The Significance of Differences Interval:
Assessing the Statistical and 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.

Keywords: significance of differences interval, SDI, comparing two quantities of interest

[*] Replication files are available in the *JOP* Dataverse (`https://dataverse.harvard.edu/dataverse/jop`). The empirical analysis has been successfully replicated by the *JOP* replication analyst. Supplementary material for this article is available in the appendix in the online edition.

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

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 $\left[\text{e.g., } \frac{\partial \Pr(y)}{\partial k}\right]$, 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., $\big[\Pr(y|(x+1), z=1) - \Pr(y|x, z=1)\big] - \big[\Pr(y|(x+1), z=0) - \Pr(y|x, z=0)\big]$.

hensive solution that relaxes theses 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

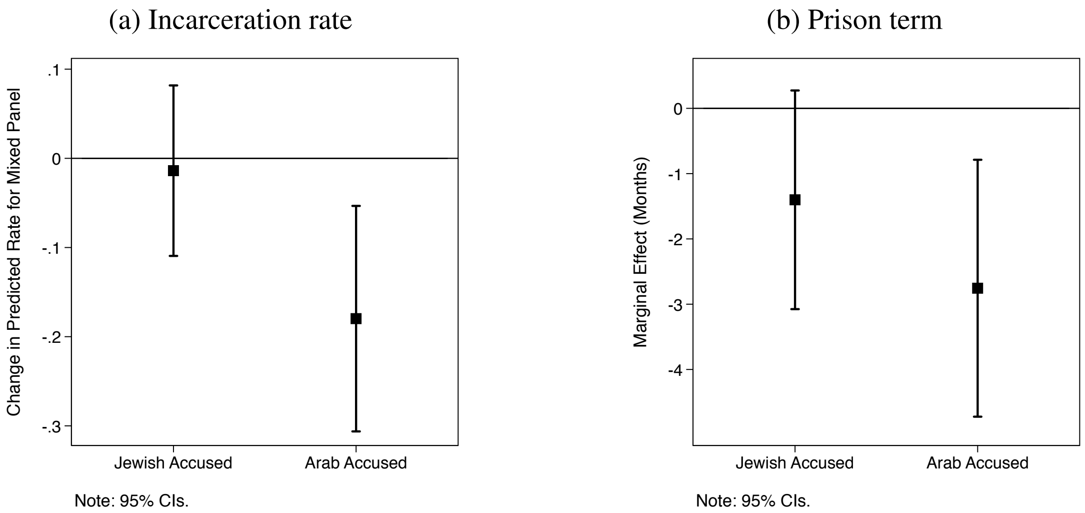
Theoretically, the limitation of overlapping CIs is straightforward: absent additional significance of differences tests, we cannot reach any conclusion. Put differently, we can neither accept nor reject the research hypothesis. Practically, however, this ambiguity leads many practitioners to incorrectly 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



(a) Incarceration rate

(b) Prison term

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

$$\begin{aligned}\big(Q_1 - Q_2\big) - z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2} >& 0 \\ \big(Q_1 - Q_2\big) >& z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1 SE_2}, \qquad (1)\end{aligned}$$

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

$$\begin{aligned}Q_1 - zSE_1 >& Q_2 + zSE_2 \\ \big(Q_1 - Q_2\big) >& z\big(SE_1 + SE_2\big). \qquad (2)\end{aligned}$$

The difference between standard tests and the overlap method is reflected in the unequal width of the respective $\big(Q_1 - Q_2\big)$ 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

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

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

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.

