Cambridge Medium (xxxx), **xx**, 1–8

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Decoupling Visualization and Testing when Presenting Confidence Intervals

David A. Armstrong II[*†] and William Poirier[‡]

†Professor, Department of Political Science, Western University, London, Ontario, Canada
‡Ph.D. Student, Department of Political Science, Western University, London, Ontario, Canada
[*]Corresponding author. Email: dave.armstrong@uwo.ca

Abstract

Confidence intervals are ubiquitous in the presentation of social science models, data, and effects. When several intervals are plotted together, one natural inclination is to ask whether the estimates represented by those intervals are significantly different from each other. Unfortunately, there is no general rule or procedure that would allow us to answer this question from the confidence intervals alone. It is well known that using the overlaps in 95% confidence intervals to perform significance tests at the 0.05 level does not work. Recent scholarship has developed and refined a set of tools for *inferential confidence intervals* that permit inference on confidence intervals with the appropriate type I error rate in many different bivariate contexts. These are all based on the same underlying idea of identifying the multiple of the standard error (i.e., a new confidence level) such that the overlap in confidence intervals matches the desired type I error rate. These procedures remain stymied by multiple simultaneous comparisons. We propose an entirely new procedure for developing inferential confidence intervals that decouples the testing and visualization that can overcome many of these problems in any visual testing scenario. We provide software in R and Stata to accomplish this goal.

Keywords: Confidence Intervals, Visual Testing, Statistical Inference

1. The Problem

Confidence intervals are ubiquitous in the presentation of data, models and effects in the social sciences. Consider, for example, Gibson's (2024) Figure 1(b) where he shows the proportion of people agreeing with the proposition that it might be better to do away with the US Supreme Court when it starts making decisions that most people disagree with. We reproduce this figure below in Figure 1. While we can see how the proportion agreeing decreases from mid to late 2020 and then increases into July 2022, we might wonder which of these estimates are different from the others. For example, is the change from July 2020 to March 2021 significant? Ideally, we could look at whether the confidence intervals for the two estimates overlap – if they do, the difference between the two estimates is not significant, if they do not, the difference is significant. Unfortunately, this will not always lead us to the right conclusion.[1] The 95% confidence intervals for July 2020 and March 2021 do overlap and as it turns out, the difference between those two proportions is not significant. However, the confidence intervals for July 2020 and July 2022 also overlap, but those two estimates are statistically different from each other.[2] What we know is that when two 95% confidence intervals do not overlap, the difference between the two estimates is significant, but when the intervals overlap somewhat, we cannot necessarily conclude that the two estimates represented by the two confidence intervals are statistically indistinguishable from each other (Browne 1979; Schenker and Gentleman 2001; Radean 2023). How then can we allow readers to visually compare confidence intervals while retaining their statistical properties?

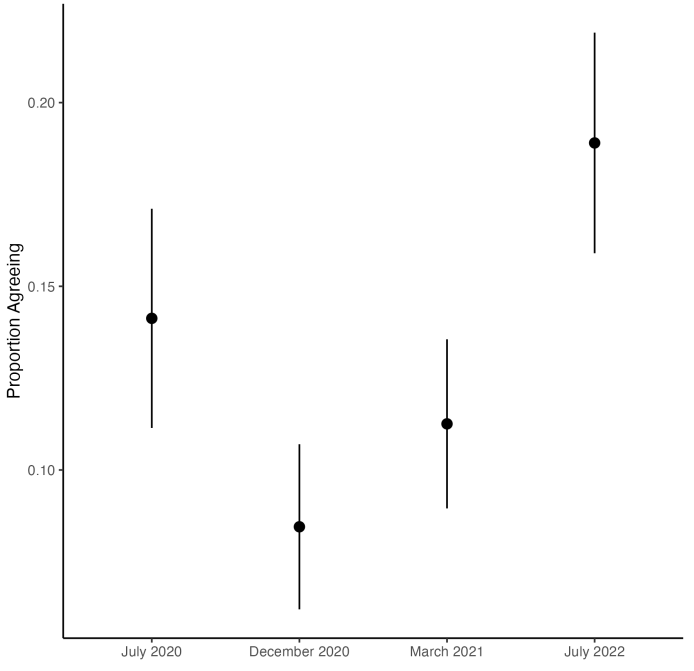

Figure 1. Proportion Agreeing - Do Away with the Supreme Court for Unpopular Decisions

