at the minimum and maximum values of *Number of Ethnic Groups*, we need 99.2% SDIs. The SDI overlap in Figure D1d indicates that the effect of electoral permissiveness is not meaningfully different between societies with high and low ethnic homogeneity. This information is unique to the difference in estimates with SDIs graph, Figure D1d, and cannot be gleaned from the difference with 95% CI plot, Figure D1c.

and v is the modifying variable (*Number of Ethnic Groups*).

Online Appendix E: The list of surveyed articles

Table E1: The 2016 AJPS (Volume 60) & JOP (Volume 78) Articles

| ID # | Title | Author(s) | Journal (Issue) |
|------|--|--|-----------------|
| 1 | The Influence of News Media on Political Elites: Investigating Strategic Responsiveness in Congress | Kevin Arceneaux, Martin Johnson, René Lindstädt, and Ryan J. Vander Wielen | AJPS (1) |
| 2 | Courting the President: How Circuit Court Judges Alter Their Behavior for Promotion to the Supreme Court | Ryan C. Black and Ryan J. Owens | AJPS (1) |
| 3 | Descriptive Representation and Judicial Outcomes in Multiethnic Societies | Guy Grossman, Oren Gazal-Ayal, Samuel D. Pimentel, and Jeremy M. Weinstein | AJPS (1) |
| 4 | Leader Incentives and Civil War Outcomes | Alyssa K. Prorok | AJPS (1) |
| 5 | Candidates or Districts? Reevaluating the Role of Race in Voter Turnout | Bernard L. Fraga | AJPS (1) |
| 6 | What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior | Ryan D. Enos | AJPS (1) |
| 7 | Greater Expectations: A Field Experiment to Improve Accountability in Mali | Jessica Gottlieb | AJPS (1) |
| 8 | Voting Against Your Constituents? How Lobbying Affects Representation | Nathalie Giger and Heike Küver | AJPS (1) |
| 9 | A Closer Look at Reporting Bias in Conflict Event Data | Nils B. Weidmann | AJPS (1) |
| 10 | The Timeline of Elections: A Comparative Perspective | Will Jennings and Christopher Wlezien | AJPS (1) |
| 11 | Decomposing Audience Costs: Bringing the Audience Back into Audience Cost Theory | Joshua D. Kertzer and Ryan Brutger | AJPS (1) |
| 12 | The Primacy of Race in the Geography of Income-Based Voting: New Evidence from Public Voting Records | Eitan D. Hersh and Clayton Nall | AJPS (2) |
| 13 | IMF Conditionality, Government Partisanship, and the Progress of Economic Reforms | Quintin H. Beazer and Byungwon Woo | AJPS (2) |
| 14 | Terrorism, Dynamic Commitment Problems, and Military Conflict | Navin A. Bapat and Sean Zeigler | AJPS (2) |
| 15 | Making Young Voters: The Impact of Preregistration on Youth Turnout | John B. Holbein and D. Sunshine Hillygus | AJPS (2) |
| 16 | Sources of Authoritarian Responsiveness: A Field Experiment in China | Jidong Chen, Jennifer Pan, and Yiqing Xu | AJPS (2) |
| 17 | Performance Federalism and Local Democracy: Theory and Evidence from School Tax Referenda | Vladimir Kogan, Stéphane Lavertu, and Zachary Peskowitz | AJPS (2) |
| 18 | Voter Buying: Shaping the Electorate through Clientelism | F. Daniel Hidalgo and Simeon Nichter | AJPS (2) |