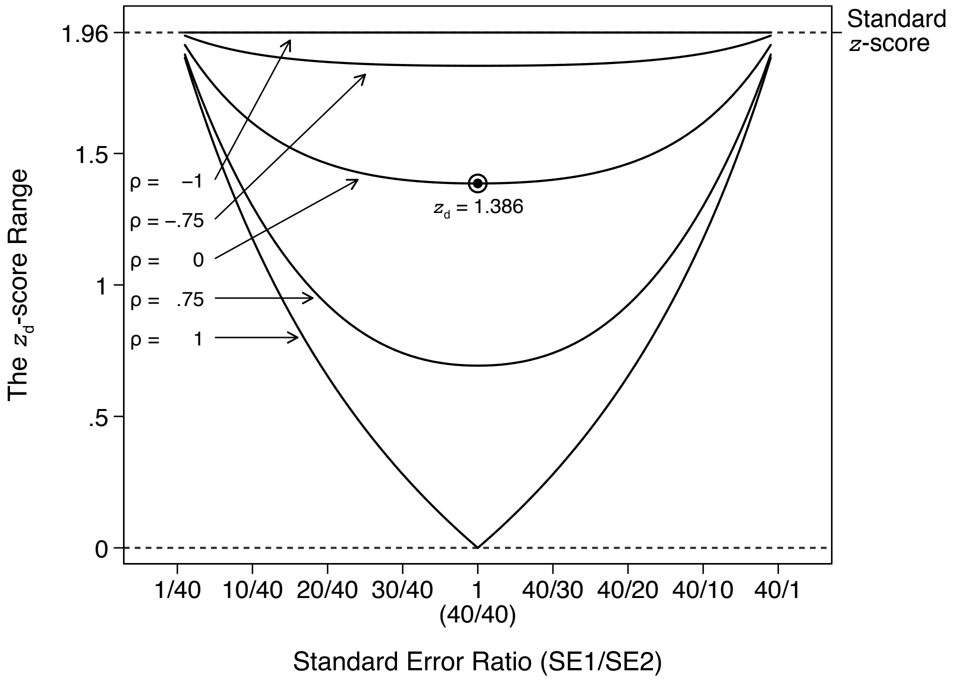
The flat solid line at the top of the graph indicates that, irrespective of the standard error ratio, when $\rho = -1$, z_d equals the standard z-score. This is z_d's highest possible value. Conversely, z_d reaches its minimum value, 0, when $\rho = 1$ and the two quantities of interest have identical 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.

More generally, the possible range for the z_d-score is $[0, z]$.[7] When z_d = 0 no SDIs are

[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.

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

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.

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).

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, $Q_{\mathit{diff}} = (Q_1 - Q_2)$. 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

$$\begin{aligned} P\big(\mathrm{L}_{Q_{\mathit{diff}}[1,\ldots,\infty]} < Q_{\mathit{diff}} < \mathrm{U}_{Q_{\mathit{diff}}[1,\ldots,\infty]}\big) &= 1 - \alpha \\ P\big((\mathrm{L}^{\mathrm{SDI}}_{Q_1[1,\ldots,\infty]} - \mathrm{U}^{\mathrm{SDI}}_{Q_2[1,\ldots,\infty]}) < Q_{\mathit{diff}} < (\mathrm{U}^{\mathrm{SDI}}_{Q_1[1,\ldots,\infty]} - \mathrm{L}^{\mathrm{SDI}}_{Q_2[1,\ldots,\infty]})\big) &= 1 - \alpha, \end{aligned} \tag{4}$$

where $\mathrm{L}_{Q_{\mathit{diff}}}$ and $\mathrm{U}_{Q_{\mathit{diff}}}$ indicate the lower and upper CI limits of Q_{diff}. $[1, \ldots, \infty]$ indicates that the (1 - α)% CI is calculated repeatedly in an (infinitely) long sequence of valid applications. $\mathrm{L}^{\mathrm{SDI}}_{Q_*}$ and $\mathrm{U}^{\mathrm{SDI}}_{Q_*}$ 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 - α)% of the $\big[(\mathrm{L}^{\mathrm{SDI}}_{Q_1} - \mathrm{U}^{\mathrm{SDI}}_{Q_2}), (\mathrm{U}^{\mathrm{SDI}}_{Q_1} - \mathrm{L}^{\mathrm{SDI}}_{Q_2})\big]$ intervals will contain, on average, the true value of the difference in estimates Q_{diff}.