1. Appendix 1 describes the problem in greater detail.
2. We are calculating statistical significance with a pairwise t-test: $t = \frac{\bar{x}_2 - \bar{x}_1}{\sqrt{\text{var}(\bar{x}_2) + \text{var}(\bar{x}_1) - 2\text{cov}(\bar{x}_2, \bar{x}_1)}}$

1.1 Previous Attempts at Visual Testing

We refer to the procedure of judging the statistical significance of the difference between two estimates by whether their confidence intervals overlap as *visual testing*. Below, we use the term *inferential confidence interval*, coined by Tyron (2001), to refer to $(1-\gamma) \times 100\%$ confidence intervals whose (non-)overlaps correspond with tests at the level α, where generally $\gamma \neq \alpha$. For nearly a century, scholars have grappled with the idea that readers will try to make inferences about differences in estimates using confidence intervals (Dice and Laraas 1936; Simpson and Roe 1939).[3] The first systematic scholarship here identified 84% confidence intervals as useful to represent tests at the 95% level under a restrictive set of circumstances (Browne 1979; Tukey 1991).[4] The ensuing decades saw several attempts to generalize this procedure to take account of differences in sample size, uncertainty, covariance and the functional form of the distribution of the difference (Afshartous and Preston 2010; Goldstein and Healy 1995; Payton, Greenstone, and Schenker 2003; Radean 2023).

The main takeaway from the line of research discussed above is that we can identify an appropriate multiplier for the standard error such that two confidence intervals overlap with a given probability. Following Afshartous and Preston (2010) and Radean (2023), we compute the inferential confidence level required for any pair of estimates.

$$Z_\gamma = \left[\frac{F^{-1}\left(\frac{\alpha}{2}\right)}{\frac{\theta}{\sqrt{\theta^2+1-2\rho\theta}} + \frac{\frac{1}{\theta}}{\sqrt{1+\theta^{-2}-2\rho\theta^{-1}}}} \right] \tag{1}$$

$$\Pr(\text{Overlap}) = 2\left(1 - F\left(Z_\gamma \frac{\theta}{\sqrt{\theta^2+1-2\rho\theta}} + \frac{\frac{1}{\theta}}{\sqrt{1+\theta^{-2}-2\rho\theta^{-1}}}\right)\right) \tag{2}$$

where θ is the ratio of standard errors for the two estimates, ρ is the correlation between the two estimates, Z_γ is the multiplier for the standard error for the inferential confidence interval, $F()$ is the CDF of the appropriate t distribution and $F^{-1}()$ is its quantile function.

The potential problem with this method is that the appropriate value of Z_γ will change for each pair of estimates. If multiple pairs are present, Afshartous and Preston (2010) would have us average over the values of Z_γ. Our main problem with this procedure is that in all but the most optimistic cases, the tests produced will not all have the same type I error rate. This leaves the user in essentially the same situation in which she started – not knowing whether estimates are different at a particular confidence level. Below, we develop a procedure that 1) works on an arbitrarily large collection of intervals, 2) directly identifies tests that are not appropriately captured by the inferential confidence intervals, and 3) is agnostic to inferential paradigm.[5]

2. Optimal Visual Testing Intervals

The main goal of existing research is to identify the intervals that produces the appropriate type I error for a particular test. Our innovation is to decouple the testing and the visualization aspects of the display. For most quantities of interest in social science it is easy to compute pairwise tests of difference. Rather than baking the inference into the confidence intervals, we use the appropriate

3. This bears some resemblance to the so-called reference category problem. We discuss the similarities and differences in Appendix 2.

4. Appendix 3 demonstrates why 84% confidence intervals may not always produces the desired result.

5. This article is not a defense of the NHST, it only acknowledges that what may be a logically flawed practice still dominates statistical decision-making in our field and others—see (Gill 1999) for a comprehensive critique. The procedure we propose works equally well to find the credible intervals that correspond with one-sided Bayesian p-values of differences.

tests to make pairwise inferences and then attempt to identify the inferential confidence level (or levels) such that overlapping intervals correspond with insignificant differences and non-overlapping intervals correspond with significant differences. Since we are not using the confidence intervals to do the test, the probability with which any pair of intervals overlaps is incidental.

In any situation where visual testing is desirable, the following algorithm may be implemented.