Continued on next page

Table E1 – continued from previous page

| ID # | Title | Author(s) | Journal (Issue) |
|------|---|--|-----------------|
| 19 | The Externalities of Inequality: Fear of Crime and Preferences for Redistribution in Western Europe | David Rueda and Daniel Stegmueller | AJPS (2) |
| 20 | Signaling and Counter-Signaling in the Judicial Hierarchy: An Empirical Analysis of En Banc Review | Deborah Beim, Alexander V. Hirsch, and Jonathan P. Kastellec | AJPS (2) |
| 21 | When Governments Regulate Governments | David M. Konisky and Manuel P. Teodoro | AJPS (3) |
| 22 | Do Constitutional Rights Make a Difference? | Adam S. Chilton and Mila Versteeg | AJPS (3) |
| 23 | Voting Rights for Whom? Examining the Effects of the Voting Rights Act on Latino Political Incorporation | Melissa J. Marschall and Amanda Rutherford | AJPS (3) |
| 24 | Opinion Backlash and Public Attitudes: Are Political Advances in Gay Rights Counterproductive? | Benjamin G. Bishin, Thomas J. Hayes, Matthew B. Incantalupo, and Charles Anthony Smith | AJPS (3) |
| 25 | Legislative Capacity and Executive Unilateralism | Alexander Bolton and Sharece Thrower | AJPS (3) |
| 26 | Are Voters Equal under Proportional Representation? | Orit Kedar, Liran Harsgor, and Raz A. Sheinerman | AJPS (3) |
| 27 | Segregation and Inequality in Public Goods | Jessica Trounstine | AJPS (3) |
| 28 | Loose Cannons or Loyal Foot Soldiers? Toward a More Complex Theory of Interest Group Advertising Strategies | Michael M. Franz, Erika Franklin Fowler, and Travis N. Ridout | AJPS (3) |
| 29 | Does Compulsory Voting Increase Support for Leftist Policy? | Michael M. Bechtel, Dominik Hangartner, and Lukas Schmid | AJPS (3) |
| 30 | Endogenous Credible Commitment and Party Competition over Redistribution under Alternative Electoral Institutions | Michael Becher | AJPS (3) |
| 31 | The Company You Keep: How Voters Infer Party Positions on European Integration from Governing Coalition Arrangements | James Adams, Lawrence Ezrow, and Christopher Wlezien | AJPS (4) |
| 32 | Conspiracy Endorsement as Motivated Reasoning: The Moderating Roles of Political Knowledge and Trust | Joanne M. Miller, Kyle L. Saunders, and Christina E. Farhart | AJPS (4) |
| 33 | Veto Rhetoric and Legislative Riders | Hans J. G. Hassell and Samuel Kernell | AJPS (4) |
| 34 | Who Inherits the State? Colonial Rule and Postcolonial Conflict | Julian Wucherpfennig, Philipp Hunziker, and Lars-Erik Cederman | AJPS (4) |
| 35 | The Dynamics of State Policy Liberalism | Devin Caughey and Christopher Warshaw | AJPS (4) |
| 36 | Experiential Learning and Presidential Management of the U.S. Federal Bureaucracy: Logic and Evidence from Agency Leadership Appointments | George A. Krause and Anne Joseph O'Connell | AJPS (4) |
| 37 | Death and Turnout: The Human Costs of War and Voter Participation in Democracies | Michael T. Koch and Stephen P. Nicholson | AJPS (4) |
| 38 | Electoral Backlash against Climate Policy: A Natural Experiment on Retrospective Voting and Local Resistance to Public Policy | Leah C. Stokes | AJPS (4) |