2.2 Conveying statistical significance via SDIs

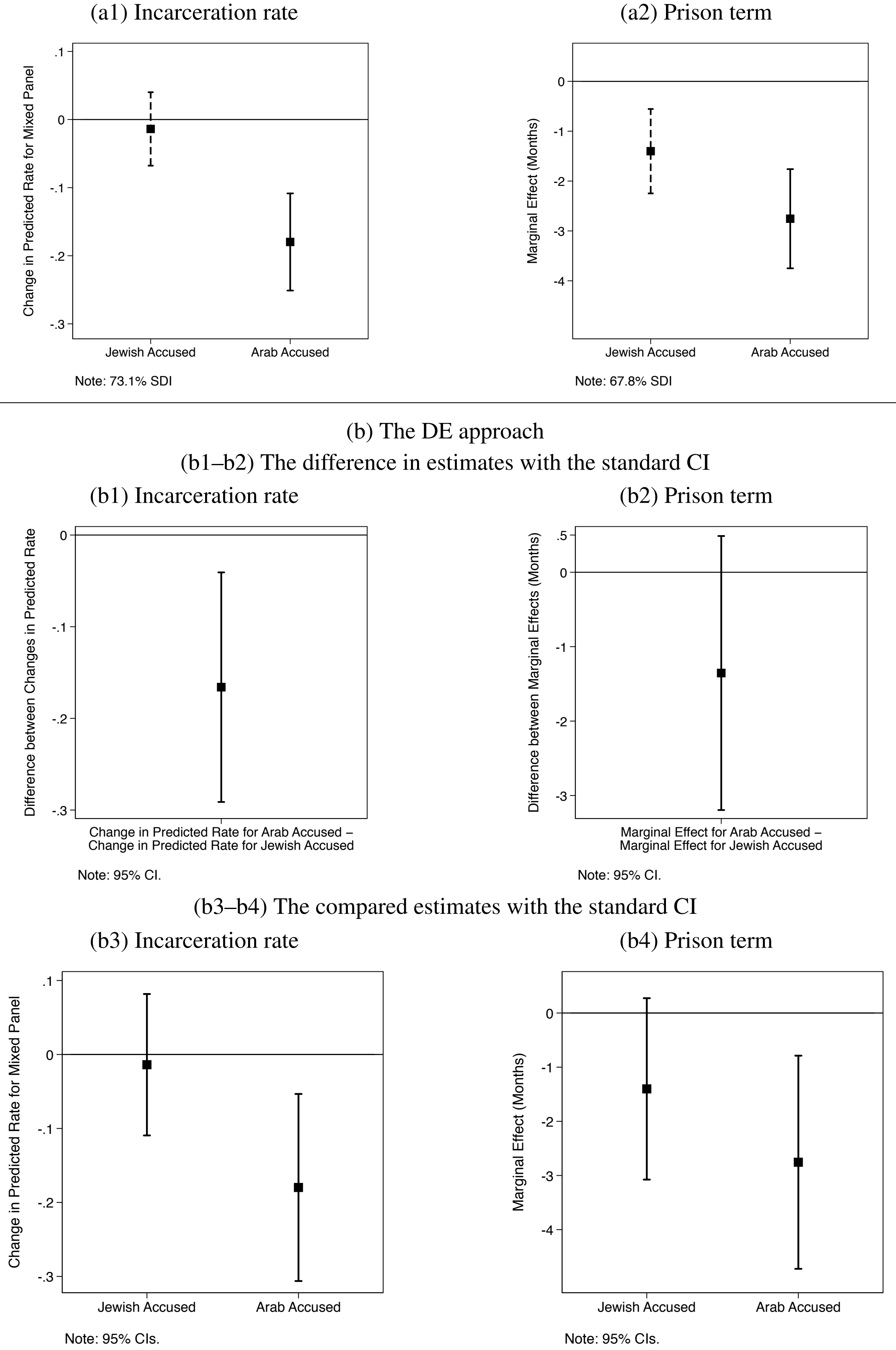
Before looking at practical applications, it is important to consider the performance of the SDI method. To check whether SDIs perform in practice as the theory implies, I conduct a series of simulations detailed in Online Appendix B. The Monte Carlo simulations vary the sample size and the difference 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

(a1) Incarceration rate

(a2) Prison term

(b) The DE approach

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

(b1) Incarceration rate

(b2) Prison term

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

(b3) Incarceration rate

(b4) Prison term

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 that 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 CI[9] – 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]

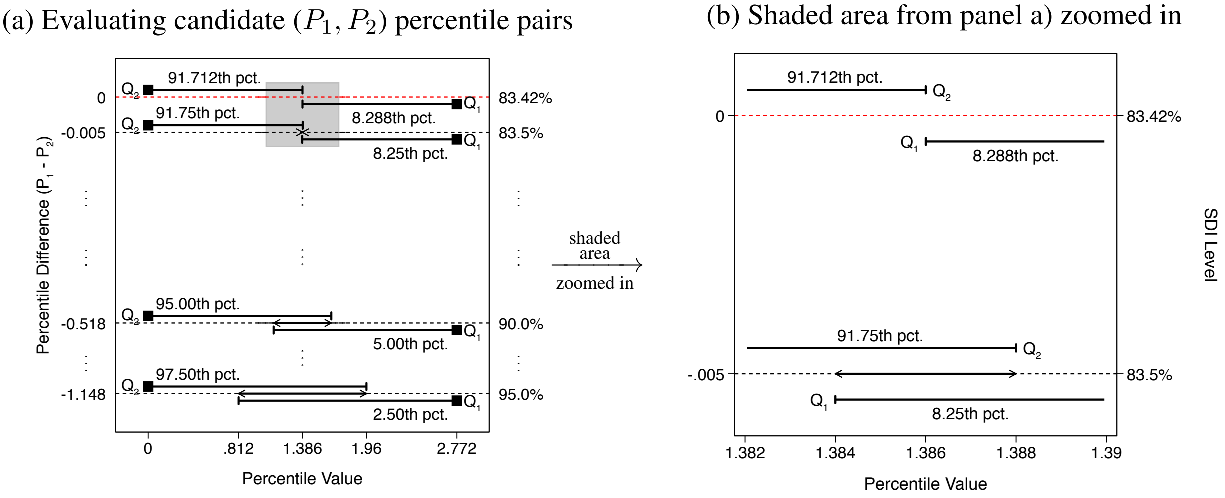
[10] Take the simple case of two independent normal distributions with equal standard errors, for which the required SDI level is 83.5%. For this interval level, the associated z_d-score is 1.386,