1. **Conduct all pairwise tests between estimates.** This could be done with or without a multiplicity adjustment, clustered/robust standard errors, etc... The software we provide allows for these kinds of adjustments. The tests could include a reference estimate of zero, with sampling variability equal to zero and with zero covariance with all other estimates. This would ensure that all univariate tests against zero are also respected by the procedure. Once the appropriate p-values (p_{ij}'s) are calculated for $b_j - b_i \; \forall \; i < j$, we define s_{ij} as 1 if $p_{ij} < \alpha$ (the desired type I error rate for the test) and 0 otherwise.[6]

2. **Find the optimal inferential confidence levels.** Once the baseline results of pairwise tests are computed, we find γ (the optimal visual testing confidence interval) as the solution to the following optimization:

$$\arg\max_{\gamma} \sum_{j=2}^{J} \sum_{i=1}^{j-1} I(s_{ij} = s_{ij}^{*}) \tag{3}$$

   where s_{ij}^{*} is 0 $(1 - \gamma) \times 100\%$ confidence intervals for b_i and b_j overlap and 1 if they do not (where **b** is ordered from largest to smallest). That is, we find the value(s) of γ that maximizes the agreement between s_{ij} the "correct" indicator of significance for the difference between b_i and b_j based on a pairwise test and s_{ij}^{*}, the indicator of (non-)overlapping of the inferential confidence intervals for b_i and b_j.

3. **Pick the inferential confidence level that is cognitively easiest.** If there are multiple levels identified by step 2, we should try to identify which one is "best". Since all the optimal levels appropriately account for the same number of tests, that cannot be a criterion for discrimination. Instead, we think about the cognitive load that is required to make the relevant comparisons. For estimates that are significantly different from each other, cognitive ease increases as the distance between their confidence bounds increases (i.e., as it is easier to see that the bounds do not overlap). For insignificant tests, cognitive ease increases as the overlap of their confidence bounds increases. We choose the value that is easiest in the aggregate.

   Assuming $b_1, b_2, \ldots, b_k = \mathbf{b}$ is ordered from largest to smallest, a significant difference between b_i and b_j will have $L_i^{(\gamma)} > U_j^{(\gamma)}$, for $i < j$, assuming we have accounted for all the pairwise tests accurately. Hence, it will be easier to identify significant differences the more distant $L_i^{(\gamma)}$ is from $U_j^{(\gamma)}$. For insignificant tests, the converse is true; where $U_j^{(\gamma)} > L_i^{(\gamma)}$. We can then define a measure of "cognitive easiness" that starts with the difference between the upper and lower bounds$(d_{ij}^{(\gamma)})$.

6. Think of bi and b_j as any estimates for which pairwise differences can be calculated.

$$d_{ij}^{(\gamma)} = \begin{cases} L_i^{(\gamma)} - U_j^{(\gamma)} & \text{if } s_{ij} = 1 \\ U_j^{(\gamma)} - L_i^{(\gamma)} & \text{if } s_{ij} = 0 \end{cases} \tag{4}$$

$d_{ij}^{(\gamma)}$ increases as much when two very close bounds move apart (which increases the readability of the display) as when two already distant bounds move apart (which does not change the display's readability). We solve this problem by imposing a threshold τ on the difference such that no $d_{ij}^{(\gamma)}$ can be bigger than τ times the range of the $(1-\gamma) \times 100\%$ confidence intervals. We define the adjusted distance, $\tilde{d}_{ij}^{(\gamma)}$, as follows:

$$r^{(\gamma)} = \max(U^{(\gamma)}) - \min(L^{(\gamma)}) \tag{5}$$

$$\tilde{d}_{ij}^{(\gamma)} = \begin{cases} \min(\tau \times r^{(\gamma)}, d_{ij}) & \text{if } d_{ij} > 0 \\ \max(-\tau \times r^{(\gamma)}, d_{ij}) & \text{if } d_{ij} < 0 \end{cases} \tag{6}$$

We can then formally define cognitive easiness as the sum of these adjusted differences over all pairs of tests.

$$\text{CE}_\gamma = \sum_{j=2}^{J} \sum_{i=1}^{j-1} \tilde{d}_{ij}^{(\gamma)} \tag{7}$$

These values are not meaningful in the absolute sense, but they are useful when compared to other easiness values for the same problem.

If we can find a level such that *all* the pairwise tests are appropriately represented by whether or not the inferential confidence intervals overlap, then using that interval in a coefficient plot or similar display would produce the desired result – readers could easily identify whether pairs of estimates are different from each other based on whether or not the intervals overlap. We produced software in R and Stata to perform these calculations easily after most models. The vignettes and help pages for the software provide examples and guides to use and interpretation.[7]

3. Case Studies: Iyengar and Westwood (2014)

Below, we describe a case where the optimal confidence intervals provide much more clarity in testing.[8] Iyengar and Westwood (2015) provide implicit, explicit, and behavioral indicators of affective polarization in the US. Of interest here is their second experiment where they explicitly asked respondents to sort between two high school seniors' applications. In each case, one application mentioned that the high school senior was either the president of the Young Republicans or the Young Democrats, randomly varying their GPA (3.5 or 4.0). This results in a four-cell design where either candidate 1 is more qualified, candidate 2 is more qualified, both are equally qualified with a 3.5 GPA, or both are equally qualified with a 4.0 GPA. The authors construct a binary independent variable for whether the respondent chose the Democrat candidate (0) or the Republican candidate (1), a treatment condition variable with three levels (Democrat more qualified, both equally qualified, Republican more qualified), and the respondent's partisan identification with five levels (Democrat, Lean Democrat, Independent, Lean Republican, Republican). Figure 6 of Iyengar and Westwood

7. A software demonstration for both R and Stata can be found in Appendix 5.

8. In the interest of space, we present one case study here. Appendix 4 demonstrates another example where we apply the methodology we describe to a Bayesian result.

(2015) presents the predicted probabilities computed from a logistic regression model where the respondent's choice is regressed on an interaction between the respondent's partisan identification and her received treatment condition.

The top panels of Figure 2 reproduces Iyengar and Westwood (2015)'s results with their original 95% confidence intervals. There are five estimates and 10 possible pairwise tests in each panel resulting in 30 pairwise tests of interest. Using 95% confidence intervals to do these tests, we would get seven of them wrong. Using our method we find that there are a range of confidence intervals that perfectly account for all 30 tests (Equally Qualified: [0.7, 0.863], Republican More Qualified: [0.817, 0.878], Democrat Most Qualified: [0.744, 0.869]). These ranges all include the 84% confidence interval that is often used, but other values (EQ: 0.766, RMQ: 0.859, DMQ: 0.823) enhance the readability of the display without degrading its inferential quality by maximally highlighting overlapping intervals for insignificant tests and non-overlapping regions for significant tests. Further, the justification for 84% confidence intervals is that they presumably overlap 5% of the time under the null hypothesis. In the example above, using 84% confidence intervals to do visual testing would have type I error rates between 4.7% and 7.9% depending on the comparison. Our procedure ensures the appropriate type I error rate then optimizes the display to correspond with the tests.

By way of example, we can see from the optimal confidence intervals (OCI) that the estimates for Republicans and Independents are statistically different when the Republican was most qualified, but the 95% intervals overlap.

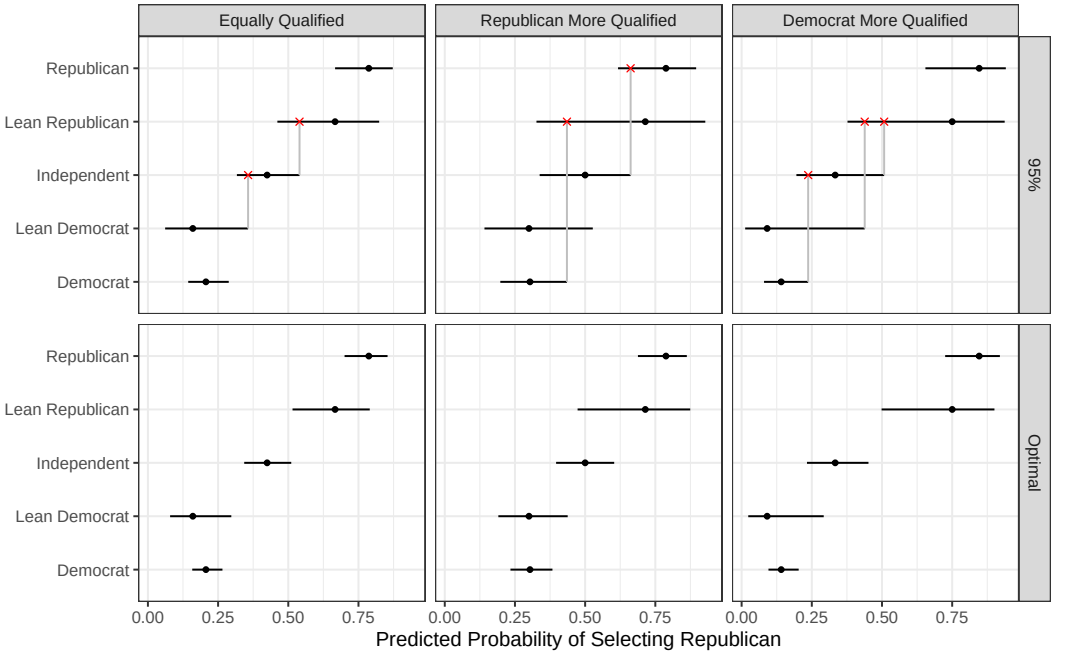

Figure 2. Iyengar and Westwood (2014)'s Predicted Probabilities for Partisan Winner Selection

4. Conclusion

There is a long literature in statistics and more recently in political science that identifies inferential confidence intervals—intervals meant to permit visual testing with the desired type I error rate rather. Despite continued development and refinement, all attempts continue to suffer from an important flaw—the optimal inferential intervals are only defined for a single pair and vary (perhaps interestingly) across pairs of estimates from the same analysis. Rather than focusing on individual tests and aggregating across the results in the hope that the variance in the type I error rate is minimal, our approach is to focus on the full set of tests to identify the inferential confidence level that maximally corresponds with properly done pairwise tests. Our approach is sufficiently flexible that "properly done" could take on any number of meanings across inferential paradigms, multiplicity corrections and operationalizations of the variance-covariance matrix of the estimates. When tests are inappropriately characterized by our procedure, those tests are directly identified and can be flagged in various different ways by the analyst.

5. Acknowledgements

This project was funded by the Social Sciences and Humanities Research Council of Canada, grant CRC-2022-00299.

References

Afshartous, David, and Richard A. Preston. 2010. Confidence intervals for dependent data: equating non-overlap with statistical significance. *Computational Statistics and Data Analysis* 54:2296–2305.

Browne, Richard H. 1979. On visual assessment of the significance of a mean difference. *Biometrics* 35 (3): 657–665.

Dice, L.R., and H.J. Laraas. 1936. A graphic method for comparing several sets of measurements. *Contributions from the Lab of Vertebrate Genetics,* no. 3, 1–3.

Gibson, James L. 2024. Losing legitimacy: the challenges of the dobbs ruling to conventional legitimacy theory. *American Journal of Political Science,* https://doi.org/https://doi.org/10.1111/ajps.12834. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12834. https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12834.

Gill, Jeff. 1999. The insignificance of null hypothesis significance testing. *Political Research Quarterly* 52 (3): 647–674.

Goldstein, Harvey, and Michael J.R. Healy. 1995. The graphical presentation of a collection of means. *Journal of the Royal Statistical Society, Series A* 158 (1): 175–177.

Iyengar, Shanto, and Sean J Westwood. 2015. Fear and loathing across party lines: new evidence on group polarization. *American Journal of Political Science* 59 (3): 690–707.

Payton, Mark E., Matthew H. Greenstone, and Nathaniel Schenker. 2003. Overlapping confidence intervals or standard error intervals: What do they mean in terms of statistical significance? *Journal of Insect Science* 3 (1): 34.

Radean, Marius. 2023. The significance of differences interval: assessing the statistical and substantive difference between two quantities of interest. *Journal of Politics* 85 (3): 969–983.

Schenker, Nathanial, and Jane F. Gentleman. 2001. On judging the significance of differences by examining the overlap between confidence intervals. *The American Statistician* 55 (3).

Simpson, George G., and Anne Roe. 1939. *Quantitative zoology, revised edition.* New York, NY: McGraw-Hill.

Tukey, John. 1991. The philosophy of multiple comparisons. *Statistical Science* 6 (1): 100–116.

Tyron, Warren W. 2001. Evaluating statistical difference, equivalend and indeterminacy using inferential confidence intervals: an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods* 6 (4): 371–386.