Continued on next page

Table E1 – continued from previous page

| ID # | Title | Author(s) | Journal (Issue) |
|------|--|--|-----------------|
| 39 | How Politicians Discount the Opinions of Constituents with Whom They Disagree | Daniel M. Butler and Adam M. Dynes | AJPS (4) |
| 40 | The Politics of Judicial Procedures: The Role of Public Oral Hearings in the German Constitutional Court | Jay N. Krehbiel | AJPS (4) |
| 41 | Breaking the Glass Ceiling: Local Gender-Based Earnings Inequality and Women's Belief in the American Dream | Benjamin J. Newman | AJPS (4) |
| 42 | The Logic of Collective Inaction: Senatorial Delay in Executive Nominations | Ian Ostrander | AJPS (4) |
| 43 | Preaching to the Choir: Americans Prefer Communicating to Copartisan Elected Officials | David E. Broockman and Timothy J. Ryan | AJPS (4) |
| 44 | Redistricting and the Causal Impact of Race on Voter Turnout | Bernard L. Fraga | JOP (1) |
| 45 | Party Control of Party Primaries: Party Influence in Nominations for the US Senate | Hans J. G. Hassell | JOP (1) |
| 46 | The Distributional Impact of Greater Responsiveness: Evidence from New York Towns | Michael W. Sances | JOP (1) |
| 47 | Democratization and Linguistic Complexity: The Effect of Franchise Extension on Parliamentary Discourse, 1832–1915 | Arthur Spirling | JOP (1) |
| 48 | Unilateral Powers, Public Opinion, and the Presidency | Andrew Reeves and Jon C. Rogowski | JOP (1) |
| 49 | How Activists Perceive the Utility of International Law | Emilie M. Hafner-Burton, Brad L. LeVeen, and David G. Victor | JOP (1) |
| 50 | Capitol Gains: The Returns to Elected Office from Corporate Board Directorships | Maxwell Palmer and Benjamin Schneer | JOP (1) |
| 51 | The Dynamics of Partisan Identification When Party Brands Change: The Case of the Workers Party in Brazil | Andy Baker, Barry Ames, Anand E. Sokhey, and Lucio R. Renno | JOP (1) |
| 52 | Office Insecurity and Electoral Manipulation | Arturas Rozenas | JOP (1) |
| 53 | Economic Sanctions, Transnational Terrorism, and the Incentive to Misrepresent | Navin A. Bapat, Luis De la Calle, Kaisa H. Hinkkainen, and Elena V. McLean | JOP (1) |
| 54 | Did Georgian Voters Desire Military Escalation in 2008? Experiments and Observations | Jesse Driscoll and Daniel Maliniak | JOP (1) |
| 55 | Ideological Donors, Contribution Limits, and Polarization of American Legislatures | Michael J. Barber | JOP (1) |
| 56 | Anti-Americanism in the Middle East: Evidence from a Field Experiment in Lebanon | Daniel Corstange | JOP (1) |
| 57 | Marriages of Convenience: Explaining Party Mergers in Europe | Raimondas Ibenskas | JOP (2) |

Continued on next page

Table E1 – continued from previous page

| ID # | Title | Author(s) | Journal (Issue) |
|------|---|---|-----------------|
| 58 | Education and Voting Conservative: Evidence from a Major Schooling Reform in Great Britain | John Marshall | JOP (2) |
| 59 | The Intrastate Contagion of Ethnic Civil War | Matthew Lane | JOP (2) |
| 60 | Diffusion of International Border Agreements | K. Chad Clay and Andrew P. Owsiak | JOP (2) |
| 61 | Economists and Public Opinion: Expert Consensus and Economic Policy Judgments | Christopher D. Johnston and Andrew O. Ballard | JOP (2) |
| 62 | Disgust Sensitivity and Public Demand for Protection | Cindy D. Kam and Beth A. Estes | JOP (2) |
| 63 | Birth Legacies, State Making, and War | Douglas Lemke and Jeff Carter | JOP (2) |
| 64 | Power Sharing, Protection, and Peace | Scott Gates, Benjamin A. T. Graham, Yonatan Lupu, Håvard Strand, and Kaare W. Strøm | JOP (2) |
| 65 | Does Immigration Induce Terrorism? | Vincenzo Bove and Tobias Böhmelt | JOP (2) |
| 66 | Presidential Success and the World Economy | Daniela Campello and Cesar Zucco Jr. | JOP (2) |
| 67 | It Is All about Political Incentives: Democracy and the Renewable Feed-in Tariff | Patrick Bayer and Johannes Urpelainen | JOP (2) |
| 68 | The Political Legacy of American Slavery | Avidit Acharya, Matthew Blackwell, and Maya Sen | JOP (3) |
| 69 | Mediating the Electoral Connection: The Information Effects of Voter Signals on Legislative Behavior | John Henderson and John Brooks | JOP (3) |
| 70 | Making Space for Women: Explaining Citizen Support for Legislative Gender Quotas in Latin America | Tiffany D. Barnes and Abby Córdova | JOP (3) |
| 71 | Priming Racial Resentment without Stereotypic Cues | LaFleur Stephens-Dougan | JOP (3) |
| 72 | Reconsidering the Neighborhood Effect: Does Exposure to Residential Unemployment Influence Voters? Perceptions of the National Economy? | Martin Bisgaard, Peter Thisted Dinesen, and Kim Mannemar Sønderskov | JOP (3) |
| 73 | Holding Individual Representatives Accountable: The Role of Electoral Systems | Lukas Rudolph and Thomas Däubler | JOP (3) |
| 74 | News from the Other Side: How Topic Relevance Limits the Prevalence of Partisan Selective Exposure | Jonathan Mummolo | JOP (3) |
| 75 | A Better Life for All? Democratization and Electrification in Post-Apartheid South Africa | Verena Kroth, Valentino Larcinese, and Joachim Wehner | JOP (3) |
| 76 | Personality, Interpersonal Disagreement, and Electoral Information | Jeffrey Lyons, Anand E. Sokhey, Scott D. McClurg, and Drew Seib | JOP (3) |
| 77 | Democracy and the Concept of Personal Integrity Rights | Daniel W. Hill Jr. | JOP (3) |
| 78 | Intergroup Strategies as Contextually Determined: Experimental Evidence from Israel | Ryan D. Enos and Noam Gidron | JOP (3) |
| 79 | Moderate as Necessary: The Role of Electoral Competitiveness and Party Size in Explaining Parties? Policy Shifts | Tarik Abou-Chadi and Matthias Orlowski | JOP (3) |

Continued on next page

Table E1 – continued from previous page

| ID # | Title | Author(s) | Journal (Issue) |
|------|---|--|-----------------|
| 80 | Economic Hardship Triggers Identification with Disadvantaged Minorities | Gábor Simonovits and Gábor Kézdi | JOP (3) |
| 81 | Group Empathy Theory: The Effect of Group Empathy on US Intergroup Attitudes and Behavior in the Context of Immigration Threats | Cigdem V. Sirin, Nicholas A. Valentino, and José D. Villalobos | JOP (3) |
| 82 | Are Costly Signals More Credible? Evidence of Sender-Receiver Gaps | Kai Quek | JOP (3) |
| 83 | Political Context, Government Redistribution, and the Public's Response to Growing Economic Inequality | William W. Franko | JOP (4) |
| 84 | Money in Exile: Campaign Contributions and Committee Access | Eleanor Neff Powell and Justin Grimmer | JOP (4) |
| 85 | Do All Good Things Go Together? Development Assistance and Insurgent Violence in Civil War | Michael Weintraub | JOP (4) |
| 86 | Bureaucratic Investments in Expertise: Evidence from a Randomized Controlled Field Trial | Simon Calmar Andersen and Donald P. Moynihan | JOP (4) |
| 87 | Sharing the Risk? Households, Labor Market Vulnerability, and Social Policy Preferences in Western Europe | Silja Häusermann, Thomas Kurer, and Hanna Schwander | JOP (4) |
| 88 | Strategic Retirements of Elected and Appointed Justices: A Hazard Model Approach | Todd A. Curry and Mark S. Hurwitz | JOP (4) |
| 89 | Inequality and Electoral Accountability: Class-Biased Economic Voting in Comparative Perspective | Timothy Hicks, Alan M. Jacobs, and J. Scott Matthews | JOP (4) |
| 90 | Are We in the Same Boat or Not? The Opposite Effects of Absolute and Relative Income Shifts on Redistributive Preferences | Vittorio Mérola and Agnar Freyr Helgason | JOP (4) |
| 91 | Mayoral Partisanship and Municipal Fiscal Policy | Justin de Benedictis-Kessner and Christopher Warshaw | JOP (4) |
| 92 | Estimating Vote-Specific Preferences from Roll-Call Data Using Conditional Autoregressive Priors | Benjamin E. Lauderdale and Tom S. Clark | JOP (4) |
| 93 | Partisanship by Invitation: Immigrants Respond to Political Campaigns | James A. McCann and Katsuo A. Nishikawa Chávez | JOP (4) |
| 94 | Interpreting Regression Discontinuity Designs with Multiple Cutoffs | Matias D. Cattaneo, Rocío Titiunik, and Gonzalo Vazquez-Bare, and Luke Keele | JOP (4) |
| 95 | The Contraction Effect: How Proportional Representation Affects Mobilization and Turnout | Gary W. Cox, Jon H. Fiva, and Daniel M. Smith | JOP (4) |
| 96 | Let's Just Agree to Disagree: Dispute Resolution Mechanisms in Coalition Agreements | Shaun Bowler, Thomas Bräuninger, Marc Debus, and Indridi H. Indridason | JOP (4) |
| 97 | Time Is Power: The Noninstitutional Sources of Stability in Autocracies | Scott Abramson and Carlos Velasco Rivera | JOP (4) |