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, $Q_{\textit{diff}}$. 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 $Q_{\textit{diff}}$ distribution first. We then calculate its 2.5th percentile, which is $Q_{\textit{diff}}$'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$-k$)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$-k$)th percentile value of Q_2 matches the value of $Q_{\textit{diff}}$'s 2.5th

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 $(Q_1 - 1.386 \times SE_1) - (Q_2 + 1.386 \times SE_2) = Q_{\textit{diff}} - 1.96 \times SE_{Q_{\textit{diff}}}$ by definition, the difference between the respective percentiles does not necessarily satisfy even the weaker condition $(8.25^{\text{th}}\ \text{pct.}_{Q_1} - 91.75^{\text{th}}\ \text{pct.}_{Q_2}) \leq Q_{\textit{diff}} - 1.96 \times SE_{Q_{\textit{diff}}}$. 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 $Q_{\textit{diff}}$), 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 $\big(Q_1 - Q_2\big)$.

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: $(Q_1, Q_2) \sim \mathcal{N}\big((1.96 \times \sqrt{2}), 0, 1, 1; 0\big)$. 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] 2.5th pct.$_{Q_{\mathit{diff}}} = Q_{\mathit{diff}} - zSE(Q_{\mathit{diff}}) = [(1.96 \times \sqrt{2}) - 0] - 1.96\sqrt{1^2 + 1^2 - 2 \times 0 \times 1 \times 1} = 0.$ $(Q_1, Q_2) \sim \mathcal{N}(\mu_{Q_1}, \mu_{Q_2}, \sigma_{Q_1}, \sigma_{Q_2}; \rho(Q_1, Q_2))$ 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

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

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

where $T\big(df, \frac{1}{2}\alpha\big)$ 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}, $Q_{\mathit{diff}} - T\Big(df_{\mathit{diff}}, \frac{1}{2} \times 0.05\Big)SE_{\mathit{diff}}$. 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 $\overline{Q}_{\mathit{diff}} = \overline{Q}_1 - \overline{Q}_2$, and standard error $SE_{\mathit{diff}} = \left\{ \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} \right\}^{1/2} \times \left(\frac{1}{n_1} + \frac{1}{n_2} \right)^{1/2}$ (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 %
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
3.800	2.50%	9.483	2.50%	6.279	97.50%	3.204	95.0 %
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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 %
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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.

$k^{\text{th}}_{Q_1}$ and (100$-k$)$^{\text{th}}_{Q_2}$ 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 $Q_{\textit{diff}}$'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 $\left(2.5^{\text{th}}_{Q_1}, 97.5^{\text{th}}_{Q_2}\right)$ 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 $Q_{\textit{diff}}$'s lower CI bound is greater than m, or $\left(Q_{\textit{diff}} - zSE_{Q_{\textit{diff}}}\right) - m > 0$. 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 satisfies the following equality

$$(Q_1 - z_{d|m}SE_1) - (Q_2 + z_{d|m}SE_2) = \left[(Q_1 - Q_2) - z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1SE_2}\right] - m$$

$$z_{d|m} = \frac{z\sqrt{SE_1^2 + SE_2^2 - 2\rho SE_1SE_2} + m}{SE_1 + SE_2}. \tag{7}$$

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

	Standard CI			*Narrower Intervals*	*No Intervals*	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.

are also excluded. Lastly, the AJPS Workshop articles are not included because they focus on methodological issues rather than theory building and hypothesis testing. See Online Appendix E for the full list of surveyed articles.

In the third column are the studies that compute and report the difference in estimates, about 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

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

having had conducted formal significance of differences tests, the articles that describe estimates as either "statistically" or "significantly" different are also excluded.

are statistically different. The large share of problematic studies aside, it is worth noting that 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.

Acknowledgments

I thank Ryan Bakker, 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.

For additional information about the sdii software (including the user's manual as well as any future updates), visit www.mariusradean.com/software

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: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.

About author

Marius Radean is a Lecturer in the Department of Government at University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom.